

**Rolf Steyer and Werner Nagel**

# **Probability and Conditional Expectation**

**Fundamentals for the Empirical Sciences**

April 4, 2016

University of Jena



# Preface

## Why another book on probability?

This book has two titles. The subtitle, 'Fundamentals for the Empirical Sciences' reflects the intentions and the motivation of the first author for writing this book. He received his academic training in psychology, but considers himself a methodologist. His scientific interest is in explicating fundamental concepts of empirical research (such as causal effects and latent variables) in terms of a language that is *precise* and at the same time is *compatible with the statistical models* used in the analysis of empirical data. Applying statistical models aims at estimating and testing hypotheses about parameters such as expectations, variances, covariances, etc. or of functions of these parameters such as differences between expectations, ratios of variances, or regression coefficients, etc., all of which are terms of probability theory. Precision is necessary for securing logical consistency of theories, whereas compatibility of substantive theories with statistical models is crucial for probing the empirical validity of theoretical propositions via statistical inference.

Much empirical research uses some kind of regression in order to investigate how the expectation of one random variable depends on the values of one or more other random variables. This is true for analysis of variance, regression analysis, applications of the general linear model and the generalized linear model, factor analysis, structural equation modeling, hierarchical linear modeling, and for the analysis of qualitative data. Using these methods we aim at learning about specific regressions. A regression is a synonym for what, in probability theory, is called a factorization of a conditional expectation, provided that the regressor is numerical. This explains the main title of this book: 'Probability and Conditional Expectation'.

## What is it about?

Since the seminal book of Kolmogoroff (1933/1977) the fundamental concepts of probability theory are considered to be special concepts of measure theory. A probability measure is a special finite measure, random variables are special measurable mappings, and expectations of random variables are integrals of measurable mappings with respect to a probability measure. This motivates Part I of this book with three chapters on the measure-theoretical foundations of probability theory. Although at first sight this part seems to be far-off from prac-

tical applications, the contrary is true. This part is indispensable for probability theory and for its applications in empirical sciences. This does not only apply to the concepts of a measure and an integral but in particular to the concept of a measurable mapping, although we concede that the full relevance of this concept will become apparent only in the chapters on conditional expectations. The relevance of measurable mappings is also the reason why this chapter is more detailed than the corresponding chapters in other books on measure theory.

Part II of the book is fairly conventional. The material covered, probability, random variable, expectation, variance, covariance, and some distributions, is found in many books on probability and statistics.

Part III is not only the longest, it is also the core of the book that distinguishes it from other books on probability or on probability and statistics. Only few of these other books contain detailed chapters on conditional expectations. Exceptions are Billingsley (1995), Fristedt and Gray (1997), and Hoffmann-Jørgensen (1994). Our book does not cover any statistical model. However, we treat in much detail *what* we are estimating and *about what* we test or evaluate hypotheses using statistical models. *How* we are estimating is important but *what* we are estimating is of most interest from the empirical scientist point of view and this point is typically neglected in books on statistics and in books on probability theory such as Bauer (1996) or Klenke (2013). A simple example in case is the meaning of the coefficient  $\beta_1$  in the equation  $E(Y|X, Z) = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 ZX$ . Oftentimes, this coefficient is misinterpreted as the ‘main effect’ of  $X$ . However, sometimes  $\beta_1$  has no autonomous meaning at all, e. g. if  $P(Z=0) = 0$ . In general, this coefficient is just a component of the function  $g_1(Z) = \beta_1 + \beta_3 Z$  that can be used to compute the conditional effects of  $X$  on  $Y$  for various values  $z$  of  $Z$  (see chapter 15 for more details). The crucial point is that such concepts can be treated most clearly within probability theory, without referring to a statistical model, sample, estimation or testing. This also includes exemplifying the limitations of conditional expectations. Simple examples show that conditional expectations do not necessarily serve the purpose of the empirical researcher, namely to evaluate the effects of an intervention on an outcome variable. But even in these situations conditional expectations are indispensable for the definition of the parameters of substantive interest (see, e. g., chapter 14).

There is much overlap of Parts II and III with Steyer (2003). However, that book is written in German and the mathematics is considerably less rigorous. Aside from mathematical precision, the two books also differ in the definition of an important concept: In Steyer (2003) the term regression is used as a synonym of a conditional expectation, whereas in this book we use it as a synonym for the factorization of a conditional expectation.

In the first chapter of Part III we gently introduce conditional expectation values and discrete conditional expectations. In the next chapter of this part we then present the general theory of conditional expectations that has been introduced by Kolmogoroff (1933/1977) and is since that time treated in many books on probability theory — although much too briefly in order to be intelligible for researchers in empirical sciences. Our chapter on conditional expectations con-

tains many more details and is supplemented by a number of other chapters on important special aspects and special cases. Such a special aspect is the concept of a residual with respect to a conditional expectation. Residuals have many interesting properties and they are used in order to introduce the concepts of conditional variance and covariance, as well as the notion of a partial correlation. We then turn to specific parameterizations of a conditional expectation, including the concepts of a linear regression and a linear logistic regression. Note that these concepts are introduced as probabilistic concepts. As mentioned above, they are what we aim at estimating in applying the corresponding statistical models. The next two chapters provide the probabilistic foundations of the analysis of conditional and average effects of treatments, interventions, or expositions to potentially harmful or beneficial environments. To our knowledge this material is not found in any other text book. Note, however, that although these two chapters provide important concepts, they do not cover the theory of causal effects, which is another book project of the first author.

Part IV uses conditional expectations in order to introduce conditional independence and conditional distributions. Although these two chapters are more extensive than comparable chapters or sections in other books, the material is found in other books on probability theory as well.

### **For Whom is it?**

This book has been written for two kinds of readers. The first are applied statisticians and empirical researchers who want to understand in a proper language, i. e., in terms of probability theory, what they estimate and test in their empirical studies. The second kind of readers are mathematicians who want to understand in terms of probability theory what applied statisticians and empirical researchers estimate and test in their research. Both kinds of readers are potential contributors to the methodology of empirical sciences.

Many exercises and their solutions provide extensive material for assignments in courses, but they also facilitate independent learning. At the same time these exercises and their solutions help streamlining the main text.

Note that we do not provide all proofs, in particular in the chapters on measure, integral, and distributions. In these cases we refer to other textbooks, instead. We decided to include only those proofs that may help to increase understanding of the background and to learn important mathematical procedures. Of course, we provide proofs of all propositions for which we did not find an appropriate reference.

### **Prerequisites**

We assume that the reader is familiar with the elementary concepts of logic, sets, functions, sequences, and matrices as presented, e. g., in chapters 1 and 2 of Rosen (2012). We try to stick to his notation as close as possible. One of the exceptions is the symbol for the implication for which we use  $\Rightarrow$  instead of  $\rightarrow$ . Another

exception is the symbol for the equivalence for which we use  $\Leftrightarrow$  instead of  $\leftrightarrow$ . Box 0.1 summarizes the most important notation to start with. The concepts referred to by these symbols are defined, e. g., in Rosen (2012) or in Ellis and Gulick (2006). For a rich collection of mathematical formulae we recommend the handbooks of Harris and Stocker (1998) and Bronshtein, Semendyayev, Musiol, and Mühlig (2015).

### **Acknowledgements**

This book could not have been written without the help of many. First of all we thank Ivailo Partchev who prepared the LaTeX framework, many of the figures, tables, and boxes. Some of the figures have been produced by Désirée Thielemann, who also cared for references, read some of the chapters, and hinted at errors. For supporting us with respect to LaTeX, finding errors or suggesting other improvements we also thank Karoline Bading, Marcel Bauer, Sonja Hahn, Gregor Kappler, Christoph Kiefer, Andreas Neudecker, Axel Mayer, Erik Sengewald, Jan Plötner, and Tom Landes. Thanks are also due to Ernesto San Martin for suggesting section 1.7 and proposition (iv) of Theorem 16.37. The proof of Lemma 12.38 is due to Peter Vogel. Finally, we would like to thank our students who kept us thinking on how to improve the text.

**Box 0.1 A First List of Symbols**

$\neg$	not
$\wedge$	and
$\vee$	or
$\Rightarrow$	implies
$\Leftrightarrow$	equivalent to
$\exists$	there is (there exists)
$\forall$	for all
$a \in A$	$a$ is an element of the set $A$
$\emptyset$	empty set
$I$	nonempty finite, countable, or uncountable index set
$(A_i, i \in I)$	family of sets $A_i, i \in I$
$A \cup B$	union of the sets $A$ and $B$
$\bigcup_{i \in I} A_i$	union of the sets $A_i, i \in I$
$A \cap B$	intersection of the sets $A$ and $B$
$\bigcap_{i \in I} A_i$	intersection of the sets $A_i, i \in I$
$A \setminus B$	set difference of the set $A$ and the set $B$
$A^c := \Omega \setminus A$	complement of a set $A \subset \Omega$ with respect to a set $\Omega$
$A \subset B$	$A$ is a subset of the set $B$ ; $A \subset B$ includes $A = B$
$A \times B$	Cartesian product (product set) of $A$ and $B$
$\prod_{i=1}^n A_i$	Cartesian product of the sets $A_i, i = 1, \dots, n$
$f: A \rightarrow B$	mapping $f$ assigning to each $a \in A$ (the domain) one and only one element $b \in B$ (the codomain)
$\sum_{i=1}^n a_i$	sum of the real numbers $a_1, \dots, a_n$
$\prod_{i=1}^n a_i$	product of the real numbers $a_1, \dots, a_n$
$\lim_{n \rightarrow \infty} a_n$	limit of a sequence $a_1, a_2, \dots$ of real numbers
$\sum_{i=1}^{\infty} a_i$	$\lim_{n \rightarrow \infty} \sum_{i=1}^n a_i$ , where $a_1, a_2, \dots$ and therefore $\sum_{i=1}^1 a_i, \sum_{i=1}^2 a_i, \dots$ are sequences of real numbers



# Contents

## Part I Measure-Theoretical Foundations of Probability Theory

<b>1 Measure</b> .....	3
1.1 Introductory Examples .....	3
1.2 $\sigma$ -Algebra and Measurable Space .....	4
1.2.1 $\sigma$ -Algebra Generated by a Set System .....	9
1.2.2 $\sigma$ -Algebra of Borel Sets on $\mathbb{R}^n$ .....	12
1.2.3 $\sigma$ -Algebra on a Cartesian Product .....	13
1.2.4 $\cap$ -Stable Set Systems That Generate a $\sigma$ -Algebra .....	15
1.3 Measure and Measure Space .....	16
1.3.1 $\sigma$ -Additivity and Related Properties .....	17
1.3.2 Other Properties .....	18
1.4 Specific Measures .....	20
1.4.1 Dirac Measure and Counting Measure .....	21
1.4.2 Lebesgue Measure .....	22
1.4.3 Other Examples of a Measure .....	23
1.4.4 Finite and $\sigma$ -Finite Measures .....	23
1.4.5 Product Measure .....	24
1.5 Continuity of a Measure .....	25
1.6 Specifying a Measure via a Generating System .....	27
1.7 $\sigma$ -Algebra That is Trivial With Respect to a Measure .....	28
1.8 Proofs .....	28
1.9 Exercises .....	31
<b>2 Measurable Mapping</b> .....	41
2.1 Image and Inverse Image .....	41
2.2 Introductory Examples .....	42
2.2.1 Example 1: Rectangles .....	42
2.2.2 Example 2: Flipping two Coins .....	44
2.3 Measurable Mapping .....	46
2.3.1 Measurable Mapping .....	46
2.3.2 $\sigma$ -Algebra Generated by a Mapping .....	51
2.3.3 Final $\sigma$ -Algebra .....	54
2.3.4 Multivariate Mapping .....	54
2.3.5 Projection Mapping .....	56
2.3.6 Measurability With Respect to a Mapping .....	56

2.4	Theorems on Measurable Mappings .....	58
2.4.1	Measurability of a Composition .....	59
2.4.2	Theorems on Measurable Functions .....	61
2.5	Equivalence of Two Mappings With Respect to a Measure .....	64
2.6	Image Measure .....	67
2.7	Proofs .....	70
2.8	Exercises .....	75
<b>3</b>	<b>Integral</b> .....	<b>83</b>
3.1	Definition .....	83
3.1.1	Integral of a Nonnegative Step Function .....	83
3.1.2	Integral of a Nonnegative Measurable Function .....	88
3.1.3	Integral of a Measurable Function .....	93
3.2	Properties .....	96
3.2.1	Integral of $\mu$ -Equivalent Functions .....	98
3.2.2	Integral With Respect to a Weighted Sum of Measures .....	100
3.2.3	Integral With Respect to an Image Measure .....	102
3.2.4	Convergence Theorems .....	103
3.3	Lebesgue and Riemann Integral .....	104
3.4	Density .....	106
3.5	Absolute Continuity and the Radon-Nikodym Theorem .....	108
3.6	Integral With Respect to a Product Measure .....	110
3.7	Proofs .....	111
3.8	Exercises .....	120
<b>Part II</b>	<b>Probability, Random Variable and its Distribution</b>	
<b>4</b>	<b>Probability Measure</b> .....	<b>127</b>
4.1	Probability Measure and Probability Space .....	127
4.1.1	Definition .....	127
4.1.2	Formal and Substantive Meaning of Probabilistic Terms .....	128
4.1.3	Properties of a Probability Measure .....	128
4.1.4	Examples .....	130
4.2	Conditional Probability .....	132
4.2.1	Definition .....	132
4.2.2	Filtration and Time Order Between Events and Sets of Events .....	133
4.2.3	Multiplication Rule .....	135
4.2.4	Examples .....	136
4.2.5	Theorem of Total Probability .....	137
4.2.6	Bayes' Theorem .....	138
4.2.7	Conditional-Probability Measure .....	139
4.3	Independence .....	143
4.3.1	Independence of Events .....	143
4.3.2	Independence of Set Systems .....	144
4.4	Conditional Independence Given an Event .....	145

4.4.1	Conditional Independence of Events Given an Event . . . . .	145
4.4.2	Conditional Independence of Set Systems Given an Event . . . . .	146
4.5	Proofs . . . . .	148
4.6	Exercises . . . . .	150
<b>5</b>	<b>Random Variable, Distribution, Density, and Distribution Function . . . . .</b>	<b>155</b>
5.1	Random Variable and its Distribution . . . . .	155
5.2	Equivalence of Two Random Variables With Respect to a Probability Measure . . . . .	161
5.2.1	Identical and $P$ -Equivalent Random Variables . . . . .	161
5.2.2	$P$ -Equivalence, $P^B$ -Equivalence, and Absolute Continuity . . . . .	164
5.3	Multivariate Random Variable . . . . .	167
5.4	Independence of Random Variables . . . . .	169
5.5	Probability Function of a Discrete Random Variable . . . . .	175
5.6	Probability Density With Respect to a Measure . . . . .	178
5.6.1	General Concepts and Properties . . . . .	178
5.6.2	Density of a Discrete Random Variable . . . . .	180
5.6.3	Density of a Bivariate Random Variable . . . . .	180
5.7	Uni- or Multivariate Real-Valued Random Variable . . . . .	182
5.7.1	Distribution Function of a Univariate Real-Valued Random Variable . . . . .	182
5.7.2	Distribution Function of a Multivariate Real-Valued Random Variable . . . . .	184
5.7.3	Density of a Continuous Univariate Real-Valued Random Variable . . . . .	185
5.7.4	Density of a Continuous Multivariate Real-Valued Random Variable . . . . .	187
5.8	Proofs . . . . .	188
5.9	Exercises . . . . .	196
<b>6</b>	<b>Expectation, Variance, and Other Moments . . . . .</b>	<b>199</b>
6.1	Expectation . . . . .	199
6.1.1	Definition . . . . .	199
6.1.2	Expectation of a Discrete Random Variable . . . . .	200
6.1.3	Computing the Expectation Using a Density . . . . .	202
6.1.4	Transformation Theorem . . . . .	203
6.1.5	Rules of Computation . . . . .	206
6.2	Moments, Variance, and Standard Deviation . . . . .	207
6.3	Proofs . . . . .	212
6.4	Exercises . . . . .	213
<b>7</b>	<b>Linear Quasi-Regression, Covariance, and Correlation . . . . .</b>	<b>217</b>
7.1	Linear Quasi-Regression . . . . .	217
7.2	Covariance . . . . .	220
7.3	Correlation . . . . .	224
7.4	Expectation Vector and Covariance Matrix . . . . .	227

7.4.1	Random Vector and Random Matrix .....	227
7.4.2	Expectation of a Random Vector and a Random Matrix ....	228
7.4.3	Covariance Matrix of two Multivariate Random Variables..	229
7.5	Multiple Linear Quasi-Regression .....	231
7.6	Proofs .....	233
7.7	Exercises .....	237
<b>8</b>	<b>Some Distributions .....</b>	<b>245</b>
8.1	Some Distributions of Discrete Random Variables .....	245
8.1.1	Discrete Uniform Distribution .....	245
8.1.2	Bernoulli Distribution .....	246
8.1.3	Binomial Distribution .....	247
8.1.4	Poisson Distribution .....	250
8.1.5	Geometric Distribution .....	252
8.2	Some Distributions of Continuous Random Variables .....	254
8.2.1	Continuous Uniform Distribution .....	254
8.2.2	Normal Distribution.....	256
8.2.3	Multivariate Normal Distribution.....	259
8.2.4	Central $\chi^2$ -Distribution .....	262
8.2.5	Central $t$ -Distribution .....	264
8.2.6	Central $F$ -Distribution.....	266
8.3	Proofs .....	267
8.4	Exercises .....	271
<b>Part III Conditional Expectation and Regression</b>		
<b>9</b>	<b>Conditional Expectation Value and Discrete Conditional Expectation</b>	<b>277</b>
9.1	Conditional Expectation Value .....	277
9.2	Transformation Theorem .....	280
9.3	Other Properties .....	282
9.4	Discrete Conditional Expectation .....	283
9.5	Discrete Regression .....	285
9.6	Examples .....	287
9.7	Proofs .....	291
9.8	Exercises .....	291
<b>10</b>	<b>Conditional Expectation .....</b>	<b>295</b>
10.1	Assumptions and Definitions .....	295
10.2	Existence and Uniqueness .....	297
10.2.1	Uniqueness With Respect to a Probability Measure .....	298
10.2.2	A Necessary and Sufficient Condition of Uniqueness .....	299
10.2.3	Examples .....	300
10.3	Rules of Computation and Other Properties .....	301
10.3.1	Rules of Computation .....	301
10.3.2	Monotonicity .....	302
10.3.3	Convergence Theorems .....	302

10.4	Factorization, Regression, and Conditional Expectation Value . . . .	306
10.4.1	Existence of a Factorization . . . . .	306
10.4.2	Conditional Expectation and Mean-Squared Error . . . . .	307
10.4.3	Uniqueness of a Factorization . . . . .	308
10.4.4	Conditional Expectation Value . . . . .	309
10.5	Characterizing a Conditional Expectation by the Joint Distribution	312
10.6	Conditional Mean Independence . . . . .	313
10.7	Proofs . . . . .	318
10.8	Exercises . . . . .	321
<b>11</b>	<b>Residual, Conditional Variance, and Conditional Covariance</b> . . . . .	<b>329</b>
11.1	Residual With Respect to a Conditional Expectation . . . . .	329
11.2	Coefficient of Determination and Multiple Correlation . . . . .	333
11.3	Conditional Variance and Covariance Given a $\sigma$ -Algebra . . . . .	338
11.4	Conditional Variance and Covariance Given a Value of a Random Variable . . . . .	339
11.5	Properties of Conditional Variances and Covariances . . . . .	342
11.6	Partial Correlation . . . . .	345
11.7	Proofs . . . . .	347
11.8	Exercises . . . . .	348
<b>12</b>	<b>Linear Regression</b> . . . . .	<b>357</b>
12.1	Basic Ideas . . . . .	357
12.2	Assumptions and Definitions . . . . .	359
12.3	Examples . . . . .	361
12.4	Linear Quasi-Regression . . . . .	366
12.5	Uniqueness and Identification of Regression Coefficients . . . . .	367
12.6	Linear Regression . . . . .	369
12.7	Parametrizations of a Discrete Conditional Expectation . . . . .	370
12.8	Invariance of Regression Coefficients . . . . .	374
12.9	Proofs . . . . .	375
12.10	Exercises . . . . .	377
<b>13</b>	<b>Linear Logistic Regression</b> . . . . .	<b>381</b>
13.1	Logit Transformation of a Conditional Probability . . . . .	381
13.2	Linear Logistic Parametrization . . . . .	383
13.3	A Parametrization of a Discrete Conditional Probability . . . . .	385
13.4	Identification of Coefficients of a Linear Logistic Parametrization .	387
13.5	Linear Logistic Regression and Linear Logit Regression . . . . .	388
13.6	Proofs . . . . .	394
13.7	Exercises . . . . .	396

<b>14</b>	<b>Conditional Expectation With Respect to a Conditional-Probability Measure</b>	399
14.1	Introductory Examples	399
14.2	Assumptions and Definitions	404
14.3	Properties	411
14.4	Partial Conditional Expectation	412
14.5	Factorization	413
14.5.1	Conditional Expectation Value With Respect to $P^B$	414
14.5.2	Uniqueness of Factorizations	415
14.6	Uniqueness	415
14.6.1	A Necessary and Sufficient Condition of Uniqueness	415
14.6.2	Uniqueness w.r.t. $P$ and Other Probability Measures	417
14.6.3	Necessary and Sufficient Conditions of $P$ -Uniqueness	418
14.6.4	Properties Related to $P$ -Uniqueness	420
14.7	Conditional Mean Independence With Respect to $P^{Z=z}$	424
14.8	Proofs	426
14.9	Exercises	431
<b>15</b>	<b>Conditional Effect Functions of a Discrete Regressor</b>	437
15.1	Assumptions and Definitions	437
15.2	Conditional Intercept Function and Effect Functions	438
15.3	Implications of Independence of $X$ and $Z$ for Regression Coefficients	441
15.4	Adjusted Conditional Effect Functions	443
15.5	Conditional Logit Effect Functions	447
15.6	Implications of Independence of $X$ and $Z$ for the Logit Regression Coefficients	450
15.7	Proofs	452
15.8	Exercises	454
<b>Part IV Conditional Independence and Conditional Distribution</b>		
<b>16</b>	<b>Conditional Independence</b>	459
16.1	Assumptions and Definitions	459
16.1.1	Two Events	459
16.1.2	Two Sets of Events	461
16.1.3	Two Random Variables	462
16.2	Properties	463
16.3	Conditional Independence and Conditional Mean Independence	470
16.4	Families of Events	473
16.5	Families of Set Systems	473
16.6	Families of Random Variables	475
16.7	Proofs	478
16.8	Exercises	486

<b>17 Conditional Distribution</b> .....	491
17.1 Conditional Distribution Given a $\sigma$ -Algebra or a Random Variable	491
17.2 Conditional Distribution Given a Value of a Random Variable ....	494
17.3 Existence and Uniqueness .....	497
17.3.1 Existence .....	497
17.3.2 Uniqueness of the Functions $P_{Y G}(\cdot, A')$ .....	498
17.3.3 Common Null Set (CNS) Uniqueness of a Conditional Distribution .....	499
17.4 Conditional-Probability Measure Given a Value of a Random Variable .....	502
17.5 Decomposing the Joint Distribution of Random Variables .....	504
17.6 Conditional Independence and Conditional Distributions .....	506
17.7 Expectations With Respect to a Conditional Distribution .....	511
17.8 Conditional Distribution Function and Probability Density .....	513
17.9 Conditional Distribution and Radon-Nikodym Density .....	516
17.10 Proofs .....	520
17.11 Exercises .....	536
<b>References</b> .....	541



## List of Figures

1.1	A Venn diagram of two sets and their intersection	4
1.2	Example of a tree representation of a Cartesian product	7
1.3	A half-open rectangle in the plane $\mathbb{R}^2$	12
1.4	Approximation of an open egg-shaped set $O$ from below	25
1.5	Approximation of an open egg-shaped set $O$ from above	26
2.1	Rectangles and their images under a function	42
2.2	Rectangles and their inverse images under a function	43
2.3	A set and its image under a function	44
2.4	A set and its inverse image under a function	45
2.5	A composition of two mappings	57
2.6	The two step functions $f$ and $h$ in Example 2.55	60
2.7	Positive and negative parts of a function, and its sign function	63
3.1	A partition and a subset of $\Omega$	87
3.2	Increasing nonnegative step functions	90
3.3	Integral of nonnegative step functions w.r.t. the Lebesgue measure	92
3.4	Lebesgue integral of a function from $-5$ to $5$	95
3.5	Illustrating the construction of the Riemann integral 3.22	105
3.6	Integral of a density for two intervals	108
4.1	A filtration with $T = \{1, 2, 3\}$	134
4.2	Probability tree illustrating the multiplication rule	137
4.3	Venn diagram illustrating a partition of a set	138
5.1	Two random variables that are $P^B$ -equivalent if $P(\{\omega_3\}) = 0$	163
6.1	Variance of an indicator of an event as a function of its probability	209
7.1	Linear quasi-regression	218
7.2	The regressor $X$ , the linear quasi-regression $f$ and their composition $Q_{lin}(Y X) = f(X)$	219
7.3	Intercept and slope of a linear function $f: \mathbb{R} \rightarrow \mathbb{R}$	220
8.1	Probability function and distribution function of a binomial distribution	248
8.2	Probability function of the sum of i. i. d. Bernoulli variables	249

8.3	Probability functions of binomial and Poisson distributions .....	251
8.4	Probability functions of three geometric distributions .....	254
8.5	Density of a bivariate uniform distribution on a rectangle .....	255
8.6	Densities of three normal distributions .....	257
8.7	Density and distribution function of the standard normal distribution .....	258
8.8	Density of a bivariate normal distribution .....	260
8.9	Densities of three central $\chi^2$ -distributions .....	263
8.10	Densities of the standard normal and three $t$ -distributions .....	265
8.11	Densities of four central $F$ -distributions .....	266
9.1	The conditional expectation $E(Y X)$ as the composition of $X$ and its factorization $g$ .....	286
11.1	Conditional expectation of the residual on its regressor .....	331
12.1	$E(Y X)$ as the composition of $X$ and the linear regression $g$ .....	369
12.2	A regression with a parametrization that is linear in $(X, X^2)$ but not in $X$ .....	374
13.1	Graph of the logit transformation of $p$ .....	383
13.2	Graphs of three logistic functions .....	385
13.3	$P(Y=1 X)$ as the composition of $X$ , the linear logit regression $f$ , and the logistic function $h$ . .....	389
13.4	Graphs of the $\xi$ -conditional probabilities of items $Y_i$ satisfying the Rasch model .....	394
16.1	Likelihood functions for the response patterns of three variables $Y_i$ satisfying the Rasch model .....	478

## List of Tables

2.1	Example of Measurable Sets Represented by a Mapping $X$ . . . . .	45
2.2	Joe and Ann With Random Assignment and Measurable Mappings . .	53
4.1	Joe and Ann With Random Assignment: Probability Measures . . . . .	131
5.1	Tom, Jim, and Kate . . . . .	159
5.2	No Treatment for Joe . . . . .	166
9.1	Joe and Ann With Random Assignment: Conditional Expectations . .	286
9.2	No Treatment for Joe With Conditional Expectations . . . . .	289
10.1	Joe and Ann With no Treatment Effects . . . . .	314
11.1	No Treatment for Joe With Conditional Expectations and Residuals .	333
11.2	Joe and Ann With Self-Selection and Residuals . . . . .	341
12.1	Joe and Ann With Self-Selection: Conditional Expectations . . . . .	359
13.1	Joe and Ann With a Latent Ability Variable . . . . .	392
14.1	Joe and Ann With Random Assignment: Conditional Expectations With Respect to $P^{X=x}$ . . . . .	400
14.2	Joe and Ann With Self-Selection: Conditional Expectations With Respect to $P^{X=x}$ . . . . .	402
14.3	No Treatment for Joe: Conditional Expectations With Respect to $P^{X=x}$	403
15.1	Four Persons With Self-Selection to Treatment . . . . .	445
15.2	Joe and Ann: Reversed Average Logit Effect . . . . .	452
16.1	Joe and Ann With no Individual Treatment Effects . . . . .	466
16.2	$Z$ -Conditional Independence of $X$ and $U$ . . . . .	471
17.1	Joe and Ann With Self-Selection: Conditional Distribution $P_{Y X}$ . . . . .	493



## List of Boxes

0.1	A First List of Symbols	IX
1.1	Rules of Computation for Measures	19
4.1	Rules of Computation for Probabilities	129
4.2	Independence and Conditional Independence of Events	147
6.1	Rules of Computation for Expectations	207
6.2	Rules of Computation for Variances	209
7.1	Rules of Computation for Covariances	222
7.2	Rules of Computation for Expectations of Random Matrices	229
7.3	Rules of Computation for Covariance Matrices	231
9.1	Rules of Computation for $B$ -Conditional Expectation Values	283
9.2	Rules of Computation for $(X=x)$ -Conditional Expectation Values	284
10.1	Rules of Computation for $\mathcal{C}$ -Conditional Expectations	303
10.2	Rules of Computation for $X$ -Conditional Expectations	304
10.3	Monotonicity of Conditional Expectations	305
11.1	Rules of Computation for a Residual	332
11.2	Rules of Computation for $\mathcal{C}$ -Conditional Covariances	343
11.3	Rules of Computation for Conditional Variances	344
14.1	$P$ -Uniqueness of $E^B(Y \mathcal{C})$	421
16.1	Notation	465
16.2	Conditional Independence of Set Systems	469
16.3	Conditional Independence of Random Variables	470



**Part I**  
**Measure-Theoretical Foundations of**  
**Probability Theory**



# Chapter 1

## Measure

In this chapter, we introduce the concept of a measure and other closely related notions. We start with some examples and then introduce the concept of a  $\sigma$ -algebra, which is crucial in measure theory and probability theory. At first glance this concept seems to be a pure technical construction, which is usually not dealt with in textbooks on ‘Probability and Statistics’ for empirical sciences. However, a  $\sigma$ -algebra turned out to be the natural domain for a measure, including probability measures. Moreover, in probability theory, a  $\sigma$ -algebra is not only the domain of probability measures. The  $\sigma$ -algebra generated by a random variable can be interpreted as the set of events that is represented by this random variable. This is treated in more detail in chapter 2 on measurable mappings, which provides the general theory of random variables because random variables are measurable mappings. The virtues of  $\sigma$ -algebras will become fully apparent in chapter 10 on conditional expectations and its subsequent chapters. The pair  $(\Omega, \mathcal{A})$  consisting of a nonempty set  $\Omega$  and a  $\sigma$ -algebra  $\mathcal{A}$  on  $\Omega$  is called a *measurable space*. Such a measurable space is crucial for the definition of a *measure*. Next, we treat some important examples of measures, including the *counting measure*, the *Dirac measure*, and the *Lebesgue measure*. Finally, we turn to *continuity* and *uniqueness* properties of a measure.

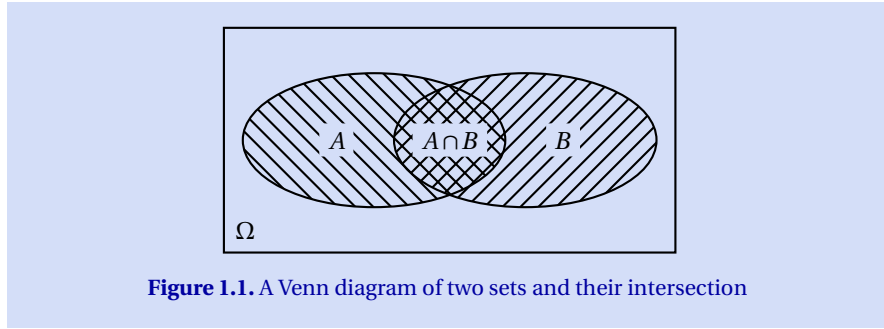
### 1.1 Introductory Examples

Consider Figure 1.1 showing the set  $\Omega$  of all points  $(x, y)$  inside the rectangle and the sets  $A$  and  $B$  of all points  $(x, y)$  inside the two ellipses, respectively. These three sets are subsets of the plane  $\mathbb{R}^2 := \mathbb{R} \times \mathbb{R}$ , where  $\mathbb{R}$  denotes the *set of all real numbers*, and  $\mathbb{R} \times \mathbb{R} := \{(a, b) : a, b \in \mathbb{R}\}$  is the set of all ordered pairs  $(a, b)$  with  $a, b \in \mathbb{R}$ , called the *Cartesian product* or *product set* of  $\mathbb{R}$  with itself. In Figure 1.1, the sets  $A$  and  $B$  have a nonempty intersection. Now let  $area(A)$  and  $area(B)$  denote their areas and  $area(A \cap B)$  the area of their intersection. Inspecting this figure reveals:

$$area(A \cup B) = area(A) + area(B) - area(A \cap B).$$

This example illustrates three important points:

- (a) A measure such as *area* is a function on a *set system on  $\Omega$* , i. e., on a *set of subsets* of a set  $\Omega$  such as  $A$ ,  $B$ , and  $A \cap B$ .



- (b) If *area* is defined for the *subsets*  $A, B \subset \Omega$ , then it is also defined for their *intersection*  $A \cap B$  and for their *union*  $A \cup B$ .
- (c) Measures are *additive*. In other words, if  $A$  and  $B$  are *disjoint* subsets of  $\Omega$ , i. e., if  $A \cap B = \emptyset$ , then  $\text{area}(A \cup B) = \text{area}(A) + \text{area}(B)$ .

Note that, in the example presented in Figure 1.1, the sets  $A$  and  $B$  are *not disjoint*, and this is why  $\text{area}(A \cap B)$  has to be subtracted in the equation displayed above. Points (a) to (c) also apply to other measures such as *length* and *volume* as well as to *probability measures*. Therefore, we adopt a more general language and talk about subsets  $A, B$  of a set  $\Omega$  (or *measurable sets*  $A, B$ ) and their *measure*  $\mu$  instead of lines and their lengths, rectangles and their areas, cubes and their volume, or events and their probabilities.

For example, if  $\Omega = \{1, \dots, 6\}$  denotes the set of possible outcomes of tossing a fair dice,  $A = \{1, 6\}$  and  $B = \{2, 4, 6\}$  denote the events of tossing a 1 or a 6 and tossing an *even number*, respectively. Furthermore,  $A \cap B = \{6\}$  and the probability of tossing a 1 or a 6 or an *even number* — the event  $A \cup B$  — is

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{2}{6} + \frac{3}{6} - \frac{1}{6} = \frac{4}{6}.$$

In the first example, the measure *area* assigns a real number to a subset of  $\mathbb{R}^2$ . In the second example, the measure  $P$  assigns a real number to a subset of  $\Omega = \{1, \dots, 6\}$ . This suggests that a measure should be defined such that it assigns a real number *to all subsets* of a set, i. e., to all elements of the power set. Unfortunately, this may lead to contradictions (see, e. g., Georgii, 2008, p. 9-10). In contrast, defining a measure on a  $\sigma$ -algebra such contradictions can be avoided.

## 1.2 $\sigma$ -Algebra and Measurable Space

In the following definition, we consider a set system  $\mathcal{A}$  on  $\Omega$ , a sequence  $A_1, A_2, \dots$  of subsets of  $\Omega$ , and their countable union. Remember, a *set system on a set*  $\Omega$  is a set of subsets of  $\Omega$  presuming that  $\Omega$  is not empty. A *sequence of subsets of a set*  $\Omega$  is a function from the set  $\mathbb{N}_0 = \{0, 1, 2, \dots\}$  or  $\mathbb{N} = \{1, 2, \dots\}$  or a subset of these sets

to  $\mathcal{P}(\Omega)$ , the *power set* of  $\Omega$ . Furthermore, the *finite union* of the sets  $A_1, \dots, A_n$  and the *countable union* of the sets  $A_1, A_2, \dots$  are defined by

$$\bigcup_{i=1}^n A_i := \{a \in \Omega : \exists i \in \{1, \dots, n\} : a \in A_i\} \quad (1.1)$$

and

$$\bigcup_{i=1}^{\infty} A_i := \{a \in \Omega : \exists i \in \mathbb{N} : a \in A_i\}, \quad (1.2)$$

respectively. Hence, by definition,  $\bigcup_{i=1}^n A_i$  is the set of all elements that are an element of at least one of the sets  $A_i$ ,  $i = 1, \dots, n$ , and  $\bigcup_{i=1}^{\infty} A_i$  is the set of all elements that are an element of at least one of the sets  $A_i$ ,  $i \in \mathbb{N}$ . Finally,  $A^c := \Omega \setminus A$  denotes the complement of  $A$  (with respect to  $\Omega$ ).

**Definition 1.1 ( $\sigma$ -Algebra)**

A set  $\mathcal{A}$  of subsets of a nonempty set  $\Omega$  is called a  $\sigma$ -algebra (or  $\sigma$ -field) on  $\Omega$ , if the following three conditions hold:

- (a)  $\Omega \in \mathcal{A}$ .
- (b) If  $A \in \mathcal{A}$ , then  $A^c \in \mathcal{A}$ .
- (c) If  $A_1, A_2, \dots \in \mathcal{A}$ , then  $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$ .

An element of a  $\sigma$ -algebra is called a *measurable set*.

**Remark 1.2 (Closure With Respect to Set Operations)** Condition (c) postulates that  $\sigma$ -algebras are closed with respect to *countable* unions of sets  $A_1, A_2, \dots \in \mathcal{A}$ . However, in conjunction with (a) and (b) this implies that a  $\sigma$ -algebra is also closed with respect to *finite* unions of sets  $A_1, \dots, A_n \in \mathcal{A}$ , because every finite union of sets  $A_1, \dots, A_n \in \mathcal{A}$  can be represented as a countable union of the sets that are elements of  $\mathcal{A}$ , e. g.,

$$\bigcup_{i=1}^n A_i = A_1 \cup \dots \cup A_n \cup \emptyset \cup \emptyset \cup \dots \quad (1.3)$$

Note that (a) and (b) imply  $\emptyset \in \mathcal{A}$ , because  $\Omega^c = \emptyset$ .

Furthermore, although condition (c) only requires explicitly that  $\sigma$ -algebras are closed with respect to countable unions, Definition 1.1 implies that a  $\sigma$ -algebra is closed also with respect to intersections such as  $A_1 \cap A_2$  and set differences  $A_1 \setminus A_2$ . In other words, if  $A_1$  and  $A_2$  are elements of  $\mathcal{A}$ , then  $A_1 \cup A_2$ ,  $A_1 \cap A_2$ , and  $A_1 \setminus A_2$  are elements of  $\mathcal{A}$  as well, provided that  $\mathcal{A}$  is a  $\sigma$ -algebra. The same is true for countable intersections  $A_1 \cap A_2 \cap \dots$  of elements of  $\mathcal{A}$ . In more formal terms: If  $\mathcal{A}$  is a  $\sigma$ -algebra, then

$$A_1, A_2, \dots \in \mathcal{A} \quad \Rightarrow \quad \bigcap_{i=1}^{\infty} A_i \in \mathcal{A} \quad (1.4)$$

(see Exercise 1-1), where  $\bigcap_{i=1}^{\infty} A_i = A_1 \cap A_2 \cap \dots$  is defined by

$$\bigcap_{i=1}^{\infty} A_i := \{a \in \Omega : \forall i \in \mathbb{N} : a \in A_i\}. \quad (1.5)$$

Because

$$\bigcap_{i=1}^n A_i = A_1 \cap \dots \cap A_n \cap \Omega \cap \Omega \cap \dots, \quad (1.6)$$

we can also conclude

$$A_1, \dots, A_n \in \mathcal{A} \Rightarrow \bigcap_{i=1}^n A_i \in \mathcal{A}, \quad (1.7)$$

where  $\bigcap_{i=1}^n A_i$ , the *finite intersection* of the sets  $A_1, \dots, A_n$ , is defined by

$$\bigcap_{i=1}^n A_i := \{a \in \Omega : \forall i \in \{1, \dots, n\} : a \in A_i\}. \quad (1.8)$$

◁

**Remark 1.3 (Countable and Uncountable Unions)** Defining a  $\sigma$ -algebra we use the symbol  $\sigma$  in order to emphasize that unions of finitely or countably many sets are considered, *but not other unions of sets*. For example, the *closed interval*  $[a, b] := \{x \in \mathbb{R} : a \leq x \leq b, a, b \in \mathbb{R}\}$  on the real axis is identical to the union of singletons  $\{x\}$  that contain only one single element  $x \in \mathbb{R}$ , i. e.,

$$[a, b] = \bigcup_{a \leq x \leq b} \{x\}. \quad (1.9)$$

This union is neither finite nor countable. Hence, condition (c) of Definition 1.1 does *not* imply that this union is necessarily an element of a  $\sigma$ -algebra  $\mathcal{A}$  on  $\mathbb{R}$ , even if all singletons  $\{x\}$ ,  $x \in \mathbb{R}$ , are elements of  $\mathcal{A}$ . ◁

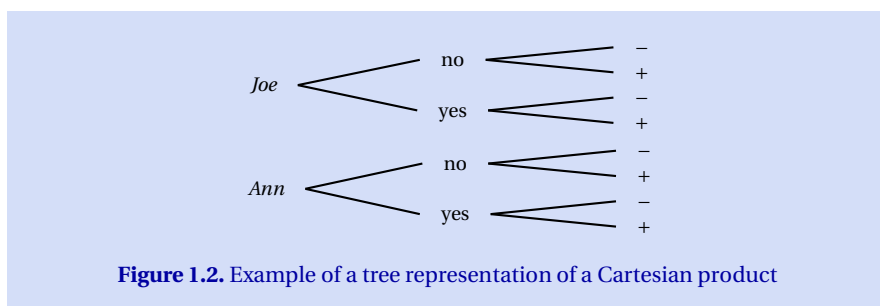
The following notion of a *measurable space* proves to be convenient in measure theory.

**Definition 1.4 (Measurable Space)**

If  $\Omega$  is a nonempty set and  $\mathcal{A}$  a  $\sigma$ -algebra on  $\Omega$ , then the pair  $(\Omega, \mathcal{A})$  is called a *measurable space*.

**Example 1.5 (The Smallest  $\sigma$ -Algebra)** The smallest  $\sigma$ -algebra on a set  $\Omega$  is  $\mathcal{A} = \{\Omega, \emptyset\}$ . It contains only the elements  $\Omega$  and the empty set  $\emptyset$ . As is easily seen,  $\Omega \cup \emptyset = \Omega$ ,  $\Omega^c = \emptyset$ , and  $\emptyset^c = \Omega$  are elements of  $\mathcal{A}$ . This shows that  $\mathcal{A} = \{\Omega, \emptyset\}$  is closed with respect to union and complement. ◁

**Example 1.6 (Power Set)** The *power set*  $\mathcal{P}(\Omega)$  of  $\Omega$ , i. e., the *set of all subsets* of  $\Omega$ , is always a  $\sigma$ -algebra on  $\Omega$ . It is the largest  $\sigma$ -algebra on a set  $\Omega$ . All other  $\sigma$ -algebras on  $\Omega$  are subsets of  $\mathcal{P}(\Omega)$ . ◁



**Example 1.7 (A Small  $\sigma$ -Algebra)** If  $A$  is a subset of  $\Omega$ , then  $\mathcal{A} = \{\Omega, \emptyset, A, A^c\}$  is always a  $\sigma$ -algebra on  $\Omega$  (see Exercise 1-2). Again, it is easily seen that this set system is closed with respect to union and complement.  $\triangleleft$

**Remark 1.8 (Motivation for  $\sigma$ -Algebras)** These examples show that there can be many different  $\sigma$ -algebras on a set  $\Omega$ . Why not simply always use the largest one, the power set  $\mathcal{P}(\Omega)$ ? In fact, this would be possible as long as  $\Omega$  is finite or countable. There are at least three reasons for using  $\sigma$ -algebras. *First*, there are important sets  $\Omega$  (e.g.,  $\Omega = \mathbb{R}$ ) such that measures of interest (e.g., *length* — which is the Lebesgue measure pertaining to  $\Omega = \mathbb{R}$ ) cannot be defined on  $\mathcal{P}(\Omega)$  (see e.g., Wise & Hall, 1993, counterexample 1.25). These measures can be defined, however, on other  $\sigma$ -algebras, such as the Borel- $\sigma$ -algebra [see Eq. (1.18)]. (For an example in which the power set is ‘too large’ see Georgii, 2008, p. 9-10). *Second*, in some sense,  $\sigma$ -algebras contain those elements of a larger set system that are relevant for a particular question. In probability theory, together with  $\Omega$  and a probability measure, each  $\sigma$ -algebra on  $\Omega$  represents a random experiment that is in some sense contained in a (often larger) random experiment. For example, if we consider the random experiment of tossing a dice, then we may focus on whether or not the number is even. Together with  $\Omega$  and the probability measure, the corresponding  $\sigma$ -algebra represents a ‘new’ random experiment contained in the random experiment of tossing a dice (see Exercise 1-3). *Third*, using different  $\sigma$ -algebras is indispensable for introducing conditional expectations, conditional independence, and conditional distributions (see chs. 9 to 17).  $\triangleleft$

**Example 1.9 (Joe and Ann)** Consider the following random experiment: *First*, we sample a unit  $u$  from the set  $\Omega_U := \{Joe, Ann\}$ . *Second*, each unit receives (*yes*) or does not receive a treatment (*no*). *Third*, it is observed whether (+) or not (−) a success criterion is reached (see Fig. 1.2). Defining  $\Omega_X := \{yes, no\}$  and  $\Omega_Y := \{+, -\}$ , the Cartesian product

$$\Omega := \Omega_U \times \Omega_X \times \Omega_Y = \{ (Joe, no, -), (Joe, no, +), \dots, (Ann, yes, +) \}$$

is the set of possible outcomes  $\omega$  of this random experiment. It has eight elements, namely the triples  $(Joe, no, -)$ ,  $(Joe, no, +)$ ,  $\dots$ ,  $(Ann, yes, +)$  (see all eight leaves of Fig. 1.2 for a complete list of these elements).

In this example, a first  $\sigma$ -algebra  $\mathcal{A}$  we may consider is *the set of all subsets of*  $\Omega$ , the power set  $\mathcal{P}(\Omega)$ . This set has  $2^8 = 256$  elements, where 8 is the number of elements, i. e., the cardinality of  $\Omega$  (see Kheyfits, 2010, Th. 1.1.37, p. 22). Among these elements is the set

$$A := \{(Joe, no, -), (Joe, no, +), (Joe, yes, -), (Joe, yes, +)\} = \{Joe\} \times \Omega_X \times \Omega_Y.$$

In the context of probability theory it is also called the event that *Joe is drawn*. Other elements of  $\mathcal{A}$  are the sets

$$B := \{(Joe, yes, -), (Joe, yes, +), (Ann, yes, -), (Ann, yes, +)\} = \Omega_U \times \{yes\} \times \Omega_Y.$$

that the *drawn person is treated*, and

$$C := \{(Joe, no, +), (Joe, yes, +), (Ann, no, +), (Ann, yes, +)\} = \Omega_U \times \Omega_X \times \{+\}$$

that  $\{+\}$  (*success*) occurs, irrespective of which person is drawn and whether or not the person is treated.

Aside from the power set of  $\Omega$  we could also consider the  $\sigma$ -algebras  $\mathcal{A}_1 := \{\Omega, \emptyset, A, A^c\}$ ,  $\mathcal{A}_2 := \{\Omega, \emptyset, B, B^c\}$ , and  $\mathcal{A}_3 := \{\Omega, \emptyset, C, C^c\}$ , to name just three. (For another one see Exercise 1-4). In a sense,  $\mathcal{A}_1$  represents the information which person is drawn. In contrast,  $\mathcal{A}_2$  contains the information whether or not the drawn person is treated, and  $\mathcal{A}_3$  whether or not the drawn person is successful. Of course, all these  $\sigma$ -algebras are subsets of  $\mathcal{P}(\Omega)$ , the power set of  $\Omega$ .  $\triangleleft$

**Example 1.10 (Trace of a Set System and Trace  $\sigma$ -Algebra)** If  $\mathcal{E}$  is a set system on  $\Omega$  and  $\Omega_0 \subset \Omega$ , then

$$\mathcal{E}|_{\Omega_0} := \{\Omega_0 \cap A : A \in \mathcal{E}\}$$

is a set system on  $\Omega_0$ . It is called the *trace of  $\mathcal{E}$  in  $\Omega_0$* . Furthermore, if  $\mathcal{A}$  is a  $\sigma$ -algebra on  $\Omega$  and  $\Omega_0 \subset \Omega$ , then the set system

$$\mathcal{A}|_{\Omega_0} := \{\Omega_0 \cap A : A \in \mathcal{A}\}$$

is a  $\sigma$ -algebra on  $\Omega_0$  (see Exercise 1-5). If  $\Omega \neq \Omega_0$ , then the trace  $\mathcal{A}|_{\Omega_0}$  is a  $\sigma$ -algebra on  $\Omega_0$ , but not on  $\Omega$ , because  $\Omega \notin \mathcal{A}|_{\Omega_0}$ .  $\triangleleft$

**Example 1.11 (Joe and Ann – continued)** In Example 1.9 we defined the event  $A$  that Joe is drawn, the event  $B$  that the drawn person is treated, and the  $\sigma$ -algebra  $\mathcal{A}_2 = \{\Omega, \emptyset, B, B^c\}$ . The trace of  $\mathcal{A}_2$  in  $A$  is

$$\mathcal{A}_2|_A = \{A, \emptyset, A \cap B, A \cap B^c\}.$$

Obviously, just like all elements of  $\mathcal{A}_2$  are subsets of  $\Omega$ , all elements of  $\mathcal{A}_2|_A$  are subsets of  $A$ . From an application point of view, considering  $\mathcal{A}_2|_A$  means to presume that Joe is drawn and consider the events that he is treated or not treated, respectively.  $\triangleleft$

### 1.2.1 $\sigma$ -Algebra Generated by a Set System

The concept of a  $\sigma$ -algebra generated by a set system is useful in order to define important  $\sigma$ -algebras. It is also useful for specifying certain measures (see section 1.6). The following theorem prepares Definition 1.13. Reading this theorem, remember that a  $\sigma$ -algebra on a set  $\Omega$  is itself a set (of subsets of  $\Omega$ ), so that we can consider the intersection of  $\sigma$ -algebras.

**Theorem 1.12 (Intersection of  $\sigma$ -Algebras is a  $\sigma$ -Algebra)**

Let  $I$  be a nonempty (finite, countable, or uncountable) index set and let all  $\mathcal{A}_i, i \in I$ , be  $\sigma$ -algebras on  $\Omega$ . Then  $\bigcap_{i \in I} \mathcal{A}_i$  is also a  $\sigma$ -algebra on  $\Omega$ .

(Proof p. 28)

This theorem allows us to define the  $\sigma$ -algebra generated by a set system on  $\Omega$ .

**Definition 1.13 ( $\sigma$ -Algebra Generated by a Set System)**

Let  $\mathcal{E}$  be a set system on  $\Omega$  and let  $(\mathcal{A}_i, i \in I)$  be the family of all  $\sigma$ -algebras on  $\Omega$  that contain  $\mathcal{E}$  as a subset. Then we define

$$\sigma(\mathcal{E}) := \bigcap_{i \in I} \mathcal{A}_i \quad (1.10)$$

and call it the  $\sigma$ -algebra generated by  $\mathcal{E}$ . The set  $\mathcal{E}$  is also called a generating system of  $\sigma(\mathcal{E})$ .

**Remark 1.14 (Smallest  $\sigma$ -Algebra Containing  $\mathcal{E}$  as a Subset)** According to Theorem 1.12, every set system  $\mathcal{E}$  on  $\Omega$  generates a uniquely defined  $\sigma$ -algebra  $\sigma(\mathcal{E})$  on  $\Omega$ . Note that the  $\sigma$ -algebra  $\sigma(\mathcal{E})$  is the *smallest*  $\sigma$ -algebra on  $\Omega$  containing  $\mathcal{E}$  as a subset, i. e.,

$$\mathcal{C} \text{ is a } \sigma\text{-algebra on } \Omega \text{ and } \mathcal{E} \subset \mathcal{C} \Rightarrow \sigma(\mathcal{E}) \subset \mathcal{C}. \quad (1.11)$$

Furthermore,

$$\sigma[\sigma(\mathcal{E})] = \sigma(\mathcal{E}). \quad (1.12)$$

◁

The following lemma immediately follows from (1.11). It can be used in proofs of the identity of two  $\sigma$ -algebras.

**Lemma 1.15 (Smallest  $\sigma$ -Algebra Containing  $\mathcal{E}$  as a Subset)**

Let  $(\Omega, \mathcal{A})$  be a measurable space and  $\mathcal{E}$  a set system on  $\Omega$  with  $\sigma(\mathcal{E}) = \mathcal{A}$ . If  $\mathcal{C}$  is a  $\sigma$ -algebra on  $\Omega$  with  $\mathcal{E} \subset \mathcal{C} \subset \mathcal{A}$ , then  $\mathcal{C} = \mathcal{A}$ .

(Proof p. 29)

**Remark 1.16 ( $\sigma$ -Algebra Generated by Unions of Set Systems)** Let  $\mathcal{D}, \mathcal{E}, \mathcal{F}$  be set systems on a nonempty set  $\Omega$ . Then

$$\sigma(\mathcal{D} \cup \mathcal{E} \cup \mathcal{F}) = \sigma[\mathcal{D} \cup \sigma(\mathcal{E} \cup \mathcal{F})] \quad (1.13)$$

(see Exercise 1-6). ◁

**Example 1.17 (Several Set Systems May Generate the Same  $\sigma$ -Algebra)** If  $A$  is a subset of  $\Omega$ , then the set system  $\{A\}$  generates the  $\sigma$ -algebra  $\{\Omega, \emptyset, A, A^c\}$ . Note that  $\{\Omega, \emptyset, A, A^c\}$  is also generated by the set systems  $\{A^c\}$  and  $\{A, A^c\}$ , for instance. Hence,

$$\sigma(\{A\}) = \sigma(\{A^c\}) = \sigma(\{A, A^c\}) = \sigma(\{\Omega, \emptyset, A, A^c\}) = \{\Omega, \emptyset, A, A^c\}.$$

In contrast, if  $\emptyset \neq A \neq \Omega$ , then the  $\sigma$ -algebra  $\{\Omega, \emptyset, A, A^c\}$  is neither generated by the set system  $\{A\}$  nor by  $\{\Omega, \emptyset\}$ . Instead,

$$\sigma(\{\emptyset\}) = \sigma(\{\Omega\}) = \sigma(\{\Omega, \emptyset\}) = \{\Omega, \emptyset\},$$

i. e.,  $\{\Omega\}$ ,  $\{\emptyset\}$ , and  $\{\Omega, \emptyset\}$  generate the  $\sigma$ -algebra  $\{\Omega, \emptyset\}$ . ◁

**Example 1.18 (A Generator of the Power Set)** Let  $\Omega$  be finite or countable and let  $\mathcal{E} := \{\{\omega\} : \omega \in \Omega\}$ . Then  $\sigma(\mathcal{E}) = \mathcal{P}(\Omega)$  (see Exercise 1-7). ◁

This example is generalized in Lemma 1.20.

**Remark 1.19 (Partition)** Reading the following lemma, remember that a set system  $\mathcal{E}$  on  $\Omega$  is called a *partition* of  $\Omega$  if

- (a)  $\forall B \in \mathcal{E} : B \neq \emptyset$ .
  - (b)  $\forall B, C \in \mathcal{E} : B \neq C \Rightarrow B \cap C = \emptyset$ .
  - (c)  $\bigcup_{B \in \mathcal{E}} B = \Omega$ .
- ◁

**Lemma 1.20 (An Element of a  $\sigma$ -Algebra Generated by a Partition)**

Let  $\mathcal{E} := \{B_1, \dots, B_n\}$  or  $\mathcal{E} := \{B_1, B_2, \dots\}$  be a finite or countable partition of  $\Omega$ , respectively. Then for all  $C \in \sigma(\mathcal{E})$  there is an  $I(C) \subset \mathbb{N}$  such that

$$C = \bigcup_{i \in I(C)} B_i = \bigcup_{B_i \subset C} B_i, \quad (1.14)$$

where, by convention,  $\bigcup_{i \in \emptyset} B_i := \emptyset$ .

(Proof p. 29)

**Remark 1.21 (Constructing a  $\sigma$ -Algebra)** If  $\mathcal{E} = \{A_1, \dots, A_m\}$  is a finite set of subsets of  $\Omega$ , then there is a finite partition  $\mathcal{F} = \{B_1, \dots, B_n\}$  of  $\Omega$  with  $\sigma(\mathcal{E}) = \sigma(\mathcal{F})$ . Furthermore, if  $\mathcal{E}$  is a *finite* set of subsets of  $\Omega$ , then each element of  $\sigma(\mathcal{E})$  is obtained by finitely many unions, intersections, or complements of elements of  $\mathcal{E}$  (see Exercise 1-8). ◁

**Example 1.22 (Joe and Ann – continued)** In Example 1.11, we already considered the event  $A$  that Joe is drawn and noted that the trace of the  $\sigma$ -algebra  $\mathcal{A}_2 = \{\Omega, \emptyset, B, B^c\}$  in  $A$  is  $\mathcal{A}_2|_A = \{A, \emptyset, A \cap B, A \cap B^c\}$ . In contrast, the  $\sigma$ -algebra on  $\Omega$  generated by the trace  $\mathcal{A}_2|_A$  is

$$\sigma(\mathcal{A}_2|_A) = \{\Omega, \emptyset, A, A^c, A \cap B, A \cap B^c, (A \cap B) \cup A^c, (A \cap B^c) \cup A^c\},$$

where  $(A \cap B) \cup A^c = A^c \cup B$  and  $(A \cap B^c) \cup A^c = A^c \cup B^c$ .  $\triangleleft$

**Remark 1.23 (Monotonicity of Generated  $\sigma$ -Algebras)**

Let  $\mathcal{E}_1, \mathcal{E}_2$  be set systems on  $\Omega$  with  $\mathcal{E}_1 \subset \mathcal{E}_2$ . Then  $\sigma(\mathcal{E}_1) \subset \sigma(\mathcal{E}_2)$  (see Exercise 1-9).  $\triangleleft$

An important kind of  $\sigma$ -algebras are those for which there is a countable set system that generates them.

**Definition 1.24 (Countably Generated  $\sigma$ -Algebra)**

Let  $(\Omega, \mathcal{A})$  be a measurable space. Then  $\mathcal{A}$  is called countably generated if there is a finite or countable set  $\mathcal{E} \subset \mathcal{A}$  such that  $\sigma(\mathcal{E}) = \mathcal{A}$ .

**Example 1.25 (Some Countably Generated  $\sigma$ -Algebras)** Examples of countably generated  $\sigma$ -algebras are:

- (a) All  $\sigma$ -algebras on a finite set  $\Omega$ .
- (b)  $\mathcal{P}(\mathbb{N}_0^n)$ ,  $n \in \mathbb{N}$ .

(For a proof see Exercise 1-10). For another example, see Remark 1.28.  $\triangleleft$

**Remark 1.26 (A Caveat)** Note that there are countably generated  $\sigma$ -algebras for which not all of their elements can be constructed by countably many unions, intersections, or complements of elements of the generating system. An example in case are Borel  $\sigma$ -algebras on  $\mathbb{R}$  or  $\mathbb{R}^n$  (see Rem. 1.28 and Michel, 1978, sect. I.4).  $\triangleleft$

**Lemma 1.27 ( $\sigma$ -Algebra Generated by the Trace of a Set System)**

Let  $A \subset \Omega$  be nonempty,  $\mathcal{E} \subset \mathcal{P}(\Omega)$ , and  $\mathcal{E}|_A := \{C \cap A : C \in \mathcal{E}\}$ . Then

$$\sigma(\mathcal{E}|_A) = \sigma(\mathcal{E})|_A, \quad (1.15)$$

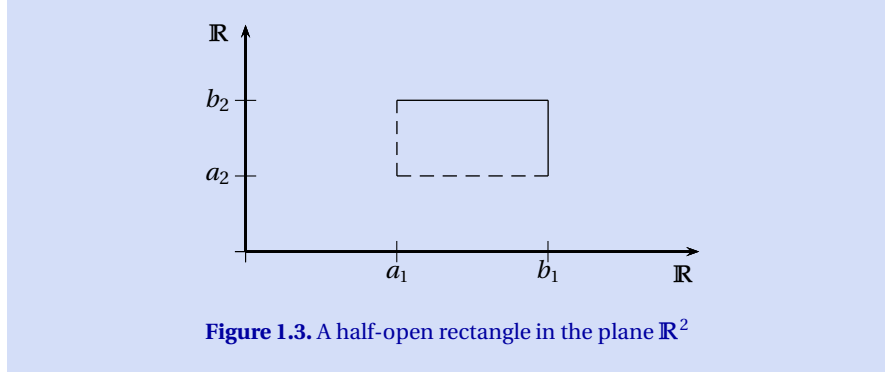
where  $\sigma(\mathcal{E}|_A)$  denotes the  $\sigma$ -algebra generated on  $A$ , whereas  $\sigma(\mathcal{E})$  is a  $\sigma$ -algebra on  $\Omega$ . Furthermore, if  $\mathcal{C}$  is a  $\sigma$ -algebra on  $\Omega$  and  $A \in \mathcal{C}$  such that

$$\forall C \in \mathcal{C}: C \neq A \Rightarrow A \cap C = \emptyset, \quad (1.16)$$

i. e.,  $A$  does not intersect with any other element of  $\mathcal{C}$ , then

$$\sigma(\mathcal{C} \cup \mathcal{E})|_A = \mathcal{C}|_A. \quad (1.17)$$

(Proof p. 30)



Hence, according to Equation (1.15), the  $\sigma$ -algebra generated by the trace of a set system  $\mathcal{E}$  is the trace of the  $\sigma$ -algebra generated by  $\mathcal{E}$  and according to Equation (1.17), the trace of the  $\sigma$ -algebra  $\sigma(\mathcal{C} \cup \mathcal{E})$  in the set  $A$  is identical to the trace of the  $\sigma$ -algebra  $\mathcal{C}$  in  $A$ , if 1.16 holds.

### 1.2.2 $\sigma$ -Algebra of Borel Sets on $\mathbb{R}^n$

For  $a, b \in \mathbb{R}$  with  $a < b$ , let us consider a *half-open interval*  $]a, b]$  in  $\mathbb{R}$ , which is defined by

$$]a, b] := \{x \in \mathbb{R} : a < x \leq b\},$$

and the *set system*

$$\mathcal{I}_1 := \{]a, b] : a, b \in \mathbb{R} \text{ and } a < b\}$$

of all half-open intervals in  $\mathbb{R}$ . The  $\sigma$ -algebra generated by this set system is called the *Borel  $\sigma$ -algebra* on  $\mathbb{R}$ . It is denoted by  $\mathcal{B}$ . The elements of  $\mathcal{B}$  are called the *Borel sets* of  $\mathbb{R}$ . In formal terms,

$$\mathcal{B} := \mathcal{B}_1 := \sigma(\mathcal{I}_1). \quad (1.18)$$

Note that there are several sets systems generating the Borel  $\sigma$ -algebra (see, e.g., Klenke, 2013, Th. 1.23, p. 9). In particular,

$$\mathcal{B}_1 = \sigma(\{]-\infty, b] : b \in \mathbb{R}\}) \quad (1.19)$$

(see Georgii, 2008, p. 12). Similarly, we define the *Borel  $\sigma$ -algebra on  $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$*  to be the  $\sigma$ -algebra generated by the set system  $\mathcal{I}_2$  of all half-open *rectangles* in  $\mathbb{R}^2$ , whose sides are parallel to the axes (see Fig. 1.3). These rectangles are defined by

$$]a_1, b_1] \times ]a_2, b_2] = \{(x_1, x_2) \in \mathbb{R}^2 : a_1 < x_1 \leq b_1, a_2 < x_2 \leq b_2\}.$$

The  $\sigma$ -algebra  $\sigma(\mathcal{I}_2)$  is denoted by  $\mathcal{B}_2$ , i. e.,  $\mathcal{B}_2 := \sigma(\mathcal{I}_2)$ , and its elements are called the *Borel sets of  $\mathbb{R}^2$* .

This definition is easily generalized: The *Borel  $\sigma$ -algebra on  $\mathbb{R}^n$*  is defined by  $\mathcal{B}_n := \sigma(\mathcal{I}_n)$ ,  $n \in \mathbb{N}$ , where  $\mathcal{I}_n$  is the system of all half-open *cuboids* in  $\mathbb{R}^n$ , whose sides are parallel to the axes. Such a cuboid is a set

$$\begin{aligned} & ] a_1, b_1] \times \dots \times ] a_n, b_n] \\ & = \{(x_1, \dots, x_n) \in \mathbb{R}^n : a_1 < x_1 \leq b_1, \dots, a_n < x_n \leq b_n\}, \end{aligned} \quad (1.20)$$

where  $a_1, \dots, a_n, b_1, \dots, b_n \in \mathbb{R}$ . Just like  $\mathcal{B}_1$ , the  $\sigma$ -algebra  $\mathcal{B}_n$  has several generating systems, one of which is

$$\mathcal{B}_n = \sigma(\{ ] -\infty, b_1] \times \dots \times ] -\infty, b_n] : b_1, \dots, b_n \in \mathbb{R} \}) \quad (1.21)$$

(see Exercise 1-11).

Note that not every subset of  $\mathbb{R}^n$  is a Borel set. In other words,  $\mathcal{B}_n$  is not the power set of  $\mathbb{R}^n$  (see Rem. 1.60). However, for each  $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ , the singleton  $\{x\}$  is a Borel set of  $\mathbb{R}^n$ , i. e.,

$$\{x\} \in \mathcal{B}_n, \quad \forall x \in \mathbb{R}^n \quad (1.22)$$

(see Exercise 1-12).

Furthermore, if  $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$  denotes the *extended set of real numbers*, then

$$\bar{\mathcal{B}} := \sigma(\mathcal{B} \cup \{-\infty, +\infty\})$$

is a  $\sigma$ -algebra on  $\bar{\mathbb{R}}$  and it is called the *Borel  $\sigma$ -algebra on  $\bar{\mathbb{R}}$* . Similarly,  $\bar{\mathcal{B}}_n$  is called the *Borel  $\sigma$ -algebra on  $\bar{\mathbb{R}}^n$* . It is defined as the product of the  $\sigma$ -algebra  $\mathcal{B}$  with itself ( $n$  times) (see Def. 1.31). Finally, we may sometimes consider  $\bar{\mathcal{B}}_n|_{\Omega_0}$ , the *trace of the Borel  $\sigma$ -algebra on  $\bar{\mathbb{R}}^n$  in  $\Omega_0 \subset \bar{\mathbb{R}}^n$* .

**Remark 1.28 (The Borel  $\sigma$ -Algebra is Countably Generated)** Note that

$$\mathcal{B} = \sigma(\{ ] a, b] : a, b \in \mathbb{Q}, a < b \}),$$

where  $\mathbb{Q}$  is the set of rational numbers. Because  $\mathbb{Q}$  is countable, the set of intervals  $\{ ] a, b] : a, b \in \mathbb{Q}, a < b \}$  is countable as well. Therefore, the Borel  $\sigma$ -algebra  $\mathcal{B}$  is countably generated. This also holds for  $\mathcal{B}_n$ ,  $n \in \mathbb{N}$  (see Klenke, 2013, Th. 1.23, p. 9).  $\triangleleft$

**Remark 1.29 (Trace of the Borel  $\sigma$ -Algebra in a Countable Subset of  $\mathbb{R}$ )** Let  $\mathcal{B}$  denote the Borel  $\sigma$ -algebra on  $\mathbb{R}$ . If  $\Omega_0 \subset \mathbb{R}$  is finite or countable, then  $\mathcal{B}|_{\Omega_0} = \mathcal{P}(\Omega_0)$ , where  $\mathcal{B}|_{\Omega_0}$  is the trace of the Borel  $\sigma$ -algebra on  $\mathbb{R}$  in  $\Omega_0 \subset \mathbb{R}$  (see Exercise 1-13).  $\triangleleft$

### 1.2.3 $\sigma$ -Algebra on a Cartesian Product

In section 1.2.2 we defined a  $\sigma$ -algebra on  $\mathbb{R}^n = \mathbb{R} \times \dots \times \mathbb{R}$  ( $n$ -times). Now we consider  $\sigma$ -algebras on general Cartesian products. We start with an example.

**Example 1.30 (Joe and Ann – continued)** In Example 1.9 we already considered the Cartesian product

$$\Omega := \Omega_U \times \Omega_X \times \Omega_Y,$$

which consists of the eight triples  $(Joe, no, -), (Joe, no, +), \dots, (Ann, yes, +)$  (see again Fig. 1.2). Now consider the  $\sigma$ -algebras  $\mathcal{A}_1 := \mathcal{P}(\Omega_U)$ ,  $\mathcal{A}_2 := \mathcal{P}(\Omega_X)$ , and  $\mathcal{A}_3 := \mathcal{P}(\Omega_Y)$ , as well as the set

$$\mathcal{E} := \{A_1 \times A_2 \times A_3 : A_1 \in \mathcal{A}_1, A_2 \in \mathcal{A}_2, A_3 \in \mathcal{A}_3\},$$

which is a set system on  $\Omega$  consisting of  $4 \cdot 4 \cdot 4 = 64$  elements. For example, the set system  $\mathcal{E}$  contains the elements

$$A := \{Joe\} \times \{no\} \times \{-\} = \{(Joe, no, -)\}$$

and

$$B := \{Ann\} \times \{yes\} \times \{+\} = \{(Ann, yes, +)\}.$$

However,  $\mathcal{E}$  does not contain

$$A \cup B = \{(Joe, no, -), (Ann, yes, +)\}$$

as an element. The only product set  $A_1 \times A_2 \times A_3$  with  $A_1 \in \mathcal{A}_1, A_2 \in \mathcal{A}_2, A_3 \in \mathcal{A}_3$  that contains  $A \cup B$  as a subset is  $\Omega_U \times \Omega_X \times \Omega_Y = \Omega$ . However,  $A \cup B \neq \Omega$ . Therefore,  $\mathcal{E}$  is not a  $\sigma$ -algebra [cf. condition (c) of Rem. 1.2]. In this example, the  $\sigma$ -algebra generated by  $\mathcal{E}$  is the power set of  $\Omega$ , i. e.,  $\sigma(\mathcal{E}) = \mathcal{P}(\Omega)$ . It consists of  $2^8 = 256$  elements. According to the following definition,  $\sigma(\mathcal{E})$  is denoted by  $\mathcal{A}_1 \otimes \mathcal{A}_2 \otimes \mathcal{A}_3$  and called the *product  $\sigma$ -algebra of  $\mathcal{A}_1, \mathcal{A}_2,$  and  $\mathcal{A}_3$* .  $\triangleleft$

**Definition 1.31 (Product  $\sigma$ -Algebra)**

Let  $(\Omega_1, \mathcal{A}_1), \dots, (\Omega_n, \mathcal{A}_n)$  be measurable spaces and  $\Omega := \Omega_1 \times \dots \times \Omega_n$ . Then

$$\mathcal{A}_1 \otimes \dots \otimes \mathcal{A}_n := \bigotimes_{i=1}^n \mathcal{A}_i := \sigma\left(\left\{\prod_{i=1}^n A_i : A_i \in \mathcal{A}_i, i = 1, \dots, n\right\}\right) \quad (1.23)$$

is called the *product  $\sigma$ -algebra of the  $\sigma$ -algebras  $\mathcal{A}_i, i = 1, \dots, n$* .

To emphasize, the product  $\sigma$ -algebra  $\mathcal{A}_1 \otimes \dots \otimes \mathcal{A}_n$  is *not* the Cartesian product  $\mathcal{A}_1 \times \dots \times \mathcal{A}_n$ . Instead, the product  $\sigma$ -algebra is generated by the set system of all Cartesian products of elements of the  $\sigma$ -algebras  $\mathcal{A}_1, \dots, \mathcal{A}_n$ . In Lemma 2.42 we give an equivalent specification of a product  $\sigma$ -algebra, using projection mappings.

The following lemma provides a relationship between the generating systems of the  $\sigma$ -algebras  $\mathcal{A}_i, i = 1, \dots, n$ , and the generating system of the product  $\sigma$ -algebra.

**Lemma 1.32 (Generating System of a Product  $\sigma$ -Algebra)**

For  $i = 1, \dots, n$ , let  $(\Omega_i, \mathcal{A}_i)$  be measurable spaces and  $\mathcal{E}_i \subset \mathcal{A}_i$  with  $\sigma(\mathcal{E}_i) = \mathcal{A}_i$ .

Then

$$\bigotimes_{i=1}^n \mathcal{A}_i = \sigma\left(\left\{\prod_{i=1}^n A_i : A_i \in \mathcal{E}_i, i = 1, \dots, n\right\}\right). \quad (1.24)$$

For a proof see Klenke (2013, Th. 14.12 (i), p. 276).

This lemma implies

$$\mathcal{B}_n = \bigotimes_{i=1}^n \mathcal{B} = \mathcal{B} \otimes \dots \otimes \mathcal{B} \quad (n\text{-times})$$

for the Borel  $\sigma$ -algebra on  $\mathbb{R}^n$ . This lemma also implies the following corollary.

**Corollary 1.33 (Countable Generating System of a Product  $\sigma$ -Algebra)**

Let  $(\Omega_i, \mathcal{A}_i)$ ,  $i = 1, \dots, n$ , be measurable spaces, where all  $\mathcal{A}_i$  are countably generated. Then  $\bigotimes_{i=1}^n \mathcal{A}_i$  is countably generated as well.

**Example 1.34 (Countable Sets and Product  $\sigma$ -Algebra)** Let  $\Omega_1, \dots, \Omega_n$  be finite or countable sets and  $\mathcal{A}_1, \dots, \mathcal{A}_n$  be their power sets. Then

$$\bigotimes_{i=1}^n \mathcal{A}_i = \mathcal{P}\left(\prod_{i=1}^n \Omega_i\right),$$

i. e.,  $\bigotimes_{i=1}^n \mathcal{A}_i$  is the power set on  $\Omega := \Omega_1 \times \dots \times \Omega_n$  (see Exercise 1-14).  $\triangleleft$

**Remark 1.35 (Complement of a Cartesian Product)** Let  $(\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2)$  be a measurable space,  $A \in \mathcal{A}_1$ , and  $B \in \mathcal{A}_2$ . Then  $(A \times B)^c \in \mathcal{A}_1 \otimes \mathcal{A}_2$ , and this set can be written

$$(A \times B)^c = (A^c \times B) \cup (\Omega_1 \times B^c), \quad (1.25)$$

which is a union of disjoint sets (see Exercise 1-15).  $\triangleleft$

**1.2.4  $\cap$ -Stable Set Systems That Generate a  $\sigma$ -Algebra**

For many proofs, generating set systems are useful, which have the property of  $\cap$ -stability.

**Definition 1.36 ( $\cap$ -Stability)**

Let  $\Omega$  denote a nonempty set. A set  $\mathcal{E}$  of subsets of  $\Omega$  is called  $\cap$ -stable (or  $\cap$ -closed) if  $A \cap B \in \mathcal{E}$  for all  $A, B \in \mathcal{E}$ .

**Example 1.37 (Set System With One Single Element)** A set system  $\{A\}$  that has only a single element  $A \subset \Omega$  is  $\cap$ -stable (cf. Example 1.17).  $\triangleleft$

**Example 1.38 (Partition and  $\cap$ -Stability)** If  $\mathcal{E}$  is a partition of the set  $\Omega$ , then  $\mathcal{D} := \mathcal{E} \cup \{\emptyset\}$  is  $\cap$ -stable.  $\triangleleft$

**Example 1.39 (A  $\cap$ -Stable Generating System of a Product  $\sigma$ -Algebra)** Consider the measurable spaces  $(\Omega_i, \mathcal{A}_i)$ ,  $i = 1, \dots, n$ . The set

$$\{A_1 \times \dots \times A_n : A_i \in \mathcal{A}_i, i = 1, \dots, n\},$$

is a  $\cap$ -stable generating system of  $\bigotimes_{i=1}^n \mathcal{A}_i$  (see Exercise 1-16).  $\triangleleft$

Another type of a set system is a Dynkin system. It can be used in order to show that a specific set system is a  $\sigma$ -algebra.

**Definition 1.40 (Dynkin System)**

A set  $\mathcal{D}$  of subsets of a set  $\Omega$  is called a Dynkin system on  $\Omega$ , if the following three conditions hold:

- (a)  $\Omega \in \mathcal{D}$ .
- (b) If  $A \in \mathcal{D}$ , then  $A^c \in \mathcal{D}$ .
- (c) If  $A_1, A_2, \dots \in \mathcal{D}$  and they are pairwise disjoint, then  $\bigcup_{i=1}^{\infty} A_i \in \mathcal{D}$ .

In the definition of a  $\sigma$ -algebra  $\mathcal{A}$  we require  $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$  for all sequences  $A_1, A_2, \dots \in \mathcal{A}$ , whereas for a Dynkin system the corresponding requirement is only made for all sequences  $A_1, A_2, \dots \in \mathcal{D}$  of pairwise disjoint sets. Analogously to Definition 1.13, for a set system  $\mathcal{E}$  on  $\Omega$ ,  $\delta(\mathcal{E})$  is defined as the Dynkin system generated by  $\mathcal{E}$ , i. e., as the intersection of all Dynkin systems containing  $\mathcal{E}$ . According to the following theorem a Dynkin system is also a  $\sigma$ -algebra if and only if it is  $\cap$ -stable.

**Theorem 1.41 (Dynkin System and  $\sigma$ -Algebra)**

Let  $\Omega$  be a nonempty set.

- (i) A Dynkin system  $\mathcal{D}$  on  $\Omega$  is a  $\sigma$ -algebra if and only if it is  $\cap$ -stable.
- (ii) If  $\mathcal{E}$  is a  $\cap$ -stable set of subsets of  $\Omega$ , then  $\delta(\mathcal{E}) = \sigma(\mathcal{E})$ .

For a proof see Bauer (2001, Ths. 2.3 and 2.4, p. 6, 7). According to proposition (i) of this theorem we can prove that a set system is a  $\sigma$ -algebra by showing that it is a  $\cap$ -stable Dynkin system, and proposition (ii) can be applied to show that the Dynkin system generated by a  $\cap$ -stable set system is a  $\sigma$ -algebra.

### 1.3 Measure and Measure Space

A measure assigns to all elements of a  $\sigma$ -algebra an element of the closed interval

$$[0, \infty] := \{x \in \mathbb{R} : 0 \leq x\} \cup \{\infty\},$$

i. e., a nonnegative real number or the element  $\infty$ .

**Example 1.42 (A First Example)** Let  $\Omega = \mathbb{R}$  and assume that the closed interval  $[3, 9] = \{x \in \mathbb{R} : 3 \leq x \leq 9\}$  as well as the union  $[3, 9] \cup [10, 12]$  are elements of a  $\sigma$ -algebra on  $\Omega$ . If the measure is *length*, then

$$\text{length}([3, 9]) = 9 - 3 = 6$$

and

$$\begin{aligned} \text{length}([3, 9] \cup [10, 12]) &= \text{length}([3, 9]) + \text{length}([10, 12]) \\ &= (9 - 3) + (12 - 10) = 6 + 2 = 8, \end{aligned}$$

because the two intervals are disjoint, i. e., their intersection is the empty set  $\emptyset$ . In this case the lengths of the intervals  $[3, 9]$  and  $[10, 12]$  add up to the length of their union  $[3, 9] \cup [10, 12]$ . In Definition 1.43 (c) we do not only require additivity but  $\sigma$ -additivity.  $\triangleleft$

Reading the following definition, remember that  $\sum_{i=1}^{\infty} a_i$  is defined by

$$\sum_{i=1}^{\infty} a_i := \lim_{n \rightarrow \infty} \sum_{i=1}^n a_i.$$

#### Definition 1.43 (Measure and Measure Space)

Let  $(\Omega, \mathcal{A})$  be a measurable space. A function  $\mu: \mathcal{A} \rightarrow \bar{\mathbb{R}}$  is called a *measure* and the triple  $(\Omega, \mathcal{A}, \mu)$  is called a *measure space*, if

- (a)  $\mu(\emptyset) = 0$ .
- (b)  $\mu(A) \geq 0, \forall A \in \mathcal{A}$ . (*nonnegativity*)
- (c) If  $A_1, A_2, \dots \in \mathcal{A}$  are pairwise disjoint, then  $\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$ . ( *$\sigma$ -additivity*)

### 1.3.1 $\sigma$ -Additivity and Related Properties

**Remark 1.44 ( $\sigma$ -Additivity Implies Finite Additivity)** Note that  $\sigma$ -additivity of a measure implies finite additivity, i. e., it implies

$$\mu\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n \mu(A_i), \quad \text{if } A_1, \dots, A_n \in \mathcal{A} \text{ are pairwise disjoint} \quad (1.26)$$

[see Rule (ii) of Box 1.1 and its proof in Exercise 1-18].  $\triangleleft$

**Remark 1.45 ( $\sigma$ -Additivity)** Using the term  *$\sigma$ -additivity* signalsizes that unions of finitely or countably many sets are considered, but not other unions of sets. If, instead of  $\sigma$ -additivity, we would require additivity for *any kind of unions*, including uncountable unions, then the Lebesgue measure  $\lambda$  on  $(\mathbb{R}, \mathcal{B})$  — the measure of *length* — could not be constructed any more. This is explained in more detail in Remark 1.71.  $\triangleleft$

**Remark 1.46 (Representation of a Union as a Union of Pairwise Disjoint Sets)**

Let  $(\Omega, \mathcal{A})$  be a measurable space. If  $A_1, A_2, \dots \in \mathcal{A}$  is a sequence of subsets of  $\Omega$ , then there is a sequence  $B_1, B_2, \dots \in \mathcal{A}$  of pairwise disjoint sets with

$$\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i. \quad (1.27)$$

One way to construct  $B_1, B_2, \dots$  is to define  $B_1 := A_1$  and

$$B_i := A_i \setminus \left( \bigcup_{j=1}^{i-1} A_j \right), \quad \text{for } i > 1, \quad (1.28)$$

(see Exercise 1-17). ◁

**Remark 1.47 (Additivity of Measures for Partitions)** Let  $(\Omega, \mathcal{A}, \mu)$  be a measure space,  $B \in \mathcal{A}$ , and assume

- (a)  $A_1, \dots, A_n \in \mathcal{A}$  are pairwise disjoint,
- (b)  $B \subset \bigcup_{i=1}^n A_i$ .

Then

$$\mu(B) = \sum_{i=1}^n \mu(B \cap A_i). \quad (1.29)$$

Analogously, if

- (c)  $A_1, A_2, \dots \in \mathcal{A}$  are pairwise disjoint,
- (d)  $B \subset \bigcup_{i=1}^{\infty} A_i$ ,

then

$$\mu(B) = \sum_{i=1}^{\infty} \mu(B \cap A_i). \quad (1.30)$$

(see Exercise 1-19). ◁

**1.3.2 Other Properties**

Other important properties of a measure are displayed in Box 1.1. Some of these properties can intuitively be understood inspecting the Venn diagram presented in Figure 1.1. These properties always hold with the conventions  $+\infty + \infty = +\infty$  and  $\alpha + \infty = +\infty$ , for  $\alpha \in \mathbb{R}$ . However, note that the term  $+\infty - \infty$  cannot meaningfully be defined. Therefore, properties (vi) and (vii) only hold if we assume  $\mu(A \cap B) < \infty$ . For proofs of all these properties see Exercise 1-18.

**Remark 1.48 (Finite Additivity and  $\sigma$ -Additivity Applied to Singletons)** If  $\Omega$  is finite or countable, then each  $A \subset \Omega$  is finite or countable as well. Hence, for any measure  $\mu$  on the measurable space  $(\Omega, \mathcal{P}(\Omega))$ ,

$$\mu(A) = \mu\left(\bigcup_{\omega \in A} \{\omega\}\right) = \sum_{\omega \in A} \mu(\{\omega\}), \quad \forall A \subset \Omega. \quad (1.31)$$

**Box 1.1 Rules of Computation for Measures**

Let  $(\Omega, \mathcal{A}, \mu)$  be a measure space.

If  $A_1, A_2, \dots \in \mathcal{A}$  are pairwise disjoint, then

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i). \quad (\sigma\text{-additivity}) \quad (\text{i})$$

$$\mu\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n \mu(A_i), \quad \forall n \in \mathbb{N}. \quad (\text{finite additivity}) \quad (\text{ii})$$

If  $A, B \in \mathcal{A}$ , then:

$$\mu(A) = \mu(A \cap B) + \mu(A \setminus B). \quad (\text{iii})$$

$$\mu(\Omega) = \mu(B) + \mu(B^c). \quad (\text{iv})$$

$$\mu(A) \leq \mu(B), \quad \text{if } A \subset B. \quad (\text{monotonicity}) \quad (\text{v})$$

$$\mu(A \setminus B) = \mu(A) - \mu(A \cap B), \quad \text{if } \mu(A \cap B) < \infty. \quad (\text{vi})$$

$$\mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B), \quad \text{if } \mu(A \cap B) < \infty. \quad (\text{vii})$$

$$\mu(A) = \mu(\Omega) < \infty \Rightarrow \mu(A \cap B) = \mu(B). \quad (\text{viii})$$

$$\mu(A) = 0 \Rightarrow \mu(A \cup B) = \mu(B). \quad (\text{ix})$$

Let  $A \in \mathcal{A}$  and let  $\Omega_0 \subset \Omega$  and be finite or countable with  $\mu(\Omega \setminus \Omega_0) = 0$ .

If, for all  $\omega \in \Omega_0$ ,  $\{\omega\} \in \mathcal{A}$ , then

$$\mu(A) = \sum_{\omega \in A \cap \Omega_0} \mu(\{\omega\}). \quad (\text{x})$$

If  $A_1, A_2, \dots \in \mathcal{A}$ , then

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \mu(A_i). \quad (\sigma\text{-subadditivity}) \quad (\text{xi})$$

This means that a measure on  $(\Omega, \mathcal{P}(\Omega))$  is already uniquely defined if its values  $\mu(\{\omega\})$  are uniquely defined for all  $\omega \in \Omega$ , provided that  $\Omega$  is finite or countable. Rule (x) of Box 1.1 extends this result to a more general measure space  $(\Omega, \mathcal{A}, \mu)$ . This rule shows that a measure on  $(\Omega, \mathcal{A})$  is already uniquely defined if its values  $\mu(\{\omega\})$  are uniquely defined for all  $\omega \in \Omega_0$ , provided that  $\Omega_0$  is finite or countable with  $\mu(\Omega \setminus \Omega_0) = 0$  and  $\{\omega\} \in \mathcal{A}$  for all  $\omega \in \Omega_0$ .  $\triangleleft$

## 1.4 Specific Measures

Now we consider some examples of measures, all of which are used later on in order to introduce still other measures. For some of these examples we use the *indicator* of a set  $A$ .

**Definition 1.49 (Indicator)**

Let  $\Omega$  be a set and  $A \subset \Omega$ . Then the function  $1_A: \Omega \rightarrow \mathbb{R}$  defined by

$$1_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A, \\ 0, & \text{if } \omega \notin A, \end{cases} \quad (1.32)$$

is called the *indicator* of  $A$ .

**Remark 1.50 (Sums and Products of Indicators)** If  $1_A, 1_B: \Omega \rightarrow \mathbb{R}$  are the indicators of two sets  $A, B \subset \Omega$ , then

$$1_A \cdot 1_B = 1_{A \cap B} \quad (1.33)$$

and

$$1_A + 1_B - 1_{A \cap B} = 1_A + 1_B - 1_A \cdot 1_B = 1_{A \cup B}. \quad (1.34)$$

Equation (1.33) immediately implies

$$1_A + 1_B = 1_{A \cup B}, \quad \text{if } A \cap B = \emptyset. \quad (1.35)$$

More generally, if  $A_1, \dots, A_n$  is a finite sequence of pairwise disjoint subsets of  $\Omega$ , then

$$\sum_{i=1}^n 1_{A_i} = 1_{\bigcup_{i=1}^n A_i}, \quad (1.36)$$

i. e., then the sum of the indicators of the sets  $A_1, \dots, A_n$  is the indicator of the union  $\bigcup_{i=1}^n A_i$ . Finally, if  $A_1, A_2, \dots$  is a sequence of pairwise disjoint subsets of  $\Omega$ , then

$$\sum_{i=1}^{\infty} 1_{A_i} = 1_{\bigcup_{i=1}^{\infty} A_i}. \quad (1.37)$$

◁

**Remark 1.51 (Indicators of Products Sets)** Let  $\Omega_1, \Omega_2$  be nonempty sets,  $A \subset \Omega_1$  and  $B \subset \Omega_2$ . Then

$$1_A(\omega_1) \cdot 1_B(\omega_2) = 1_{A \times B}(\omega_1, \omega_2), \quad \forall (\omega_1, \omega_2) \in \Omega_1 \times \Omega_2. \quad (1.38)$$

This equation follows from the definitions of the product set and the indicator. ◁

### 1.4.1 Dirac Measure and Counting Measure

**Example 1.52 (Dirac Measure)** Let  $(\Omega, \mathcal{A})$  be a measurable space, let  $\omega \in \Omega$ , and consider the function  $\delta_\omega: \mathcal{A} \rightarrow \{0, 1\}$  defined by

$$\delta_\omega(A) := 1_A(\omega), \quad \forall A \in \mathcal{A}. \quad (1.39)$$

Then  $\delta_\omega$  is a measure on  $(\Omega, \mathcal{A})$  (see Exercise 1-20).  $\triangleleft$

**Definition 1.53 (Dirac Measure)**

The function  $\delta_\omega$  defined by Equation (1.39) is called the *Dirac measure at (point)  $\omega$* .

**Example 1.54 (Counting Measure)** Let  $(\Omega, \mathcal{A})$  be a measurable space and define the function  $\mu_\#: \mathcal{A} \rightarrow \bar{\mathbb{R}}$  by

$$\mu_\#(A) := \begin{cases} \sum_{\omega \in \Omega} 1_A(\omega), & \text{if } A \text{ is finite,} \\ \infty, & \text{if } A \text{ is infinite,} \end{cases} \quad \forall A \in \mathcal{A}. \quad (1.40)$$

Then  $\mu_\#$  is a measure on  $(\Omega, \mathcal{A})$  (see Exercise 1-21).  $\triangleleft$

**Definition 1.55 (Counting Measure)**

The function  $\mu_\#$  defined by Equation (1.40) is called the *counting measure on  $(\Omega, \mathcal{A})$* .

**Remark 1.56 (Cardinality of a Set)** If  $A$  is finite, then  $\mu_\#(A)$  is called the *cardinality* of  $A$ , i. e.,  $\mu_\#(A)$  simply counts the number of elements  $\omega$  of the set  $A$ . Furthermore, for finite or countable  $\Omega$  and  $A \subset \Omega$ ,

$$\mu_\#(A) = \sum_{\omega \in \Omega} 1_A(\omega) = \sum_{\omega \in \Omega} \delta_\omega(A). \quad (1.41)$$

$\triangleleft$

**Example 1.57 (Sum of Dirac Measures)** Let  $(\Omega, \mathcal{A})$  be a measurable space. If  $B \subset \Omega$  is finite or countable and  $\delta_\omega$  is the Dirac measure on  $(\Omega, \mathcal{A})$  at point  $\omega$ , then  $\sum_{\omega \in B} \delta_\omega: \mathcal{A} \rightarrow [0, \infty]$  defined by

$$\left( \sum_{\omega \in B} \delta_\omega \right)(A) := \sum_{\omega \in B} \delta_\omega(A), \quad \forall A \in \mathcal{A}, \quad (1.42)$$

is a measure on  $(\Omega, \mathcal{A})$  (see Exercise 1-22). Hence, if  $\Omega$  itself is finite or countable, then  $\sum_{\omega \in \Omega} \delta_\omega$  is a measure on  $(\Omega, \mathcal{A})$ , and it is identical to the counting measure defined in Example 1.54, because, for  $A \in \mathcal{A}$ ,

$$\begin{aligned}
\left( \sum_{\omega \in \Omega} \delta_{\omega} \right) (A) &= \sum_{\omega \in \Omega} \delta_{\omega}(A) && [(1.42)] \\
&= \sum_{\omega \in \Omega} 1_A(\omega) && [(1.39)] \\
&= \mu_{\#}(A). && [(1.41)]
\end{aligned} \tag{1.43}$$

&lt;

### 1.4.2 Lebesgue Measure

Consider the *half-open interval*  $]a, b]$ . Then

$$\lambda_1(]a, b]) = b - a \tag{1.44}$$

is the *length* of the interval  $]a, b]$ . Next consider a *rectangle*  $]a_1, b_1] \times ]a_2, b_2]$  in  $\mathbb{R}^2$  with  $a_1 < b_1$  and  $a_2 < b_2$ . This set can be visualized by the set of all points inside the rectangle presented in Figure 1.3 (excluding the lower and left boundary). Obviously,

$$\lambda_2(]a_1, b_1] \times ]a_2, b_2]) = (b_1 - a_1) \cdot (b_2 - a_2) \tag{1.45}$$

is the *area* of this rectangle.

According to the following theorem, there is one and only one measure on  $(\mathbb{R}, \mathcal{B})$  satisfying (1.44) for all such intervals. This measure is called the *Lebesgue measure on  $(\mathbb{R}, \mathcal{B})$*  and is denoted by  $\lambda$  or  $\lambda_1$ . Similarly, there is one and only one measure on  $(\mathbb{R}^2, \mathcal{B}_2)$  satisfying (1.45) for all such rectangles. It is called the *Lebesgue measure on  $(\mathbb{R}^2, \mathcal{B}_2)$*  and is denoted by  $\lambda_2$ . The following theorem deals with the general case.

**Theorem 1.58 (Existence and Uniqueness of the Lebesgue Measure)**

For all  $n \in \mathbb{N}$ , there is a uniquely defined measure  $\lambda_n$  on  $(\mathbb{R}^n, \mathcal{B}_n)$  satisfying

$$\begin{aligned}
\lambda_n(]a_1, b_1] \times \dots \times ]a_n, b_n]) &= \prod_{i=1}^n (b_i - a_i), \\
\forall a_i, b_i \in \mathbb{R} \text{ with } a_i < b_i, \quad i &= 1, \dots, n.
\end{aligned} \tag{1.46}$$

For a proof see Klenke (2013, Th. 1.55, p. 24).

**Definition 1.59 (Lebesgue Measure)**

The measure  $\lambda_n$  satisfying Equation (1.46) is called the *Lebesgue measure on  $(\mathbb{R}^n, \mathcal{B}_n)$* .

**Remark 1.60 (Sets of Real Numbers That are not Lebesgue Measurable)** Hence, the Lebesgue measure  $\lambda_n$  is defined on  $(\mathbb{R}^n, \mathcal{B}_n)$ . Note, however, that this measure

space  $(\mathbb{R}^n, \mathcal{B}_n, \lambda_n)$  can be completed by extending the Borel  $\sigma$ -algebra by including all the subsets of sets  $A \in \mathcal{B}_n$  with  $\lambda_n(A) = 0$ . In Wise and Hall (1993, counterexample 1.25) it is shown for  $n = 1$  that there are subsets  $B \subset \mathbb{R}$  that are not elements of the completed  $\sigma$ -algebra. Therefore,  $B \notin \mathcal{B}$ , and this implies  $\mathcal{B} \neq \mathcal{P}(\mathbb{R})$ .  $\triangleleft$

### 1.4.3 Other Examples of a Measure

**Example 1.61 (Restriction of a Measure to a Sub- $\sigma$ -Algebra)** Suppose  $(\Omega, \mathcal{A}, \mu)$  is a measure space and  $\mathcal{C} \subset \mathcal{A}$  a  $\sigma$ -algebra. Then the function  $\nu: \mathcal{C} \rightarrow \bar{\mathbb{R}}$  defined by

$$\nu(A) := \mu(A), \quad \forall A \in \mathcal{C}, \quad (1.47)$$

is a measure on  $(\Omega, \mathcal{C})$  (see Exercise 1-23).  $\triangleleft$

**Example 1.62 (Weighted Sum of Measures)** If  $\mu_1, \mu_2, \dots$  are measures on  $(\Omega, \mathcal{A})$  and  $0 \leq \alpha_1, \alpha_2, \dots \in \mathbb{R}$ , then  $\sum_{i=1}^{\infty} \alpha_i \mu_i: \mathcal{A} \rightarrow [0, \infty]$  defined by

$$\left( \sum_{i=1}^{\infty} \alpha_i \mu_i \right) (A) := \sum_{i=1}^{\infty} \alpha_i \mu_i(A), \quad \forall A \in \mathcal{A}, \quad (1.48)$$

is again a measure on  $(\Omega, \mathcal{A})$  (see Exercise 1-24). For  $0 = \alpha_{n+1} = \alpha_{n+2} = \dots$  this implies: If  $\mu_1, \dots, \mu_n$  are measures on  $(\Omega, \mathcal{A})$  and  $\alpha_1, \dots, \alpha_n$  are nonnegative, then the function  $\sum_{i=1}^n \alpha_i \mu_i$  defined by

$$\left( \sum_{i=1}^n \alpha_i \mu_i \right) (A) := \sum_{i=1}^n \alpha_i \mu_i(A), \quad \forall A \in \mathcal{A}, \quad (1.49)$$

is also a measure on  $(\Omega, \mathcal{A})$ .  $\triangleleft$

### 1.4.4 Finite and $\sigma$ -Finite Measures

A measure  $\mu$  on a measurable space  $(\Omega, \mathcal{A})$  is called *finite* if  $\mu(\Omega) < \infty$ . Otherwise it is called *infinite*. Within the class of infinite measures there is a subclass with an important property, called  *$\sigma$ -finiteness*. Many fundamental propositions of measure and integration theory only hold for measures that are  $\sigma$ -finite.

#### Definition 1.63 ( $\sigma$ -Finite Measure)

Let  $\mu$  be a measure on a measurable space  $(\Omega, \mathcal{A})$ . Then  $\mu$  is called  *$\sigma$ -finite* if there is a sequence  $A_1, A_2, \dots \in \mathcal{A}$  with  $\bigcup_{i=1}^{\infty} A_i = \Omega$  and, for all  $i = 1, 2, \dots$ ,  $\mu(A_i) < \infty$ .

To emphasize, even if  $\mu(\Omega) = \infty$ , the measure  $\mu$  can be  $\sigma$ -finite (see Examples 1.64 and 1.65). Note that any finite measure is also  $\sigma$ -finite.

**Example 1.64 ( $\sigma$ -Finiteness of the Lebesgue-Measure)** The Lebesgue measure  $\lambda$  on  $(\mathbb{R}, \mathcal{B})$  is  $\sigma$ -finite, because  $\mathbb{R} = \bigcup_{i=1}^{\infty} [-i, i]$  and  $\lambda([-i, i]) = 2 \cdot i < \infty$ , for all  $i \in \mathbb{N}$ .  $\triangleleft$

**Example 1.65 (A  $\sigma$ -Finite Counting Measure)** Consider the measurable space  $(\mathbb{R}, \mathcal{B})$  and the measure  $\mu: \mathcal{B} \rightarrow [0, \infty]$ , where  $\mu = \sum_{i=0}^{\infty} \delta_i$  and  $\delta_i$  denotes the Dirac measure at  $i$  on  $(\mathbb{R}, \mathcal{B})$  with  $\delta_i(A) = 1_A(i)$ ,  $A \in \mathcal{B}$ ,  $i \in \mathbb{N}_0$  (see Example 1.57). Then  $\mu$  is  $\sigma$ -finite because  $\mathbb{R} = \bigcup_{n=1}^{\infty} [-n, n]$  and  $\mu([-n, n]) = n + 1$ , for all  $n \in \mathbb{N}_0$ . This measure simply counts the number of elements  $i \in \mathbb{N}_0$  in a Borel set  $A$ . In other words, for all finite  $A \in \mathcal{B}$ ,  $\mu(A)$  is the cardinality of the set  $A \cap \mathbb{N}_0$ .  $\triangleleft$

### 1.4.5 Product Measure

In section 1.4.2 we considered the Lebesgue measure on  $(\mathbb{R}^n, \mathcal{B}_n)$  that is specified for  $n$ -dimensional cuboids by Equation (1.46) using the *product* of one-dimensional Lebesgue measures on  $(\mathbb{R}, \mathcal{B})$ . Now we introduce the general concept of a product measure. The following lemma shows that  $\sigma$ -finiteness of measures is sufficient for the existence and uniqueness of such a measure. Hence, this lemma shows that presuming finite measures is sufficient but not necessary for the definition of the product measure.

#### Lemma 1.66 (Existence and Uniqueness)

Let  $(\Omega_i, \mathcal{A}_i, \mu_i)$  be measure spaces with  $\sigma$ -finite measures  $\mu_i$ ,  $i = 1, \dots, n$ . Then there is a uniquely defined measure, denoted  $\mu_1 \otimes \dots \otimes \mu_n$ , on the product space

$$\left( \prod_{i=1}^n \Omega_i, \bigotimes_{i=1}^n \mathcal{A}_i \right),$$

satisfying

$$\begin{aligned} \forall (A_1, \dots, A_n) \in \mathcal{A}_1 \times \dots \times \mathcal{A}_n: \\ \mu_1 \otimes \dots \otimes \mu_n(A_1 \times \dots \times A_n) = \mu_1(A_1) \cdot \dots \cdot \mu_n(A_n). \end{aligned} \quad (1.50)$$

This measure is  $\sigma$ -finite as well.

For a proof see Bauer (2001, Th. 23.9, p. 143). Hence,  $\mu := \mu_1 \otimes \dots \otimes \mu_n$  is a measure on the product space  $(\prod_{i=1}^n \Omega_i, \bigotimes_{i=1}^n \mathcal{A}_i)$  with

$$\mu(A_1 \times \dots \times A_n) := \mu_1(A_1) \cdot \dots \cdot \mu_n(A_n), \quad \forall (A_1, \dots, A_n) \in (\mathcal{A}_1 \times \dots \times \mathcal{A}_n). \quad (1.51)$$

#### Definition 1.67 (Product Measure)

The measure  $\mu_1 \otimes \dots \otimes \mu_n$  defined by Equation (1.50) is called the *product measure* of  $\mu_1, \dots, \mu_n$ .

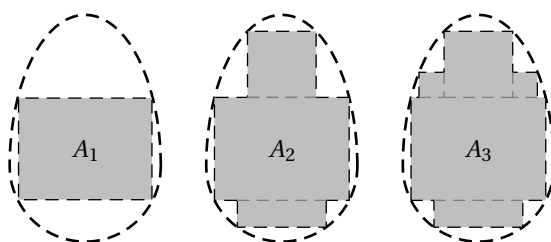


Figure 1.4. Approximation of an open egg-shaped set  $O$  from below

## 1.5 Continuity of a Measure

The term  $\sigma$ -additivity refers to *countable* unions of pairwise disjoint sets and it implies finite additivity, which involves *finite* unions of pairwise disjoint sets. Furthermore,  $\sigma$ -additivity implies the following continuity properties of a measure, which is essential for the definition of the integral (see ch. 3).

### Theorem 1.68 (Continuity of a Measure)

Let  $(\Omega, \mathcal{A}, \mu)$  be a measure space and let  $A_1, A_2, \dots \in \mathcal{A}$ .

(i) If  $A_1 \subset A_2 \subset \dots$ , then

$$\lim_{i \rightarrow \infty} \mu(A_i) = \mu\left(\bigcup_{i=1}^{\infty} A_i\right). \quad (\text{continuity from below})$$

(ii) If  $A_1 \supset A_2 \supset \dots$  and there is an  $n \in \mathbb{N}$  with  $\mu(A_i) < \infty$ , then

$$\lim_{i \rightarrow \infty} \mu(A_i) = \mu\left(\bigcap_{i=1}^{\infty} A_i\right). \quad (\text{continuity from above})$$

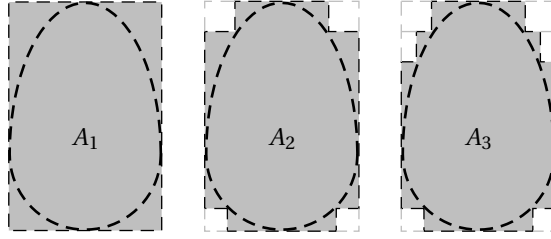
For a proof see Klenke (2013, Theorem 1.36, p. 15, 16).

**Remark 1.69 (Finite Case)** If  $A_1, \dots, A_n \in \mathcal{A}$  is a finite sequence with  $A_1 \subset \dots \subset A_n$ , then  $\bigcup_{i=1}^n A_i = A_n$  and

$$\mu\left(\bigcup_{i=1}^n A_i\right) = \mu(A_n). \quad (1.52)$$

This is a trivial case of Theorem 1.68 (with  $A_n = A_{n+1} = A_{n+2} = \dots$ ).  $\triangleleft$

**Example 1.70 (Geometric Examples)** Figures 1.4 and 1.5 illustrate this theorem for the Lebesgue measure  $\lambda_2$  on  $(\mathbb{R}^2, \mathcal{B}_2)$ , the *area* of a set  $O$  and the sets  $A_i, i \in \mathbb{N}$ . In this example,  $A_1$  is the open rectangle in the open (i.e., the set without its boundary) egg-shaped set  $O$  displayed in Figure 1.4,  $A_2$  the union of  $A_1$  with two



**Figure 1.5.** Approximation of an open egg-shaped set  $O$  from above

other rectangles in the middle figure, and  $A_3$  the union of  $A_2$  with two additional rectangles in the right figure. Adding more and more rectangles it is plausible that  $A_1 \subset A_2 \subset \dots \subset O$  and that their union approximates  $O$ , i.e.,  $\bigcup_{i=1}^{\infty} A_i = O$ . Under these premises Theorem 1.68 (i) yields the conclusion  $\lim_{i \rightarrow \infty} \lambda_2(A_i) = \lambda_2(\bigcup_{i=1}^{\infty} A_i) = \lambda_2(O)$ . Figure 1.5 illustrates the same principle. However, now the area of the egg-shaped set  $O$  is approximated from above by subtracting the areas of appropriate rectangles.

As a second example consider the Lebesgue measure  $\lambda$  on  $(\mathbb{R}, \mathcal{B})$  and the intervals  $A_i = ]x - \frac{1}{i}, x]$ ,  $i \in \mathbb{N}$ . Obviously,  $A_1 \supset A_2 \supset \dots$  and  $\lambda(A_i) = \frac{1}{i} < \infty$ , for all  $i \in \mathbb{N}$  (see also Exercise 1-12). Hence, for all  $x \in \mathbb{R}$ ,

$$\lambda(\{x\}) = \lambda\left(\bigcap_{i=1}^{\infty} ]x - \frac{1}{i}, x]\right) = \lim_{i \rightarrow \infty} \lambda(]x - \frac{1}{i}, x]) = \lim_{i \rightarrow \infty} \frac{1}{i} = 0. \quad (1.53)$$

This is an implication of continuity from above, and it implies

$$\forall a, b \in \mathbb{R}: a < b \Rightarrow \lambda(]a, b]) = \lambda([a, b]) = \lambda([a, b[) = \lambda(]a, b[) = b - a. \quad (1.54)$$

◁

**Remark 1.71 (A Motivation for  $\sigma$ -Additivity)** As already mentioned in Remark 1.45,  $\sigma$ -additivity refers to unions of finitely or countably many sets. Now consider  $\bigcup_{1 \leq x \leq 2} \{x\} = [1, 2] \in \mathcal{B}$  [see Eq. (1.9)]. According to Equation (1.53),  $\lambda(\{x\}) = 0$ , for all  $x \in [1, 2]$ , and hence  $\lambda(\{x \in [1, 2]: x \in \mathbb{Q}\}) = 0$ , because the set of rational numbers is countable. In other words, the Lebesgue measure  $\lambda$  of the set of all rational number in the closed interval  $[1, 2]$  is zero, and this is not a contradiction to

$$\lambda\left(\bigcup_{1 \leq x \leq 2} \{x\}\right) = \lambda([1, 2]) = 2 - 1 = 1,$$

because  $\bigcup_{1 \leq x \leq 2} \{x\}$  is an uncountable union. This illustrates that additivity for uncountable unions can be meaningless. ◁

## 1.6 Specifying a Measure via a Generating System

Given a measurable space  $(\Omega, \mathcal{A})$ , a measure is a function that is defined on  $\mathcal{A}$ . In many situations, e. g., when  $\mathcal{A} = \sigma(\mathcal{E})$  can only be described by a generating set system  $\mathcal{E}$  (such as the set system  $\mathcal{I}_1$  generating the Borel  $\sigma$ -algebra on  $\mathbb{R}$ ), it is important to answer the following questions:

- (a) *Existence*: If there is a set function  $\tilde{\mu}: \mathcal{E} \rightarrow \bar{\mathbb{R}}$ , is there also a measure  $\mu: \sigma(\mathcal{E}) \rightarrow \bar{\mathbb{R}}$  such that  $\mu(A) = \tilde{\mu}(A), \forall A \in \mathcal{E}$ ?
- (b) *Uniqueness*: Is a measure  $\mu$  on  $(\Omega, \sigma(\mathcal{E}))$  already uniquely defined by its values  $\mu(A), A \in \mathcal{E}$ ?

(Sufficient conditions for the existence of such a measure  $\mu$  are formulated in Klenke, 2013, Theorem 1.53, p. 23.)

The following uniqueness theorem for finite measures provides an answer to these questions, which suffices for our purposes. (A more general formulation for  $\sigma$ -finite measures with additional assumptions and a proof of Theorem 1.72 is found in Klenke, 2013, Lemma 1.42, p. 18.)

### Theorem 1.72 (Generating System and Uniqueness of a Measure)

Let  $(\Omega, \mathcal{A})$  be a measurable space and let  $\mathcal{E} \subset \mathcal{A}$ , where  $\mathcal{E}$  is  $\cap$ -stable and  $\sigma(\mathcal{E}) = \mathcal{A}$ . If  $\mu_1$  and  $\mu_2$  are finite measures on  $(\Omega, \mathcal{A})$ , i. e., measures with  $\mu_1(\Omega), \mu_2(\Omega) < \infty$ , then

$$\forall A \in \mathcal{E}: \mu_1(A) = \mu_2(A) \quad \Rightarrow \quad \forall A \in \mathcal{A}: \mu_1(A) = \mu_2(A).$$

**Example 1.73 (Countable  $\Omega$ )** Let  $\Omega$  be a finite or countable set and let  $\mathcal{A} = \mathcal{P}(\Omega)$ . Then the set system

$$\mathcal{E}_1 = \{\emptyset\} \cup \{\{\omega\}: \omega \in \Omega\}$$

is  $\cap$ -stable and  $\sigma(\mathcal{E}_1) = \mathcal{A}$ . As already noted in Remark 1.48, a finite measure  $\mu$  on  $(\Omega, \mathcal{A})$  is uniquely defined by its values  $\mu(\{\omega\}), \omega \in \Omega$ .  $\triangleleft$

**Example 1.74 (Measures on  $(\mathbb{R}, \mathcal{B})$ )** The set system

$$\mathcal{E}_2 = \{]a, b[: a < b, a, b \in \mathbb{R}\} \cup \{\emptyset\}$$

is  $\cap$ -stable and  $\sigma(\mathcal{E}_2) = \mathcal{B}$  [see Eq. (1.18) and section 1.2.4]. Another  $\cap$ -stable set system  $\mathcal{E}_3$  with  $\sigma(\mathcal{E}_3) = \mathcal{B}$  is

$$\mathcal{E}_3 = \{]-\infty, b]: b \in \mathbb{R}\}$$

(cf. Klenke, 2013, p. 9). This set system is crucial for the definition of a cumulative distribution function (see section 5.7.1).  $\triangleleft$

## 1.7 $\sigma$ -Algebra That is Trivial With Respect to a Measure

All  $\sigma$ -algebras treated in section 1.2 have been defined without reference to a measure. Now we define the concept of a *trivial  $\sigma$ -algebra*, which is defined referring to a measure. We start with a lemma about the set of all subsets of a set  $\Omega$  with  $\mu(A) = 0$  or  $\mu(A) = \mu(\Omega)$ , i. e., the set of all sets that are *trivial* with respect to the measure  $\mu$ . Hence, the set of  $\mu$ -trivial sets includes all *null sets*, i. e., all sets  $A \subset \Omega$  with  $\mu(A) = 0$ , and all sets  $A \subset \Omega$  with  $\mu(A) = \mu(\Omega)$ .

### Lemma 1.75 (The Set of all Trivial Sets is a $\sigma$ -Algebra)

Let  $(\Omega, \mathcal{A}, \mu)$  be a measure space and assume that  $\mu$  is finite. Then

$$\mathcal{T}_\mu := \{A \in \mathcal{A} : \mu(A) = 0 \text{ or } \mu(A) = \mu(\Omega)\} \quad (1.55)$$

is a  $\sigma$ -algebra.

(Proof p. 30)

This lemma allows for the following definition:

### Definition 1.76 (Trivial $\sigma$ -Algebra With Respect to a Measure)

Let  $(\Omega, \mathcal{A}, \mu)$  be a measure space, assume that  $\mu$  is finite, and let  $\mathcal{T}_\mu$  be defined by (1.55). Then each  $\sigma$ -algebra  $\mathcal{C} \subset \mathcal{T}_\mu$  is called a  $\mu$ -trivial  $\sigma$ -algebra and its elements  $\mu$ -trivial sets.

Obviously,  $\{\Omega, \emptyset\}$  is a trivial  $\sigma$ -algebra with respect to all measures on  $\{\Omega, \emptyset\}$ . Hence, we can call it a *trivial  $\sigma$ -algebra*.

## 1.8 Proofs

### Proof of Theorem 1.12

(a)

$$\begin{aligned} \forall i \in I: \mathcal{A}_i \text{ is a } \sigma\text{-algebra on } \Omega &\Rightarrow \forall i \in I: \Omega \in \mathcal{A}_i && [\text{Def. 1.1 (a)}] \\ &\Rightarrow \Omega \in \bigcap_{i \in I} \mathcal{A}_i. \end{aligned}$$

(b)

$$\begin{aligned} A \in \bigcap_{i \in I} \mathcal{A}_i &\Rightarrow \forall i \in I: A \in \mathcal{A}_i \\ &\Rightarrow \forall i \in I: A^c \in \mathcal{A}_i && [\text{Def. 1.1 (b)}] \\ &\Rightarrow A^c \in \bigcap_{i \in I} \mathcal{A}_i. \end{aligned}$$

(c)

$$\begin{aligned}
A_1, A_2, \dots \in \bigcap_{i \in I} \mathcal{A}_i &\Rightarrow \forall i \in I: A_1, A_2, \dots \in \mathcal{A}_i \\
&\Rightarrow \forall i \in I: \bigcup_{j=1}^{\infty} A_j \in \mathcal{A}_i \quad [\text{Def. 1.1 (c)}] \\
&\Rightarrow \bigcup_{j=1}^{\infty} A_j \in \bigcap_{i \in I} \mathcal{A}_i.
\end{aligned}$$

### **Proof of Lemma 1.15**

If  $\mathcal{C}$  is a  $\sigma$ -algebra with  $\mathcal{E} \subset \mathcal{C}$  and  $\mathcal{A} = \sigma(\mathcal{E})$ , then (1.11) and the assumption  $\mathcal{C} \subset \mathcal{A}$  imply  $\mathcal{A} = \sigma(\mathcal{E}) \subset \mathcal{C} \subset \mathcal{A}$ . Hence,  $\mathcal{C} = \mathcal{A}$ .

### **Proof of Lemma 1.20**

Define  $\mathcal{D} := \{C = \bigcup_{i \in I(C)} B_i : I(C) \subset \mathbb{N}\}$ .

$\mathcal{E} \subset \mathcal{D}$ : For  $B_j \in \mathcal{E}$  choose  $I(B_j) = \{j\}$ . Then  $B_j = \bigcup_{i \in I(B_j)} B_i$ .

$\mathcal{D} \subset \mathcal{E}$ : Because  $\mathbb{N}$  is countable, any  $I(C) \subset \mathbb{N}$  is finite or countable, and this implies that  $C = \bigcup_{i \in C} B_i$  is an element of  $\sigma(\mathcal{E})$  [see Def. 1.1 (c), (1.3)].

Checking the three conditions defining a  $\sigma$ -algebra (see Def. 1.1), we show that  $\mathcal{D}$  is a  $\sigma$ -algebra.

(a)

$$\Omega = \begin{cases} \bigcup_{i=1}^n B_i, & \text{if } \mathcal{E} = \{B_1, \dots, B_n\} \\ \bigcup_{i=1}^{\infty} B_i, & \text{if } \mathcal{E} = \{B_1, B_2, \dots\}, \end{cases}$$

because  $\mathcal{E}$  is assumed to be a partition. This shows that  $\Omega \in \mathcal{D}$ .

(b) The equation for  $\Omega$  in (a) also implies  $I(C^c) = I(C)^c$ . Therefore,  $C^c \in \mathcal{D}$  if  $C \in \mathcal{D}$ .

(c) If  $C_1, C_2, \dots \in \mathcal{D}$ , then

$$\bigcup_{j=1}^{\infty} C_j = \bigcup_{j=1}^{\infty} \bigcup_{i \in I(C_j)} B_i = \bigcup_{i \in \bigcup_{j=1}^{\infty} I(C_j)} B_i \in \mathcal{D},$$

because  $\bigcup_{j=1}^{\infty} I(C_j) \subset \mathbb{N}$ .

Finally, we prove the second equation in (1.14). If  $j \in I(C)$  and  $C = \bigcup_{i \in I(C)} B_i$ , then  $B_j \subset C$ , which implies

$$\bigcup_{i \in I(C)} B_i \subset \bigcup_{B_i \subset C} B_i.$$

Vice versa, if  $B_j \subset C$ , then  $j \in I(C)$ , because for any  $\omega \in B_j$ , there is no  $i \neq j$  such that  $\omega \in B_i$  [see condition (b) of Rem. 1.19]. Hence,

$$\bigcup_{B_i \subset C} B_i \subset \bigcup_{i \in I(C)} B_i,$$

which proves the second equation in (1.14).

### **Proof of Lemma 1.27**

In this proof we use  $\sigma_\Omega(\mathcal{E})$  to denote the  $\sigma$ -algebra on  $\Omega$  generated by  $\mathcal{E} \subset \mathcal{P}(\Omega)$ . Similarly,  $\sigma_A(\mathcal{D})$  denotes the  $\sigma$ -algebra on  $A$  generated by  $\mathcal{D} \subset \mathcal{P}(A)$ .

(1.15).  $\sigma_\Omega(\mathcal{E})$  is a  $\sigma$ -algebra on  $\Omega$  and  $\mathcal{E} \subset \sigma_\Omega(\mathcal{E})$ , by definition of  $\sigma_\Omega(\mathcal{E})$ . Hence,  $\mathcal{E}|_A \subset \sigma_\Omega(\mathcal{E})|_A$ , and  $\sigma_\Omega(\mathcal{E})|_A$  is a  $\sigma$ -algebra on  $A$  (see Exercise 1-5). Therefore, the definition (1.10) yields

$$\sigma_A(\mathcal{E}|_A) \subset \sigma_\Omega(\mathcal{E})|_A.$$

Furthermore,  $\mathcal{E} \subset \sigma_\Omega(\mathcal{E}|_A \cup \mathcal{E}|_{A^c})$ , which implies

$$\sigma_\Omega(\mathcal{E}) \subset \sigma_\Omega(\mathcal{E}|_A \cup \mathcal{E}|_{A^c}) \quad [\text{Rem. 1.23}]$$

$$\subset \sigma_\Omega(\sigma_A(\mathcal{E}|_A) \cup \sigma_{A^c}(\mathcal{E}|_{A^c})) \quad [\text{Rem. 1.23}]$$

$$= \{C \cup D: C \in \sigma_A(\mathcal{E}|_A), D \in \sigma_{A^c}(\mathcal{E}|_{A^c})\}. \quad [\text{this set system is a } \sigma\text{-algebra}]$$

Therefore,

$$\begin{aligned} \sigma_\Omega(\mathcal{E})|_A &\subset \{C \cup D: C \in \sigma_A(\mathcal{E}|_A), D \in \sigma_{A^c}(\mathcal{E}|_{A^c})\}|_A \\ &= \{(C \cup D) \cap A: C \in \sigma_A(\mathcal{E}|_A), D \in \sigma_{A^c}(\mathcal{E}|_{A^c})\} \\ &= \{C \cap A: C \in \sigma_A(\mathcal{E}|_A)\} \quad [D \subset A^c] \\ &= \sigma_A(\mathcal{E}|_A). \quad [C \subset A] \end{aligned}$$

Hence, we have shown  $\sigma_A(\mathcal{E}|_A) \subset \sigma_\Omega(\mathcal{E})|_A$  and  $\sigma_\Omega(\mathcal{E})|_A \subset \sigma_A(\mathcal{E}|_A)$ , which is equivalent to  $\sigma_A(\mathcal{E}|_A) = \sigma_\Omega(\mathcal{E})|_A$ .

(1.17).

$$\begin{aligned} \sigma_\Omega(\mathcal{C} \cup \mathcal{E})|_A &= \sigma_A(\mathcal{C} \cup \mathcal{E}|_A) \quad [(1.15)] \\ &= \sigma_A(\mathcal{C}|_A \cup \mathcal{E}|_A) \quad [\text{see def. of the trace in Example 1.10}] \\ &= \sigma_A(\mathcal{C}|_A \cup \{\emptyset, A\}) \quad [(1.16)] \\ &= \sigma_A(\mathcal{C}|_A) \quad [\{\emptyset, A\} \subset \mathcal{C}|_A] \\ &= \mathcal{C}|_A. \quad [\text{Exercise 1-5, (1.12)}] \end{aligned}$$

### **Proof of Lemma 1.75**

(a)  $\Omega \in \mathcal{T}_\mu$  by definition of  $\mathcal{T}_\mu$ .

(b) If  $A \in \mathcal{T}_\mu$ , then Rules (iv) and (v) of Box 1.1 and finiteness of  $\mu$  yield

$$\mu(A^c) = \mu(\Omega) - \mu(A) = \begin{cases} \mu(\Omega), & \text{if } \mu(A) = 0, \\ 0, & \text{if } \mu(A) = \mu(\Omega), \end{cases}$$

which implies  $A^c \in \mathcal{T}_\mu$ .

(c) Let  $A_1, A_2, \dots \in \mathcal{A}$ . We consider two cases. *First*, if  $\mu(A_i) = 0$ , for all  $A_i, i \in \mathbb{N}$ , then Rule (xi) of Box 1.1 yields  $\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \mu(A_i) = 0$ , i. e.,  $\bigcup_{i=1}^{\infty} A_i \in \mathcal{T}_\mu$ . *Second*, if there is a  $j \in \mathbb{N}$  such that  $\mu(A_j) = \mu(\Omega)$ , then Rule (v) of Box 1.1 yields

$$\mu(\Omega) = \mu(A_j) \leq \mu\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \mu(\Omega),$$

which implies  $\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \mu(\Omega)$ . Therefore,  $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}_\mu$ .

## 1.9 Exercises

▷ **Exercise 1-1** Let  $\mathcal{A}$  be a  $\sigma$ -algebra of subsets of a nonempty set  $\Omega$  and let  $A_1, A_2, \dots \in \mathcal{A}$ . Show: (a)  $A_1 \cap A_2 \cap \dots \in \mathcal{A}$ , (b)  $A_1 \cap A_2 \in \mathcal{A}$ , and (c)  $A_1 \setminus A_2 \in \mathcal{A}$ .

▷ **Exercise 1-2** Show that the set system  $\mathcal{A} = \{\Omega, \emptyset, A, A^c\}$  is stable (closed) with respect to union of elements of  $\mathcal{A}$ .

▷ **Exercise 1-3** Consider the set  $\Omega = \{\omega_1, \dots, \omega_6\}$  representing the set of all possible outcomes of tossing a dice and the power set  $\mathcal{P}(\Omega)$ , which, in probability theory, represents the set of all possible events (including the ‘impossible’ event  $\emptyset$ ) in this random experiment. Specify the  $\sigma$ -algebra on  $\Omega$  that represents all possible events if we only distinguish between even and uneven number of points.

▷ **Exercise 1-4** Consider the random experiment that has been described in Example 1.9. Aside from the power set of  $\Omega$  we already considered the  $\sigma$ -algebras  $\mathcal{A}_1 = \{\Omega, \emptyset, A, A^c\}$ ,  $\mathcal{A}_2 = \{\Omega, \emptyset, B, B^c\}$ , and  $\mathcal{A}_3 = \{\Omega, \emptyset, C, C^c\}$ . Define another  $\sigma$ -algebra not yet mentioned.

▷ **Exercise 1-5** Prove: If  $\mathcal{A}$  is a  $\sigma$ -algebra on  $\Omega$  and  $\Omega_0 \subset \Omega$ , then  $\mathcal{A}|_{\Omega_0} = \{\Omega_0 \cap A : A \in \mathcal{A}\}$  is a  $\sigma$ -algebra on  $\Omega_0$ .

▷ **Exercise 1-6** Prove the proposition of Remark 1.16.

▷ **Exercise 1-7** Show that  $\sigma(\mathcal{E}) = \mathcal{P}(\Omega)$  if  $\Omega$  is finite or countable and  $\mathcal{E} := \{\{\omega\} : \omega \in \Omega\}$ .

▷ **Exercise 1-8** Prove the proposition of Remark 1.21.

▷ **Exercise 1-9** Let  $\mathcal{E}_1, \mathcal{E}_2$  be set systems on  $\Omega$  with  $\mathcal{E}_1 \subset \mathcal{E}_2$ . Show that  $\sigma(\mathcal{E}_1) \subset \sigma(\mathcal{E}_2)$ .

▷ **Exercise 1-10** Prove propositions (a) and (b) of Example 1.25.

▷ **Exercise 1-11** Prove Equation (1.21).

▷ **Exercise 1-12** Show that  $\{x\} \in \mathcal{B}_n$  for all  $x \in \mathbb{R}^n$ , where  $\mathcal{B}_n$  is the Borel  $\sigma$ -algebra on  $\mathbb{R}^n$ .

▷ **Exercise 1-13** Let  $\mathcal{B}$  be the Borel  $\sigma$ -algebra on  $\mathbb{R}$  and let  $\Omega_0 \subset \mathbb{R}$  be finite or countable. Show that  $\mathcal{B}|_{\Omega_0} = \mathcal{P}(\Omega_0)$ .

▷ **Exercise 1-14** Prove the proposition of Example 1.34.

▷ **Exercise 1-15** Prove the proposition of Remark 1.35.

▷ **Exercise 1-16** Let  $(\Omega_i, \mathcal{A}_i)$ ,  $i = 1, \dots, n$ , be measurable spaces. Show that the set system  $\mathcal{E} := \{A_1 \times \dots \times A_n : A_i \in \mathcal{A}_i, i = 1, \dots, n\}$  is  $\cap$ -stable.

▷ **Exercise 1-17** Prove the proposition of Remark 1.46.

▷ **Exercise 1-18** Prove the rules of Box 1.1.

- ▷ **Exercise 1-19** Prove the propositions of Remark 1.47.
- ▷ **Exercise 1-20** Show that  $\delta_\omega: \mathcal{A} \rightarrow \{0, 1\}$  in Example 1.52 is a measure.
- ▷ **Exercise 1-21** Prove that the function defined by Equation (1.40) is a measure on  $(\Omega, \mathcal{A})$ .
- ▷ **Exercise 1-22** Show that  $\sum_{\omega \in B} \delta_\omega$  in Example 1.57 is a measure.
- ▷ **Exercise 1-23** Show that  $\nu: \mathcal{C} \rightarrow \bar{\mathbf{R}}$  defined in Example 1.61 is a measure on  $(\Omega, \mathcal{C})$ .
- ▷ **Exercise 1-24** Prove that the function  $\sum_{i=1}^{\infty} \alpha_i \mu_i$  defined in Example 1.62 is a measure on  $(\Omega, \mathcal{A})$ .

## Solutions

- ▷ **Solution 1-1** (a) If  $A_1, A_2, \dots \in \mathcal{A}$ , then  $A_1^c, A_2^c, \dots \in \mathcal{A}$  [see Def. 1.1 (b)]. Hence,

$$\bigcap_{i=1}^{\infty} A_i = \left[ \left( \bigcap_{i=1}^{\infty} A_i \right)^c \right]^c = \left[ \bigcup_{i=1}^{\infty} A_i^c \right]^c \quad \begin{array}{l} \text{[de Morgan]} \\ \text{[Def. 1.1 (c), (b)]} \end{array}$$

$\in \mathcal{A}.$

- (b) Let  $A_1, A_2 \in \mathcal{A}$  and choose  $A_3, A_4, \dots$  such that  $\Omega = A_i$ , for all  $i \geq 3$ ,  $i \in \mathbf{N}$ . Then, according to Definition 1.1 (a),

$$A_1 \cap A_2 = A_1 \cap A_2 \cap \Omega = \bigcap_{i=1}^{\infty} A_i \in \mathcal{A}.$$

- (c)  $A_1 \setminus A_2 = A_1 \cap A_2^c \in \mathcal{A}$  [see (b) and Def. 1.1 (b)].

- ▷ **Solution 1-2** The unions  $\Omega \cup A = \Omega$ ,  $\Omega \cup A^c = \Omega$ , and  $\Omega \cup \emptyset = \Omega$  are all elements of  $\mathcal{A}$  and the same is true for  $\emptyset \cup A = A$ ,  $\emptyset \cup A^c = A^c$ , and  $A \cup A^c = \Omega$ . Furthermore,  $B \cup B = B$  for all  $B \in \mathcal{A}$ .

- ▷ **Solution 1-3** The  $\sigma$ -algebra on  $\Omega$  that only distinguishes between even and uneven number of points is  $\mathcal{A}_1 := \{\{\omega_1, \omega_3, \omega_5\}, \{\omega_2, \omega_4, \omega_6\}, \Omega, \emptyset\}$ . This is a sub- $\sigma$ -algebra of  $\mathcal{P}(\Omega)$ . Therefore,  $\mathcal{A}_1$  represents the set of all possible events of a random experiment that is, in a sense, contained in the original random experiment.

- ▷ **Solution 1-4** Consider the set system that contains as elements  $A, A^c, B, B^c, \Omega, \emptyset$ , all unions and all intersections of these sets as well as the unions and intersections of the resulting sets such as  $(A^c \cup B^c) \cap (A \cup B)$  and  $(A^c \cup B^c) \cup (A \cup B)$ . Altogether these are 16 sets. This is  $\sigma(\mathcal{A}_1 \cup \mathcal{A}_2)$ , the  $\sigma$ -algebra generated by  $\mathcal{A}_1 \cup \mathcal{A}_2 = \{A, A^c, B, B^c, \Omega, \emptyset\}$  (see Def. 1.13 and Rem. 1.21).

- ▷ **Solution 1-5** (a)  $\Omega_0 \cap \Omega = \Omega_0$ . This implies  $\Omega_0 \in \mathcal{A}|_{\Omega_0}$ .
- (b)

$$A^* \in \mathcal{A}|_{\Omega_0} \Rightarrow \exists A \in \mathcal{A}: A^* = \Omega_0 \cap A.$$

With this set  $A$  and using  $B^c$  for the complement of a set  $B$  with respect to  $\Omega$ ,

$$\Omega_0 \setminus A^* = \Omega_0 \setminus (\Omega_0 \cap A)$$

$$\begin{aligned}
&= \Omega_0 \cap (\Omega_0 \cap A)^c \\
&= \Omega_0 \cap (\Omega_0^c \cup A^c) \\
&= (\Omega_0 \cap \Omega_0^c) \cup (\Omega_0 \cap A^c) \\
&= \Omega_0 \cap A^c \in \mathcal{A}|_{\Omega_0}.
\end{aligned}$$

(c)

$$A_1^*, A_2^*, \dots \in \mathcal{A}|_{\Omega_0} \Rightarrow \exists A_1, A_2, \dots \in \mathcal{A}: A_i^* = \Omega_0 \cap A_i, i \in \mathbf{N}.$$

Hence,

$$A_1^* \cup A_2^* \cup \dots = (\Omega_0 \cap A_1) \cup (\Omega_0 \cap A_2) \cup \dots = \Omega_0 \cap (A_1 \cup A_2 \cup \dots) \in \mathcal{A}|_{\Omega_0}.$$

▷ **Solution 1-6** If  $\mathcal{G}$  is a  $\sigma$ -algebra on  $\Omega$ , then

$$\mathcal{E} \cup \mathcal{F} \subset \mathcal{G} \Leftrightarrow \sigma(\mathcal{E} \cup \mathcal{F}) \subset \mathcal{G}. \quad [(1.11)] \quad (1.56)$$

Furthermore, for three sets  $A, B, C$ ,

$$A \cup B \subset C \Leftrightarrow A \subset C \wedge B \subset C. \quad (1.57)$$

Hence,

$$\mathcal{D} \cup \mathcal{E} \cup \mathcal{F} \subset \mathcal{G} \Leftrightarrow (\mathcal{D} \subset \mathcal{G}) \wedge (\mathcal{E} \cup \mathcal{F} \subset \mathcal{G}) \quad [(1.57)]$$

$$\Leftrightarrow (\mathcal{D} \subset \mathcal{G}) \wedge (\sigma(\mathcal{E} \cup \mathcal{F}) \subset \mathcal{G}) \quad [(1.56)]$$

$$\Leftrightarrow \mathcal{D} \cup \sigma(\mathcal{E} \cup \mathcal{F}) \subset \mathcal{G}. \quad [(1.57)]$$

Now Definition 1.13 yields the proposition.

▷ **Solution 1-7** If  $\Omega$  is finite or countable, then each of its subsets  $A$  is finite or countable as well. Therefore,

$$\forall A \subset \Omega: A = \bigcup_{\omega \in A} \{\omega\} \in \sigma(\mathcal{E}). \quad [\text{Def. 1.1 (c), Rem. 1.2}]$$

Because each element  $A$  of  $\mathcal{P}(\Omega)$  is a union  $\bigcup_{\omega \in A} \{\omega\}$  of singletons  $\{\omega\}$ ,  $\omega \in A$ , this implies  $\mathcal{P}(\Omega) \subset \sigma(\mathcal{E})$ . Hence,  $\mathcal{E} \subset \mathcal{P}(\Omega) \subset \sigma(\mathcal{E})$ . Therefore, Lemma 1.15 implies  $\sigma(\mathcal{E}) = \mathcal{P}(\Omega)$ .▷ **Solution 1-8** Suppose that  $\mathcal{E} = \{A_1, \dots, A_m\}$  and  $A_j^1 := A_j$  and let  $A_j^c$  denote the complement of  $A_j$ . Then, for all  $(k_1, \dots, k_m) \in \{1, c\}^m$  define

$$B_{(k_1, \dots, k_m)} := \bigcap_{j=1}^m A_j^{k_j}.$$

Then

$$\mathcal{F} := \{B_{(k_1, \dots, k_m)} : (k_1, \dots, k_m) \in \{1, c\}^m, B_{(k_1, \dots, k_m)} \neq \emptyset\}$$

is a finite partition of  $\Omega$ . Note that  $\mathcal{F}$  contains all nonempty intersections of sets  $A_j$  or their complements, respectively, where  $j = 1, \dots, m$ . Now Lemma 1.20 implies the proposition.▷ **Solution 1-9** If  $\mathcal{E}_1 \subset \mathcal{E}_2 \subset \mathcal{P}(\Omega)$ , then for any  $\sigma$ -algebra  $\mathcal{A}$  on  $\Omega$  with  $\mathcal{E}_2 \subset \mathcal{A}$  also  $\mathcal{E}_1 \subset \mathcal{A}$ . Remember, if  $J \subset I$ , then  $\bigcap_{i \in J} B_i \subset \bigcap_{i \in I} B_i$ , for any sets  $B_i, i \in I$ . Therefore,  $\sigma(\mathcal{E}_1)$ , which is the intersection of all  $\sigma$ -algebras containing  $\mathcal{E}_1$ , is a subset of the intersection of all  $\sigma$ -algebras containing  $\mathcal{E}_2$ , which is  $\sigma(\mathcal{E}_2)$ .

▷ **Solution 1-10** (a) If  $\Omega$  is finite, then  $\mathcal{P}(\Omega)$  is a finite set system. Therefore, each  $\sigma$ -algebra  $\mathcal{A}$  on  $\Omega$  is a finite set system. Because  $\mathcal{A} = \sigma(\mathcal{A})$ , this  $\sigma$ -algebra is countably generated.

(b) The set  $\mathbb{N}_0$  is countable and therefore also  $\mathbb{N}_0^n$  for  $n \in \mathbb{N}$ . Example 1.18 then implies that  $\mathcal{P}(\mathbb{N}_0^n)$  is countably generated.

▷ **Solution 1-11** Let  $\mathcal{H}_n = \{ ]-\infty, b_1] \times \dots \times ]-\infty, b_n] : b_1, \dots, b_n \in \mathbb{R} \}$ .

(i) For all  $(b_1, \dots, b_n) \in \mathbb{R}^n$  and all  $m \in \mathbb{N}$  with  $m < b_i, i = 1, \dots, n$ ,

$$B_m := ]-m, b_1] \times \dots \times ]-m, b_n] \in \mathcal{H}_n.$$

According to Definition 1.1 (c) this implies

$$\bigcup_{\substack{m \in \mathbb{N} \\ m < b_i, i=1, \dots, n}} B_m = ]-\infty, b_1] \times \dots \times ]-\infty, b_n] \in \sigma(\mathcal{H}_n).$$

Hence,  $\mathcal{H}_n \subset \sigma(\mathcal{H}_n)$ , which, according to (1.11) and (1.12), implies

$$\sigma(\mathcal{H}_n) \subset \sigma(\mathcal{H}_n) = \mathcal{B}_n.$$

(ii) For all  $a_1, \dots, a_n, b_1, \dots, b_n \in \mathbb{R}$ , with  $a_i < b_i, i = 1, \dots, n$ ,

$$]a_1, b_1] \times \dots \times ]a_n, b_n] = ]-\infty, b_1] \times \dots \times ]-\infty, b_n] \setminus \left( \bigcup_{j=1}^n H_j \right),$$

where  $H_j := ]-\infty, b_1] \times \dots \times ]-\infty, b_{j-1}] \times ]-\infty, a_j] \times ]-\infty, b_{j+1}] \times \dots \times ]-\infty, b_n]$ . Hence, according to Remark 1.2,  $]a_1, b_1] \times \dots \times ]a_n, b_n] \in \sigma(\mathcal{H}_n)$  and  $\mathcal{H}_n \subset \sigma(\mathcal{H}_n)$ , which, according to (1.11) and (1.12), implies

$$\mathcal{B}_n = \sigma(\mathcal{H}_n) \subset \sigma(\mathcal{H}_n).$$

▷ **Solution 1-12** If  $x \in \mathbb{R}$ , then  $\{x\} = \bigcap_{i=1}^{\infty} ]x - 1/i, x]$ . According to Equation (1.18), the intervals  $]x - 1/i, x]$  are elements of the generating set system of  $\mathcal{B}$ , the Borel  $\sigma$ -algebra on  $\mathbb{R}$ . Therefore, their countable intersection is an element of  $\mathcal{B}$ . If  $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ , then

$$\{x\} = \bigcap_{i=1}^{\infty} \left( \bigtimes_{j=1}^n ]x_j - \frac{1}{i}, x_j] \right).$$

According to Equation (1.20), the cuboids  $\bigtimes_{j=1}^n ]x_j - \frac{1}{i}, x_j]$  are elements of the set system  $\mathcal{H}_n$  and  $\sigma(\mathcal{H}_n) = \mathcal{B}_n$ .

▷ **Solution 1-13** Because  $\{x\} \in \mathcal{B}$  for all  $x \in \mathbb{R}$  (see Exercise 1-12) we can conclude:  $\{x\} \in \mathcal{B}|_{\Omega_0}$  for all  $x \in \Omega_0$ . Hence, if  $\Omega_0$  is finite or countable, Example 1.18 implies  $\mathcal{B}|_{\Omega_0} = \mathcal{P}(\Omega_0)$ .

▷ **Solution 1-14** Let  $\Omega_1, \dots, \Omega_n$  be finite or countable sets and let  $\mathcal{A}_1, \dots, \mathcal{A}_n$  be their power sets. Then  $\omega_1 \in \Omega_1, \dots, \omega_n \in \Omega_n$  implies  $\{\omega_1\} \in \mathcal{A}_1, \dots, \{\omega_n\} \in \mathcal{A}_n$ . Therefore,

$$\{(\omega_1, \dots, \omega_n)\} = \{\omega_1\} \times \dots \times \{\omega_n\} \in \left\{ \bigtimes_{i=1}^n A_i : A_i \in \mathcal{A}_i, i \in \{1, \dots, n\} \right\}.$$

Hence,

$$\sigma(\{(\omega_1, \dots, \omega_n)\} : \omega_1 \in \Omega_1, \dots, \omega_n \in \Omega_n) \subset \bigotimes_{i=1}^n \mathcal{A}_i.$$

With  $\Omega_i$  being finite or countable,  $\Omega = \Omega_1 \times \dots \times \Omega_n$  is finite or countable. Therefore,

$$\sigma(\{(\omega_1, \dots, \omega_n) : \omega_1 \in \Omega_1, \dots, \omega_n \in \Omega_n\}) = \mathcal{P}(\Omega)$$

(see Example 1.18). Because  $\bigotimes_{i=1}^n \mathcal{A}_i \subset \mathcal{P}(\Omega)$ , we can conclude

$$\bigotimes_{i=1}^n \mathcal{A}_i = \mathcal{P}(\Omega_1 \times \dots \times \Omega_n) = \mathcal{P}\left(\prod_{i=1}^n \Omega_i\right).$$

▷ **Solution 1-15**

$$\begin{aligned} (A \times B)^c &= \{(\omega_1, \omega_2) \in \Omega_1 \times \Omega_2 : \omega_1 \notin A \text{ or } \omega_2 \notin B\} \\ &= \{(\omega_1, \omega_2) \in \Omega_1 \times \Omega_2 : (\omega_1 \notin A, \omega_2 \in B) \text{ or } \omega_2 \notin B\} \\ &= (A^c \times B) \cup (\Omega_1 \times B^c) \end{aligned}$$

and

$$\begin{aligned} (A^c \times B) \cap (\Omega_1 \times B^c) &= \{(\omega_1, \omega_2) \in \Omega_1 \times \Omega_2 : \omega_1 \notin A, \omega_2 \in B, \omega_2 \notin B\} \\ &= \{(\omega_1, \omega_2) \in \Omega_1 \times \Omega_2 : \omega_1 \notin A, \omega_2 \in B \cap B^c = \emptyset\} \\ &= \emptyset. \end{aligned}$$

▷ **Solution 1-16** Remember that  $(a \in A, b \in B)$  means  $(a \in A \text{ and } b \in B)$  and that  $(a \in A \text{ and } b \in B)$  and  $(b \in B \text{ and } a \in A)$  are equivalent. Let  $A_1, B_1 \in \mathcal{A}_1, \dots, A_n, B_n \in \mathcal{A}_n$ . Then  $A_1 \cap B_1 \in \mathcal{A}_1, \dots, A_n \cap B_n \in \mathcal{A}_n$ . Hence,  $A_1 \times \dots \times A_n \in \mathcal{E}, B_1 \times \dots \times B_n \in \mathcal{E}$  and  $(A_1 \cap B_1) \times \dots \times (A_n \cap B_n) \in \mathcal{E}$ . Furthermore,

$$\begin{aligned} (A_1 \times \dots \times A_n) \cap (B_1 \times \dots \times B_n) &= \{(\omega_1, \dots, \omega_n) : \omega_1 \in A_1, \dots, \omega_n \in A_n, \omega_1 \in B_1, \dots, \omega_n \in B_n\} \\ &= \{(\omega_1, \dots, \omega_n) : \omega_1 \in (A_1 \cap B_1), \dots, \omega_n \in (A_n \cap B_n)\} \\ &= (A_1 \cap B_1) \times \dots \times (A_n \cap B_n) \in \mathcal{E}. \end{aligned}$$

▷ **Solution 1-17** Let  $B_i$  denote the sets defined in Remark 1.46.

(i)  $B_1 = A_1 \in \mathcal{A}$ . For all  $i \in \mathbb{N}, i > 1, B_i \in \mathcal{A}$ :

$$B_i = A_i \setminus \left( \bigcup_{j=1}^{i-1} A_j \right) = A_i \cap \left( \bigcup_{j=1}^{i-1} A_j \right)^c \in \mathcal{A}. \quad [\text{Def. 1.1 (b), Rem. 1.2}]$$

(ii) For any sequence  $C_1, C_2, \dots \subset \Omega$ , define

$$\bigcup_{j=m}^n C_j := \emptyset, \quad \text{if } m > n, \quad \text{and} \quad \bigcap_{j=m}^n C_j := \Omega, \quad \text{if } m > n.$$

Then, using associativity and commutativity of the intersection, for  $1 \leq k < l$ ,

$$\begin{aligned} B_k \cap B_l &= \left[ A_k \setminus \left( \bigcup_{j=1}^{k-1} A_j \right) \right] \cap \left[ A_l \setminus \left( \bigcup_{j=1}^{l-1} A_j \right) \right] \\ &= A_k \cap \left( \bigcup_{j=1}^{k-1} A_j \right)^c \cap A_l \cap \left( \bigcup_{j=1}^{l-1} A_j \right)^c && [A \setminus B = A \cap B^c] \\ &= A_k \cap \left( \bigcap_{j=1}^{k-1} A_j^c \right) \cap A_l \cap \left( \bigcap_{j=1}^{l-1} A_j^c \right) && [\text{de Morgan}] \\ &= A_k \cap A_l \cap \left( \bigcap_{j=1}^{k-1} A_j^c \right) \cap \left( \bigcap_{j=1}^{l-1} A_j^c \right) \cap A_k^c \cap \left( \bigcap_{j=k+1}^{l-1} A_j^c \right) \\ &= \emptyset. && [A_k \cap A_k^c = \emptyset] \end{aligned}$$

(iii) The sets  $B_i$  are defined such that  $B_i \subset A_i$ , for all  $i \in I$ . Therefore,  $\bigcup_{i=1}^{\infty} B_i \subset \bigcup_{i=1}^{\infty} A_i$ . Furthermore, for all  $\omega \in \Omega$ ,

$$\begin{aligned} \omega \in \bigcup_{i=1}^{\infty} A_i &\Rightarrow \exists i \in \mathbf{N}: \omega \in A_i \wedge (\forall j < i: \omega \notin A_j) \\ &\Rightarrow \exists i \in \mathbf{N}: \omega \in A_1^c \cap \dots \cap A_{i-1}^c \cap A_i = B_i \\ &\Rightarrow \omega \in \bigcup_{i=1}^{\infty} B_i. \end{aligned}$$

Hence,  $\bigcup_{i=1}^{\infty} A_i \subset \bigcup_{i=1}^{\infty} B_i$ , and this implies  $\bigcup_{i=1}^{\infty} B_i = \bigcup_{i=1}^{\infty} A_i$ .

▷ **Solution 1-18** (i) This is condition (c) of Definition 1.43.

(ii) If  $A_1, \dots, A_n \in \mathcal{A}$  are pairwise disjoint, then  $A_1, A_2, \dots$  with  $\emptyset = A_{n+1} = A_{n+2} = \dots$  is a sequence of pairwise disjoint measurable sets. Therefore, conditions (a) and (c) of Def. 1.43 imply

$$\mu\left(\bigcup_{i=1}^n A_i\right) = \mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i) = \sum_{i=1}^n \mu(A_i) + \sum_{i=n+1}^{\infty} \mu(\emptyset) = \sum_{i=1}^n \mu(A_i).$$

(iii) For  $A, B \subset \Omega$ ,

$$A = (A \cap B) \cup (A \cap B^c) = (A \cap B) \cup (A \setminus B)$$

and

$$(A \cap B) \cap (A \cap B^c) = A \cap B \cap B^c = \emptyset.$$

Hence, for sets  $A, B \in \mathcal{A}$ , Rule (ii) (finite additivity of  $\mu$ ) implies proposition (iii).

(iv) This proposition is a special case of (iii) with  $A = \Omega$ .

(v) Exchanging the roles of  $A$  and  $B$  in (iii) we obtain

$$\mu(B) = \mu(A \cap B) + \mu(B \setminus A).$$

If  $A \subset B$ , then  $A \cap B = A$  and, because  $\mu(B \setminus A) \geq 0$ ,

$$\mu(A) = \mu(A \cap B) \leq \mu(A \cap B) + \mu(B \setminus A) = \mu(B).$$

(vi) This rule immediately follows from proposition (iv) for  $\mu(A \cap B) < \infty$ . [Note that  $\mu(A) - \mu(A \cap B)$  is not defined if  $\mu(A) = \mu(A \cap B) = \infty$ .]

(vii) For  $A, B \subset \Omega$ ,

$$A \cup B = (A \setminus B) \cup (A \cap B) \cup (B \setminus A).$$

Because the right-hand side is a union of pairwise disjoint sets, finite additivity of  $\mu$  yields

$$\begin{aligned} \mu(A \cup B) + \mu(A \cap B) &= \mu(A \setminus B) + \mu(A \cap B) + \mu(B \setminus A) + \mu(A \cap B) \\ &= \mu(A) + \mu(B). \end{aligned} \quad [\text{Box 1.1 (iii)}]$$

(viii)  $\mu(\Omega) = \mu(A \cup A^c) = \mu(A) + \mu(A^c)$ . Hence, if  $\mu(\Omega) = \mu(A) < \infty$ , then  $\mu(A^c) = 0$ . Therefore, for all  $B \in \mathcal{A}$ , (v) implies  $\mu(A^c \cap B) = 0$ . Furthermore,  $B = (A \cap B) \cup (A^c \cap B)$  and  $(A \cap B) \cap (A^c \cap B) = \emptyset$ . Hence,  $\mu(B) = \mu(A \cap B) + \mu(A^c \cap B) = \mu(A \cap B)$ . Note that, in general,  $\mu(A) = \mu(\Omega)$  does not imply  $A = \Omega$ .

(ix)  $\mu(A) = 0$  implies

$$\begin{aligned} \mu(B) &= \mu(A) + \mu(B) \\ &\geq \mu(A \cup B) \quad [(xi)] \\ &\geq \mu(B). \quad [(v)] \end{aligned}$$

Note that, in general,  $\mu(A) = 0$  does not imply  $A = \emptyset$ .

(x) Let  $B := \Omega \setminus \Omega_0$ . Then  $\mu(B) = 0$  as well as  $\mu(A \cap B) = 0$  for all  $A \in \mathcal{A}$  [see Box 1.1 (v)]. Furthermore, for  $A \in \mathcal{A}$ :  $A = (A \cap \Omega_0) \cup (A \cap B)$ , where  $A \cap \Omega_0$  and  $A \cap B$  are disjoint. Now, the sets  $A \cap \Omega_0$ ,  $A \in \mathcal{A}$ , are the elements of the trace  $\sigma$ -algebra and  $(\Omega_0, \mathcal{A}|_{\Omega_0}) = (\Omega_0, \mathcal{P}(\Omega_0))$ . Therefore we can apply Equation (1.31). Hence, for all  $A \in \mathcal{A}$ ,

$$\begin{aligned} \mu(A) &= \mu(A \cap \Omega_0) + \mu(A \cap B) && \text{[Box 1.1 (ii)]} \\ &= \sum_{\omega \in A \cap \Omega_0} \mu(\{\omega\}) + \mu(A \cap B) && \text{[(1.31)]} \\ &= \sum_{\omega \in A \cap \Omega_0} \mu(\{\omega\}). && [\mu(A \cap B) = 0] \end{aligned}$$

(xi) Let  $A_1, A_2, \dots \in \mathcal{A}$  and define  $B_1, B_2, \dots \in \mathcal{A}$  by  $B_1 = A_1$ , and  $B_i = A_i \setminus \bigcup_{j=1}^{i-1} B_j$  for  $i > 1$  (see Rem. 1.46). Then  $B_1, B_2, \dots$  is a sequence of pairwise disjoint sets with  $B_i \subset A_i$  for all  $i \in \mathbb{N}$  and  $\bigcup_{i=1}^{\infty} B_i = \bigcup_{i=1}^{\infty} A_i$ . Hence,

$$\begin{aligned} \mu\left(\bigcup_{i=1}^{\infty} A_i\right) &= \mu\left(\bigcup_{i=1}^{\infty} B_i\right) \\ &= \sum_{i=1}^{\infty} \mu(B_i) && \text{[Def. 1.43 (c)]} \\ &\leq \sum_{i=1}^{\infty} \mu(A_i). && \text{[Box 1.1 (v)]} \end{aligned}$$

▷ **Solution 1-19** If the  $A_1, \dots, A_n \in \mathcal{A}$  are pairwise disjoint and  $B \in \mathcal{A}$ , then, for  $i \neq j$ ,  $i, j = 1, \dots, n$ ,

$$(B \cap A_i) \cap (B \cap A_j) = B \cap (A_i \cap A_j) = B \cap \emptyset = \emptyset.$$

Hence, the sets  $B \cap A_1, \dots, B \cap A_n$  are pairwise disjoint. Furthermore, condition (b) of Remark 1.47 implies

$$\bigcup_{i=1}^n (B \cap A_i) = B \cap \bigcup_{i=1}^n A_i = B.$$

Therefore, additivity of  $\mu$  yields

$$\mu(B) = \mu\left(\bigcup_{i=1}^n (B \cap A_i)\right) = \sum_{i=1}^n \mu(B \cap A_i),$$

which is Equation (1.29). The proof of Equation (1.30) is literally the same except for replacing  $\bigcup_{i=1}^n$  by  $\bigcup_{i=1}^{\infty}$ ,  $\sum_{i=1}^n$  by  $\sum_{i=1}^{\infty}$ , and additivity of  $\mu$  by  $\sigma$ -additivity.

▷ **Solution 1-20** Let  $\omega \in \Omega$ .

(a) According to Equation (1.32),  $\delta_{\omega}(\emptyset) = 1_{\emptyset}(\omega) = 0$ .

(b) According to Equation (1.32),  $\delta_{\omega}(A) = 1_A(\omega) \in \{0, 1\}$ , for all  $A \in \mathcal{A}$ , and this implies  $\delta_{\omega}(A) \geq 0$ , for all  $A \in \mathcal{A}$ .

(c) If  $A_1, A_2, \dots \in \mathcal{A}$  are pairwise disjoint, then

$$\begin{aligned} \delta_{\omega}\left(\bigcup_{i=1}^{\infty} A_i\right) &= 1_{\bigcup_{i=1}^{\infty} A_i}(\omega) && \text{[(1.32)]} \\ &= \sum_{i=1}^{\infty} 1_{A_i}(\omega) && \text{[(1.37)]} \\ &= \sum_{i=1}^{\infty} \delta_{\omega}(A_i). && \text{[(1.32)]} \end{aligned}$$

- ▷ **Solution 1-21** (a) According to Equation (1.40),  $\mu_{\#}(\emptyset) = \sum_{\omega \in \Omega} 1_{\emptyset}(\omega) = 0$ .  
 (b) According to Equation (1.40),  $\mu_{\#}(A) = \sum_{\omega \in \Omega} 1_A(\omega)$ , for all finite  $A \in \mathcal{A}$ , and  $\mu_{\#}(A) = \infty$ , if  $A$  is infinite. This implies  $\mu_{\#}(A) \geq 0$ , for all  $A \in \mathcal{A}$ .  
 (c) If  $A_1, A_2, \dots \in \mathcal{A}$  are pairwise disjoint and all  $A_i$  are finite, then

$$\begin{aligned} \mu_{\#}\left(\bigcup_{i=1}^{\infty} A_i\right) &= \sum_{\omega \in \Omega} 1_{\bigcup_{i=1}^{\infty} A_i}(\omega) \quad [(1.40)] \\ &= \sum_{\omega \in \Omega} \sum_{i=1}^{\infty} 1_{A_i}(\omega) \quad [(1.37)] \\ &= \sum_{i=1}^{\infty} \sum_{\omega \in \Omega} 1_{A_i}(\omega) \\ &= \sum_{i=1}^{\infty} \mu_{\#}(A_i). \quad [(1.40)] \end{aligned}$$

Note that the set  $\bigcup_{i=1}^{\infty} A_i$  can be countably infinite even if all  $A_i$  are finite. In this case  $\mu_{\#}\left(\bigcup_{i=1}^{\infty} A_i\right) = \infty = \sum_{i=1}^{\infty} \mu_{\#}(A_i)$ . If at least one of the  $A_i$  is infinite, then  $\bigcup_{j=1}^{\infty} A_j \supset A_i$  is an infinite set and  $\mu_{\#}\left(\bigcup_{j=1}^{\infty} A_j\right) \geq \mu_{\#}(A_i)$  is infinite as well.

- ▷ **Solution 1-22** (a) Using Equations (1.42) and (1.39),

$$\left(\sum_{\omega \in B} \delta_{\omega}\right)(\emptyset) = \sum_{\omega \in B} \delta_{\omega}(\emptyset) = \sum_{\omega \in B} 1_{\emptyset}(\omega) = \sum_{\omega \in B} 0 = 0.$$

- (b) Using Equations (1.42) and (1.39),

$$\forall A \in \mathcal{A}: \left(\sum_{\omega \in B} \delta_{\omega}\right)(A) = \sum_{\omega \in B} \delta_{\omega}(A) = \sum_{\omega \in B} 1_A(\omega) \geq 0.$$

- (c) If  $A_1, A_2, \dots \in \mathcal{A}$  are pairwise disjoint, then

$$\begin{aligned} \left(\sum_{\omega \in B} \delta_{\omega}\right)\left(\bigcup_{i=1}^{\infty} A_i\right) &= \sum_{\omega \in B} \delta_{\omega}\left(\bigcup_{i=1}^{\infty} A_i\right) \quad [(1.42)] \\ &= \sum_{\omega \in B} 1_{\bigcup_{i=1}^{\infty} A_i}(\omega) \quad [(1.39)] \\ &= \sum_{\omega \in B} \sum_{i=1}^{\infty} 1_{A_i}(\omega) \quad [(1.37)] \\ &= \sum_{\omega \in B} \sum_{i=1}^{\infty} \delta_{\omega}(A_i) \quad [(1.39)] \\ &= \sum_{i=1}^{\infty} \left(\sum_{\omega \in B} \delta_{\omega}\right)(A_i). \quad [(1.42)] \end{aligned}$$

- ▷ **Solution 1-23** (a) Equation (1.47) yields:  $\nu(\emptyset) = \mu(\emptyset) = 0$ .  
 (b) Equation (1.47) also yields:  $\nu(A) = \mu(A) \geq 0$ , for all  $A \in \mathcal{C}$ .  
 (c) If  $A_1, A_2, \dots \in \mathcal{C}$  are pairwise disjoint, then

$$\begin{aligned} \nu\left(\bigcup_{i=1}^{\infty} A_i\right) &= \mu\left(\bigcup_{i=1}^{\infty} A_i\right) \quad [\text{Def. 1.1 (c), (1.47)}] \\ &= \sum_{i=1}^{\infty} \mu(A_i) \quad [\text{Def. 1.43 (c)}] \\ &= \sum_{i=1}^{\infty} \nu(A_i). \quad [(1.47)] \end{aligned}$$

▷ **Solution 1-24** (a) Using Equation (1.48) and Definition Def. 1.43 (a) yields

$$\left( \sum_{i=1}^{\infty} \alpha_i \mu_i \right) (\emptyset) = \sum_{i=1}^{\infty} \alpha_i \mu_i (\emptyset) = \sum_{i=1}^{\infty} 0 = 0.$$

(b) Similarly, using Equation (1.48) yields, for all  $A \in \mathcal{A}$ ,

$$\left( \sum_{i=1}^{\infty} \alpha_i \mu_i \right) (A) = \sum_{i=1}^{\infty} \alpha_i \mu_i (A) = \lim_{n \rightarrow \infty} \sum_{i=1}^n \alpha_i \mu_i (A) \geq 0,$$

because  $\mu_i(A) \geq 0$  and we assume  $\alpha_i \geq 0$ .

(c) If  $A_1, A_2, \dots \in \mathcal{A}$  are pairwise disjoint, then

$$\begin{aligned} \left( \sum_{i=1}^{\infty} \alpha_i \mu_i \right) \left( \bigcup_{j=1}^{\infty} A_j \right) &= \sum_{i=1}^{\infty} \alpha_i \mu_i \left( \bigcup_{j=1}^{\infty} A_j \right) && [(1.48)] \\ &= \sum_{i=1}^{\infty} \alpha_i \sum_{j=1}^{\infty} \mu_i (A_j) && [\text{Def. 1.43 (c)}] \\ &= \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} \alpha_i \mu_i (A_j) \\ &= \sum_{j=1}^{\infty} \left( \sum_{i=1}^{\infty} \alpha_i \mu_i \right) (A_j). && [(1.48)] \end{aligned}$$

Note that the last but one equation holds, because rearranging summands does not change the sum if the terms  $\alpha_i$  and  $\mu_i(A_j)$  are nonnegative.



## Chapter 2

# Measurable Mapping

In chapter 1 we treated the concepts of a  $\sigma$ -algebra and a  $\sigma$ -algebra generated by a set system on a set  $\Omega$ . An element  $A$  of a  $\sigma$ -algebra  $\mathcal{A}$  has been called a *measurable set*. We also introduced the concept of a *measure*, which assigns a nonnegative real number or  $\infty$  to all elements of a  $\sigma$ -algebra. This chapter is devoted to the concept of a *measurable mapping*, related concepts such as the  $\sigma$ -algebra generated by a mapping, and the *image measure* of  $\mu$  under  $f$ , the measure induced by a measurable mapping  $f$  on its codomain space. All these concepts play an important role in integration and probability theory. In probability theory, a measurable set is called an event, a measurable mapping  $f$  is called a *random variable* and the image measure of the probability measure  $P$  under  $f$  is called the *distribution of  $f$* .

### 2.1 Image and Inverse Image

Two key concepts of this chapter are the *image* of a set  $A \subset \Omega$  and the *inverse image* of a set  $A' \subset \Omega'$  under a mapping  $f: \Omega \rightarrow \Omega'$ . We start with the formal definitions and then illustrate these concepts in section 2.2.

**Definition 2.1 (Image and Inverse Image)**

Let  $\Omega, \Omega'$  denote two sets and  $f: \Omega \rightarrow \Omega'$  a mapping. Then we call

$$f(A) := \{f(\omega) : \omega \in A\}, \quad A \subset \Omega, \quad (2.1)$$

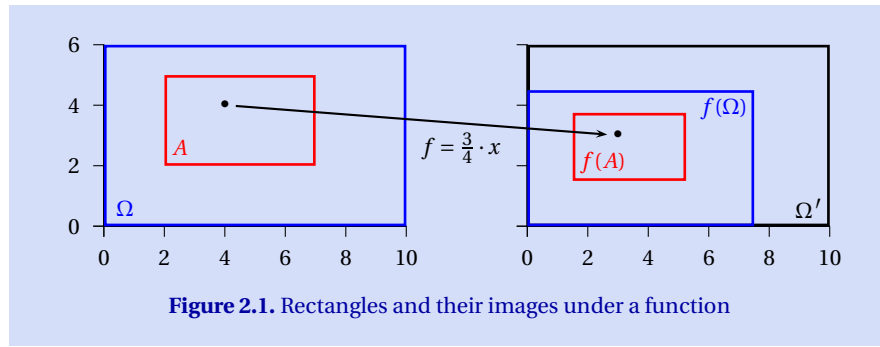
the *image of  $A$  under  $f$* , and

$$f^{-1}(A') := \{\omega \in \Omega : f(\omega) \in A'\}, \quad A' \subset \Omega', \quad (2.2)$$

the *inverse image of  $A'$  under  $f$* .

Whereas the image  $f(A)$  is a subset of  $\Omega'$ , the inverse image  $f^{-1}(A')$  is the set of all elements of the *domain*  $\Omega$  for which  $f$  takes on a value in the subset  $A'$  of its *codomain*  $\Omega'$ . For convenience, we also use the notation

$$\{f \in A'\} := f^{-1}(A') \quad \text{and} \quad \{f = \omega'\} := f^{-1}(\{\omega'\}). \quad (2.3)$$



**Remark 2.2 (Properties of Inverse Images)** Let  $f: \Omega \rightarrow \Omega'$  be a mapping,  $I$  be an index set,  $A' \subset \Omega'$ , and  $(A'_i, i \in I)$  a family of subsets  $A'_i$  of  $\Omega'$ . Then

$$f^{-1}[(A')^c] = [f^{-1}(A')]^c, \quad (2.4)$$

$$f^{-1}\left(\bigcap_{i \in I} A'_i\right) = \bigcap_{i \in I} f^{-1}(A'_i), \quad (2.5)$$

$$f^{-1}\left(\bigcup_{i \in I} A'_i\right) = \bigcup_{i \in I} f^{-1}(A'_i) \quad (2.6)$$

(see Exercise 2-1). Note that, in general, the corresponding properties do not necessarily hold for the image  $f(A)$ ,  $A \subset \Omega$ .  $\triangleleft$

## 2.2 Introductory Examples

### 2.2.1 Example 1: Rectangles

Our first example deals with rectangles, their *images*, and their *inverse images* under a mapping  $f$ .

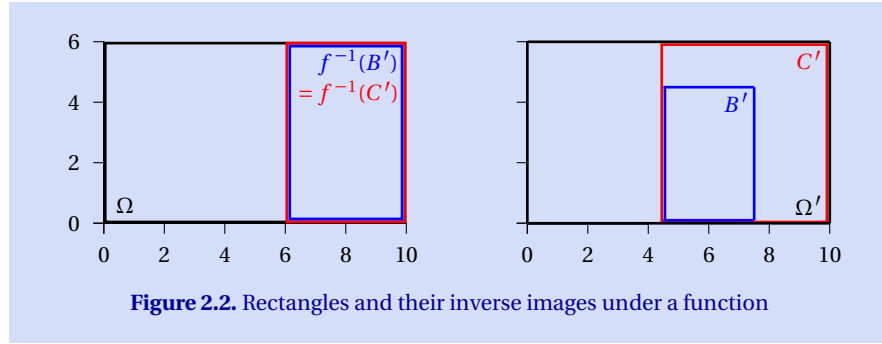
#### The Measurable Space

Let  $[a, b]$ ,  $a, b \in \mathbb{R}$ , denote the closed interval between  $a$  and  $b$ , inclusively, and consider the two rectangles

$$\Omega = [0, 10] \times [0, 6] \quad \text{and} \quad A = [2, 7] \times [2, 5]$$

depicted on the left-hand side of Figure 2.1. The elements of  $\Omega$  and  $A$  are points  $x = (x_1, x_2)$  in these rectangles with coordinates  $x_1$  on the horizontal axis and  $x_2$  on the vertical axis. Furthermore, let us consider a  $\sigma$ -algebra on  $\Omega$ ,

$$\mathcal{A} = \{\Omega, \emptyset, A, A^c\}.$$



### The Mapping and the Image

Consider the set  $\Omega' = \Omega$  and the function  $f: \Omega \rightarrow \Omega'$  defined by

$$f(x) = \frac{3}{4} \cdot x = \left( \frac{3}{4} \cdot x_1, \frac{3}{4} \cdot x_2 \right), \quad \forall x \in \Omega. \quad (2.7)$$

Hence,  $f$  maps all points  $x = (x_1, x_2) \in \Omega$  to the points  $f(x_1, x_2) \in \Omega'$ . This is illustrated by Figure 2.1 for the point  $x = (4, 4)$ , which is mapped to  $f(x) = (3, 3)$ . The right-hand side of Figure 2.1 also depicts the *image of A under f*, i. e.,  $f(A) = \{f(x): x \in A\}$ , as well as the image  $f(\Omega)$  of  $\Omega$  under  $f$ .

### The Inverse Images

We specify the  $\sigma$ -algebra

$$\mathcal{A}' = \{\Omega', \emptyset, B', (B')^c\}$$

on  $\Omega'$ , where

$$B' = ]4.5, 7.5] \times ]0, 4.5]$$

is the rectangle depicted on the right-hand side of Figure 2.2, and  $(B')^c = \Omega' \setminus B'$  is its complement.

Now we consider the *inverse image* of  $B'$  under  $f$  [see Eq. (2.7)], i. e.,

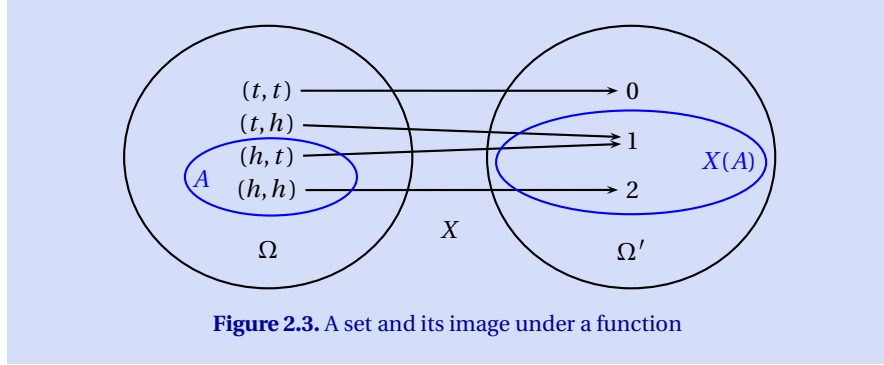
$$f^{-1}(B') = ]6, 10] \times ]0, 6]$$

(see Fig. 2.2). It is the rectangle on the right side of  $\Omega$ . For further examples see Exercises 2-2 and 2-3.

Also consider the inverse image of the rectangle

$$C' = ]4.5, 10] \times ]0, 6]$$

(see Fig. 2.2). Its inverse image under  $f$  is identical to the inverse image  $f^{-1}(B')$ , i. e.,



$$f^{-1}(C') = f^{-1}(B'),$$

which follows from

$$\begin{aligned} f^{-1}(C') &= f^{-1}[B' \cup (C' \setminus B')] && [B' \subset C', \text{ Fig. 2.2}] \\ &= f^{-1}(B') \cup f^{-1}(C' \setminus B') && [(2.6)] \\ &= f^{-1}(B') \cup \emptyset = f^{-1}(B'). \end{aligned}$$

Note that  $f^{-1}(C' \setminus B') = \emptyset$ , because  $f$  has been defined on  $\Omega = [0, 10] \times [0, 6]$ . If we would define  $f$  on  $\Omega = \mathbb{R}^2$ , then the set  $f^{-1}(C' \setminus B')$  would *not* be empty. (See also Exercise 2-4.)

### 2.2.2 Example 2: Flipping two Coins

Now we consider the random experiment of *flipping two coins*.

#### The Measurable Space

In this random experiment, the set of possible outcomes is

$$\Omega = \{(h, h), (h, t), (t, h), (t, t)\}.$$

This set consists of four elements (pairs). For example, the first component of the pair  $(h, t)$  represents the outcome of flipping  $h = \text{heads}$  with the first coin and the second component represents the outcome of flipping  $t = \text{tails}$  with the second coin. As a  $\sigma$ -algebra on  $\Omega$  we consider the power set  $\mathcal{A} = \mathcal{P}(\Omega)$ .

#### The Mapping

Consider the function  $X: \Omega \rightarrow \Omega' = \{0, 1, 2\}$  defined by

$$X[(t, t)] = 0, \quad X[(t, h)] = 1, \quad X[(h, t)] = 1, \quad \text{and} \quad X[(h, h)] = 2.$$

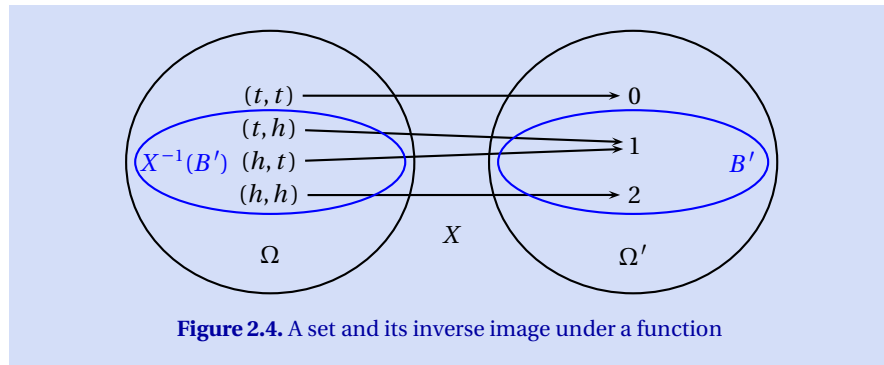


Figure 2.4. A set and its inverse image under a function

Table 2.1. Example of Measurable Sets Represented by a Mapping  $X$

$\{X \in \Omega'\}$	$= X^{-1}(\Omega')$	$= \Omega$	0, 1, or 2 heads are flipped
$\{X \in \emptyset\}$	$= X^{-1}(\emptyset)$	$= \emptyset$	neither 0, 1, nor 2 heads are flipped
$\{X = 0\}$	$= X^{-1}(\{0\})$	$= \{(t, t)\}$	no heads are flipped.
$\{X = 1\}$	$= X^{-1}(\{1\})$	$= \{(h, t), (t, h)\}$	heads are flipped exactly once
$\{X = 2\}$	$= X^{-1}(\{2\})$	$= \{(h, h)\}$	two heads are flipped
$\{X \in \{0, 1\}\}$	$= X^{-1}(\{0, 1\})$	$= \{(h, t), (t, h), (t, t)\}$	not more than one heads are flipped.
$\{X \in \{0, 2\}\}$	$= X^{-1}(\{0, 2\})$	$= \{(h, h), (t, t)\}$	either two heads or no heads at all are flipped
$\{X \in \{1, 2\}\}$	$= X^{-1}(\{1, 2\})$	$= \{(h, h), (h, t), (t, h)\}$	at least one heads is flipped

Looking at this assignment rule shows that this function may be called *number of flipping heads*. Again, we consider the *image of a set*  $A \subset \Omega$  under  $X$ , i. e.,  $X(A) = \{X(\omega) : \omega \in A\}$ ,  $A \subset \Omega$ . For example, for  $A = \{(h, h), (h, t)\}$ , the image under  $X$  is  $X(A) = \{1, 2\}$  (see Fig. 2.3).

### The Inverse Images

Suppose  $\mathcal{A}' = \mathcal{P}(\Omega')$  is the power set of  $\Omega' = \{0, 1, 2\}$ . In this example there are also  $2^3 = 8$  inverse images  $X^{-1}(A') = \{\omega \in \Omega : X(\omega) \in A'\}$ ,  $A' \in \mathcal{A}'$ . Three of these eight inverse images are:

$$X^{-1}(\{0\}) = \{(t, t)\}, \quad X^{-1}(\{1\}) = \{(h, t), (t, h)\}, \quad X^{-1}(\{2\}) = \{(h, h)\}.$$

These are the events that  $X$  takes on the value 0, 1, and 2, respectively. (In order to identify the inverse images listed above, trace back the arrows from right to left in Figure 2.4.) Furthermore, consider the inverse images

$$\begin{aligned} X^{-1}(\{0,1\}) &= \{(t,t), (h,t), (t,h)\}, \\ X^{-1}(\{0,2\}) &= \{(t,t), (h,h)\}, \\ X^{-1}(\{1,2\}) &= \{(h,t), (t,h), (h,h)\}. \end{aligned}$$

These are the events that  $X$  takes on a value in the sets  $\{0,1\}$ ,  $\{0,2\}$ , and  $\{1,2\}$ , respectively. One of these inverse images, namely  $X^{-1}(B')$ , with  $B' := \{1,2\}$ , is represented in Figure 2.4. Finally,

$$X^{-1}(\Omega') = \Omega \quad \text{and} \quad X^{-1}(\emptyset) = \emptyset.$$

Hence, we listed all eight inverse images  $X^{-1}(A')$ ,  $A' \in \mathcal{A}'$ . They are the eight measurable sets that can be represented by the mapping  $X$  and the  $\sigma$ -algebra  $\mathcal{A}' = \mathcal{P}(\Omega')$ . These sets are listed in Table 2.1, using the notation  $\{X \in A'\} := X^{-1}(A')$ ,  $A' \in \mathcal{A}'$ , and  $\{X = x\} := X^{-1}(\{x\})$ ,  $\{x\} \in \mathcal{A}'$  [see Eq. (2.3)].

### 2.3 Measurable Mapping

Now we define the concept of a *measurable mapping* and related concepts such as the  *$\sigma$ -algebra generated by a mapping* and *measurability of a mapping with respect to a mapping*.

**Remark 2.3 (Mapping)** Remember, a mapping  $f: \Omega \rightarrow \Omega'$  assigns to *all*  $\omega \in \Omega$  a unique  $f(\omega) \in \Omega'$ . Hence,  $f$  is, by definition, a subset of the Cartesian product  $\Omega \times \Omega'$ , i. e.,  $f = \{(\omega, f(\omega)): \omega \in \Omega\}$ . This implies that, instead of  $f: \Omega \rightarrow \Omega'$ , we can also write  $f: \Omega \rightarrow \Omega''$  for the *same mapping*, provided that  $f(\Omega) \subset \Omega''$ .  $\triangleleft$

**Remark 2.4 (Identical Mappings)** If  $f, g: \Omega \rightarrow \Omega'$  are two mappings, then

$$f = g \quad \Leftrightarrow \quad \{(\omega, f(\omega)): \omega \in \Omega\} = \{(\omega, g(\omega)): \omega \in \Omega\}. \quad (2.8)$$

If  $f = g$  we say that the two mappings are *identical*. Hence, even if  $f: \Omega \rightarrow \Omega'$  and  $g: \Omega \rightarrow \Omega''$  are mappings with  $\Omega' \neq \Omega''$ , it is still possible that  $f$  and  $g$  are identical. Note that (2.8) also implies: If, for  $f: \Omega \rightarrow \Omega'$  and  $g: \Omega \rightarrow \Omega''$ , we write  $f, g: \Omega \rightarrow \Omega'''$  with  $\Omega''' := \Omega' \cup \Omega''$ , then  $f$  and  $g$  remain unchanged.  $\triangleleft$

#### 2.3.1 Measurable Mapping

Now the core concept of this chapter is defined as follows:

**Definition 2.5 (Measurable Mapping)**

Let  $(\Omega, \mathcal{A})$ ,  $(\Omega', \mathcal{A}')$  be measurable spaces and let  $f: \Omega \rightarrow \Omega'$  be a mapping. Then  $f$  is called  $(\mathcal{A}, \mathcal{A}')$ -measurable if

$$f^{-1}(A') \in \mathcal{A}, \quad \forall A' \in \mathcal{A}'.$$

**Remark 2.6 (Notation)** We use the notation

$$f: (\Omega, \mathcal{A}) \rightarrow (\Omega', \mathcal{A}')$$

to express that the mapping  $f: \Omega \rightarrow \Omega'$  is  $(\mathcal{A}, \mathcal{A}')$ -measurable. If there is no ambiguity about  $\mathcal{A}'$ , we also say that  $f$  is  $\mathcal{A}$ -measurable or measurable with respect to  $\mathcal{A}$ .  $\triangleleft$

### Examples

**Example 2.7 (Rectangles – continued)** In Example 2.2.1, we considered the mapping  $f: \Omega \rightarrow \Omega' = \Omega$  defined by  $f(x) = \frac{3}{4}x$ . Furthermore, we considered the rectangle  $B' = ]4.5, 7.5] \times [0, 4.5]$  and the inverse image

$$f^{-1}(B') = ]6, 10] \times [0, 6].$$

If  $A = [2, 7] \times [2, 5]$ , then the inverse image  $f^{-1}(B')$  is not an element of the  $\sigma$ -algebra  $\mathcal{A} = \{\Omega, \emptyset, A, A^c\}$ . In this example, we also specified the  $\sigma$ -algebra  $\mathcal{A}' = \{\Omega', \emptyset, B', (B')^c\}$ . Hence,  $f$  is not  $(\mathcal{A}, \mathcal{A}')$ -measurable. However, if we specify a  $\sigma$ -algebra  $\mathcal{C}$  such that  $f^{-1}(B') \in \mathcal{C}$ , then  $f$  is  $(\mathcal{C}, \mathcal{A}')$ -measurable. As we see later on, this condition is sufficient for  $f$  to be  $(\mathcal{C}, \mathcal{A}')$ -measurable (see Th. 2.20).  $\triangleleft$

**Example 2.8 (Flipping two Coins – continued)** In Example 2.2.2, we considered the mapping  $X = \text{number of flipping heads}$  and in Table 2.1 (p. 45) we listed all inverse images  $X^{-1}(A')$ ,  $A' \in \mathcal{A}' = \mathcal{P}(\{0, 1, 2\})$ . Of course,  $\mathcal{A} = \mathcal{P}(\Omega)$  ensures that all inverse images  $X^{-1}(A')$ ,  $A' \in \mathcal{A}'$ , are elements of  $\mathcal{A}$ .

However, instead of  $\mathcal{A} = \mathcal{P}(\Omega)$ , we might consider the  $\sigma$ -algebra

$$\mathcal{A}_0 = \{\Omega, \emptyset, \{(h, h), (h, t)\}, \{(t, h), (t, t)\}\}.$$

The element  $\{(h, h), (h, t)\}$  represents the event that *heads* are flipped in the first flip and  $\{(t, h), (t, t)\}$  is the event that *tails* are flipped in the first flip. Hence, the  $\sigma$ -algebra  $\mathcal{A}_0$  contains the events that refer to the outcome of the *first flip* only, whereas  $X$  represents the number of heads in *both coin flips*. If we choose  $\mathcal{A}'$  to be the power set of  $\Omega' = \{0, 1, 2\}$ , then it is *not* true that all eight inverse images  $X^{-1}(A')$ ,  $A' \in \mathcal{A}'$ , are elements of  $\mathcal{A}_0$ . The inverse image  $X^{-1}(\{2\}) = \{(h, h)\}$ , e. g., is not an element of  $\mathcal{A}_0$ . Hence, if we consider the measurable spaces  $(\Omega, \mathcal{A}_0)$  and  $(\Omega', \mathcal{P}(\Omega'))$ , then the mapping  $X$  is *not*  $(\mathcal{A}_0, \mathcal{P}(\Omega'))$ -measurable. Hence, in some sense  $\mathcal{A}_0$  is ‘not well-adapted’ to  $X$ .  $\triangleleft$

**Example 2.9 (Two Trivial Cases)** If (a)  $\mathcal{A} = \mathcal{P}(\Omega)$  is the power set of  $\Omega$  or if (b)  $\mathcal{A}' = \{\Omega', \emptyset\}$ , then every mapping  $f: \Omega \rightarrow \Omega'$  is  $(\mathcal{A}, \mathcal{A}')$ -measurable. This is easily seen as follows: (a) If  $\mathcal{A} = \mathcal{P}(\Omega)$  is the power set of  $\Omega$ , then all inverse images  $f^{-1}(A')$ ,  $A' \subset \Omega'$ , are elements in  $\mathcal{A} = \mathcal{P}(\Omega)$ , because it is the set of *all* subsets of  $\Omega$ . (b) If  $\mathcal{A}' = \{\Omega', \emptyset\}$ , then every mapping  $f: \Omega \rightarrow \Omega'$  is  $(\mathcal{A}, \mathcal{A}')$ -measurable, because  $f^{-1}(\Omega') = \Omega$  and  $f^{-1}(\emptyset) = \emptyset$ . Again, the inverse images  $\Omega$  and  $\emptyset$  are both elements in every  $\sigma$ -algebra on  $\Omega$ . Hence, in both cases, (a) and (b), every mapping  $f: \Omega \rightarrow \Omega'$  is  $(\mathcal{A}, \mathcal{A}')$ -measurable.  $\triangleleft$

**Example 2.10 (Constant Mapping)** A constant mapping  $f: \Omega \rightarrow \Omega'$  is defined by

$$f(\omega) = \omega', \quad \forall \omega \in \Omega,$$

where  $\omega'$  is a fixed element of  $\Omega'$ . Such a constant mapping is  $(\mathcal{A}, \mathcal{A}')$ -measurable for any  $\sigma$ -algebra  $\mathcal{A}$  on  $\Omega$  and any  $\sigma$ -algebra  $\mathcal{A}'$  on  $\Omega'$ . This is true, because for all subsets  $A'$  of  $\Omega'$ : If  $\omega' \in A'$ , then  $f^{-1}(A') = \Omega$ . If, in contrast,  $\omega' \notin A'$ , then  $f^{-1}(A') = \emptyset$ . However,  $\Omega$  and  $\emptyset$  are elements of *all*  $\sigma$ -algebras on  $\Omega$ .  $\triangleleft$

**Example 2.11 (Identity Mapping)** The identity mapping  $id: \Omega \rightarrow \Omega$  defined by

$$id(\omega) = \omega, \quad \forall \omega \in \Omega,$$

is  $(\mathcal{A}, \mathcal{A}_0)$ -measurable for any pair of  $\sigma$ -algebras on  $\Omega$  with  $\mathcal{A}_0 \subset \mathcal{A}$ . This is easily seen as follows:

$$id^{-1}(A) = A, \quad \forall A \in \mathcal{A}_0.$$

Because we assume  $\mathcal{A}_0 \subset \mathcal{A}$ , we can conclude that  $id$  is  $(\mathcal{A}, \mathcal{A}_0)$ -measurable.  $\triangleleft$

**Example 2.12 (Indicator of a Measurable Set)** Let  $(\Omega, \mathcal{A}), (\Omega', \mathcal{A}')$  be two measurable spaces, where  $\mathcal{A}'$  is any  $\sigma$ -algebra on  $\Omega' \subset \mathbb{R}$  with  $\{0\}, \{1\} \in \mathcal{A}'$ . Then the indicator  $1_A: \Omega \rightarrow \Omega'$  is  $(\mathcal{A}, \mathcal{A}')$ -measurable *if and only if*  $A \in \mathcal{A}$ . Note that the requirement  $\{0\}, \{1\} \in \mathcal{A}'$  is not only satisfied by  $(\Omega', \mathcal{A}') := (\{0, 1\}, \mathcal{P}(\{0, 1\}))$ , but also by  $(\Omega', \mathcal{A}') = (\mathbb{R}, \mathcal{B})$  and by  $(\Omega', \mathcal{A}') = (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ , where  $\mathcal{B}$  denotes the Borel  $\sigma$ -algebra on  $\mathbb{R}$  and  $\overline{\mathcal{B}}$  the Borel  $\sigma$ -algebra on  $\overline{\mathbb{R}}$ .  $\triangleleft$

**Example 2.13 (Indicators of Unions and Intersections)** If  $(\Omega, \mathcal{A})$  is a measurable space and  $A, B \in \mathcal{A}$ , then  $1_{A \cap B}$  and  $1_{A \cup B}$  are  $(\mathcal{A}, \mathcal{B})$ -measurable. This follows from the fact that  $A \cap B \in \mathcal{A}$  and  $A \cup B \in \mathcal{A}$ . For the same reason,  $A_1, A_2, \dots \in \mathcal{A}$  implies that  $1_{\bigcup_{i=1}^{\infty} A_i}$  is  $(\mathcal{A}, \mathcal{B})$ -measurable.  $\triangleleft$

**Example 2.14 (Constant Function)** Assume that  $(\Omega, \mathcal{A})$  and  $(\Omega', \mathcal{A}')$  are measurable spaces such that  $\{\omega'\} \in \mathcal{A}'$ , for all  $\omega' \in \Omega'$ . Furthermore, let  $f: \Omega \rightarrow \Omega'$ . If  $\mathcal{A} = \{\Omega, \emptyset\}$ , then  $f$  is  $(\mathcal{A}, \mathcal{A}')$ -measurable if and only if  $f$  is a constant function, i. e., if and only if there is an  $\omega' \in \Omega'$  such that  $f(\omega) = \omega'$ , for all  $\omega \in \Omega$  (see Exercise 2-5). Note that for  $(\Omega', \mathcal{A}') = (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ ,  $\{x\} \in \overline{\mathcal{B}}$ , for all  $x \in \overline{\mathbb{R}}$ .  $\triangleleft$

**Example 2.15 (Dichotomous Function)** If  $\mathcal{A} = \{\Omega, \emptyset, A, A^c\}$  with  $A \subset \Omega$ , then  $f: \Omega \rightarrow \mathbb{R}$  is  $(\mathcal{A}, \mathcal{B})$ -measurable if and only if  $f = \alpha_1 1_A + \alpha_2 1_{A^c}$  for  $\alpha_1, \alpha_2 \in \mathbb{R}$  (see Exercise 2-6).  $\triangleleft$

### Step Function

Another important example of a measurable function is a step function, which is defined as follows:

**Definition 2.16 (Step Function)**

Let  $A_1, \dots, A_n$ ,  $n \in \mathbb{N}$ , be a finite sequence of subsets of a set  $\Omega$ . Then a finite linear combination

$$f = \sum_{i=1}^n \alpha_i 1_{A_i}, \quad \alpha_1, \dots, \alpha_n \in \mathbb{R}, \quad (2.9)$$

is called a step function.

**Remark 2.17 (Step Function and a Partition of  $\Omega$ )** If the sets  $A_1, \dots, A_n$  are pairwise disjoint, if we define  $A_{n+1} := \Omega \setminus (\bigcup_{i=1}^n A_i)$ , then  $\{A_1, \dots, A_n, A_{n+1}\}$  is a finite partition of  $\Omega$ . Furthermore, if  $f$  satisfies (2.9), and  $\alpha_{n+1} := 0$ , then, for all  $A' \subset \mathbb{R}$ ,

$$f^{-1}(A') = \bigcup_{\substack{i=1, \dots, n+1, \\ \alpha_i \in A'}} A_i \quad (2.10)$$

(see Exercise 2-7). ◁

**Remark 2.18 (Measurability of a Step Function)** If  $(\Omega, \mathcal{A})$  is a measurable space and  $A_1, \dots, A_n \in \mathcal{A}$ , then the step function  $f: \Omega \rightarrow \mathbb{R}$  defined by Equation (2.9) is  $(\mathcal{A}, \mathcal{B})$ -measurable (see Exercise 2-8). ◁

**Lemma 2.19 (Measurability if  $\mathcal{A}$  is Countably Generated)**

Let  $(\Omega, \mathcal{A})$  be a measurable space and let  $\mathcal{A} = \sigma(\mathcal{E})$ , where  $\mathcal{E}$  is a finite (i. e.,  $\mathcal{E} = \{A_1, \dots, A_n\}$ ) or countable (i. e.,  $\mathcal{E} = \{A_1, A_2, \dots\}$ ) partition of  $\Omega$ . Then  $f: \Omega \rightarrow \overline{\mathbb{R}}$  is  $(\mathcal{A}, \overline{\mathcal{B}})$ -measurable if and only if there are  $\alpha_1, \alpha_2, \dots \in \overline{\mathbb{R}}$  such that  $f = \sum_i \alpha_i 1_{A_i}$ , where  $i \in \{i = 1, \dots, n\}$  or  $i \in \mathbb{N}$ , respectively.

*(Proof p. 70)*

**A Necessary and Sufficient Condition of Measurability**

Let  $(\Omega', \mathcal{A}')$  be a measurable space and  $\mathcal{E}' \subset \mathcal{A}'$ . Then we denote

$$f^{-1}(\mathcal{E}') := \{f^{-1}(A') : A' \in \mathcal{E}'\}. \quad (2.11)$$

This notation is used in the following theorem, which can be utilized for proving  $(\mathcal{A}, \mathcal{A}')$ -measurability of a mapping  $f: \Omega \rightarrow \Omega'$ .

**Theorem 2.20 (Measurable Mapping and Generating Systems)**

Let  $(\Omega, \mathcal{A}), (\Omega', \mathcal{A}')$  denote measurable spaces, let  $\mathcal{E}' \subset \mathcal{A}'$ , and  $f: \Omega \rightarrow \Omega'$ . Then

$$\sigma[f^{-1}(\mathcal{E}')] = f^{-1}[\sigma(\mathcal{E}')]. \quad (2.12)$$

Furthermore, if  $\sigma(\mathcal{E}') = \mathcal{A}'$ , then  $f$  is  $(\mathcal{A}, \mathcal{A}')$ -measurable if and only if  $f^{-1}(A') \in \mathcal{A}$ , for all  $A' \in \mathcal{E}'$ .

For a proof see Klenke (2013, Theorem 1.81, p. 33).

Now consider a finite or countable set  $\Omega'$ . Then Theorem 2.20 and Example 1.18 immediately imply the following corollary:

**Corollary 2.21 (Finite or Countable Generating Systems)**

Let  $(\Omega, \mathcal{A})$ ,  $(\Omega', \mathcal{P}(\Omega'))$  be measurable spaces, where  $\Omega'$  is finite or countable, and let  $\mathcal{E}' = \{\{\omega'\}: \omega' \in \Omega'\}$ . Then a mapping  $f: \Omega \rightarrow \Omega'$  is  $(\mathcal{A}, \mathcal{P}(\Omega'))$ -measurable if and only if  $f^{-1}(\{\omega'\}) \in \mathcal{A}$ , for all  $\omega' \in \Omega'$ .

**Example 2.22 (Rectangles – continued)** In Example 2.2.1 we considered the mapping  $f: \Omega \rightarrow \Omega' = \Omega$  defined by  $f(x) = \frac{3}{4}x$ . Furthermore, we considered the rectangle  $B' = ]4.5, 7.5] \times [0, 4.5]$ . The set system

$$\mathcal{E}' = \{B'\},$$

which contains  $B'$  as the only element, generates the  $\sigma$ -algebra

$$\mathcal{A}' = \{\Omega', \emptyset, B', (B')^c\}.$$

Hence according to Theorem 2.20, the mapping  $f$  is  $(\mathcal{A}, \mathcal{A}')$ -measurable provided that  $f^{-1}(B') \in \mathcal{A}$ .  $\triangleleft$

**Example 2.23 (Flipping two Coins – continued)** In Example 2.2.2, we defined the mapping  $X = \text{number of flipping heads}$  with codomain  $\Omega' = \{0, 1, 2\}$ . Now consider the system

$$\mathcal{E}' = \{\{0\}, \{1\}\}$$

of subsets of  $\Omega'$ . First, note that  $\sigma(\mathcal{E}') = \mathcal{P}(\Omega')$ . Therefore, Theorem 2.20 implies that  $X$  is  $(\mathcal{A}, \mathcal{P}(\Omega'))$ -measurable for each  $\sigma$ -algebra  $\mathcal{A}$  on

$$\Omega = \{(h, h), (h, t), (t, h), (t, t)\}$$

for which

$$X^{-1}(\{0\}) = \{(t, t)\} \in \mathcal{A} \quad \text{and} \quad X^{-1}(\{1\}) = \{(h, t), (t, h)\} \in \mathcal{A}.$$

This does not only hold for  $\mathcal{A}_1 = \mathcal{P}(\Omega)$ , but also for the  $\sigma$ -algebra

$$\mathcal{A}_2 = \{\Omega, \emptyset, \{(t, t)\}, \{(h, t), (t, h)\}, \{(h, h)\}, \\ \{(h, h), (h, t), (t, h)\}, \{(h, h), (t, t)\}, \{(h, t), (t, h), (t, t)\}\}.$$

As mentioned before,  $\mathcal{A}_2$  contains all events that can be represented by  $X$  (see Table 2.1, p. 45). In contrast, this does not hold for the  $\sigma$ -algebra

$$\mathcal{A}_0 = \{\Omega, \emptyset, \{(h, h), (h, t)\}, \{(t, h), (t, t)\}\}$$

(see Example 2.8). Hence,  $X$  is measurable with respect to  $\mathcal{A}_1$  and  $\mathcal{A}_2$ , but it is not measurable with respect to  $\mathcal{A}_0$ . In this application, this means that the events  $\{(h, h), (h, t)\}$  and  $\{(t, h), (t, t)\}$  cannot be formulated in terms of  $X$ . Furthermore, some of the events that *can* be formulated in terms of  $X$  are not elements of  $\mathcal{A}_0$ . For example,  $X^{-1}(\{0\}) = \{(t, t)\}$  is *not* an element of  $\mathcal{A}_0$ .  $\triangleleft$

### 2.3.2 $\sigma$ -Algebra Generated by a Mapping

Let us consider again Example 2.2.2 and the mapping  $X = \text{number of flipping heads}$ . The set that consists of the eight inverse images  $X^{-1}(A')$ ,  $A' \in \mathcal{A}'$ , is again a  $\sigma$ -algebra on  $\Omega$ . In a sense, this  $\sigma$ -algebra carries the information associated with the mapping  $X$ ; it contains all events that can be represented by  $X$  (see Table 2.1). In the following theorem we formulate the general proposition.

**Theorem 2.24 ( $\sigma$ -Algebra Generated by a Mapping)**

Let  $f: \Omega \rightarrow \Omega'$  be a mapping and let  $(\Omega', \mathcal{A}')$  be a measurable space. Then

$$f^{-1}(\mathcal{A}') := \{f^{-1}(A') : A' \in \mathcal{A}'\} \quad (2.13)$$

is a  $\sigma$ -algebra on  $\Omega$ .

For a proof see Klenke (2013, Theorem 1.81, p. 33)).

**Remark 2.25 (Smallest  $\sigma$ -Algebra)** Note that  $f^{-1}(\mathcal{A}')$  is the smallest  $\sigma$ -algebra  $\mathcal{C}$  on  $\Omega$  such that  $f$  is  $(\mathcal{C}, \mathcal{A}')$ -measurable, i. e.,

$$\mathcal{C} \text{ is a } \sigma\text{-algebra on } \Omega \text{ and } f \text{ is } (\mathcal{C}, \mathcal{A}')\text{-measurable} \Rightarrow f^{-1}(\mathcal{A}') \subset \mathcal{C}.$$

◁

The set  $f^{-1}(\mathcal{A}')$  contains all sets in  $\mathcal{A}$  that can be represented by  $f$  and elements of  $\mathcal{A}'$ . Because  $f^{-1}(\mathcal{A}')$  is important, it has an own name and an alternative notation, which is sometimes more convenient.

**Definition 2.26 ( $\sigma$ -Algebra Generated by a Mapping)**

The set  $f^{-1}(\mathcal{A}')$  defined by Equation (2.13) is called the  $\sigma$ -algebra generated by  $f$  and  $\mathcal{A}'$ . If there is no ambiguity about  $\mathcal{A}'$ , then we also say that  $f^{-1}(\mathcal{A}')$  is generated by  $f$  and use the notation

$$\sigma(f) := f^{-1}(\mathcal{A}'). \quad (2.14)$$

**Remark 2.27 (Monotonicity)** Note that, for two set systems  $\mathcal{C}' \subset \mathcal{A}'$ ,

$$f^{-1}(\mathcal{C}') \subset f^{-1}(\mathcal{A}'), \quad (2.15)$$

because  $f^{-1}(\mathcal{C}') = \{f^{-1}(A') : A' \in \mathcal{C}'\} \subset \{f^{-1}(A') : A' \in \mathcal{A}'\} = f^{-1}(\mathcal{A}')$ . ◁

The following corollary immediately follows from Definition 2.26 and the definition of  $(\mathcal{A}, \mathcal{A}')$ -measurability (see Def. 2.5).

**Corollary 2.28 (A Condition Equivalent to Measurability)**

Let  $f: \Omega \rightarrow \Omega'$  be a mapping and let  $(\Omega', \mathcal{A}')$  be a measurable space. Then  $f$  is  $(\mathcal{A}, \mathcal{A}')$ -measurable if and only if  $\sigma(f) \subset \mathcal{A}$ .

In the following lemma and the subsequent remark, we treat a  $\cap$ -stable generating system (see Def. 1.36). For a  $\sigma$ -algebra  $\mathcal{C}$  and a measurable function  $f: (\Omega, \mathcal{A}) \rightarrow (\Omega', \mathcal{A}')$  we use the notation  $\sigma(\mathcal{C}, f) := \sigma(\mathcal{C} \cup f^{-1}(\mathcal{A}'))$ .

**Lemma 2.29 ( $\cap$ -Stable Generating System)**

Let  $(\Omega, \mathcal{A})$  be a measurable space and let  $\mathcal{C} \subset \mathcal{A}$  be a  $\sigma$ -algebra. Furthermore, assume that  $\Omega'$  is finite or countable and let  $f: (\Omega, \mathcal{A}) \rightarrow (\Omega', \mathcal{P}(\Omega'))$  be a measurable mapping. Then the set

$$\mathcal{D} := \{C \cap f^{-1}(\{\omega'\}) : \omega' \in \Omega' \text{ and } C \in \mathcal{C}\}$$

is a  $\cap$ -stable generating system of  $\sigma(\mathcal{C}, f) := \sigma(\mathcal{C} \cup f^{-1}[\mathcal{P}(\Omega')])$ .

(Proof p. 71)

**Remark 2.30 (A Special Case)** Let us consider the special case in which  $\mathcal{C} = \{\Omega, \emptyset\}$ . In this case, Lemma 2.29 simplifies as follows: Let  $(\Omega, \mathcal{A})$  be a measurable space and let  $f: (\Omega, \mathcal{A}) \rightarrow (\Omega', \mathcal{P}(\Omega'))$  be a measurable mapping, where  $\Omega'$  is finite or countable. Then the set  $\{f^{-1}(\{\omega'\}) : \omega' \in \Omega'\} \cup \{\emptyset\}$  is a  $\cap$ -stable generating system of  $\sigma(f) := f^{-1}[\mathcal{P}(\Omega')]$ .  $\triangleleft$

**Example 2.31 ( $\sigma$ -Algebra Generated by an Indicator)** Let  $1_A: (\Omega, \mathcal{A}) \rightarrow (\mathbb{R}, \mathcal{B})$  be the indicator of  $A \in \mathcal{A}$ . Then  $\sigma(1_A) = \{\Omega, \emptyset, A, A^c\}$ . The same  $\sigma$ -algebra is generated by  $1_A: (\Omega, \mathcal{A}) \rightarrow (\{0, 1\}, \mathcal{P}(\{0, 1\}))$  (see Remark 2.33 for the general proposition).  $\triangleleft$

**Example 2.32 (Flipping two Coins – continued)** In Example 2.2.2, we considered flipping two coins and the measurable mapping  $X = \text{number of flipping heads}$  with codomain  $\Omega' = \{0, 1, 2\}$ . In this example, all elements of  $X^{-1}[\mathcal{P}(\Omega')]$  have been listed in Table 2.1 (p. 45) as the inverse images  $X^{-1}(A')$  of the eight sets  $A' \in \mathcal{P}(\Omega')$ . Furthermore,  $X^{-1}[\mathcal{P}(\Omega')] = \mathcal{A}_2$ , where  $\mathcal{A}_2$  is the  $\sigma$ -algebra defined in Example 2.23.

Instead of choosing  $\Omega' = \{0, 1, 2\}$  as the codomain of  $X$ , we may also choose the set  $\mathbb{R}$  of real numbers, i. e.,  $X: \Omega \rightarrow \mathbb{R}$  is then considered to be a function into  $\mathbb{R}$ . In this case we use the *Borel  $\sigma$ -algebra*  $\mathcal{B}$  on  $\mathbb{R}$ . However, according to the following remark, the  $\sigma$ -algebra  $X^{-1}(\mathcal{B})$  generated by  $X$  and  $\mathcal{B}$  is the same as the  $\sigma$ -algebra  $X^{-1}[\mathcal{P}(\Omega')]$  generated by  $X$  and the power set of  $\Omega' = \{0, 1, 2\}$ .  $\triangleleft$

**Remark 2.33 ( $\sigma$ -Algebra Generated by a Function Into a Countable Set)** Let us consider a function  $f: \Omega \rightarrow \Omega' \subset \mathbb{R}$  and let  $\mathcal{B}$  denote the Borel  $\sigma$ -algebra on  $\mathbb{R}$ . If  $\Omega'$  is finite or countable, then  $f^{-1}[\mathcal{P}(\Omega')] = f^{-1}(\mathcal{B})$  (see Exercise 2-9).  $\triangleleft$

**Table 2.2.** Joe and Ann With Random Assignment and Measurable Mappings

Elements of $\Omega$				Measurable mappings		
Unit	Treatment	Success	$P(\{\omega\})$	Person variable $U$	Treatment variable $X$	Outcome variable $Y$
(Joe, no, -)			.09	Joe	0	0
(Joe, no, +)			.21	Joe	0	1
(Joe, yes, -)			.04	Joe	1	0
(Joe, yes, +)			.16	Joe	1	1
(Ann, no, -)			.24	Ann	0	0
(Ann, no, +)			.06	Ann	0	1
(Ann, yes, -)			.12	Ann	1	0
(Ann, yes, +)			.08	Ann	1	1

**Example 2.34 (Joe and Ann – continued)** Table 2.2 (p. 53) displays mappings on  $(\Omega, \mathcal{A})$ , all components of which have already been specified in Example 1.9. The *first* mapping displayed in Table 2.2 is the *person variable*  $U$  that assigns to each possible outcome  $\omega \in \Omega$  the value *Joe* if  $\omega \in \{\text{Joe}\} \times \Omega_X \times \Omega_Y$  and the value *Ann* if  $\omega \in \{\text{Ann}\} \times \Omega_X \times \Omega_Y$ . Hence,  $U: \Omega \rightarrow \Omega_U$  is a mapping with domain  $\Omega = \Omega_U \times \Omega_X \times \Omega_Y$  and codomain  $\Omega_U$ . It projects the first component  $u$  of  $\omega = (u, \omega_X, \omega_Y)$  onto the set  $\Omega_U$ . Therefore, it is also called the *first projection mapping*.

The *second* mapping in this table is the *treatment variable*  $X$ . It assigns to each possible outcome  $\omega \in \Omega$  the value 0 if  $\omega \in \Omega_U \times \{\text{no}\} \times \Omega_Y$  and the value 1 if  $\omega \in \Omega_U \times \{\text{yes}\} \times \Omega_Y$ . Hence,  $X: \Omega \rightarrow \Omega'$  is a function with domain  $\Omega$  and codomain  $\Omega' = \{0, 1\}$ .

The *third* mapping is the *outcome variable*  $Y$ . It assigns to each  $\omega \in \Omega$  the value 0 if  $\omega \in \Omega_U \times \Omega_X \times \{-\}$  and the value 1 if  $\omega \in \Omega_U \times \Omega_X \times \{+\}$ . Therefore,  $Y: \Omega \rightarrow \Omega'$  is a function with domain  $\Omega$  and codomain  $\Omega' = \{0, 1\}$ . Hence, all three mapping  $U$ ,  $X$ , and  $Y$  have the same domain  $\Omega$ .

Considering  $U: \Omega \rightarrow \Omega_U$  and the  $\sigma$ -algebra  $\mathcal{A}_U := \{\Omega_U, \emptyset, \{\text{Joe}\}, \{\text{Ann}\}\}$ , the  $\sigma$ -algebra  $U^{-1}(\mathcal{A}_U)$  consists of the following four inverse images: the event

$$U^{-1}(\{\text{Joe}\}) = \{(\text{Joe}, \text{no}, -), (\text{Joe}, \text{no}, +), (\text{Joe}, \text{yes}, -), (\text{Joe}, \text{yes}, +)\}$$

that *Joe is drawn*, the event

$$U^{-1}(\{\text{Ann}\}) = \{(\text{Ann}, \text{no}, -), (\text{Ann}, \text{no}, +), (\text{Ann}, \text{yes}, -), (\text{Ann}, \text{yes}, +)\}$$

that *Ann is drawn*, the sure event  $U^{-1}(\Omega_U) = \Omega$  that *Joe or Ann are drawn*, and the impossible event  $U^{-1}(\emptyset) = \emptyset$  that *neither Joe nor Ann are drawn*.  $\triangleleft$

### 2.3.3 Final $\sigma$ -Algebra

Consider the mapping  $f: \Omega \rightarrow \Omega'$ . As noted in Remark 2.25, for a  $\sigma$ -algebra  $\mathcal{A}'$  on  $\Omega'$ ,  $\sigma(f) = f^{-1}(\mathcal{A}')$  is the smallest  $\sigma$ -algebra on  $\Omega$  for which  $f$  is measurable. In contrast, now we consider a  $\sigma$ -algebra  $\mathcal{C}$  on  $\Omega$  and look for the largest  $\sigma$ -algebra  $\mathcal{C}'$  on  $\Omega'$  such that  $f$  is  $(\mathcal{C}, \mathcal{C}')$ -measurable. This  $\sigma$ -algebra is specified in the following lemma. It is called the *final  $\sigma$ -algebra*.

**Lemma 2.35 (Final  $\sigma$ -Algebra)**

Let  $f: \Omega \rightarrow \Omega'$  be a mapping and  $\mathcal{C}$  a  $\sigma$ -algebra on  $\Omega$ .

(i) Then

$$\mathcal{C}'_f := \{A' \subset \Omega' : f^{-1}(A') \in \mathcal{C}\} \quad (2.16)$$

is a  $\sigma$ -algebra on  $\Omega'$ .

(ii) Furthermore, if  $f: (\Omega, \mathcal{C}) \rightarrow (\Omega', \mathcal{A}')$  is a measurable mapping, then  $\mathcal{A}' \subset \mathcal{C}'_f$ .

(Proof p. 71)

Note that (ii) is a formal way of saying that  $\mathcal{C}'_f$  is the largest  $\sigma$ -algebra on  $\Omega'$  such that  $f$  is  $\mathcal{C}$ -measurable.

**Definition 2.36 (Final  $\sigma$ -Algebra)**

The  $\sigma$ -algebra  $\mathcal{C}'_f$  defined by Equation (2.16) is called the *final  $\sigma$ -algebra of  $\mathcal{C}$  under  $f$* .

### 2.3.4 Multivariate Mapping

Now consider the measurable space  $(\prod_{i=1}^n \Omega'_i, \otimes_{i=1}^n \mathcal{A}'_i)$  and note that the definitions of measurable mappings and of the  $\sigma$ -algebra generated by a mapping also apply to  $n$ -variate mappings  $f: \Omega \rightarrow \Omega'_1 \times \dots \times \Omega'_n$  and in particular to functions for which  $\Omega'_1 \times \dots \times \Omega'_n = \mathbb{R}^n$ .

**Lemma 2.37 ( $\sigma$ -Algebra Generated by a Multivariate Mapping)**

Let  $\Omega$  be a nonempty set,  $(\Omega'_i, \mathcal{A}'_i)$ ,  $i = 1, \dots, n$ ,  $n \in \mathbb{N}$ , be measurable spaces, and  $f = (f_1, \dots, f_n)$  be a multivariate mapping with  $f_i: \Omega \rightarrow \Omega'_i$ ,  $i = 1, \dots, n$ , i. e.,  $f: \Omega \rightarrow \prod_{i=1}^n \Omega'_i$ . Then

$$\sigma(f_1, \dots, f_n) := \sigma(f) = f^{-1} \left( \otimes_{i=1}^n \mathcal{A}'_i \right) = \sigma \left( \bigcup_{i=1}^n \sigma(f_i) \right). \quad (2.17)$$

(Proof p. 72)

According to the following theorem, a multivariate mapping is measurable if and only if all its components are measurable.

**Theorem 2.38 (Measurability of Multivariate Mappings)**

Under the assumptions of Lemma 2.37, the following two propositions are equivalent to each other:

- (a)  $f: (\Omega, \mathcal{A}) \rightarrow \left( \prod_{i=1}^n \Omega'_i, \otimes_{i=1}^n \mathcal{A}'_i \right)$  is a measurable mapping.  
 (b)  $\forall i = 1, \dots, n: f_i: (\Omega, \mathcal{A}) \rightarrow (\Omega'_i, \mathcal{A}'_i)$  is a measurable mapping.

(Proof p. 72)

**Remark 2.39 ( $\sigma$ -Algebra Generated by a Family of Mappings)** Let  $I$  be a (finite, countable, or uncountable) index set and let  $(f_i, i \in I)$  be a family of mappings  $f_i: (\Omega, \mathcal{A}) \rightarrow (\Omega'_i, \mathcal{A}'_i)$ . The  $\sigma$ -algebra generated by this family is defined as

$$\sigma(f_i, i \in I) := \sigma\left(\bigcup_{i \in I} \sigma(f_i)\right). \quad (2.18)$$

Then Equation (2.17) implies

$$\sigma(f) = \sigma(f_i, i \in I), \quad \text{where } I = \{i = 1, \dots, n\}. \quad (2.19)$$

◁

**Example 2.40 (Joe and Ann – continued)** In Example 2.34 we already considered the function  $X: \Omega \rightarrow \mathbb{R}$  indicating with its values 1 and 0 whether or not the drawn person is treated and the function  $Y: \Omega \rightarrow \mathbb{R}$  indicating with its values 1 and 0 whether or not the drawn person is successful. If we specify the  $\sigma$ -algebra  $\mathcal{A}$  on  $\Omega$  such that  $X$  and  $Y$  are both  $(\mathcal{A}, \mathcal{B}_2)$ -measurable, then the bivariate function  $(X, Y): \Omega \rightarrow \mathbb{R}^2$  is  $(\mathcal{A}, \mathcal{B}_2)$ -measurable. And vice versa, if we specify the  $\sigma$ -algebra  $\mathcal{A}$  on  $\Omega$  such that the bivariate function  $(X, Y): \Omega \rightarrow \mathbb{R}^2$  is  $(\mathcal{A}, \mathcal{B}_2)$ -measurable then  $X$  and  $Y$  are both  $(\mathcal{A}, \mathcal{B})$ -measurable. In this example  $X$ ,  $Y$ , and  $(X, Y)$  are measurable with respect to  $\mathcal{A}$  whenever the two inverse images  $X^{-1}(\{1\})$  and  $Y^{-1}(\{1\})$  are elements of  $\mathcal{A}$  (see Exercise 2-10). ◁

**Remark 2.41 (Lower Dimension Multivariate Mappings)** Lemma 2.37 and Remark 1.23 imply

$$\sigma(f_i, i \in J) \subset f^{-1}\left(\otimes_{i=1}^n \mathcal{A}'_i\right), \quad \forall J \subset \{1, \dots, n\}.$$

Furthermore, Theorem 2.38 implies: If

$$f = (f_1, \dots, f_n): (\Omega, \mathcal{A}) \rightarrow \left( \prod_{i=1}^n \Omega'_i, \otimes_{i=1}^n \mathcal{A}'_i \right)$$

is a measurable mapping and  $J = \{i_1, \dots, i_k\} \subset \{1, \dots, n\}$ ,  $k \leq n$ , then

$$f_J := (f_{i_1}, \dots, f_{i_k}): (\Omega, \mathcal{A}) \rightarrow \left( \prod_{j=1}^k \Omega'_{i_j}, \otimes_{j=1}^k \mathcal{A}'_{i_j} \right)$$

is measurable as well. ◁

### 2.3.5 Projection Mapping

In Definition 1.31 we introduced the product  $\sigma$ -algebra  $\otimes_{i=1}^n \mathcal{A}_i$  for a finite number of measurable spaces  $(\Omega_i, \mathcal{A}_i)$ . Now we give an equivalent characterization. Let  $(\Omega_i, \mathcal{A}_i)$ ,  $i = 1, \dots, n$ , be measurable spaces. Then, for  $j = 1, \dots, n$ , the  $j$ th *projection mapping*  $\pi_j: \times_{i=1}^n \Omega_i \rightarrow \Omega_j$  is defined by

$$\pi_j(\omega_1, \dots, \omega_n) = \omega_j, \quad \forall (\omega_1, \dots, \omega_n) \in \times_{i=1}^n \Omega_i. \quad (2.20)$$

The inverse images are

$$\pi_j^{-1}(A_j) = \Omega_1 \times \dots \times \Omega_{j-1} \times A_j \times \Omega_{j+1} \times \dots \times \Omega_n, \quad \text{for } A_j \subset \Omega_j. \quad (2.21)$$

#### Lemma 2.42 (Product $\sigma$ -Algebra)

If  $(\Omega_i, \mathcal{A}_i)$ ,  $i = 1, \dots, n$ , are measurable spaces, then

$$\otimes_{i=1}^n \mathcal{A}_i = \sigma(\pi_1, \dots, \pi_n). \quad (2.22)$$

(Proof p. 73)

### 2.3.6 Measurability With Respect to a Mapping

In the next definition, we consider two mappings and the concept of a mapping being measurable with respect to another mapping.

#### Definition 2.43 (Measurability With Respect to a Mapping)

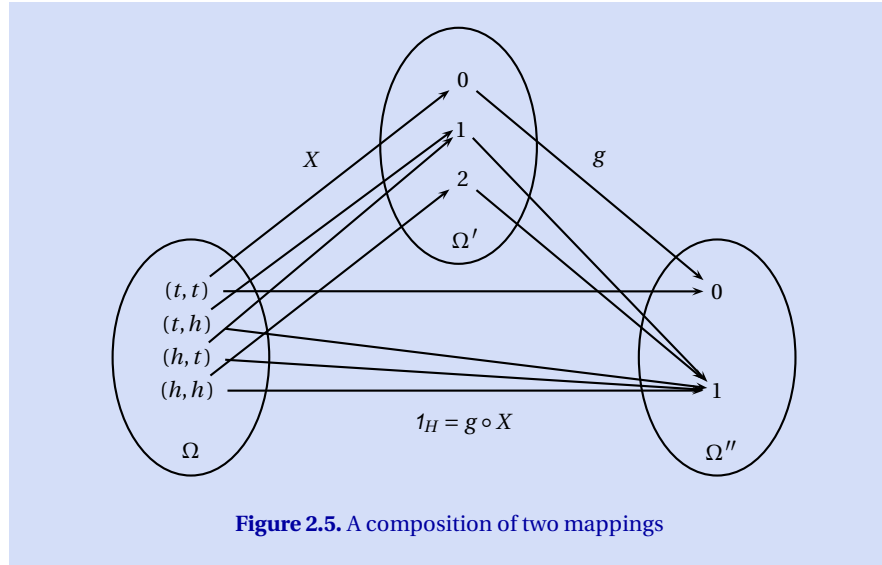
Let  $f: \Omega \rightarrow \Omega'$  and  $h: \Omega \rightarrow \Omega''$  be mappings, and let  $(\Omega', \mathcal{A}')$  and  $(\Omega'', \mathcal{A}'')$  be measurable spaces. Then  $h$  is called *measurable w.r.t.  $f$*  (or  *$f$ -measurable*) if

$$h^{-1}(\mathcal{A}'') \subset f^{-1}(\mathcal{A}'). \quad (2.23)$$

If  $\Omega'$  is finite or countable, then the following corollary provides a representation for all functions that are measurable with respect to  $f$ .

#### Corollary 2.44 (Measurability With Respect to a Discrete Function)

Let  $f: \Omega \rightarrow \Omega'$  be a mapping, let  $(\Omega', \mathcal{P}(\Omega'))$  be a measurable space, where  $\Omega'$  is finite or countable, and let  $h: \Omega \rightarrow \bar{\mathbb{R}}$  be a function. Then  $h$  is measurable w.r.t.  $f$  if and only if for all  $\omega' \in \Omega'$  there are  $\alpha_{\omega'} \in \bar{\mathbb{R}}$  such that



$$h = \sum_{\omega' \in \Omega'} \alpha_{\omega'} \cdot 1_{f^{-1}(\{\omega'\})}. \quad (2.24)$$

(Proof p. 73)

**Example 2.45 (Flipping Two Coins – continued)** Consider the mapping  $X = \text{number of flipping heads}$  with codomain  $\Omega' = \{0, 1, 2\}$ , let  $H := \{(h, t), (h, h), (t, h)\}$ , and let  $1_H: \Omega \rightarrow \Omega''$  denote the indicator of  $H$ , with  $\Omega'' = \{0, 1\}$ . Hence,  $1_H$  indicates with its values 1 and 0 whether or not at least one *heads* is flipped. If we consider the  $\sigma$ -algebra  $\mathcal{A}' = \mathcal{P}(\Omega')$  on  $\Omega'$  and the  $\sigma$ -algebra  $\mathcal{A}'' = \mathcal{P}(\Omega'')$  on  $\Omega''$ , then

$$X^{-1}(\mathcal{A}') = \{ \Omega, \emptyset, \{(h, h)\}, \{(h, t), (t, h)\}, \{(t, t)\}, \\ \{(h, h), (h, t), (t, h)\}, \{(h, h), (t, t)\}, \{(h, t), (t, h), (t, t)\} \}$$

and

$$1_H^{-1}(\mathcal{A}'') = \{ \Omega, \emptyset, \{(h, h), (h, t), (t, h)\}, \{(t, t)\} \}.$$

Obviously,  $1_H^{-1}(\mathcal{A}'') \subset X^{-1}(\mathcal{A}')$ . Therefore,  $1_H$  is measurable with respect to  $X$ , but not vice versa. That is,  $X$  represents a more detailed information about the outcome of the random experiment than  $1_H$ . Hence, if the value of  $X$  is known, then we can compute the value of  $1_H$ , but not vice versa. In our example, Figure 2.5 shows: if  $X(\omega) = 1$ , then  $1_H(\omega) = 1$ . However, if  $1_H(\omega) = 1$ , then  $X(\omega) = 1$  or  $X(\omega) = 2$ . (For a more general presentation of this property see Lemma 2.52.)  $\triangleleft$

## 2.4 Theorems on Measurable Mappings

In this section we consider compositions of mappings, which are defined as follows: Let  $\Omega$ ,  $\Omega'$ , and  $\Omega''$  be nonempty sets and let  $f: \Omega \rightarrow \Omega'$  and  $g: \Omega' \rightarrow \Omega''$  be mappings. Then the *composition of  $f$  and  $g$*  is the mapping  $g \circ f: \Omega \rightarrow \Omega''$  defined by:

$$g \circ f(\omega) := g[f(\omega)], \quad \forall \omega \in \Omega, \quad (2.25)$$

(see Fig. 2.5), where  $g \circ f(\omega)$  denotes the value of the mapping  $g \circ f$  for the argument  $\omega$ . Instead of  $g \circ f$ , we often use the notation  $g(f)$  and say that  $g(f)$  is a *function of  $f$* . Using this notation, Equation (2.25) can be written

$$g(f)(\omega) = g[f(\omega)], \quad \forall \omega \in \Omega. \quad (2.26)$$

### Lemma 2.46 (Compositions With a Finite or Countable Number of Values)

Let  $f: \Omega \rightarrow \Omega'$  be a mapping, where  $\Omega'$  is finite or countable, and let  $g: \Omega' \rightarrow \mathbb{R}$  be a function. Furthermore, for  $\omega' \in \Omega'$ , define  $1_{f=\omega'} := 1_{f^{-1}(\{\omega'\})}$ . Then

$$g \circ f = g(f) = \sum_{\omega' \in \Omega'} g(\omega') \cdot 1_{f^{-1}(\{\omega'\})} = \sum_{\omega' \in \Omega'} g(\omega') \cdot 1_{f=\omega'}. \quad (2.27)$$

(Proof p. 73)

Hence, under the assumptions of Lemma 2.46, for all  $\omega \in \Omega$ ,

$$g \circ f(\omega) = g[f(\omega)] = \sum_{\omega' \in \Omega'} g(\omega') \cdot 1_{f^{-1}(\{\omega'\})}(\omega) = \sum_{\omega' \in \Omega'} g(\omega') \cdot 1_{f=\omega'}(\omega). \quad (2.28)$$

**Example 2.47 (Flipping two Coins – continued)** Let us consider  $X = \text{number of flipping heads}$  and the mapping  $g: \Omega' \rightarrow \Omega''$  defined by  $g(x) := 1_{\{1,2\}}(x)$ , for all  $x \in \Omega'$  (see Fig. 2.5). Then the composition  $g \circ X$  defines a new mapping  $g \circ X: \Omega \rightarrow \Omega''$ , where  $\Omega'' = \{0, 1\}$ . In this example, the composition  $g \circ X$  is identical to the indicator  $1_H$  of the event  $H = \{(h, h), (t, h), (h, t)\}$  that heads are flipped at least once.  $\triangleleft$

**Example 2.48 (Joe and Ann – continued)** In Example 2.34 we already considered the mapping  $U: \Omega \rightarrow \Omega_U = \{\text{Joe}, \text{Ann}\}$  showing which person is drawn and the mapping  $X: \Omega \rightarrow \Omega' = \{0, 1\}$  indicating whether or not the drawn person is treated. Now we can consider the bivariate random variable  $(U, X): \Omega \rightarrow \Omega_U \times \Omega'$  and we can write

$$X = g \circ (U, X) = g(U, X)$$

as the composition of  $(U, X)$  and a (projection) mapping  $g$ ,

$$g[(u, x)] = x, \quad \forall (u, x) \in \Omega_U \times \Omega'. \quad \triangleleft$$

### 2.4.1 Measurability of a Composition

The following theorem shows that measurability is preserved by the composition of mappings.

**Theorem 2.49 (Measurability of a Composition)**

If  $f: (\Omega, \mathcal{A}) \rightarrow (\Omega', \mathcal{A}')$  and  $g: (\Omega', \mathcal{A}') \rightarrow (\Omega'', \mathcal{A}'')$  are measurable mappings, then the composition  $g \circ f$  is  $(\mathcal{A}, \mathcal{A}'')$ -measurable.

(Proof p. 74)

**Remark 2.50 ( $\sigma$ -Algebra Generated by a Composition)** Note that

$$(g \circ f)^{-1}(\mathcal{A}'') = f^{-1}[g^{-1}(\mathcal{A}'')] \quad (2.29)$$

(see the proof of Theorem 2.49).  $\triangleleft$

**Example 2.51 (Flipping two Coins – continued)** Figure 2.5 illustrates Theorem 2.49. If (a)  $X$  is  $(\mathcal{A}, \mathcal{A}')$ -measurable and (b)  $g$  is  $(\mathcal{A}', \mathcal{A}'')$ -measurable, then  $1_H = g \circ X$  is  $(\mathcal{A}, \mathcal{A}'')$ -measurable. Suppose  $\mathcal{A}' = \mathcal{P}(\Omega')$  and  $\mathcal{A}'' = \mathcal{P}(\Omega'')$ , where  $\Omega' = \{0, 1, 2\}$  and  $\Omega'' = \{0, 1\}$ . Then the premise '(a) and (b)' is satisfied if  $\mathcal{A}$  is such that  $X^{-1}(\mathcal{A}') \subset \mathcal{A}$ . If the premise '(a) and (b)' is *not* satisfied, then we cannot conclude that  $1_H$  is  $(\mathcal{A}, \mathcal{A}'')$ -measurable. Note that in this example  $1_H$  can be  $(\mathcal{A}, \mathcal{A}'')$ -measurable even if (a) and (b) do not hold. A sufficient requirement is that  $\{(t, t)\}$  and  $\{(t, h), (h, t), (h, h)\}$ , the inverse images of  $\{0\}$  and  $\{1\}$  under  $1_H$ , respectively, are elements of  $\mathcal{A}$  (see Cor. 2.21).  $\triangleleft$

If a mapping  $h$  is measurable with respect to a mapping  $f$ , then each element in the  $\sigma$ -algebra generated by  $h$  is an element in the  $\sigma$ -algebra generated by  $f$ . If  $h$  is measurable with respect to  $f$ , then in a sense, the information represented by  $h$  is already contained in  $f$  (cf. section 2.3.2). This is expressed in more formal terms in the following lemma, which is crucial, e. g., in the general definition of conditional expectations  $E(Y|X=x)$  (see ch. 10).

**Lemma 2.52 (Factorization Lemma of Measurable Functions)**

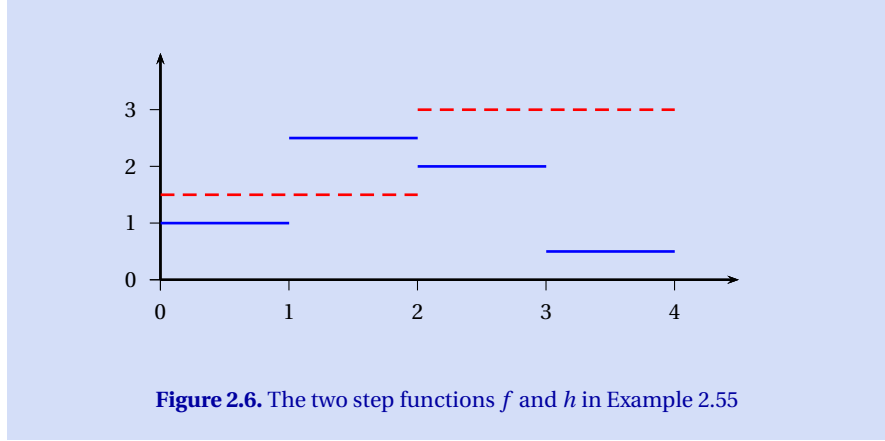
Let  $f: \Omega \rightarrow \Omega'$  be a mapping, let  $(\Omega', \mathcal{A}')$  be a measurable space, and let  $h: \Omega \rightarrow \overline{\mathbb{R}}$  be a function. Then  $h$  is measurable w.r.t.  $f$ , i. e.,  $h^{-1}(\overline{\mathcal{B}}) \subset f^{-1}(\mathcal{A}')$ , if and only if there is a measurable function  $g: (\Omega', \mathcal{A}') \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  such that

$$h = g \circ f \quad (2.30)$$

is the composition of  $f$  and  $g$ . We call  $g$  a factorization of  $h$  w.r.t.  $f$ .

For a proof see Klenke (2013, Corollary 1.97, pp. 38).

If, instead of  $(\overline{\mathbb{R}}, \overline{\mathcal{B}})$  we consider a measurable space  $(\Omega'', \mathcal{P}(\Omega''))$ , where  $\Omega''$  is finite or countable, then the elements  $\omega'' \in \Omega''$  can be renamed by real numbers such as 1, 2, etc. Renaming is a one-to-one measurable function, because



**Figure 2.6.** The two step functions  $f$  and  $h$  in Example 2.55

the  $\sigma$ -algebra on  $\Omega''$  is the power set of  $\Omega''$  (see Example 2.9). Hence, Lemma 2.52 implies the following corollary:

**Corollary 2.53 (Factorization of a Mapping Into a Finite or Countable Set)**

Let  $f: \Omega \rightarrow \Omega'$  be a mapping,  $(\Omega', \mathcal{A}')$  a measurable space, and  $h: \Omega \rightarrow \Omega''$  a mapping, where  $\Omega''$  is finite or countable. Then  $h$  is measurable w.r.t.  $f$ , i. e.,  $h^{-1}[\mathcal{P}(\Omega'')] \subset f^{-1}(\mathcal{A}')$ , if and only if there is a measurable mapping  $g: (\Omega', \mathcal{A}') \rightarrow (\Omega'', \mathcal{P}(\Omega''))$  such that  $h = g \circ f$ .

**Example 2.54 (Flipping two Coins – continued)** If we specify  $\Omega' = \{0, 1, 2\}$ , the  $\sigma$ -algebra  $\mathcal{A}' = \mathcal{P}(\Omega')$ , the set  $\Omega'' = \{0, 1\}$ , the  $\sigma$ -algebra  $\mathcal{A}'' = \mathcal{P}(\Omega'')$ , and the function  $h = 1_H$ , then the example depicted in Figure 2.5 can be used to illustrate this corollary. The mapping  $g$  in this figure is such that  $1_H = g \circ X$ .  $\triangleleft$

**Example 2.55 (Two Step Functions)** Figure 2.6 presents an example in which  $\Omega = [0, 4]$ ,  $A_1 = [0, 1]$ ,  $A_2 = ]1, 2]$ ,  $A_3 = ]2, 3]$ , and  $A_4 = ]3, 4]$ . Note that the sets  $A_1, \dots, A_4$  are pairwise disjoint. The measurable function  $f: (\Omega, \mathcal{A}) \rightarrow (\mathbb{R}, \mathcal{B})$  is defined by

$$f = \sum_{i=1}^4 \alpha_i 1_{A_i},$$

where  $\alpha_1 = 1$ ,  $\alpha_2 = 2.5$ ,  $\alpha_3 = 2$ , and  $\alpha_4 = 0.5$ . Furthermore, the function  $h: \Omega \rightarrow \mathbb{R}$  is defined by

$$h = \sum_{j \in \{1,3\}} \beta_j 1_{A_j \cup A_{j+1}}$$

with  $\beta_1 = 1.5$  and  $\beta_3 = 3$ . Note that  $\sigma(h) = \sigma(\{A_j \cup A_{j+1}: j \in \{1,3\}\})$ , whereas  $\sigma(f) = \sigma(\{A_i: i = 1, \dots, 4\})$  (see Exercise 2-11). Therefore,  $h$  is measurable with respect to  $f$ , i. e.,  $\sigma(h) \subset \sigma(f)$ .

According to Lemma 2.52, there is a function  $g: \mathbb{R} \rightarrow \mathbb{R}$  such that  $h = g \circ f$ . In fact, if we define  $g$  by

$$g(x) = \sum_{j \in \{1,3\}} \beta_j 1_{\{\alpha_j, \alpha_{j+1}\}}(x), \quad \forall x \in \mathbb{R},$$

then  $h = g \circ f$ . The function  $g$  takes on the value 1.5 if  $x = \alpha_1 = 1$  or  $x = \alpha_2 = 2.5$  and the value 3 if  $x = \alpha_3 = 2$  or  $x = \alpha_4 = 0.5$ . For all other  $x \in \mathbb{R}$  the value of  $g$  is 0.  $\triangleleft$

**Example 2.56 (Square of a Real-Valued Function)** Suppose  $f: (\Omega, \mathcal{A}) \rightarrow (\mathbb{R}, \mathcal{B})$  is a real-valued measurable function and  $f^2(\omega) := f(\omega)^2$ , for all  $\omega \in \Omega$ .

- (i) If  $f$  is *nonnegative*, i. e., if  $f(\omega) \geq 0$ , for all  $\omega \in \Omega$ , then  $f$  and  $f^2$  are measurable with respect to each other, i. e.,  $\sigma(f) = \sigma(f^2)$ .
- (ii) If there are  $\omega_1, \omega_2 \in \Omega$  with  $f(\omega_1) < 0 < f(\omega_2)$  and  $f^2(\omega_1) = f^2(\omega_2)$ , then  $\sigma(f^2) \subset \sigma(f)$ , but  $\sigma(f) \neq \sigma(f^2)$

(see Exercise 2-13.) In a sense,  $\sigma(f) = \sigma(f^2)$  means that  $f$  and  $f^2$  contain the same information, whereas  $\sigma(f^2) \subset \sigma(f)$ ,  $\sigma(f) \neq \sigma(f^2)$  means that  $f^2$  contains less information than  $f$ . If, e. g.,  $f^2(\omega) = 4$ , then  $f(\omega) = 2$  or  $f(\omega) = -2$ .  $\triangleleft$

### 2.4.2 Theorems on Measurable Functions

In the first theorem we consider sums and differences as well as products and ratios of measurable functions. The *sum* of two functions  $f, h: \Omega \rightarrow \mathbb{R}^n$  is again a function  $(f + h): \Omega \rightarrow \mathbb{R}^n$  defined by

$$(f + h)(\omega) := \begin{pmatrix} f_1 + h_1 \\ \vdots \\ f_n + h_n \end{pmatrix}(\omega) := \begin{pmatrix} f_1(\omega) + h_1(\omega) \\ \vdots \\ f_n(\omega) + h_n(\omega) \end{pmatrix}, \quad \forall \omega \in \Omega.$$

The first parentheses in the term  $(f + h)(\omega)$  are used to make clear that  $f + h$  is a symbol of a new function on  $\Omega$ . Of course, the difference  $f - h$  is defined in the same way as  $f + h$  replacing  $+$  by  $-$ .

Similarly, the *product*  $f \cdot h$  of two functions  $f, h: \Omega \rightarrow \mathbb{R}$  is again a function  $(f \cdot h): \Omega \rightarrow \mathbb{R}$  defined by

$$(f \cdot h)(\omega) := f(\omega) \cdot h(\omega), \quad \forall \omega \in \Omega.$$

Correspondingly,  $f/h: \Omega \rightarrow \mathbb{R}$  is defined by

$$(f/h)(\omega) := f(\omega)/h(\omega), \quad \forall \omega \in \Omega,$$

provided that  $h(\omega) \neq 0$  for all  $\omega \in \Omega$ .

**Theorem 2.57 (Sums and Products of Measurable Functions)**

If  $f, h: (\Omega, \mathcal{A}) \rightarrow (\mathbb{R}^n, \mathcal{B}_n)$  are measurable functions, then  $f + h$  and  $f - h$  are  $(\mathcal{A}, \mathcal{B}_n)$ -measurable as well. Furthermore, if  $f, h: (\Omega, \mathcal{A}) \rightarrow (\mathbb{R}, \mathcal{B})$  are measurable functions, then  $f \cdot h$  and  $f/h$  (with  $h(\omega) \neq 0$ , for all  $\omega \in \Omega$ ) are also  $(\mathcal{A}, \mathcal{B})$ -measurable.

For a proof see Klenke (2013, Theorem 1.91, p. 37).

**Remark 2.58 (Squared Function)** If  $f: (\Omega, \mathcal{A}) \rightarrow (\mathbb{R}, \mathcal{B})$  is a measurable function, then  $f^2 := f \cdot f$  is also  $(\mathcal{A}, \mathcal{B})$ -measurable. Obviously, this also applies to  $f^n$ ,  $n \in \mathbb{N}$ . Hence, if  $f$  is  $(\mathcal{A}, \mathcal{B})$ -measurable, then  $f^n$  is also  $(\mathcal{A}, \mathcal{B})$ -measurable.  $\triangleleft$

**Example 2.59 (Scale Transformations and Translations)** Remember that a constant real number can always be interpreted as a measurable function (see Example 2.10). Therefore, Theorem 2.57 implies that, for all  $\alpha \in \mathbb{R}$ , the functions  $f + \alpha$ ,  $f - \alpha$ , and  $\alpha \cdot f$  are  $(\mathbb{R}, \mathcal{B})$ -measurable if  $f: (\Omega, \mathcal{A}) \rightarrow (\mathbb{R}, \mathcal{B})$  is a measurable function.  $\triangleleft$

**Example 2.60 (Number of Flipping Heads)** Consider flipping a coin  $n$  times, let  $\Omega = \{h, t\}^n$ , and let  $1_{A_i}: \Omega \rightarrow \mathbb{R}$  denote the indicators of flipping heads at the  $i$ th flip of the coin. Then

$$X = \sum_{i=1}^n 1_{A_i}$$

is the number of flipping heads. If  $A_i \in \mathcal{A} = \mathcal{P}(\Omega)$ ,  $i = 1, \dots, n$ , then  $(\Omega, \mathcal{A})$  is a measurable space and  $X$  is measurable for any  $\sigma$ -algebra on  $\mathbb{R}$  (see Example 2.9). In the case  $\mathcal{A} = \mathcal{P}(\Omega)$  it is not necessary to apply Theorem 2.57.  $\triangleleft$

**Example 2.61 (Linear Combination of two Functions)** Let  $f, h: (\Omega, \mathcal{A}) \rightarrow (\mathbb{R}, \mathcal{B})$  be measurable functions and  $\alpha, \beta \in \mathbb{R}$ . Then, according to Theorem 2.57, the function  $(\alpha \cdot f + \beta \cdot h): (\Omega, \mathcal{A}) \rightarrow (\mathbb{R}, \mathcal{B})$  defined by

$$(\alpha \cdot f + \beta \cdot h)(\omega) = \alpha \cdot f(\omega) + \beta \cdot h(\omega), \quad \forall \omega \in \Omega. \quad (2.31)$$

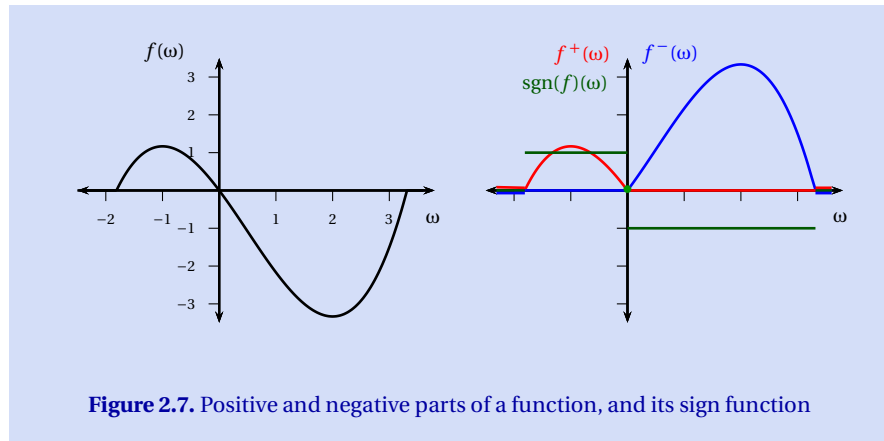
is  $(\mathcal{A}, \mathcal{B})$ -measurable.  $\triangleleft$

**Remark 2.62 (Positive and Negative Parts of a Function)** In Theorem 2.66 we consider the positive and the negative parts of a function  $f: \Omega \rightarrow \overline{\mathbb{R}}$ . The positive part  $f^+: \Omega \rightarrow \overline{\mathbb{R}}$  is defined by

$$f^+(\omega) := \max(f(\omega), 0), \quad \forall \omega \in \Omega,$$

and the negative part  $f^-: \Omega \rightarrow \overline{\mathbb{R}}$  by

$$f^-(\omega) := -\min(f(\omega), 0), \quad \forall \omega \in \Omega.$$



**Figure 2.7.** Positive and negative parts of a function, and its sign function

Hence, the value  $f^+(\omega)$  of the positive part of  $f$  is defined to be the *greater* one of the two numbers  $f(\omega)$  and 0 if they differ and  $f^+(\omega) = 0$  if  $f(\omega) = 0$ . In contrast, the value  $f^-(\omega)$  of the negative part of  $f$  is defined to be the *smaller* one of the two numbers  $f(\omega)$  and 0 *multiplied by -1* if they differ and  $f^-(\omega) = 0$  if  $f(\omega) = 0$ . Note that  $f^+$  and  $f^-$  are both nonnegative functions and that

$$f = f^+ - f^-.$$

◁

**Example 2.63 (Positive and Negative Parts of a Function)** The positive and negative parts of a function are illustrated by Figure 2.7 showing the graph of the function  $f: \mathbb{R} \rightarrow \mathbb{R}$  defined by

$$f(x) = \begin{cases} \frac{x^3}{3} - \frac{x^2}{2} - 2x, & \text{if } -1.81 < x < 3.315, \\ 0, & \text{otherwise.} \end{cases}$$

The positive part  $f^+$  takes on the value 0 if  $x \leq 0$  (see the red line on the horizontal axis), whereas negative part  $f^-$  takes on the value 0 if  $x \geq 0$  (see the blue line on the horizontal axis).

◁

**Remark 2.64 (Absolute Value Function)** Furthermore, we consider the *absolute value function*  $|f|: \Omega \rightarrow \overline{\mathbb{R}}$  defined by

$$|f|(\omega) := |f(\omega)| := \begin{cases} f(\omega), & \text{if } f(\omega) \geq 0 \\ -f(\omega), & \text{if } f(\omega) < 0. \end{cases}$$

Note that  $|f| = f^+ + f^- = \max(f^+, f^-)$ . Hence, in Figure 2.7, the absolute value function is represented by the red and blue lines *above* (if  $-1.81 < x < 3.315$ ) or *on* (if  $x < -1.81$  or  $x > 3.315$ ) the horizontal axis.

◁

**Remark 2.65 (Sign Function)** In Theorem 2.66, we also refer to  $\text{sgn}(f): \Omega \rightarrow \mathbb{R}$ , called the *sign function*, which is defined by

$$\text{sgn}(f)(\omega) = \begin{cases} 1, & \text{if } f(\omega) > 0 \\ 0, & \text{if } f(\omega) = 0 \\ -1, & \text{if } f(\omega) < 0. \end{cases}$$

In Figure 2.7, the graph of this function is represented by the green lines above and below the horizontal axis, and by the green point.  $\triangleleft$

**Theorem 2.66 (Positive and Negative Parts of a Function)**

Let  $(\Omega, \mathcal{A})$  be a measurable space. If  $f: \Omega \rightarrow \overline{\mathbb{R}}$  is  $(\mathcal{A}, \overline{\mathcal{B}})$ -measurable, then the functions  $f^+$ ,  $f^-$ ,  $|f|$ , and  $\text{sgn}(f)$  are  $(\mathcal{A}, \overline{\mathcal{B}})$ -measurable as well.

For a proof see Klenke (2013, Corollary 1.89, p. 36). The positive part  $f^+$ , the negative part  $f^-$ , and the absolute value function  $|f|$  of a function  $f$  plays an important role in integration theory (see ch. 3).

Another implication of Theorem 2.57 on the measurability of some sets that are often used is formulated in the following remark.

**Remark 2.67 (Some Important Measurable Sets)** Let  $(\Omega, \mathcal{A})$  be a measurable space and let  $f, g: (\Omega, \mathcal{A}) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  be measurable functions. Then

- (a)  $\{\omega \in \Omega: f(\omega) \geq g(\omega)\} \in \mathcal{A}$
- (b)  $\{\omega \in \Omega: f(\omega) > g(\omega)\} \in \mathcal{A}$
- (c)  $\{\omega \in \Omega: f(\omega) = g(\omega)\} \in \mathcal{A}$

(see Exercise 2-12).  $\triangleleft$

## 2.5 Equivalence of Two Mappings With Respect to a Measure

Now we study some properties of mappings  $f: \Omega \rightarrow \Omega'$  involving a measure space  $(\Omega, \mathcal{A}, \mu)$ . In this case, we use the notation

$$f: (\Omega, \mathcal{A}, \mu) \rightarrow \Omega'$$

to express that  $f: \Omega \rightarrow \Omega'$  is a mapping and that  $\mu$  is a measure on the measurable space  $(\Omega, \mathcal{A})$ . If there is also a  $\sigma$ -algebra  $\mathcal{A}'$  on  $\Omega'$ , then we use the notation

$$f: (\Omega, \mathcal{A}, \mu) \rightarrow (\Omega', \mathcal{A}')$$

to express that the mapping  $f: \Omega \rightarrow \Omega'$  is  $(\mathcal{A}, \mathcal{A}')$ -measurable and that  $\mu$  is a measure on the measurable space  $(\Omega, \mathcal{A})$ .

Remember, two mappings  $f$  and  $g$  are *identical*, i. e.,  $f = g$ , if and only if

$$\{\omega \in \Omega: f(\omega) \neq g(\omega)\} = \emptyset.$$

A less restrictive concept is their *equivalence with respect to the measure  $\mu$* .

**Definition 2.68 (Equivalence of Two Mappings With Respect to a Measure)**

Let  $f, g: (\Omega, \mathcal{A}, \mu) \rightarrow \Omega'$  be mappings. Then  $f$  and  $g$  are called  $\mu$ -equivalent, denoted by

$$f \stackrel{\mu}{=} g,$$

if there is an  $A \in \mathcal{A}$  with  $\mu(A) = 0$  and, for all  $\omega \in \Omega \setminus A$ ,  $f(\omega) = g(\omega)$ .

Because  $\mu(\emptyset) = 0$ ,  $f = g$  implies  $f \stackrel{\mu}{=} g$ .

**Remark 2.69 (A Note on Notation)** If  $f, g$  are denoted by  $f: (\Omega, \mathcal{A}, \mu) \rightarrow \Omega'_f$  and  $g: (\Omega, \mathcal{A}, \mu) \rightarrow \Omega'_g$ , then we can choose  $\Omega' = \Omega'_f \cup \Omega'_g$  and denote  $f, g: (\Omega, \mathcal{A}, \mu) \rightarrow \Omega'$  (see Rem. 2.4).  $\triangleleft$

**Remark 2.70 (An Alternative Notation)** If  $f \stackrel{\mu}{=} g$  we also say that  $f = g$ ,  $\mu$ -almost everywhere ( $\mu$ -a.e.). Furthermore, we also write

$$f(\omega) = g(\omega), \quad \text{for } \mu\text{-almost all } \omega \in \Omega, \quad (2.32)$$

and use  $f(\omega) \stackrel{\mu\text{-a.e.}}{=} g(\omega)$  as a shortcut.  $\triangleleft$

**Remark 2.71 (Singleton With a Positive Value of a Measure)** If  $f \stackrel{\mu}{=} g$  or, equivalently,  $f(\omega) \stackrel{\mu\text{-a.e.}}{=} g(\omega)$ , and  $\{\omega^*\} \in \mathcal{A}$  with  $\mu(\{\omega^*\}) > 0$ , then

$$f(\omega^*) = g(\omega^*).$$

$\triangleleft$

**Remark 2.72 ( $\mu$ -Equivalence, Restricted Functions, and Compositions)** Let  $f, g: (\Omega, \mathcal{A}, \mu) \rightarrow \Omega'$  be mappings.

(i) If  $\Omega' = \mathbb{R}$ , then

$$f \stackrel{\mu}{=} g \Rightarrow 1_A \cdot f \stackrel{\mu}{=} 1_A \cdot g, \quad \forall A \in \mathcal{A}. \quad (2.33)$$

(ii) If  $h: \Omega' \rightarrow \Omega''$  is also a mapping, then

$$f \stackrel{\mu}{=} g \Rightarrow h \circ f \stackrel{\mu}{=} h \circ g \quad (2.34)$$

(see Exercise 2-14).  $\triangleleft$

**Remark 2.73 (Equivalence Relation)** If  $\mathcal{M}$  is a set of mappings  $(\Omega, \mathcal{A}, \mu) \rightarrow \Omega'$ , then  $\stackrel{\mu}{=}$  is an *equivalence relation* on  $\mathcal{M}$  (see Exercise 2-15). In other words, if  $f, g, h \in \mathcal{M}$ , then the following propositions hold for  $\stackrel{\mu}{=}$ :

- (i)  $f \stackrel{\mu}{=} f$  (reflexivity).
- (ii)  $g \stackrel{\mu}{=} f$  if and only if  $f \stackrel{\mu}{=} g$  (symmetry).
- (iii) If  $f \stackrel{\mu}{=} g$  and  $g \stackrel{\mu}{=} h$ , then  $f \stackrel{\mu}{=} h$  (transitivity).

&lt;

**Definition 2.74 (Equivalence Class With Respect to a Measure)**

Let  $\mathcal{M}$  be a set of mappings  $(\Omega, \mathcal{A}, \mu) \rightarrow \Omega'$  and let  $f \in \mathcal{M}$ . Then

$$C(f) := \{g \in \mathcal{M} : g \stackrel{\mu}{=} f\}$$

is called the  $\mu$ -equivalence class of  $f$  in  $\mathcal{M}$  and  $f$  a representative of the class  $C(f)$ .

**Remark 2.75 (A Partition of the Set  $\mathcal{M}$ )** If  $\mathcal{M}$  is a set of mappings  $(\Omega, \mathcal{A}, \mu) \rightarrow \Omega'$ , then the set  $\{C(f) : f \in \mathcal{M}\}$  is a partition of  $\mathcal{M}$ , i. e.,

- (a)  $\forall f \in \mathcal{M} : C(f) \neq \emptyset$ .
- (b)  $\forall f, g \in \mathcal{M} : C(f) = C(g)$  or  $C(f) \cap C(g) = \emptyset$ .
- (c)  $\bigcup_{f \in \mathcal{M}} C(f) = \mathcal{M}$ .

(see Exercise 2-16).

&lt;

**Remark 2.76 (Other Properties of  $\mu$ -Equivalence)**

- (i) Let  $f, g : (\Omega, \mathcal{A}, \mu) \rightarrow \Omega'$ . If  $\mu(\Omega) > 0$ , then

$$\forall \alpha \in \Omega', \forall \beta \in \Omega' : f \stackrel{\mu}{=} \alpha \wedge g \stackrel{\mu}{=} \beta \wedge f \stackrel{\mu}{=} g \Rightarrow \alpha = \beta. \quad (2.35)$$

- (ii) If  $f, g, f^*, g^* : (\Omega, \mathcal{A}, \mu) \rightarrow \mathbb{R}$ , then

$$\begin{aligned} f \stackrel{\mu}{=} f^* \wedge g \stackrel{\mu}{=} g^* &\Rightarrow f + g \stackrel{\mu}{=} f^* + g^*, \\ f - g \stackrel{\mu}{=} f^* - g^*, & \\ f \cdot g \stackrel{\mu}{=} f^* \cdot g^*. & \end{aligned} \quad (2.36)$$

Furthermore, suppose  $\mu(\{\omega \in \Omega : g(\omega) = 0\}) = 0$ , and define  $\frac{f}{g} : \Omega \rightarrow \mathbb{R}$  by

$$\frac{f}{g}(\omega) := \begin{cases} \frac{f(\omega)}{g(\omega)}, & \text{if } g(\omega) \neq 0, \\ 0, & \text{otherwise,} \end{cases} \quad \forall \omega \in \Omega,$$

and let  $\frac{f^*}{g^*}$  be defined analogously. Then

$$f \stackrel{\mu}{=} f^* \wedge g \stackrel{\mu}{=} g^* \Rightarrow \frac{f}{g} \stackrel{\mu}{=} \frac{f^*}{g^*}. \quad (2.37)$$

(iii) If  $f_i, f_i^* : (\Omega, \mathcal{A}, \mu) \rightarrow \mathbb{R}$  and  $\alpha_i \in \mathbb{R}$ ,  $i = 1, \dots, n$ , then

$$(\forall i = 1, \dots, n: f_i \stackrel{\mu}{=} f_i^*) \Rightarrow \sum_{i=1}^n \alpha_i f_i \stackrel{\mu}{=} \sum_{i=1}^n \alpha_i f_i^*. \quad (2.38)$$

(iv) If  $f_1, f_2, \dots, f_1^*, f_2^*, \dots : (\Omega, \mathcal{A}, \mu) \rightarrow \mathbb{R}$  and  $\alpha_1, \alpha_2, \dots \in \mathbb{R}$ , then

$$(\forall i = 1, 2, \dots: f_i \stackrel{\mu}{=} f_i^*) \Rightarrow \sum_{i=1}^{\infty} \alpha_i f_i \stackrel{\mu}{=} \sum_{i=1}^{\infty} \alpha_i f_i^*, \quad (2.39)$$

provided that the limits denoted by the infinite sums (see Box 0.1) exist.

For proofs see Exercise 2-17.  $\triangleleft$

**Remark 2.77 (Order Relations for Real-Valued Functions)** For two mappings  $f, g : \Omega \rightarrow \mathbb{R}$  we write  $f < g$ , if and only if

$$\{\omega \in \Omega: f(\omega) \geq g(\omega)\} = \emptyset.$$

The notation  $f > g$ ,  $f \leq g$ , and  $f \geq g$  is used correspondingly.  $\triangleleft$

**Remark 2.78 (Order Relations With Respect to a Measure  $\mu$ )** For functions  $f, g, h : (\Omega, \mathcal{A}, \mu) \rightarrow \mathbb{R}$ , we also use the notation

$$f \stackrel{\mu}{<} g,$$

if there is an  $A \in \mathcal{A}$  with  $f(\omega) < g(\omega)$  for all  $\omega \in \Omega \setminus A$  and  $\mu(A) = 0$ . The notation  $f \stackrel{\mu}{>} g$ ,  $f \stackrel{\mu}{\leq} g$ , and  $f \stackrel{\mu}{\geq} g$  is used correspondingly. Furthermore,

$$f \stackrel{\mu}{<} g \quad \text{and} \quad g \stackrel{\mu}{=} h \quad \Rightarrow \quad f \stackrel{\mu}{<} h. \quad (2.40)$$

The analog propositions hold for  $\stackrel{\mu}{>}$ ,  $\stackrel{\mu}{\leq}$ , and  $\stackrel{\mu}{\geq}$  (see Exercise 2-18).  $\triangleleft$

## 2.6 Image Measure

In the definition of a measurable mapping  $f : (\Omega, \mathcal{A}) \rightarrow (\Omega', \mathcal{A}')$  we required: for all  $A' \in \mathcal{A}'$ :  $f^{-1}(A') \in \mathcal{A}$ . Because a measure  $\mu$  assigns a value to *all* elements  $A \in \mathcal{A}$ , the measure  $\mu$  also assigns a value to each  $f^{-1}(A') := \{\omega \in \Omega: f(\omega) \in A'\}$ . This is the reason for choosing the term *measurable mapping*: If  $\mu$  is a measure on  $\mathcal{A}$  and  $f$  is  $(\mathcal{A}, \mathcal{A}')$ -measurable, then there is a value  $\mu[f^{-1}(A')]$  for all inverse images  $f^{-1}(A')$ ,  $A' \in \mathcal{A}'$ .

According to the following theorem, a measurable mapping  $f : (\Omega, \mathcal{A}, \mu) \rightarrow (\Omega', \mathcal{A}')$  induces a measure on the codomain space  $(\Omega', \mathcal{A}')$ .

**Theorem 2.79 (Image Measure)**

Let  $f: (\Omega, \mathcal{A}, \mu) \rightarrow (\Omega', \mathcal{A}')$  be a measurable mapping. Then the function  $\mu_f: \mathcal{A}' \rightarrow \overline{\mathbb{R}}$  defined by

$$\mu_f(A') := \mu[f^{-1}(A')], \quad \forall A' \in \mathcal{A}', \quad (2.41)$$

is a measure on the measurable space  $(\Omega', \mathcal{A}')$ .

(Proof p. 74)

**Definition 2.80 (Image Measure)**

If  $f: (\Omega, \mathcal{A}, \mu) \rightarrow (\Omega', \mathcal{A}')$  is a measurable mapping, then  $\mu_f: \mathcal{A}' \rightarrow \mathbb{R}$  defined by Equation (2.41) is called the image measure of  $\mu$  under  $f$ .

**Example 2.81 (Rectangles – continued)** Now we consider a measure  $\mu: \mathcal{A} \rightarrow \mathbb{R}$  restricted to  $\mathcal{A}$ , which is specified by

$$\mu(A) = (7-2) \cdot (5-2) = 15$$

and

$$\mu(\Omega) = (10-0) \cdot (6-0) = 60.$$

This specification determines the areas of all four sets in  $\mathcal{A}$ , because  $\mu(A^c) = \mu(\Omega) - \mu(A) = 60 - 15 = 45$  and  $\mu(\emptyset) = 0$ . Hence, the measure space  $(\Omega, \mathcal{A}, \mu)$  is completely determined. Note that  $\mu$  is the restriction of the Lebesgue measure  $\lambda_2$  to the  $\sigma$ -algebra  $\mathcal{A}$ , i. e.,  $\mu(A) = \lambda_2(A)$ , for all  $A \in \mathcal{A}$ .

In Example 2.2.1 we considered the mapping  $f: \Omega \rightarrow \Omega' = \Omega$  defined by  $f(x) = \frac{3}{4}x$ . Furthermore, we considered the rectangle  $B' = ]4.5, 7.5] \times [0, 4.5]$  and the  $\sigma$ -algebra  $\mathcal{A}' = \{\Omega', \emptyset, B', (B')^c\}$ . If we specify  $\mathcal{A}$  such that  $f^{-1}(B') \in \mathcal{A}$ , then  $f$  is  $(\mathcal{A}, \mathcal{A}')$ -measurable. In this case, all inverse images  $f^{-1}(A')$  of sets  $A' \in \mathcal{A}'$  are elements of the  $\sigma$ -algebra  $\mathcal{A}$ . Therefore, the areas  $\lambda_2[f^{-1}(A')]$  of these inverse images are defined by the measure  $\lambda_2$  on  $\mathcal{A}$  that assigns the area to *all* elements of  $\mathcal{A}$ . If we specify  $\mathcal{A} = \{\Omega, \emptyset, f^{-1}(B'), f^{-1}[(B')^c]\}$ , then

$$\lambda_{2_f}(B') = \lambda_2[f^{-1}(B')] = \lambda_2[]6, 10] \times [0, 6] = (10-6) \cdot (6-0) = 24,$$

$$\lambda_{2_f}((B')^c) = \lambda_2(f^{-1}[(B')^c]) = \lambda_2([0, 6] \times [0, 6]) = (6-0) \cdot (6-0) = 36,$$

$\lambda_{2_f}(\Omega') = 60$ , and  $\lambda_{2_f}(\emptyset) = 0$ . Then the function  $\lambda_{2_f}: \mathcal{A}' \rightarrow \mathbb{R}$  defined by

$$\lambda_{2_f}(B') = \lambda_2[f^{-1}(B')], \quad \forall B' \in \mathcal{A}',$$

is again a measure, the *image measure* of  $\lambda_2$  under  $f$ . Therefore,  $(\Omega', \mathcal{A}', \lambda_{2_f})$  is a measure space.

Note that the image measure  $\lambda_{2_f}$  on the  $\sigma$ -algebra  $\mathcal{A}'$  differs from the area measure on  $\mathcal{A}'$ . In fact, the area of  $B'$  is  $(7.5-4.5) \cdot 4.5 = 13.5$  and the area of  $(B')^c$  is  $60 - 13.5 = 46.5$ .  $\triangleleft$

**Remark 2.82 (Cumulation of the Values  $\mu(\{\omega\})$ )** If  $\{\omega\} \in \mathcal{A}$ , for all  $\omega \in \Omega$ , then

$$\mu_f(\{\omega'\}) = \sum_{\omega: f(\omega)=\omega'} \mu(\{\omega\}), \quad (2.42)$$

provided that the sum is over a finite or countable number of summands. The measure  $\mu$  assigns to the singletons and other elements  $A \in \mathcal{A}$  a nonnegative number  $\mu(A)$ , and  $f$  maps each element  $\omega \in \Omega$  to an element  $\omega'$  in  $\Omega'$ . Thereby it translates the values  $\mu(A)$  of the measure  $\mu$  to their images  $f(A)$ . In particular, this applies to the singletons  $\{\omega\}$ . This is illustrated in the following example.  $\triangleleft$

**Example 2.83 (Flipping Two Coins – continued)** In this example,

$$P(\{\omega\}) = \frac{1}{4}, \quad \forall \omega \in \Omega, \quad (2.43)$$

uniquely defines a measure  $P: \mathcal{P}(\Omega) \rightarrow \mathbb{R}$  and the measure space  $(\Omega, \mathcal{P}(\Omega), P)$ . The reason is that the singletons  $\{\omega\}$  are pairwise disjoint and Rule (x) of Box 1.1 implies

$$P(A) = P\left(\bigcup_{\omega \in A} \{\omega\}\right) = \sum_{\omega \in A} P(\{\omega\}), \quad \forall A \in \mathcal{A}.$$

For instance, the set  $A = \text{flipping one and only one head}$  is the union  $A = \{(h, t)\} \cup \{(t, h)\} = \{(h, t), (t, h)\}$ . Hence,

$$P(A) = \sum_{\omega \in A} P(\{\omega\}) = P(\{(h, t)\}) + P(\{(t, h)\}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.$$

Now consider Figure 2.4 and realize that each arrow translates the value  $\mu(\{\omega\}) = \frac{1}{4}$  from left to right. According to Equation (2.42), this yields

$$P_X(\{0\}) = P[X^{-1}(\{0\})] = P(\{(t, t)\}) = \frac{1}{4},$$

$$P_X(\{1\}) = P[X^{-1}(\{1\})] = P(\{(t, h), (h, t)\}) = \frac{2}{4},$$

and

$$P_X(\{2\}) = P[X^{-1}(\{2\})] = P(\{(h, h)\}) = \frac{1}{4}.$$

$\triangleleft$

**Example 2.84 (Image Measure Under a Step Function)** If  $f: (\Omega, \mathcal{A}, \mu) \rightarrow (\mathbb{R}, \mathcal{B})$  is measurable such that  $f = \sum_{i=1}^n \alpha_i 1_{A_i}$  with pairwise different  $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ ,  $\alpha_i \neq 0$ , and pairwise disjoint  $A_i \in \mathcal{A}$ ,  $i = 1, \dots, n$ , and if we define  $A_{n+1} := \Omega \setminus (\bigcup_{i=1}^n A_i)$ , and  $\alpha_{n+1} := 0$ , then the image measure is

$$\mu_f = \sum_{i=1}^{n+1} \mu(A_i) \cdot \delta_{\alpha_i} \quad (2.44)$$

(see Exercise 2-19). Equation (2.44) generalizes Equation (2.42): For all  $\omega \in A_i$

$$f(\omega) = \alpha_i \cdot 1_{A_i}(\omega) = \alpha_i.$$

Hence,  $f$  translates the value  $\mu(A_i)$  to  $\alpha_i \in \mathbb{R}$  and  $\mu_f$  assigns the value  $\mu(A_i)$  to the singleton  $\{\alpha_i\}$ ,  $i = 1, \dots, n+1$ .  $\triangleleft$

Our next theorem deals with the image measures of  $\mu$ -equivalent measurable mappings. As a random variable is a particular measurable mapping and the distribution of a random variable a particular image measure (see section 5.1), this theorem has important implications on all concepts that in some sense describe properties of distributions of random variables such as expectations, variances, covariances, etc.

**Theorem 2.85 ( $\mu$ -Equivalence Implies Equality of Image Measures)**

If  $f, g: (\Omega, \mathcal{A}, \mu) \rightarrow (\Omega', \mathcal{A}')$  are measurable mappings, then

$$f \stackrel{\mu}{=} g \quad \Rightarrow \quad \mu_f = \mu_g. \quad (2.45)$$

(Proof p. 74)

In the following theorem we present a necessary and sufficient condition for  $\mu$ -equivalence of two compositions  $g \circ f$  and  $g^* \circ f$ .

**Theorem 2.86 ( $\mu$ -Equivalence of Compositions)**

If  $f: (\Omega, \mathcal{A}, \mu) \rightarrow (\Omega', \mathcal{A}')$  and  $g, g^*: (\Omega', \mathcal{A}') \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  are measurable mappings, then

$$g \stackrel{\mu_f}{=} g^* \quad \Leftrightarrow \quad g \circ f \stackrel{\mu}{=} g^* \circ f. \quad (2.46)$$

(Proof p. 75)

## 2.7 Proofs

**Proof of Lemma 2.19**

(a) If  $f = \sum_i \alpha_i 1_{A_i}$ , then for all  $B \in \bar{\mathcal{B}}$ ,

$$f^{-1}(B) = \bigcup_{i: \alpha_i \in B} A_i \in \mathcal{A},$$

because  $\mathcal{A}$  is closed with respect to finite and countable unions.

(b) Assume that there are no  $\alpha_1, \alpha_2, \dots \in \bar{\mathbb{R}}$  such that  $f = \sum_i \alpha_i 1_{A_i}$ . Then there are an  $i$  and elements  $\omega_1, \omega_2 \in A_i$  with  $f(\omega_1) \neq f(\omega_2)$ . Applying Equation (2.5) yields

$$f^{-1}(\{f(\omega_1)\}) \cap f^{-1}(\{f(\omega_2)\}) = \emptyset.$$

Furthermore, because  $\omega_j \in f^{-1}(\{f(\omega_j)\})$ ,  $j = 1, 2$ ,

$$f^{-1}(\{f(\omega_1)\}) \cap A_i \neq \emptyset, \quad \text{and} \quad f^{-1}(\{f(\omega_2)\}) \cap A_i \neq \emptyset.$$

Therefore, we conclude:  $f^{-1}(\{f(\omega_1)\}) \notin \mathcal{A}$  and  $f^{-1}(\{f(\omega_2)\}) \notin \mathcal{A}$ . Because  $\{f(\omega_1)\}, \{f(\omega_2)\} \in \bar{\mathcal{B}}$ , it follows that  $f$  is not  $(\mathcal{A}, \bar{\mathcal{B}})$ -measurable.

**Proof of Lemma 2.29**

(i)  $\cap$ -stability of  $\mathcal{D}$ . If  $C_1, C_2 \in \mathcal{C}$ , and  $\omega'_1, \omega'_2 \in \Omega'$ , then

$$\begin{aligned} & [C_1 \cap f^{-1}(\{\omega'_1\})] \cap [C_2 \cap f^{-1}(\{\omega'_2\})] \\ &= (C_1 \cap C_2) \cap [f^{-1}(\{\omega'_1\}) \cap f^{-1}(\{\omega'_2\})] \quad [\cap \text{ is associative and commutative}] \\ &= \begin{cases} (C_1 \cap C_2) \cap f^{-1}(\{\omega'_1\}), & \text{if } \omega'_1 = \omega'_2 \\ \emptyset, & \text{if } \omega'_1 \neq \omega'_2 \end{cases} \quad [(2.5)] \end{aligned}$$

is an element of  $\mathcal{D}$ , because  $C_1 \cap C_2 \in \mathcal{C}$  and  $\emptyset \in \mathcal{C}$ , which follows from the definition of a  $\sigma$ -algebra.

(ii) Denote  $\mathcal{A}' = \mathcal{D}(\Omega')$  and define  $\sigma(\mathcal{D}) = \sigma(\mathcal{C}, f) := \sigma[\mathcal{C} \cup f^{-1}(\mathcal{A}')]$ .

(a)  $\sigma(\mathcal{D}) \subset \sigma[\mathcal{C} \cup f^{-1}(\mathcal{A}')]$ . Obviously,  $\{f^{-1}(\{\omega'\}) : \omega' \in \Omega'\} \subset f^{-1}(\mathcal{A}')$ . Therefore,

$$\begin{aligned} \mathcal{D} &= \{C \cap f^{-1}(\{\omega'\}) : \omega' \in \Omega', C \in \mathcal{C}\} \\ &\subset \{C \cap f^{-1}(A') : A' \in \mathcal{A}', C \in \mathcal{C}\} \\ &\subset \sigma[\mathcal{C} \cup f^{-1}(\mathcal{A}')]. \quad [\text{Rem. 1.2}] \end{aligned}$$

Hence, according to Remark 1.23,  $\sigma(\mathcal{D}) \subset \sigma[\mathcal{C} \cup f^{-1}(\mathcal{A}')]$ .

(b)  $\sigma[\mathcal{C} \cup f^{-1}(\mathcal{A}')] \subset \sigma(\mathcal{D})$ . Because  $\Omega \in \mathcal{C}$  and  $\Omega \in f^{-1}(\mathcal{A}')$ , all  $C \in \mathcal{C}$  and all  $f^{-1}(A')$  are elements of  $\sigma(\mathcal{D})$  (see Def. 1.1,  $\Omega'$  is finite or countable). Therefore,  $\mathcal{C} \cup f^{-1}(\mathcal{A}') \subset \sigma(\mathcal{D})$ . Proposition (1.11) then implies  $\sigma[\mathcal{C} \cup f^{-1}(\mathcal{A}')] \subset \sigma(\mathcal{D})$ .

**Proof of Lemma 2.35**

(i) We have to show that  $\mathcal{C}'_f$  satisfies conditions (a) to (c) of Definition 1.1.

(a)

$$\Omega = f^{-1}(\Omega') \in \mathcal{C} \Rightarrow \Omega' \in \mathcal{C}'_f. \quad [(2.16)]$$

(b)

$$\begin{aligned} A' \in \mathcal{C}'_f &\Rightarrow f^{-1}(A') \in \mathcal{C} \quad [(2.16)] \\ &\Rightarrow f^{-1}(A')^c = f^{-1}[(A')^c] \in \mathcal{C} \quad [\text{Def. 1.1 (b), (2.4)}] \\ &\Rightarrow (A')^c \in \mathcal{C}'_f. \quad [(2.16)] \end{aligned}$$

(c)

$$\begin{aligned} A'_1, A'_2, \dots \in \mathcal{C}'_f &\Rightarrow f^{-1}(A'_1), f^{-1}(A'_2), \dots \in \mathcal{C} \quad [(2.16)] \\ &\Rightarrow \bigcup_{i=1}^{\infty} f^{-1}(A'_i) = f^{-1}\left(\bigcup_{i=1}^{\infty} A'_i\right) \in \mathcal{C} \quad [\text{Def. 1.1 (c), (2.6)}] \\ &\Rightarrow \bigcup_{i=1}^{\infty} A'_i \in \mathcal{C}'_f. \quad [(2.16)] \end{aligned}$$

(ii) For all  $A' \in \mathcal{A}'$ ,

$$\begin{aligned} A' \in \mathcal{A}' &\Rightarrow f^{-1}(A') \in \mathcal{C} \quad [(\mathcal{C}, \mathcal{A}')\text{-measurability of } f] \\ &\Rightarrow A' \in \mathcal{C}'_f. \quad [(2.16)] \end{aligned}$$

Hence,  $(\mathcal{C}, \mathcal{A}')$ -measurability of  $f$  implies  $\mathcal{A}' \subset \mathcal{C}'_f$ .

**Proof of Lemma 2.37**

First, note that, for  $A'_i \in \mathcal{A}'_i, i = 1, \dots, n$ ,

$$\begin{aligned}
 f^{-1}(A'_1 \times \dots \times A'_n) &= \{\omega \in \Omega: f(\omega) \in A'_1 \times \dots \times A'_n\} \\
 &= \{\omega \in \Omega: (f_1(\omega), \dots, f_n(\omega)) \in A'_1 \times \dots \times A'_n\} \\
 &= \{\omega \in \Omega: f_1(\omega) \in A'_1, \dots, f_n(\omega) \in A'_n\} \\
 &= \bigcap_{i=1}^n \{\omega \in \Omega: f_i(\omega) \in A'_i\} \\
 &= \bigcap_{i=1}^n f_i^{-1}(A'_i).
 \end{aligned} \tag{2.47}$$

Hence,

$$\begin{aligned}
 \sigma(f) &= \{f^{-1}(A'): A' \in \bigotimes_{i=1}^n \mathcal{A}'_i\} && \text{[Def. 2.26]} \\
 &= \sigma(\{f^{-1}(A'_1 \times \dots \times A'_n): A'_i \in \mathcal{A}'_i, i = 1, \dots, n\}) && \text{[Th. 2.20, Defs. 1.13, 1.31]} \\
 &= \sigma(\{f_1^{-1}(A'_1) \cap \dots \cap f_n^{-1}(A'_n): A'_i \in \mathcal{A}'_i, i = 1, \dots, n\}) && \text{[(2.47)]} \\
 &\supseteq \sigma\left(\bigcup_{i=1}^n \{f_i^{-1}(A'_i) \cap \bigcap_{j=1, j \neq i}^n f_j^{-1}(\Omega'_j): A'_i \in \mathcal{A}'_i, i = 1, \dots, n\}\right) && \text{[Rem. 1.23]} \\
 &= \sigma\left(\bigcup_{i=1}^n \{f_i^{-1}(A'_i): A'_i \in \mathcal{A}'_i, i = 1, \dots, n\}\right) && \text{[}f_j^{-1}(\Omega'_j) = \Omega\text{]} \\
 &= \sigma\left(\bigcup_{i=1}^n \sigma(f_i)\right). && \text{[Def. 2.26]}
 \end{aligned}$$

Furthermore,

$$\begin{aligned}
 &\{f_1^{-1}(A'_1) \cap \dots \cap f_n^{-1}(A'_n): A'_i \in \mathcal{A}'_i, i = 1, \dots, n\} \\
 &\subset \sigma\left(\bigcup_{i=1}^n \{f_i^{-1}(A'_i): A'_i \in \mathcal{A}'_i, i = 1, \dots, n\}\right). && \text{[Rem. 1.2, finite intersections]}
 \end{aligned}$$

Therefore,

$$\begin{aligned}
 \sigma\left(\bigcup_{i=1}^n \sigma(f_i)\right) &= \sigma\left(\bigcup_{i=1}^n \{f_i^{-1}(A'_i): A'_i \in \mathcal{A}'_i, i = 1, \dots, n\}\right) && \text{[Def. 2.26]} \\
 &\supseteq \sigma(\{f_1^{-1}(A'_1) \cap \dots \cap f_n^{-1}(A'_n): A'_i \in \mathcal{A}'_i, i = 1, \dots, n\}) && \text{[Rem. 1.23]} \\
 &= \sigma(f).
 \end{aligned}$$

Hence,  $\sigma(f) = \sigma(\bigcup_{i=1}^n \sigma(f_i))$ .

**Proof of Theorem 2.38**

(b)  $\Rightarrow$  (a): For all  $i = 1, \dots, n$ : Let  $A'_i \in \mathcal{A}'_i$ . If  $f_i$  is measurable, then  $f_i^{-1}(A'_i) \in \mathcal{A}$ . Hence,

$$\begin{aligned}
 f^{-1}(A'_1 \times \dots \times A'_n) &= \{\omega \in \Omega: f(\omega) \in A'_1 \times \dots \times A'_n\} \\
 &= \bigcap_{i=1}^n f_i^{-1}(A'_i) \in \mathcal{A}.
 \end{aligned}$$

Because  $\{A'_1 \times \dots \times A'_n : A'_i \in \mathcal{A}'_i, i = 1, \dots, n\}$  is a generating system of  $\bigotimes_{i=1}^n \mathcal{A}'_i$ , Theorem 2.20 implies that  $f$  is measurable.

(a)  $\Rightarrow$  (b): If  $f$  is measurable, then for all  $i = 1, \dots, n$ ,

$$\begin{aligned} f_i^{-1}(\mathcal{A}'_i) &= \{f_i^{-1}(A'_i) : A'_i \in \mathcal{A}'_i\} \\ &= \{f_i^{-1}(A'_i) \cap \bigcap_{j=1, j \neq i}^n f_j^{-1}(\Omega'_j) : A'_i \in \mathcal{A}'_i\} \\ &= \{f^{-1}(\Omega'_1 \times \dots \times \Omega'_{i-1} \times A'_i \times \Omega'_{i+1} \times \dots \times \Omega'_n) : A'_i \in \mathcal{A}'_i\} \quad [(2.47)] \\ &\subset f^{-1}\left(\bigotimes_{i=1}^n \mathcal{A}'_i\right) \subset \mathcal{A}. \end{aligned}$$

### **Proof of Lemma 2.42**

Consider

$$\pi = (\pi_1, \dots, \pi_n) : \left( \prod_{i=1}^n \Omega_i, \bigotimes_{i=1}^n \mathcal{A}_i \right) \rightarrow \left( \prod_{i=1}^n \Omega_i, \bigotimes_{i=1}^n \mathcal{A}_i \right).$$

Analogously to the proof of Lemma 2.37 and using Definition 1.31,

$$\begin{aligned} \bigotimes_{i=1}^n \mathcal{A}_i &= \sigma(\{A_1 \times \dots \times A_n : A_i \in \mathcal{A}_i, i = 1, \dots, n\}) \\ &= \sigma(\{\pi^{-1}(A_1 \times \dots \times A_n) : A_i \in \mathcal{A}_i, i = 1, \dots, n\}) \quad [(2.20), (2.21)] \\ &= \sigma(\pi) \quad [\text{Th. 2.20}, (2.12)] \\ &= \sigma(\pi_i, i = 1, \dots, n). \quad [\text{Lem. 2.37}, (2.18)] \end{aligned}$$

### **Proof of Corollary 2.44**

If  $\Omega'$  is finite or countable and we consider the measurable space  $(\Omega', \mathcal{P}(\Omega'))$ , then  $\sigma(f) = \sigma[\{f^{-1}(\{\omega'\}) : \omega' \in \Omega'\}]$  [see Lemma 2.29 with  $\mathcal{C} = \{\Omega, \emptyset\}$ ]. Because  $\{f^{-1}(\{\omega'\}) : \omega' \in \Omega'\}$  is a finite or countable partition of  $\Omega$ , this corollary is an immediate implication of Lemma 2.19.

### **Proof of Lemma 2.46**

For all  $\omega \in \Omega$  and all  $\omega' \in \Omega'$ ,

$$g[f(\omega)] \cdot 1_{f=\omega'}(\omega) = g(\omega') \cdot 1_{f=\omega'}(\omega) = \begin{cases} 0, & \text{if } f(\omega) \neq \omega' \\ g(\omega'), & \text{if } f(\omega) = \omega'. \end{cases} \quad (2.48)$$

This equation is equivalent to

$$g(f) \cdot 1_{f=\omega'} = g(\omega') \cdot 1_{f=\omega'}. \quad (2.49)$$

Because the set  $\{f^{-1}(\{\omega'\}) : \omega' \in \Omega'\}$  is a finite or countable partition of  $\Omega$  we can conclude:

$$1_\Omega = \sum_{\omega' \in \Omega'} 1_{f=\omega'}.$$

Therefore,

$$g(f) = g(f) \cdot 1_{\Omega} = \sum_{\omega' \in \Omega'} g(f) \cdot 1_{f=\omega'} = \sum_{\omega' \in \Omega'} g(\omega') \cdot 1_{f=\omega'},$$

and this implies Equation (2.27).

### **Proof of Theorem 2.49**

If  $f: (\Omega, \mathcal{A}) \rightarrow (\Omega', \mathcal{A}')$ ,  $g: (\Omega', \mathcal{A}') \rightarrow (\Omega'', \mathcal{A}'')$  are measurable mappings, then, according to Corollary 2.28,  $f^{-1}(\mathcal{A}') \subset \mathcal{A}$  and  $g^{-1}(\mathcal{A}'') \subset \mathcal{A}'$ . Hence, for all  $A'' \in \mathcal{A}''$ ,

$$\begin{aligned} (g \circ f)^{-1}(A'') &= \{\omega \in \Omega: g[f(\omega)] \in A''\} \\ &= \{\omega \in \Omega: f(\omega) \in g^{-1}(A'')\} \\ &= f^{-1}[g^{-1}(A'')]. \end{aligned}$$

Furthermore,

$$\begin{aligned} f^{-1}[g^{-1}(\mathcal{A}'')] &\subset f^{-1}(\mathcal{A}') && [(2.15)] \\ &\subset \mathcal{A}. && [(2.15)] \end{aligned}$$

### **Proof of Theorem 2.79**

We show that  $\mu_f$  has the properties (a) to (c) required in Definition 1.43. For each property of  $\mu_f$  we use the corresponding property of  $\mu$ .

- (a)  $\mu_f(\emptyset) = \mu[f^{-1}(\emptyset)] = \mu(\emptyset) = 0$ .
- (b) For all  $A' \in \mathcal{A}'$ :  $\mu_f(A') = \mu[f^{-1}(A')] \geq 0$ .
- (c) If  $A'_1, A'_2, \dots \in \mathcal{A}'$  are pairwise disjoint, then, according to Equation (2.5), for  $i \neq j$ ,

$$f^{-1}(A'_i) \cap f^{-1}(A'_j) = f^{-1}(A'_i \cap A'_j) = f^{-1}(\emptyset),$$

i. e., the inverse images  $f^{-1}(A'_1), f^{-1}(A'_2), \dots$  are pairwise disjoint as well. Therefore,

$$\mu_f\left(\bigcup_{i=1}^{\infty} A'_i\right) = \mu\left(f^{-1}\left(\bigcup_{i=1}^{\infty} A'_i\right)\right) \quad [(2.41)]$$

$$= \mu\left(\bigcup_{i=1}^{\infty} f^{-1}(A'_i)\right) \quad [(2.6)]$$

$$= \sum_{i=1}^{\infty} \mu(f^{-1}(A'_i)) \quad [\text{Def. 1.43 (c)}]$$

$$= \sum_{i=1}^{\infty} \mu_f(A'_i). \quad [(2.41)]$$

### **Proof of Theorem 2.85**

If  $f \stackrel{\mu}{=} g$ , then there is a set  $A$  satisfying

$$f(\omega) = g(\omega), \quad \forall \omega \in \Omega \setminus A \text{ and } \mu(A) = 0.$$

Monotonicity of  $\mu$  implies  $\mu(\{\omega \in A: f(\omega) \in A'\}) = 0 = \mu(\{\omega \in A: g(\omega) \in A'\})$  for all  $A' \in \mathcal{A}'$ . Hence, using additivity of  $\mu$ ,

$$\begin{aligned}
\mu_f(A') &= \mu[f^{-1}(A')] \\
&= \mu(\{\omega \in \Omega \setminus A: f(\omega) \in A'\}) + \mu(\{\omega \in A: f(\omega) \in A'\}) \\
&= \mu(\{\omega \in \Omega \setminus A: g(\omega) \in A'\}) + \mu(\{\omega \in A: g(\omega) \in A'\}) \\
&= \mu[g^{-1}(A')] = \mu_g(A').
\end{aligned}$$

### ***Proof of Theorem 2.86***

For measurable functions  $g, g^*: (\Omega', \mathcal{A}') \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  define  $A' := \{\omega' \in \Omega': g(\omega') \neq g^*(\omega')\}$ . Note that  $A' \in \mathcal{A}'$  [see Rem. 2.67 (c)]. Then

$$\begin{aligned}
f^{-1}(A') &= \{\omega \in \Omega: f(\omega) \in A'\} \\
&= \{\omega \in \Omega: g[f(\omega)] \neq g^*[f(\omega)]\} \\
&= \{\omega \in \Omega: (g \circ f)(\omega) \neq (g^* \circ f)(\omega)\}.
\end{aligned}$$

Hence,  $g \stackrel{\mu_f}{=} g^* \Leftrightarrow \mu_f(A') = 0 \Leftrightarrow \mu[f^{-1}(A')] = 0 \Leftrightarrow g \circ f \stackrel{\mu}{=} g^* \circ f$ .

## **2.8 Exercises**

▷ **Exercise 2-1** Prove Equations (2.4) to (2.6).

▷ **Exercise 2-2** Consider Example 2.2.1 and compute the inverse images of the sets  $\{(4.5, 0)\}$ ,  $\{(7.5, 0)\}$ ,  $\{(7.5, 4.5)\}$ , and  $\{(4.5, 4.5)\}$  under the function

$$f(x_1, x_2) = \frac{3}{4} \cdot (x_1, x_2) = \left(\frac{3}{4} \cdot x_1, \frac{3}{4} \cdot x_2\right).$$

▷ **Exercise 2-3** Consider Example 2.2.1 and specify the inverse images of the rectangles  $[8, 10] \times [0, 2]$  and  $[3, 7.5] \times [0, 3]$  under the function  $f: \Omega \rightarrow \Omega'$  defined by  $f(x_1, x_2) = \frac{3}{4} \cdot (x_1, x_2)$ .

▷ **Exercise 2-4** Consider Example 2.2.1 and use Equation (2.4) to determine the inverse image  $f^{-1}[(C')^c]$ .

▷ **Exercise 2-5** Prove the proposition of Example 2.14.

▷ **Exercise 2-6** Prove the proposition of Example 2.15.

▷ **Exercise 2-7** Prove the proposition of Example 2.17.

▷ **Exercise 2-8** Prove the proposition of Remark 2.18.

▷ **Exercise 2-9** Prove the proposition formulated in Remark 2.33.

▷ **Exercise 2-10** Consider Example 2.40 and show that  $X$ ,  $Y$ , and  $(X, Y)$  are measurable with respect to  $\mathcal{A}$  whenever the two inverse images  $X^{-1}(\{1\})$  and  $Y^{-1}(\{1\})$  are elements of  $\mathcal{A}$ .

▷ **Exercise 2-11** In Example 2.55 we considered  $\Omega = [0, 4]$ ,  $A_1 = [0, 1]$ ,  $A_2 = ]1, 2]$ ,  $A_3 = ]2, 3]$ , and  $A_4 = ]3, 4]$ . There, we also defined the functions  $f$  and  $h$ . Show that  $\sigma(h) = \sigma(\{A_j \cup A_{j+1}: j \in J\})$  and  $\sigma(f) = \sigma(\{A_i: i = 1, \dots, 4\})$ . Furthermore, show  $\sigma(h) \subset \sigma(f)$ .

- ▷ **Exercise 2-12** Prove the propositions of Remark 2-12.
- ▷ **Exercise 2-13** Prove the proposition of Example 2.56.
- ▷ **Exercise 2-14** Prove the propositions of Remark 2.72.
- ▷ **Exercise 2-15** Consider Remark 2.73 and show: If  $\mathcal{M}$  is a set of mappings  $(\Omega, \mathcal{A}, \mu) \rightarrow \Omega'$ , then  $\equiv_{\mu}$  is an equivalence relation on  $\mathcal{M}$ .
- ▷ **Exercise 2-16** Show that  $\{C(f): f \in \mathcal{M}\}$  is a partition of  $\mathcal{M}$  (see Remark 2.75).
- ▷ **Exercise 2-17** Prove the propositions of Remark 2.76.
- ▷ **Exercise 2-18** Prove proposition (2.40).
- ▷ **Exercise 2-19** Prove the proposition of Example 2.84.

## Solutions

- ▷ **Solution 2-1** Equation (2.4):

$$f^{-1}[(A')^c] = \{\omega \in \Omega: f(\omega) \in (A')^c\} = \{\omega \in \Omega: f(\omega) \notin A'\} = [f^{-1}(A')]^c.$$

Equation (2.5):

$$\begin{aligned} f^{-1}\left(\bigcap_{i \in I} A'_i\right) &= \{\omega \in \Omega: f(\omega) \in \bigcap_{i \in I} A'_i\} = \{\omega \in \Omega: f(\omega) \in A'_i, \forall i \in I\} \\ &= \bigcap_{i \in I} \{\omega \in \Omega: f(\omega) \in A'_i\} = \bigcap_{i \in I} f^{-1}(A'_i). \end{aligned}$$

Equation (2.6):

$$\begin{aligned} f^{-1}\left(\bigcup_{i \in I} A'_i\right) &= \{\omega \in \Omega: f(\omega) \in \bigcup_{i \in I} A'_i\} = \{\omega \in \Omega: \exists i \in I: f(\omega) \in A'_i\} \\ &= \bigcup_{i \in I} \{\omega \in \Omega: f(\omega) \in A'_i\} = \bigcup_{i \in I} f^{-1}(A'_i). \end{aligned}$$

- ▷ **Solution 2-2** The inverse images are the sets  $f^{-1}\{(4.5, 0)\} = \{(6, 0)\}$ ,  $f^{-1}\{(7.5, 0)\} = \{(10, 0)\}$ ,  $f^{-1}\{(7.5, 4.5)\} = \{(10, 6)\}$ , and  $f^{-1}\{(4.5, 4.5)\} = \{(6, 6)\}$ .
- ▷ **Solution 2-3**  $f^{-1}([8, 10] \times [0, 2]) = \emptyset$  and  $f^{-1}([3, 7.5] \times [0, 3]) = [4, 10] \times [0, 4]$ .
- ▷ **Solution 2-4** According to Equation (2.4), the inverse image of  $(C')^c$  under  $f$  is

$$\begin{aligned} f^{-1}[(C')^c] &= [f^{-1}(C')]^c \\ &= \Omega \setminus ([6, 10] \times [0, 6]) \\ &= ([0, 10] \times [0, 6]) \setminus ([6, 10] \times [0, 6]) \\ &= [0, 6] \times [0, 6]. \end{aligned}$$

▷ **Solution 2-5** If  $f: \Omega \rightarrow \Omega'$  is constant, then, according to Example 2.10, it is  $(\mathcal{A}, \mathcal{A}')$ -measurable for  $\mathcal{A} = \{\Omega, \emptyset\}$ . Now, assume that  $f$  is not constant, i. e.,  $\exists \omega_1, \omega_2 \in \Omega: f(\omega_1) \neq f(\omega_2)$ . According to our assumptions,

$$\{f(\omega_1)\}, \{f(\omega_2)\} \in \mathcal{A}'.$$

Furthermore,  $\omega_i \in f^{-1}\{f(\omega_i)\}$ , for  $i = 1, 2$ , i. e., the inverse images are nonempty sets. Now,  $f(\omega_1) \neq f(\omega_2)$  implies

$$\{f(\omega_1)\} \cap \{f(\omega_2)\} = \emptyset,$$

and, using Equation (2.5),

$$f^{-1}(\{f(\omega_1)\}) \cap f^{-1}(\{f(\omega_2)\}) = f^{-1}(\{f(\omega_1)\} \cap \{f(\omega_2)\}) = f^{-1}(\emptyset) = \emptyset.$$

Hence, the inverse images are nonempty disjoint sets, and therefore none of them is in  $\mathcal{A} = \{\Omega, \emptyset\}$ . This implies that  $f$  is not  $(\mathcal{A}, \mathcal{A}')$ -measurable if it is not constant.

▷ **Solution 2-6** We consider  $\{A, A^c, \Omega, \emptyset\}$ . If  $f = \alpha_1 1_A + \alpha_2 1_{A^c}$ , then for all  $A' \in \mathcal{B}$ ,

$$f^{-1}(A') = \begin{cases} \emptyset, & \text{if } \alpha_1 \notin A', \alpha_2 \notin A' \\ A, & \text{if } \alpha_1 \in A', \alpha_2 \notin A' \\ A^c, & \text{if } \alpha_1 \notin A', \alpha_2 \in A' \\ \Omega, & \text{if } \alpha_1 \in A', \alpha_2 \in A'. \end{cases}$$

Hence,  $f$  is  $(\mathcal{A}, \mathcal{B})$ -measurable. (Note that this also holds if  $A = \emptyset$  or  $A = \Omega$ , and also if  $\alpha_1 = \alpha_2$ .)

Now assume that  $f$  is an  $(\mathcal{A}, \mathcal{B})$ -measurable function.

(a) If  $f$  takes one only one single value, say  $\alpha$ , then

$$f = \alpha 1_\Omega = \alpha_1 1_A + \alpha_2 1_{A^c}, \quad \text{with } \alpha_1 = \alpha_2 = \alpha.$$

(b) If  $f$  takes on exactly two different values  $\beta_1 \neq \beta_2$ , then  $f^{-1}(\{\beta_1, \beta_2\}) = f^{-1}(\{\beta_1\}) \cup f^{-1}(\{\beta_2\}) = \Omega$ , and according to Equation (2.5),  $f^{-1}(\{\beta_1\}) \cap f^{-1}(\{\beta_2\}) = \emptyset$ , and  $f^{-1}(\{\beta_i\}) \neq \emptyset$ , for  $i = 1, 2$ . Hence,  $f$  is  $(\mathcal{A}, \mathcal{B})$ -measurable if and only if

$$f^{-1}(\{\beta_1\}) = A \quad \text{or} \quad f^{-1}(\{\beta_1\}) = A^c \quad \text{and} \quad A, A^c \neq \emptyset.$$

This implies

$$f = \beta_1 1_A + \beta_2 1_{A^c} \quad \text{or} \quad f = \beta_2 1_A + \beta_1 1_{A^c},$$

respectively.

(c) If  $f$  takes on *three or more* pairwise different values, then, using the same kind of argument as in (a), we can conclude that there are at least three pairwise disjoint and nonempty inverse images under  $f$ , say  $A_1, A_2, A_3 \subset \Omega$ . Hence, in this case  $f$  is not  $(\mathcal{A}, \mathcal{B})$ -measurable.

▷ **Solution 2-7** If  $A_1, \dots, A_n \in \mathcal{A}$  are pairwise disjoint and we define  $A_{n+1} := \Omega \setminus (\bigcup_{i=1}^n A_i)$ , and  $\alpha_{n+1} := 0$ , then

$$f = \sum_{i=1}^n \alpha_i 1_{A_i} = \sum_{i=1}^{n+1} \alpha_i 1_{A_i}.$$

Because  $A_1, \dots, A_{n+1}$  are pairwise disjoint and  $\bigcup_{i=1}^{n+1} A_i = \Omega$ , there is, for all  $\omega \in \Omega$ , exactly one  $i \in \{1, \dots, n+1\}$  such that  $\omega \in A_i$ , and therefore  $f(\omega) = \alpha_i$ . Hence, the codomain of  $f$  is  $\{\alpha_1, \dots, \alpha_{n+1}\}$ . Vice versa, for all  $\alpha_i$ ,  $i = 1, \dots, n+1$ , we obtain the inverse image

$$f^{-1}(\{\alpha_i\}) = \{\omega \in \Omega: f(\omega) = \alpha_i\} = \bigcup_{j: \alpha_j = \alpha_i} A_j \quad (2.50)$$

(If the  $\alpha_1, \dots, \alpha_{n+1}$  are pairwise different, then  $f^{-1}(\{\alpha_i\}) = A_i$ .) Now, for all  $A' \subset \mathbb{R}$ ,

$$\begin{aligned} f^{-1}(A') &= \{\omega \in \Omega: f(\omega) \in A'\} && \text{[Def. 2.1]} \\ &= \{\omega \in \Omega: f(\omega) \in \bigcup_{i: \alpha_i \in A'} \{\alpha_i\}\} && \text{[codomain of } f \text{ is } \{\alpha_1, \dots, \alpha_{n+1}\}] \\ &= f^{-1}\left(\bigcup_{i: \alpha_i \in A'} \{\alpha_i\}\right) && \text{[Def. 2.1]} \\ &= \bigcup_{i: \alpha_i \in A'} f^{-1}(\{\alpha_i\}) && \text{[(2.6)]} \\ &= \bigcup_{i: \alpha_i \in A'} A_i. && \text{[(2.50)]} \end{aligned}$$

▷ **Solution 2-8** If  $A_1, \dots, A_n \in \mathcal{A}$  are pairwise disjoint and if we define  $A_{n+1} := \Omega \setminus (\bigcup_{i=1}^n A_i)$ , and  $\alpha_{n+1} := 0$ , then

$$\forall A' \in \mathcal{B}: f^{-1}(A') = \bigcup_{\substack{i=1, \dots, n+1, \\ \alpha_i \in A'}} A_i \in \mathcal{A},$$

[see Eq. (2.10)].

If  $A_1, \dots, A_n \in \mathcal{A}$  are *not* pairwise disjoint, define the  $2^n$  sets

$$B_j := A_1^{c_1(j)} \cap \dots \cap A_n^{c_n(j)} \in \mathcal{A}, \quad j = 1, \dots, 2^n,$$

with  $(c_1(j), \dots, c_n(j)) \in \{0, 1\}^n$  and

$$A_i^0 := A_i, \quad A_i^1 := A_i^c.$$

Note that some of the sets  $B_j$  can be empty. Then

$$f = \sum_{i=1}^n \alpha_i 1_{A_i} = \sum_{j=1}^{2^n} \beta_j 1_{B_j},$$

with  $\beta_j = \sum_{i: c_i(j)=0} \alpha_i$ . Because  $B_1, \dots, B_{2^n}$  are pairwise disjoint and  $\bigcup_{j=1}^{2^n} B_j = \Omega$ , the function  $f$  is  $(\mathcal{A}, \mathcal{B})$ -measurable (see the first part of this solution).

▷ **Solution 2-9** If all values of  $f$  are elements of  $\Omega'$ , then

$$f^{-1}(B) = f^{-1}(\Omega' \cap B), \quad \forall B \in \mathcal{B}.$$

Therefore,  $f^{-1}(\mathcal{B}) = f^{-1}(\mathcal{B}|_{\Omega'})$ , where  $\mathcal{B}|_{\Omega'}$  denotes the trace of  $\mathcal{B}$  in  $\Omega'$  (see Example 1.10). Note that  $\mathcal{B}|_{\Omega'} = \mathcal{P}(\Omega')$  (see Exercise 1-13). Hence,  $f^{-1}(\mathcal{B}) = f^{-1}(\mathcal{B}|_{\Omega'}) = f^{-1}[\mathcal{P}(\Omega')]$ .

▷ **Solution 2-10** First of all note that  $X, Y$  are measurable with respect to  $\mathcal{A}$  if and only if and  $(X, Y)$  is measurable with respect to  $\mathcal{A}$  (see Theorem 2.38). Hence, it suffices to show that  $X$  is measurable with respect to  $\mathcal{A}$  if  $X^{-1}(\{1\}) \in \mathcal{A}$ . Because  $X: \Omega \rightarrow \mathbb{R}$  is an indicator,  $X^{-1}(\{0\}) = X^{-1}(\{1\}^c) = (X^{-1}(\{1\}))^c$  [see Eq. (2.5)]. Hence, if  $X^{-1}(\{1\}) \in \mathcal{A}$ , then  $(X^{-1}(\{1\}))^c = X^{-1}(\{0\}) \in \mathcal{A}$ . Furthermore, for all  $B \in \mathcal{B}$ ,

$$X^{-1}(B) = \begin{cases} \emptyset, & \text{if } 0 \notin B, 1 \notin B, \\ X^{-1}(\{1\}), & \text{if } 0 \notin B, 1 \in B \\ \Omega \setminus X^{-1}(\{1\}), & \text{if } 0 \in B, 1 \notin B \\ \Omega, & \text{if } \{0, 1\} \subset B. \end{cases}$$

(The proof for  $Y$  is analog.)

▷ **Solution 2-11** Because the codomain of  $f$  is  $\mathbb{R}$ ,

$$\begin{aligned} \sigma(f) &= \sigma\left(f^{-1}(\mathcal{P}(\{1, 2.5, 2, 0.5\}))\right) && \text{[Rem. 2.33]} \\ &= \sigma\left(\{f^{-1}(\{1\}), f^{-1}(\{2.5\}), f^{-1}(\{2\}), f^{-1}(\{0.5\})\}\right) && \text{[(2.12)]} \\ &= \sigma(\{A_1, A_2, A_3, A_4\}). \end{aligned}$$

Analogously,

$$\begin{aligned} \sigma(h) &= \sigma\left(h^{-1}(\mathcal{P}(\{1.5, 3\}))\right) && \text{[Rem. 2.33]} \\ &= \sigma\left(\{h^{-1}(\{1.5\}), h^{-1}(\{3\})\}\right) && \text{[(2.12)]} \\ &= \sigma(\{A_1 \cup A_2, A_3 \cup A_4\}). \end{aligned}$$

Because

$$\{A_1 \cup A_2, A_3 \cup A_4\} \subset \sigma(\{A_1, A_2, A_3, A_4\}) \quad \text{[Rem. 1.2]}$$

monotonicity of the generated  $\sigma$ -algebras (see Remark 1.23) implies  $\sigma(h) \subset \sigma(f)$ .

▷ **Solution 2-12** Denote  $A_\infty := \{\omega \in \Omega: f(\omega) = \infty\}$ ,  $A_{-\infty} := \{\omega \in \Omega: f(\omega) = -\infty\}$ ,  $B_\infty := \{\omega \in \Omega: g(\omega) = \infty\}$ , and  $B_{-\infty} := \{\omega \in \Omega: g(\omega) = -\infty\}$ . Because  $\{\infty, -\infty\} \in \overline{\mathcal{B}}$ , all four sets defined above are elements of  $\mathcal{A}$ . Furthermore,

$$A := \{\omega \in \Omega: -\infty < f(\omega) < \infty\} = f^{-1}(\mathbb{R}) \in \mathcal{A}$$

and

$$B := \{\omega \in \Omega: -\infty < g(\omega) < \infty\} = g^{-1}(\mathbb{R}) \in \mathcal{A}.$$

(a) Now

$$\begin{aligned} \{\omega \in \Omega: f(\omega) \geq g(\omega)\} &= A_\infty \cup \{\omega \in A \cap B: f(\omega) \geq g(\omega)\} \cup B_{-\infty} \\ &= A_\infty \cup \{\omega \in A \cap B: f(\omega) - g(\omega) \geq 0\} \cup B_{-\infty} \\ &= A_\infty \cup [1_{A \cap B} \cdot (f - g)]^{-1}([0, \infty[) \cup B_{-\infty} && \text{[Def. 2.1]} \\ &\in \mathcal{A}. && \text{[Th. 2.57]} \end{aligned}$$

(b) Analogously,

$$\begin{aligned} \{\omega \in \Omega: f(\omega) > g(\omega)\} &= (A_\infty \cap (B \cup B_{-\infty})) \cup \{\omega \in A \cap B: f(\omega) > g(\omega)\} \\ &= (A_\infty \cap (B \cup B_{-\infty})) \cup [1_{A \cap B} \cdot (f - g)]^{-1}([0, \infty[) && \text{[Def. 2.1]} \\ &\in \mathcal{A}. && \text{[Th. 2.57]} \end{aligned}$$

(c) Finally,

$$\begin{aligned} \{\omega \in \Omega: f(\omega) = g(\omega)\} &= \{\omega \in \Omega: f(\omega) \geq g(\omega)\} \setminus \{\omega \in \Omega: f(\omega) > g(\omega)\} \\ &\in \mathcal{A}. && \text{[Rem. 1.2]} \end{aligned}$$

▷ **Solution 2-13** For any real-valued measurable function  $f: (\Omega, \mathcal{A}) \rightarrow (\mathbb{R}, \mathcal{B})$ , Lemma 2.52 yields  $\sigma(f^2) \subset \sigma(f)$ , because  $f^2 = g(f)$  for  $g: (\mathbb{R}, \mathcal{B}) \rightarrow (\mathbb{R}, \mathcal{B})$  with the measurable function  $g(x) = x^2$ , for all  $x \in \mathbb{R}$ . [Note that  $g$  is a continuous function that is  $\mathcal{B}$ -measurable (see Klenke, 2013, Th. 1.88, p. 36)].

(i) We assume that  $f$  is nonnegative and measurable. Then  $f^2(\omega) = x$  if and only if  $f(\omega) = \sqrt{x}$ , for all  $x \geq 0$ . Hence, for all  $A \in \mathcal{A}$ ,

$$\begin{aligned} A \in \sigma(f) &\Rightarrow \exists B_1 \in \mathcal{B}: A = f^{-1}(B_1) \\ &\Rightarrow \exists B_2 \in \mathcal{B}: A = (f^2)^{-1}(B_2) && [\text{choose } B_2 := g^{-1}(B_1)] \\ &\Rightarrow A \in \sigma(f^2). \end{aligned}$$

This implies  $\sigma(f) \subset \sigma(f^2)$ .

(ii) Assume that there are  $\omega_1, \omega_2 \in \Omega$  with  $f(\omega_1) < 0 < f(\omega_2)$  and  $f^2(\omega_1) = f^2(\omega_2)$ . Then  $A := f^{-1}(-\infty, 0)$  implies  $A \in \sigma(f^2)$ , and  $\omega_1 \in A$  and  $\omega_2 \notin A$ . Furthermore, for all  $B \in \mathcal{B}: \{\omega_1, \omega_2\} \subset (f^2)^{-1}(B)$  if  $f^2(\omega_1) \in B$  and  $\{\omega_1, \omega_2\} \cap (f^2)^{-1}(B) = \emptyset$  if  $f^2(\omega_1) \notin B$ . Hence,  $A \notin (f^2)^{-1}(\mathcal{B})$ .

▷ **Solution 2-14** (i) If  $f \stackrel{\mu}{=} g$ , then there is a set  $B \in \mathcal{A}$  with  $\mu(B) = 0$  and  $f(\omega) = g(\omega)$  for all  $\omega \in \Omega \setminus B$ . Hence,  $1_A(\omega) \cdot f(\omega) = 1_A(\omega) \cdot g(\omega)$  for all  $\omega \in \Omega \setminus B$ . According to Definition 2.68,  $1_A f \stackrel{\mu}{=} 1_A g$ .

(ii) Note that  $A_h := \{\omega \in \Omega: h[f(\omega)] \neq h[g(\omega)]\} \subset \{\omega \in \Omega: f(\omega) \neq g(\omega)\} =: A$ . Therefore,  $\mu(A) = 0$  implies  $\mu(A_h) = 0$  [see Box 1.1 (v)].

▷ **Solution 2-15** Reflexivity.  $\mu(\{\omega \in \Omega: f(\omega) \neq f(\omega)\}) = \mu(\emptyset) = 0$ . Hence,  $f \stackrel{\mu}{=} f$ .

Symmetry. Assume that  $f, g \in \mathcal{M}$  and  $f \stackrel{\mu}{=} g$ . Then

$$\mu(\{\omega \in \Omega: g(\omega) \neq f(\omega)\}) = \mu(\{\omega \in \Omega: f(\omega) \neq g(\omega)\}) = 0.$$

Hence,  $g \stackrel{\mu}{=} f$ .

Transitivity. Assume that  $f, g, h \in \mathcal{M}$ ,  $f \stackrel{\mu}{=} g$ , and  $g \stackrel{\mu}{=} h$ . Then transitivity of  $=$  and subadditivity of  $\mu$  yield

$$\begin{aligned} \mu(\{\omega \in \Omega: f(\omega) \neq h(\omega)\}) &\leq \mu(\{\omega \in \Omega: f(\omega) \neq g(\omega)\} \cup \{\omega \in \Omega: g(\omega) \neq h(\omega)\}) \\ &\leq \mu(\{\omega \in \Omega: f(\omega) \neq g(\omega)\}) + \mu(\{\omega \in \Omega: g(\omega) \neq h(\omega)\}) \\ &= 0 + 0 \quad [f \stackrel{\mu}{=} g, g \stackrel{\mu}{=} h] \\ &= 0. \end{aligned}$$

Therefore,  $f \stackrel{\mu}{=} h$ .

▷ **Solution 2-16** (a)  $\forall f \in \mathcal{M}: f \stackrel{\mu}{=} f$  (reflexivity). This implies:  $\forall f \in \mathcal{M}: f \in C(f)$  and therefore  $\forall f \in \mathcal{M}: C(f) \neq \emptyset$ .

(b) Let  $f, g \in \mathcal{M}$ . We consider two cases,  $f \stackrel{\mu}{=} g$  and  $\neg(f \stackrel{\mu}{=} g)$ .  
 $f \stackrel{\mu}{=} g$ . Transitivity implies:  $\forall h \in \mathcal{M}: f \stackrel{\mu}{=} h$  if and only if  $g \stackrel{\mu}{=} h$ . Hence,  $\forall h \in \mathcal{M}: h \in C(f)$  if and only if  $h \in C(g)$ . This means that  $C(f) = C(g)$ .

$\neg(f \stackrel{\mu}{=} g)$ . We show  $C(f) \cap C(g) = \emptyset$  by contraposition. Assume:  $\exists h \in \mathcal{M}: h \in C(f) \cap C(g)$ . Then  $f \stackrel{\mu}{=} h$ ,  $h \stackrel{\mu}{=} g$ , and transitivity implies:  $f \stackrel{\mu}{=} g$ , which is a contradiction to  $\neg(f \stackrel{\mu}{=} g)$ .

(c) In part (a) we showed that  $\forall f \in \mathcal{M}: f \in C(f)$ . Therefore,  $\forall f \in \mathcal{M}: f \in \bigcup_{f^* \in \mathcal{M}} C(f^*)$ .

▷ **Solution 2-17** (i) Suppose that  $\mu \neq 0$ ,  $\alpha, \beta \in \Omega'$  and  $f \stackrel{\mu}{=} \alpha \wedge g \stackrel{\mu}{=} \beta \wedge f \stackrel{\mu}{=} g$ . If  $\alpha \neq \beta$ , then subadditivity and monotonicity of  $\mu$  yield

$$\begin{aligned}
0 &< \mu(\Omega) \\
&= \mu(\{\omega \in \Omega: f(\omega) = \alpha \wedge g(\omega) = \beta\}) + \mu(\{\omega \in \Omega: f(\omega) \neq \alpha \vee g(\omega) \neq \beta\}) \quad [\text{Box 1.1 (iv)}] \\
&= \mu(\{\omega \in \Omega: f(\omega) = \alpha \wedge g(\omega) = \beta\}) \quad [f \stackrel{\mu}{=} \alpha, g \stackrel{\mu}{=} \beta] \\
&\leq \mu(\{\omega \in \Omega: f(\omega) \neq g(\omega)\}) \quad [\alpha \neq \beta] \\
&= 0, \quad [f \stackrel{\mu}{=} g]
\end{aligned}$$

which proves (i) by contraposition.

(ii) If  $f \stackrel{\mu}{=} f^*$  and  $g \stackrel{\mu}{=} g^*$  and  $A_0 := \{\omega \in \Omega: f(\omega) \neq f^*(\omega) \vee g(\omega) \neq g^*(\omega)\}$ , then

$$\begin{aligned}
\mu(A_0) &= \mu(\{\omega \in \Omega: f(\omega) \neq f^*(\omega)\} \cup \{\omega \in \Omega: g(\omega) \neq g^*(\omega)\}) \\
&\leq \mu(\{\omega \in \Omega: f(\omega) \neq f^*(\omega)\}) + \mu(\{\omega \in \Omega: g(\omega) \neq g^*(\omega)\}) \quad [\text{Box 1.1 (vii), (v)}] \\
&= 0. \quad [f \stackrel{\mu}{=} f^*, g \stackrel{\mu}{=} g^*]
\end{aligned}$$

Note that  $\{\omega \in \Omega: f(\omega) + g(\omega) \neq f^*(\omega) + g^*(\omega)\} \subset A_0$ , and that this also holds for the corresponding sets for the difference, product, and ratio. This implies Equations (2.36) and (2.37).

(iii), (iv) Denote  $I := \{i = 1, \dots, n\}$  for (iii) and  $I := \mathbf{N}$  for (iv), respectively. Furthermore, define

$$A_0 := \bigcup_{i \in I} \{\omega \in \Omega: f_i(\omega) \neq f_i^*(\omega)\} = \{\omega \in \Omega: \exists i \in I: f_i(\omega) \neq f_i^*(\omega)\}.$$

Then

$$\begin{aligned}
\mu(A_0) &\leq \sum_{i \in I} \mu(\{\omega \in \Omega: f_i(\omega) \neq f_i^*(\omega)\}) \quad [\text{Box 1.1 (xi)}] \\
&= 0, \quad \text{if } f_i \stackrel{\mu}{=} f_i^*, \forall i \in I.
\end{aligned}$$

Hence,

$$\begin{aligned}
\mu\left(\left\{\omega \in \Omega: \sum_{i \in I} \alpha_i f_i(\omega) \neq \sum_{i \in I} \alpha_i f_i^*(\omega)\right\}\right) &\leq \mu(A_0) \quad [\text{Box 1.1 (v)}] \\
&= 0, \quad \text{if } f_i \stackrel{\mu}{=} f_i^*, \forall i \in I.
\end{aligned}$$

▷ **Solution 2-18**

$$\forall \omega \in \Omega: (f(\omega) < g(\omega) \wedge g(\omega) = h(\omega)) \Rightarrow f(\omega) < h(\omega),$$

which, by contraposition, is equivalent to

$$\forall \omega \in \Omega: f(\omega) \geq h(\omega) \Rightarrow (f(\omega) \geq g(\omega) \vee g(\omega) \neq h(\omega)).$$

Therefore,

$$\{\omega \in \Omega: f(\omega) \geq h(\omega)\} \subset \{\omega \in \Omega: f(\omega) \geq g(\omega)\} \cup \{\omega \in \Omega: g(\omega) \neq h(\omega)\}.$$

Now  $(f \stackrel{\mu}{<} g \wedge g \stackrel{\mu}{=} h)$  implies

$$\mu(\{\omega \in \Omega: f(\omega) \geq h(\omega)\})$$

$$\begin{aligned}
&\leq \mu(\{\omega \in \Omega: f(\omega) \geq g(\omega)\}) + \mu(\{\omega \in \Omega: g(\omega) \neq h(\omega)\}) && \text{[Box 1.1 (vii)]} \\
&= 0 + 0 = 0. && [f \stackrel{\mu}{\leq} g, g \stackrel{\mu}{=} h]
\end{aligned}$$

Because a measure is nonnegative, this implies  $\mu(\{\omega \in \Omega: f(\omega) \geq h(\omega)\}) = 0$ , which is equivalent to  $f \stackrel{\mu}{<} h$ .

▷ **Solution 2-19** For all  $A' \in \mathcal{A}'$ ,

$$\mu_f(A') = \mu[f^{-1}(A')] \quad [(2.41)]$$

$$= \mu \left( \bigcup_{\substack{i=1, \dots, n+1, \\ \alpha_i \in A'}} A_i \right) \quad \text{[Def. 2.16]}$$

$$= \sum_{\substack{i=1, \dots, n+1, \\ \alpha_i \in A'}} \mu(A_i) \quad \text{[Def. 1.43, (c)]}$$

$$= \sum_{i=1}^{n+1} \mu(A_i) \cdot \delta_{\alpha_i}(A'). \quad [(1.39)]$$

## Chapter 3

# Integral

In the preceding chapters we introduced the most important concepts of measure theory related to the concepts of a measure and a measurable mapping. In this chapter we introduce the *integral* of measurable functions. This concept is fundamental also for probability theory, because the expectation of a numerical random variable with respect to a probability measure is the *integral* of a measurable function with respect to a probability measure. In chapter 6 we shall see that this also applies to variances, covariances, and correlations. We start defining the integral of a measurable function with respect to a measure  $\mu$ . Then we study the most important rules of computation and other properties of integrals, introduce the concept of a *measure with density*, and treat the relationship between the Riemann integral and the integral with respect to the Lebesgue measure. The next section is on *absolute continuity* and the *Radon-Nikodym Theorem*. Both issues are crucial for conditional expectations (see ch. 10). A section on the integral with respect to a product measure concludes this chapter.

### 3.1 Definition

At first we define the integral for *nonnegative step functions*, then we extend the integral to *nonnegative measurable functions*, and finally we introduce the integral for *measurable functions* that may take on negative or nonnegative values.

#### 3.1.1 Integral of a Nonnegative Step Function

In this subsection we introduce the integral of a *nonnegative step function*, also called *nonnegative simple function* or *elementary function*.

#### Nonnegative Step Function

**Definition 3.1 (Nonnegative Step Function and Normal Representation)**

Let  $(\Omega, \mathcal{A})$  be a measurable space. Then  $f: \Omega \rightarrow \mathbb{R}$  is called a *nonnegative step function*, if there is a finite sequence  $A_1, \dots, A_n \in \mathcal{A}$  and a finite sequence  $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ ,  $\alpha_i \geq 0$ ,  $i = 1, \dots, n$ , such that

$$f = \sum_{i=1}^n \alpha_i 1_{A_i}. \quad (3.1)$$

If  $A_1, \dots, A_n \in \mathcal{A}$  are pairwise disjoint, then  $f = \sum_{i=1}^n \alpha_i 1_{A_i}$  is called a normal representation of  $f$ .

**Remark 3.2 (Step Functions Take on Finitely Many Values)** A nonnegative step function  $f = \sum_{i=1}^n \alpha_i 1_{A_i}$  is a measurable function  $f: (\Omega, \mathcal{A}) \rightarrow (\mathbb{R}, \mathcal{B})$  taking on only a finite number of *nonnegative* values. These values are not necessarily  $\alpha_1, \dots, \alpha_n$ . However, note:

- (i) If  $\mathcal{E} = \{A_1, \dots, A_n\}$  is a partition of  $\Omega$ , then  $\alpha_1, \dots, \alpha_n$  are the values of  $f$ .
- (ii) If  $A_1, \dots, A_n$  are pairwise disjoint but  $\mathcal{E}$  is not a partition of  $\Omega$ , i. e.,  $A_{n+1} := \Omega \setminus \bigcup_{i=1}^n A_i \neq \emptyset$ , then

$$f = \sum_{i=1}^n \alpha_i 1_{A_i} + 0 \cdot 1_{A_{n+1}}.$$

This implies:  $f(\omega) = 0$ , for all  $\omega \in A_{n+1}$ .

- (iii) If  $A_1, \dots, A_n$  are pairwise disjoint and additionally  $\alpha_1, \dots, \alpha_n$  are pairwise different and not 0, then  $A_i = f^{-1}(\{\alpha_i\})$ ,  $i = 1, \dots, n$ .
- (iv) If  $A_1, \dots, A_n$  are pairwise disjoint, then, for all  $\alpha_i \neq 0$ ,

$$f^{-1}(\{\alpha_i\}) = \bigcup_{j: \alpha_j = \alpha_i} A_j.$$

Hence in this case, the inverse image of the set  $\{\alpha_i\}$  under  $f$  is the union of all sets  $A_j$ ,  $j \in \{1, \dots, n\}$ , for which  $\alpha_j = \alpha_i$ .

◁

**Remark 3.3 (Different Representations of a Nonnegative Step Function)** Note that nonnegative step functions can have different representations and also different normal representations (see Example 3.7). ◁

**Example 3.4 (Indicator Function)** Let  $(\Omega, \mathcal{A})$  be a measurable space and  $A \in \mathcal{A}$ . The indicator function  $1_A$ , which has already been introduced in Example 2.12, can also be written  $1 \cdot 1_A + 0 \cdot 1_{A^c}$ . Hence, because  $A \in \mathcal{A}$  and 1 is a real number,  $1_A$  is a nonnegative step function. Note that  $1 \cdot 1_A$  is also a normal representation of a nonnegative step function. ◁

**Example 3.5 (Two Nonnegative Step Functions)** In Example 2.55 we already presented two nonnegative step functions  $f$  and  $h$  and an illustrating Figure (see Fig. 2.6). The representations of both functions are normal. ◁

**Example 3.6 (Tossing a Dice)** Consider the set  $\Omega = \{\omega_1, \dots, \omega_6\}$  of possible outcomes of tossing a dice. Furthermore, let  $\mathcal{A} = \mathcal{P}(\Omega)$  be the power set of  $\Omega$ , and define  $X: \Omega \rightarrow \mathbb{R}$  by

$$X(\omega_i) = i, \quad \forall \omega_i \in \Omega.$$

Hence,  $X(\omega_i)$  is the *number of dots*. Considering the elements  $\{\omega_1\}, \dots, \{\omega_6\}$  of  $\mathcal{A}$ , and

$$X = \sum_{i=1}^6 i \cdot \mathbf{1}_{\{\omega_i\}}$$

shows that  $X$  has a normal representation of a nonnegative step function. (For a related example see Exercise 3-1.)  $\triangleleft$

**Example 3.7 (Several Representations of Nonnegative Step Functions)** Consider the measurable space  $(\mathbb{R}, \mathcal{B})$  and the nonnegative function  $f: \mathbb{R} \rightarrow \mathbb{R}$  defined by

$$f(x) = \begin{cases} 2, & \text{if } x \in [0, 1[ \\ 5, & \text{if } x \in [1, 2] \\ 4, & \text{if } x \in ]2, 3] \\ 1, & \text{if } x \in ]3, 4] \\ 0, & \text{otherwise.} \end{cases}$$

This function can also be represented by

$$\begin{aligned} f &= 2 \cdot \mathbf{1}_{[0,1[} + 5 \cdot \mathbf{1}_{[1,2]} + 4 \cdot \mathbf{1}_{]2,3]} + 1 \cdot \mathbf{1}_{]3,4]} \\ &= 2 \cdot \mathbf{1}_{[0,.5]} + 2 \cdot \mathbf{1}_{].5,1]} + 5 \cdot \mathbf{1}_{[1,2]} + 4 \cdot \mathbf{1}_{]2,3]} + 1 \cdot \mathbf{1}_{]3,4]} \\ &= 2 \cdot \mathbf{1}_{[0,2]} + 3 \cdot \mathbf{1}_{[1,3]} + 1 \cdot \mathbf{1}_{[2,4]} \\ &= 1 \cdot \mathbf{1}_{[0,4]} + 1 \cdot \mathbf{1}_{[0,3]} + 2 \cdot \mathbf{1}_{[1,3]} + 1 \cdot \mathbf{1}_{[1,2]}. \end{aligned} \quad (3.2)$$

The first two representations are normal, the latter two are nonnormal representations of  $f$ .  $\triangleleft$

**Remark 3.8 (Existence of a Normal Representation)** For every nonnegative step function there exists a normal representation (see Exercise 3-2).

If  $f = \sum_{i=1}^n \alpha_i \mathbf{1}_{A_i}$  is a normal representation of a nonnegative step function, then there may be another sequence  $C_1, \dots, C_m$  of pairwise disjoint elements of  $\mathcal{A}$  and another sequence  $\gamma_1, \dots, \gamma_m$  of nonnegative real numbers such that

$$f = \sum_{i=1}^n \alpha_i \mathbf{1}_{A_i} = \sum_{i=1}^m \gamma_i \mathbf{1}_{C_i}.$$

Both sum terms are normal representations. The first two representations of  $f$  in Equation (3.2) provide an example.  $\triangleleft$

### Integral of a Nonnegative Step Function

The following uniqueness property holds for two normal representations of a nonnegative step function:

**Lemma 3.9 (A Uniqueness Property)**

Let  $(\Omega, \mathcal{A}, \mu)$  be a measure space. If  $f: \Omega \rightarrow \mathbb{R}$  is a nonnegative step function and  $f = \sum_{i=1}^n \alpha_i 1_{A_i} = \sum_{i=1}^m \gamma_i 1_{C_i}$  are two normal representations, then

$$\sum_{i=1}^n \alpha_i \mu(A_i) = \sum_{i=1}^m \gamma_i \mu(C_i). \quad (3.3)$$

For a proof see Klenke (2013, Lemma 4.1, p. 85). Note, by convention  $0 \cdot \infty = 0$ .

According to this lemma, the number  $\sum_{i=1}^n \alpha_i \mu(A_i)$  assigned to a nonnegative step function  $f$  does not depend on the specific normal representation of  $f$  (for an illustration see Exercise 3-3). This property allows us to define the *integral* of a nonnegative step function with respect to a measure  $\mu$  as follows:

**Definition 3.10 (Integral of a Nonnegative Step Function)**

Let  $(\Omega, \mathcal{A}, \mu)$  be a measure space and let  $f = \sum_{i=1}^n \alpha_i 1_{A_i}$  be a normal representation of a nonnegative step function  $f: \Omega \rightarrow \mathbb{R}$ . Then the number

$$\int f d\mu = \sum_{i=1}^n \alpha_i \mu(A_i) \quad (3.4)$$

is called the *integral of  $f$  (over  $\Omega$ ) with respect to  $\mu$* .

**Remark 3.11 (Integral of a Constant)** Let  $(\Omega, \mathcal{A}, \mu)$  be a measure space. If  $f = \alpha$ ,  $\alpha \in \mathbb{R}$ , then Equation (3.4) immediately implies

$$\int \alpha d\mu = \alpha \cdot \mu(\Omega). \quad (3.5)$$

◁

**Remark 3.12 (Integral Over a Subset of  $\Omega$ )** Let  $A \in \mathcal{A}$ . If  $f = \sum_{i=1}^n \alpha_i 1_{A_i}$  is a normal representation of a nonnegative step function, then the product  $1_A \cdot f$  is a nonnegative step function as well and can be written

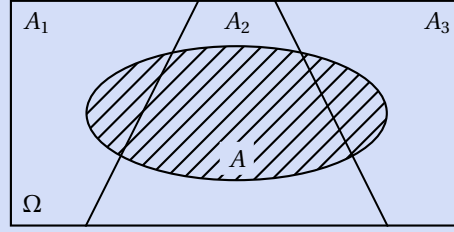
$$1_A \cdot f = \sum_{i=1}^n \alpha_i 1_{A \cap A_i}, \quad (3.6)$$

which is a normal representation of  $1_A \cdot f$  (see Fig. 3.1 and Exercise 3-4). Hence, we may also consider the integral  $\int 1_A \cdot f d\mu$  and define the *integral of  $f$  over a subset  $A$  of  $\Omega$*  by

$$\int_A f d\mu = \int 1_A \cdot f d\mu. \quad (3.7)$$

◁

The following corollary is a special case of Equation (3.7) if  $f = \alpha$ ,  $\alpha \in \mathbb{R}$ . (For proof see Exercise 3-5).



**Figure 3.1.** A partition and a subset of  $\Omega$

**Corollary 3.13 (Constants)**

Let  $(\Omega, \mathcal{A}, \mu)$  be a measure space and  $\alpha \in \mathbb{R}$ . If  $A \in \mathcal{A}$ , then

$$\int_A \alpha \, d\mu = \alpha \mu(A). \quad (3.8)$$

**Examples**

**Example 3.14 (Indicator Function)** Consider a measure space  $(\Omega, \mathcal{A}, \mu)$  and the indicator  $1_A$  of  $A \in \mathcal{A}$ . Then  $1_A = 1 \cdot 1_A$  is a normal representation of  $1_A$ . Therefore,

$$\int_A d\mu = \int_A 1_A \, d\mu = 1 \cdot \mu(A) = \mu(A). \quad (3.9)$$

◁

**Example 3.15 (Nonnegative Step Function and Dirac Measure)** Let  $(\Omega, \mathcal{A})$  be a measurable space, and for  $\omega \in \Omega$  let  $\delta_\omega$  denote the *Dirac measure* at  $\omega$  (see Example 1.52). Furthermore, consider a normal representation  $f = \sum_{i=1}^m \alpha_i 1_{A_i}$  of a nonnegative step function. Its integral with respect to the Dirac measure is

$$\int f \, d\delta_\omega = \sum_{i=1}^m \alpha_i \delta_\omega(A_i) = \sum_{i=1}^m \alpha_i 1_{A_i}(\omega) = f(\omega). \quad (3.10)$$

According to this equation, the integral of a nonnegative step function  $f$  with respect to the Dirac measure at  $\omega$  is the value of  $f$  for the argument  $\omega$ . Furthermore, if  $f = 1_A$  is the indicator of  $A \in \mathcal{A}$ , then

$$\int 1_A \, d\delta_\omega = 1_A(\omega). \quad (3.11)$$

Hence, in this special case, the integral is the value of the indicator  $1_A$  for the argument  $\omega$ . ◁

**Example 3.16 (Nonnegative Step Function and Counting Measure)** Suppose  $\Omega = \{1, \dots, n\}$ ,  $n \in \mathbb{N}$ . For the measurable space  $(\Omega, \mathcal{P}(\Omega))$ , the counting measure  $\mu_{\#}$  on the power set  $\mathcal{P}(\Omega)$  is defined by

$$\mu_{\#}(A) = \sum_{\omega=1}^n 1_A(\omega), \quad \forall A \subset \Omega, \quad (3.12)$$

(see Example 1.54). Hence,  $\mu_{\#}(A)$  is simply the number of elements, i. e., the cardinality of  $A$ . Now consider a nonnegative step function with normal representation  $f = \sum_{i=1}^m \alpha_i 1_{A_i}$ . According to Equations (3.4) and (3.12), its integral with respect to the counting measure is

$$\begin{aligned} \int f d\mu_{\#} &= \sum_{i=1}^m \alpha_i \mu_{\#}(A_i) = \sum_{i=1}^m \alpha_i \sum_{\omega=1}^n 1_{A_i}(\omega) = \sum_{\omega=1}^n \sum_{i=1}^m \alpha_i 1_{A_i}(\omega) \\ &= \sum_{\omega=1}^n f(\omega). \end{aligned} \quad (3.13)$$

Hence, the integral of a nonnegative step function  $f$  with respect to the counting measure  $\mu_{\#}$  is the sum over all values of  $f$  (see also Exercise 3-6). Using Equations (1.41) and (3.10), this integral can also be written

$$\int f d\mu_{\#} = \int f d\left(\sum_{\omega=1}^n \delta_{\omega}\right) = \sum_{\omega=1}^n \int f d\delta_{\omega}. \quad (3.14)$$

<

### 3.1.2 Integral of a Nonnegative Measurable Function

In this section we extend the concept of an integral to nonnegative measurable functions. Before introducing the definition we consider a theorem according to which every nonnegative measurable function can be represented as a limit of an increasing sequence of nonnegative step functions. We begin with an example.

**Example 3.17 (Increasing Sequence of Nonnegative Step Functions)** Consider the measurable space  $(\mathbb{R}, \mathcal{B})$  and the function  $f: \mathbb{R} \rightarrow \mathbb{R}$  defined by

$$f(x) = \begin{cases} 1 - x^2, & \forall x \in [0, 1], \\ 0, & \text{otherwise.} \end{cases} \quad (3.15)$$

Now we construct three functions  $f_i: \mathbb{R} \rightarrow \mathbb{R}$ ,  $i = 1, 2, 3$ , with  $f_1 \leq f_2 \leq f_3 \leq f$  that approximate  $f$  (see Fig. 3.2). Let us start with  $f_1$  defined by

$$f_1(x) = \begin{cases} \alpha_1 = .50, & \text{if } x \in A_1 = [0, (1 - .50)^{1/2}], \\ \alpha_2 = 0, & \text{if } x \in A_1^c, \end{cases}$$

where  $[0, (1 - .50)^{1/2}]$  denotes the closed interval between 0 and  $(1 - .50)^{1/2} \approx .707$ . Because  $A_1$  is an element of  $\mathcal{B}$  and .50 is a nonnegative real number,  $f_1 = \alpha_1 1_{A_1}$  is a nonnegative step function. Next consider  $f_2$  defined by

$$f_2(x) = \begin{cases} \beta_1 = .75, & \text{if } x \in B_1 = [0, (1 - .75)^{1/2}], \\ \beta_2 = .50, & \text{if } x \in B_2 = ](1 - .75)^{1/2}, (1 - .50)^{1/2}], \\ \beta_3 = .25, & \text{if } x \in B_3 = ](1 - .50)^{1/2}, (1 - .25)^{1/2}], \\ \beta_4 = 0, & \text{if } x \in (B_1 \cup B_2 \cup B_3)^c. \end{cases}$$

Because  $B_1, B_2, B_3$  are elements of  $\mathcal{B}$  and  $.75, .50, .25$  are nonnegative real numbers,  $f_2 = \sum_{i=1}^3 \beta_i 1_{B_i}$  is a nonnegative step function. Finally, let  $f_3$  be defined by

$$f_3(x) = \begin{cases} \gamma_1 = .875, & \text{if } x \in C_1 = [0, (1 - .875)^{1/2}], \\ \gamma_2 = .750, & \text{if } x \in C_2 = ](1 - .875)^{1/2}, (1 - .750)^{1/2}], \\ \gamma_3 = .625, & \text{if } x \in C_3 = ](1 - .750)^{1/2}, (1 - .625)^{1/2}], \\ \gamma_4 = .500, & \text{if } x \in C_4 = ](1 - .625)^{1/2}, (1 - .500)^{1/2}], \\ \gamma_5 = .375, & \text{if } x \in C_5 = ](1 - .500)^{1/2}, (1 - .375)^{1/2}], \\ \gamma_6 = .250, & \text{if } x \in C_6 = ](1 - .375)^{1/2}, (1 - .250)^{1/2}], \\ \gamma_7 = .125, & \text{if } x \in C_7 = ](1 - .250)^{1/2}, (1 - .125)^{1/2}], \\ \gamma_8 = 0, & \text{if } x \in (C_1 \cup \dots \cup C_7)^c. \end{cases}$$

Again,  $C_1, \dots, C_7$  is a sequence of elements of  $\mathcal{B}$  and  $.875, .750, .625, .500, .375, .250, .125$  is a sequence of nonnegative real numbers. Therefore,  $f_3 = \sum_{i=1}^7 \gamma_i 1_{C_i}$  is a nonnegative step function. The integral of the functions  $f_1$  and  $f_2$  are computed in Exercise 3-7.  $\triangleleft$

### Convergence of an Increasing Sequence of Nonnegative Step Functions

**Example 3.18 (Convergence)** Figure 3.2 shows that  $f_1(\omega) \leq f_2(\omega) \leq f_3(\omega) \leq f(\omega)$  for all  $\omega \in \Omega$ . Hence,  $f_1, f_2, f_3$  is a *finite increasing sequence* of nonnegative step functions. The interval  $[0, 1]$  on the vertical axis is partitioned and these partitions are refined step by step. In our example, we started with the partition  $\{[0, .50[, [.50, 1]\}$ . Then we partitioned

$$[0, .50[ \text{ to } \{[0, .25[, [.25, .50]\} \quad \text{and} \quad [.50, 1] \text{ to } \{[.50, .75[, [.75, 1]\}, \text{ etc.}$$

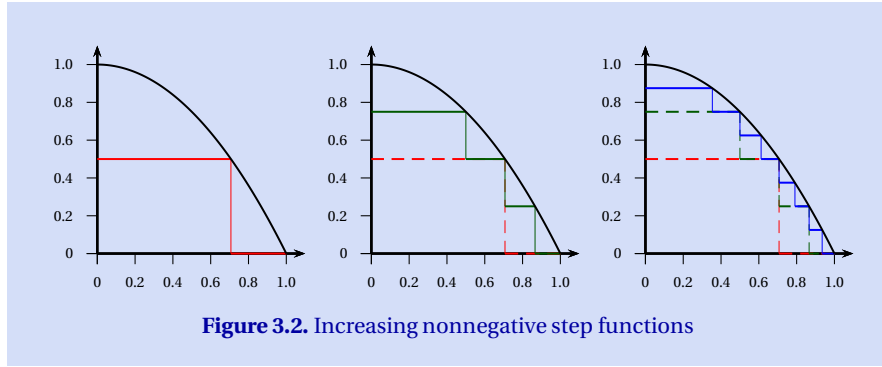
Following this idea, we can define functions  $f_4, f_5, \dots$  such that  $f_1, f_2, \dots$  is an infinite sequence of nonnegative step functions with  $f_1(\omega) \leq f_2(\omega) \leq \dots \leq f(\omega)$ , for all  $\omega \in \Omega$ , and  $\lim_{n \rightarrow \infty} f_n(\omega) = f(\omega)$ , for all  $\omega \in \Omega$ . According to the following theorem, this holds for *all* nonnegative measurable functions  $f: (\Omega, \mathcal{A}) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ .  $\triangleleft$

#### Theorem 3.19 (Approximation of Nonnegative Functions)

Let  $(\Omega, \mathcal{A})$  be a measurable space and  $f: (\Omega, \mathcal{A}) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  a nonnegative measurable function. Then:

(i) There is a sequence  $f_1, f_2, \dots$  of nonnegative step functions such that

$$f_1(\omega) \leq f_2(\omega) \leq \dots, \quad \forall \omega \in \Omega \quad (3.16)$$



and

$$\lim_{n \rightarrow \infty} f_n(\omega) = f(\omega), \quad \forall \omega \in \Omega. \quad (3.17)$$

(ii) There is a sequence of sets  $A_1, A_2, \dots \in \mathcal{A}$  and a sequence of nonnegative real numbers  $\alpha_1, \alpha_2, \dots$  such that

$$f = \sum_{i=1}^{\infty} \alpha_i 1_{A_i}. \quad (3.18)$$

For a proof see Klenke (2013, Theorem 1.96, p. 38).

**Remark 3.20 (Infinite Sums)** Equation (3.18) can be visualized by Figure 3.2. The function  $f_3$  on the right-hand side of this figure is already close to  $f$ . Partitioning the intervals *on the vertical axis* again and again leads to better approximations of  $f$ . Note that the horizontal axis does not have to be a subset of  $\mathbb{R}$ ; instead, it can be any nonempty set  $\Omega$ .

Remember, the right-hand side of Equation (3.18) is just a symbol for the corresponding limit, i. e.,

$$\sum_{i=1}^{\infty} \alpha_i 1_{A_i} = \lim_{n \rightarrow \infty} \sum_{i=1}^n \alpha_i 1_{A_i}. \quad (3.19)$$

Note that, for  $\alpha_i \geq 0$ , this limit always exists. ◁

Before turning to the definition of the integral of a nonnegative measurable function let us use the properties (3.16) and (3.17) to define the concepts *increasing sequence of nonnegative step functions* and *pointwise convergence*.

**Definition 3.21 (Increasing Sequence of Nonnegative Step Functions)**

Let  $(\Omega, \mathcal{A})$  be a measurable space and  $f: (\Omega, \mathcal{A}) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  a nonnegative measurable function. A sequence  $f_1, f_2, \dots$  of nonnegative step functions satisfying

(3.16) is called *increasing*. If it also satisfies (3.17), then we say that  $f_1, f_2, \dots$  converges pointwise to  $f$  and denote it by  $f_n \uparrow f$ .

### Uniqueness of the Limits of an Integral

In Theorem 3.19 we have seen that every nonnegative measurable function  $f$  can be represented by the limit  $\lim_{n \rightarrow \infty} f_n$  of an increasing sequence  $f_1, f_2, \dots$  of nonnegative step functions, i. e.,

$$f = \lim_{n \rightarrow \infty} f_n. \quad (3.20)$$

The definition of the integral of nonnegative step functions implies that the *integrals* of the functions  $f_n$  are increasing as well, i. e.,

$$f_n \leq f_{n+1} \Rightarrow \int f_n d\mu \leq \int f_{n+1} d\mu, \quad \forall n \in \mathbb{N},$$

[see Bauer, 2001, proposition (10.7), p. 55]. Hence, the sequence of the integrals either converges to a (finite) real number or diverges to  $+\infty$ .

In Figure 3.2 we presented the first three nonnegative step functions  $f_1, f_2$ , and  $f_3$  of such an increasing sequence  $f_1, f_2, \dots$  that approximates the function  $f: \mathbb{R} \rightarrow \mathbb{R}$  defined by Equation (3.15). Figure 3.3 visualizes the convergence of the integrals  $\int f_n d\lambda$  with respect to the Lebesgue measure  $\lambda$  on  $(\mathbb{R}, \mathcal{B})$  (see the shaded areas in Figure 3.3).

It should be noted, however, that there is not only one single increasing sequence of nonnegative step functions whose limit is  $f$ . This is illustrated in the following example.

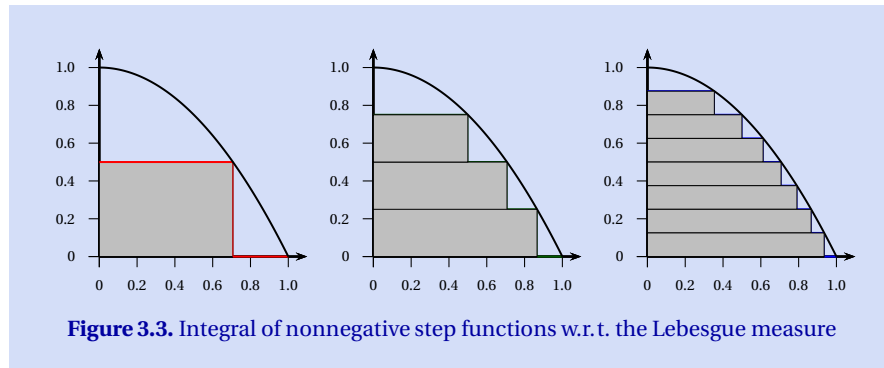
**Example 3.22 (Uniqueness)** For an example, construct a sequence  $g_1, g_2, \dots$  analogously to the sequence  $f_1, f_2, \dots$  in Example 3.17, using other partitions of the interval  $[0, 1]$  on the vertical axis, e. g.,

$$\{[0, .40[, [.40, 1]\} \quad \text{and} \quad \{[0, .20[, [.20, .40[, [.40, .80[, [.80, 1]\}, \quad \text{etc.}$$

Then  $g_1, g_2, \dots$  is a second increasing sequence that also approximates  $f$ . Figure 3.3 suggests that the specific choice of an increasing sequence of nonnegative step functions is irrelevant for the limit of their integrals. And in fact, according to the following theorem this does not only apply to our example and to the Lebesgue measure  $\lambda$  on  $\mathcal{B}$ , but to *any* nonnegative measurable function and *any* measure  $\mu$ . ◁

### Theorem 3.23 (Uniqueness of the Limits of Integrals)

If  $f_1, f_2, \dots$  and  $g_1, g_2, \dots$  are two increasing sequences of nonnegative step functions  $f_n, g_n: (\Omega, \mathcal{A}, \mu) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ , then  $\lim_{n \rightarrow \infty} f_n = \lim_{n \rightarrow \infty} g_n$  implies



$$\lim_{n \rightarrow \infty} \int f_n d\mu = \lim_{n \rightarrow \infty} \int g_n d\mu. \quad (3.21)$$

For a proof see Bauer (2001, Corollary 11.2, p. 58).

According to this theorem, if we consider two increasing sequences of nonnegative step functions with identical limits, then we know that the limits of their integrals are identical.

### Definition of the Integral of a Nonnegative Measurable Function

Based on the result of Theorem 3.23, we define the integral of any nonnegative measurable function.

#### Definition 3.24 (Integral of a Nonnegative Measurable Function)

Assume that  $f: (\Omega, \mathcal{A}, \mu) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  is a nonnegative measurable function and let  $f = \lim_{n \rightarrow \infty} f_n$  be a representation of  $f$  as the limit of an increasing sequence  $f_1, f_2, \dots$  of nonnegative step functions. Then

$$\int f d\mu := \lim_{n \rightarrow \infty} \int f_n d\mu \quad (3.22)$$

is called the *integral of  $f$  (over  $\Omega$ ) with respect to  $\mu$* .

Note that the integral of a nonnegative measurable function is either a nonnegative real number or  $+\infty$ .

**Example 3.25 (Integral With Respect to a Dirac Measure)** Suppose the assumptions of Definition 3.24 hold. Then, for  $\omega \in \Omega$ ,

$$\begin{aligned} \int f d\delta_\omega &= \lim_{n \rightarrow \infty} \int f_n d\delta_\omega && [(3.22)] \\ &= \lim_{n \rightarrow \infty} f_n(\omega) && [(3.10)] \quad (3.23) \\ &= f(\omega). && [(3.20)] \end{aligned}$$

Hence, the integral of a nonnegative measurable function  $f$  with respect to the Dirac measure at  $\omega$  is the value of  $f$  for  $\omega$ .  $\triangleleft$

We conclude this section by the following lemma on monotonicity of the integrals of nonnegative measurable functions.

**Lemma 3.26 (Monotonicity)**

If  $f, g: (\Omega, \mathcal{A}) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  are nonnegative and measurable, then

$$f \leq g \Rightarrow \int f d\mu \leq \int g d\mu. \quad (3.24)$$

For a proof see Bauer (2001, p. 59, Eq. (11.8)).

**Remark 3.27 (Bounds of the Integral of a Bounded Function)** Let  $f: (\Omega, \mathcal{A}) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  be nonnegative and measurable and  $\alpha \in \mathbb{R}$ . Then, for  $g = \alpha$ , Equations (3.24) and (3.5) imply

$$f \leq \alpha \Rightarrow \int f d\mu \leq \alpha \cdot \mu(\Omega), \quad (3.25)$$

and

$$f \geq \alpha \Rightarrow \int f d\mu \geq \alpha \cdot \mu(\Omega). \quad (3.26)$$

$\triangleleft$

### 3.1.3 Integral of a Measurable Function

Now we define the integral of a measurable function  $f: (\Omega, \mathcal{A}) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  using the positive part  $f^+$  and the negative part  $f^-$  of  $f$  that have been introduced in section 2.4.2. According to Theorem 2.66,  $f^+$  and  $f^-$  are both nonnegative measurable functions. Reading the following definition, remember the conventions:  $\infty + \infty = \infty$ ,  $-\infty - \infty = -\infty$ ,  $x + \infty = \infty$ , for all  $x \in \mathbb{R}$ ,  $x - \infty = -\infty$ , for all  $x \in \mathbb{R}$ . Also note that  $\infty - \infty$  is *not defined*, which has to be observed whenever integrals are not necessarily finite.

**Definition 3.28 (Integral of a Measurable Function)**

Let  $f: (\Omega, \mathcal{A}, \mu) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  be a measurable function. If  $\int f^+ d\mu$  or  $\int f^- d\mu$

are finite, then  $f$  is called *quasi-integrable with respect to  $\mu$* , or simply *quasi- $\mu$ -integrable*, and

$$\int f \, d\mu = \int f^+ \, d\mu - \int f^- \, d\mu \quad (3.27)$$

is called the *integral of  $f$  (over  $\Omega$ ) with respect to  $\mu$* . If  $\int f^+ \, d\mu$  and  $\int f^- \, d\mu$  are both finite, then  $f$  is called *integrable with respect to  $\mu$* , or simply  *$\mu$ -integrable*.

**Remark 3.29 (Integrability and Quasi-Integrability)** Of course, every integrable measurable function is quasi-integrable and each nonnegative function is also quasi-integrable. Furthermore, assuming that a function  $f: (\Omega, \mathcal{A}, \mu) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  is integrable or quasi-integrable includes the assumption that  $f$  is measurable. Finally, if  $f$  is  $\mu$ -integrable, then

$$-\infty < \int f \, d\mu < +\infty,$$

i. e., the integral is *finite*, taking a value in  $\mathbb{R}$ . If  $f$  is quasi- $\mu$ -integrable, then the integral may also be *infinite*, i. e., it may also take on the values  $+\infty$  or  $-\infty$ .  $\triangleleft$

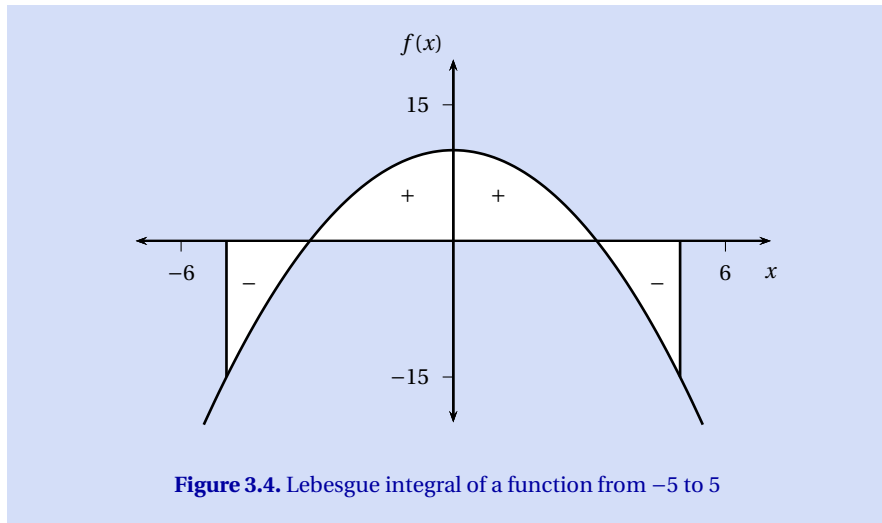
**Remark 3.30 (A Standard Method for Proofs)** The integral of a quasi-integrable function has been defined in three steps, for *nonnegative measurable step functions*, for *nonnegative measurable functions*, and for *quasi-integrable functions*. Oftentimes, these steps are also followed in proofs of propositions involving integrals. That is, in a first step it is shown that the proposition holds for nonnegative measurable step functions. In a second step, using Equation (3.22), it is proven for nonnegative measurable functions, and finally, Equation (3.27) is applied to complete the proof for all quasi-integrable functions. An example is the proof of Theorem 3.36. Oftentimes, we only detail the first step, in particular, if the remaining two steps are straightforward.  $\triangleleft$

**Example 3.31 (Integral With Respect to the Lebesgue Measure  $\lambda$ )** Figure 3.4 displays the integral of a function  $f: (\mathbb{R}, \mathcal{B}, \lambda) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  with respect to the Lebesgue measure  $\lambda$ . Because  $f^+$  and  $f^-$  are both nonnegative (see Rem. 2.62), the integrals  $\int f^+ \, d\mu$  and  $\int f^- \, d\mu$  are positive and identical to the white areas marked + and - in Figure 3.4. According to Equation (3.27), the integral of  $f$  is the difference between the area  $\int f^+ \, d\mu$  and the area  $\int f^- \, d\mu$ .  $\triangleleft$

**Remark 3.32 (An Alternative Notation)** An alternative notation for the integral of  $f$  is

$$\int f \, d\mu = \int f(\omega) \, \mu(d\omega) = \int_{\Omega} f(\omega) \, \mu(d\omega), \quad (3.28)$$

which explicitly uses the values  $f(\omega)$  of  $f$ . This notation conveys the idea that the values  $f(\omega)$  of  $f$  are weighted by the measure of  $d\omega$ . If  $\Omega = \mathbb{R}$ , then  $d\omega$  symbolizes



the length of an infinitesimal interval between two elements in  $\mathbb{R}$ . If  $\Omega$  is finite or countable, then  $\mu(d\omega)$  symbolizes the value of  $\mu$  for the singleton  $\{\omega\}$  and the integral can be written as a sum (see Example 3.16).  $\triangleleft$

**Lemma 3.33 (Integrability Carries Over to Restrictions of Functions)**

- (i) If  $f: (\Omega, \mathcal{A}, \mu) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  is quasi- $\mu$ -integrable and  $A \in \mathcal{A}$ , then  $1_A f$  is quasi- $\mu$ -integrable.
- (ii) If  $f: (\Omega, \mathcal{A}, \mu) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  is  $\mu$ -integrable and  $A \in \mathcal{A}$ , then  $1_A f$  is  $\mu$ -integrable.

*(Proof p. 111)*

**Remark 3.34 (Integral of  $1_A f$ )** Lemma 3.33 (ii) means: If  $f: (\Omega, \mathcal{A}, \mu) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  is  $\mu$ -integrable and  $A \in \mathcal{A}$ , then

$$\int f d\mu \text{ is finite} \quad \Rightarrow \quad \int 1_A f d\mu \text{ is finite.} \quad (3.29)$$

$\triangleleft$

If  $f: (\Omega, \mathcal{A}, \mu) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  is quasi- $\mu$ -integrable and  $A \in \mathcal{A}$ , then Lemma 3.33 implies that the integral  $\int 1_A f d\mu$  is well-defined. Hence, we can now introduce the integral of  $f$  over a subset  $A$  of  $\Omega$  as follows:

**Definition 3.35 (Integral Over a Subset  $A$  of  $\Omega$ )**

If  $f: (\Omega, \mathcal{A}, \mu) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  is quasi- $\mu$ -integrable and  $A \in \mathcal{A}$ , then

$$\int_A f \, d\mu := \int 1_A f \, d\mu \quad (3.30)$$

is called the integral of  $f$  over  $A$  with respect to  $\mu$ .

Because  $1_\Omega f = f$ , a special case of Equation (3.30) is

$$\int_\Omega f \, d\mu = \int 1_\Omega f \, d\mu = \int f \, d\mu. \quad (3.31)$$

**3.2 Properties**

In this section we consider some important properties and rules of computation for the integral of a measurable function  $f: (\Omega, \mathcal{A}, \mu) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ .

**Theorem 3.36 (Linearity)**

Consider the functions  $f: (\Omega, \mathcal{A}, \mu) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  and  $g: (\Omega, \mathcal{A}, \mu) \rightarrow (\mathbb{R}, \mathcal{B})$ .

(i) If  $f$  is quasi- $\mu$ -integrable and  $\alpha \in \mathbb{R}$ , then  $\alpha f$  is quasi- $\mu$ -integrable and

$$\int \alpha f \, d\mu = \alpha \int f \, d\mu. \quad (3.32)$$

(ii) If  $f$  is quasi- $\mu$ -integrable and  $g$  is  $\mu$ -integrable, then  $f + g$  is quasi- $\mu$ -integrable, and

$$\int (f + g) \, d\mu = \int f \, d\mu + \int g \, d\mu. \quad (3.33)$$

(Proof p. 111)

Combining propositions (i) and (ii) of Theorem 3.36 immediately yields the following corollary.

**Corollary 3.37 (Linearity)**

Consider the functions  $f: (\Omega, \mathcal{A}, \mu) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ ,  $g: (\Omega, \mathcal{A}, \mu) \rightarrow (\mathbb{R}, \mathcal{B})$  and let  $\alpha, \beta \in \mathbb{R}$ . If  $f$  is quasi- $\mu$ -integrable and  $g$  is  $\mu$ -integrable, then  $\alpha f + \beta g$  is quasi- $\mu$ -integrable, and

$$\int (\alpha f + \beta g) \, d\mu = \alpha \int f \, d\mu + \beta \int g \, d\mu. \quad (3.34)$$

Linearity can also be used to prove the following corollary on the equivalence of integrability of a measurable function  $f$  and finiteness of the integral of the absolute value function  $|f|$ .

**Corollary 3.38 (Integrability and Absolute Value Function)**

The function  $f : (\Omega, \mathcal{A}, \mu) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  is  $\mu$ -integrable if and only if

$$\int |f| d\mu < \infty.$$

(Proof p. 115)

**Example 3.39 (Integral Over the Union of Two Sets)** If  $f : (\Omega, \mathcal{A}, \mu) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  is  $\mu$ -integrable and  $A, B \in \mathcal{A}$ , then  $1_{A \cup B} f$  is  $\mu$ -integrable and

$$\int_{A \cup B} f d\mu = \int 1_{A \cup B} f d\mu = \int_A f d\mu + \int_B f d\mu - \int_{A \cap B} f d\mu. \quad (3.35)$$

If  $A \cap B = \emptyset$  and  $f$  is quasi- $\mu$ -integrable, then

$$\int_{A \cup B} f d\mu = \int 1_{A \cup B} f d\mu = \int_A f d\mu + \int_B f d\mu. \quad (3.36)$$

(see Exercise 3-8).

◁

**Lemma 3.40 (Measures That are Identical on a Sub- $\sigma$ -Algebra)**

Assume that  $f : (\Omega, \mathcal{A}, \mu) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  is nonnegative or  $\mu$ -integrable. Furthermore, let  $\mathcal{C} \subset \mathcal{A}$  be a  $\sigma$ -algebra, let  $f$  be  $\mathcal{C}$ -measurable, and suppose that  $\nu(A) = \mu(A)$ , for all  $A \in \mathcal{C}$ . Then  $(\Omega, \mathcal{C}, \nu)$  is a measure space and

$$\int f d\nu = \int f d\mu. \quad (3.37)$$

(Proof p. 115)

Hence, the integral  $\int f d\mu$  only depends on the values of  $\mu$  on the  $\sigma$ -algebra  $\sigma(f)$ , the  $\sigma$ -algebra generated by  $f$ .

**Lemma 3.41 (Integrable Functions are  $\mu$ -a.e. Real-Valued)**

Let  $f : (\Omega, \mathcal{A}, \mu) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  be measurable. If  $f$  is  $\mu$ -integrable, then  $f$  is real-valued  $\mu$ -almost everywhere. If  $f$  is quasi- $\mu$ -integrable, then

$$\mu(\{\omega \in \Omega : f(\omega) = \infty\}) > 0 \Rightarrow \int f d\mu = \infty, \quad (3.38)$$

$$\mu(\{\omega \in \Omega : f(\omega) = -\infty\}) > 0 \Rightarrow \int f d\mu = -\infty. \quad (3.39)$$

(Proof p. 115)

**Remark 3.42 (Integrable Functions are Assumed to be Real-Valued)** Contraposition of (3.38) and (3.39) yields: If  $\int f d\mu$  is finite, then  $f(\omega) \in \mathbb{R}$  (i. e.,  $-\infty < f(\omega) < \infty$ ), for  $\mu$ -almost all  $\omega \in \Omega$  (see Def. 2.68 and Remark 2.70). In this case, there is a *real-valued* measurable function  $f^*: (\Omega, \mathcal{A}, \mu) \rightarrow (\mathbb{R}, \mathcal{B})$  with  $f^* \stackrel{\mu}{=} f$ . (For example, define  $f^* := 1_A \cdot f + 1_{A^c} \cdot 0 = 1_A \cdot f$  for  $A := \{\omega \in \Omega: f(\omega) \in \mathbb{R}\}$ ). Therefore, without substantial loss of generality, for simplicity, we often assume that a function is real-valued if it has a finite integral.  $\triangleleft$

### 3.2.1 Integral of $\mu$ -Equivalent Functions

The concept of equivalence of two measurable functions with respect to a measure has already been introduced in section 2.5. Now we treat the relationship of this concept to the integrals of two numerical functions.

**Theorem 3.43 (A Condition Equivalent to  $f \stackrel{\mu}{=} 0$ )**

If  $f: (\Omega, \mathcal{A}, \mu) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  is a nonnegative measurable function, then

$$\int f d\mu = 0 \Leftrightarrow f \stackrel{\mu}{=} 0. \quad (3.40)$$

For a proof see Bauer (2001, Theorem 13.2, p. 71).

**Lemma 3.44 (Integral of a Positive Function)**

If  $f: (\Omega, \mathcal{A}, \mu) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  is quasi- $\mu$ -integrable and there is an  $A \in \mathcal{A}$  such that  $\mu(A) > 0$  and  $f(\omega) > 0$ , for all  $\omega \in A$ , then

$$\int 1_A \cdot f d\mu > 0. \quad (3.41)$$

(Proof p. 116)

If  $(\Omega, \mathcal{A}, \mu)$  is a measure space, then a set  $A \in \mathcal{A}$  with  $\mu(A) = 0$  is called a *null set* with respect to  $\mu$ . In the following lemma we consider the integral over such a null set (see Exercise 3-9).

**Lemma 3.45 (Integral Over a Null Set)**

Let  $f: (\Omega, \mathcal{A}, \mu) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  be measurable. If  $A \in \mathcal{A}$ , with  $\mu(A) = 0$ , then  $1_A \cdot f$  is  $\mu$ -integrable and

$$\int_A f d\mu = \int 1_A \cdot f d\mu = 0. \quad (3.42)$$

(Proof p. 116)

**Remark 3.46 (Integration Over Null Sets Can be Neglected)** The conjunction of Equations (3.36) and (3.42) implies: If  $f$  is quasi- $\mu$ -integrable and  $A \in \mathcal{A}$  with  $\mu(A) = 0$ , then

$$\int f d\mu = \int_{\Omega} f d\mu = \int_{\Omega \setminus A} f d\mu + \int_A f d\mu = \int_{\Omega \setminus A} f d\mu. \quad (3.43)$$

&lt;

**Lemma 3.47 (Integrals of  $\mu$ -Equivalent Functions)**

Let  $f, g: (\Omega, \mathcal{A}, \mu) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  be quasi- $\mu$ -integrable. Then

$$f \stackrel{\mu}{=} g \Rightarrow \int f d\mu = \int g d\mu. \quad (3.44)$$

(Proof p. 116)

The following theorem presents a condition that is necessary and sufficient for  $\mu$ -equivalence of  $f$  and  $g$ .

**Theorem 3.48 (Identity of Integrals of  $\mu$ -Equivalent Functions)**

If  $f, g: (\Omega, \mathcal{A}, \mu) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  are  $\mu$ -integrable, then

$$f \stackrel{\mu}{=} g \Leftrightarrow \int_A f d\mu = \int_A g d\mu, \quad \forall A \in \mathcal{A}. \quad (3.45)$$

(Proof p. 117)

In section 3.4 we shall see that, if  $f$  and  $g$  are  $\mu$ -integrable and nonnegative, then it is sufficient to consider the integrals over all sets  $A$  in a  $\cap$ -stable generating system of  $\mathcal{A}$  in order to show  $\mu$ -equivalence of  $f$  and  $g$  (see Th. 3.68).

**Remark 3.49 (A Counter-Example)** Note that Equation (3.45) does not hold if  $f, g$  are nonnegative but not  $\mu$ -integrable measurable functions. This is exemplified as follows: Consider  $f, g: (\mathbb{R}, \mathcal{A}, \mu) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ , where  $\mathcal{A} = \{\mathbb{R}, \emptyset\}$ ,  $\mu(\mathbb{R}) = \infty$ ,  $f = 1$ , and  $g = 2$ . Then  $\int_{\mathbb{R}} f d\mu = \int_{\mathbb{R}} g d\mu = \infty$  and  $\int_{\emptyset} f d\mu = \int_{\emptyset} g d\mu = 0$ . Hence,  $\int_A f d\mu = \int_A g d\mu$ , for all  $A \in \mathcal{A}$ , but  $f$  and  $g$  are *not equivalent* with respect to  $\mu$ .

&lt;

**Remark 3.50 (Some Special Cases)** Theorem 3.48 implies: If  $f: (\Omega, \mathcal{A}, \mu) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  is a measurable function, then  $f \stackrel{\mu}{=} \alpha$ ,  $\alpha \in \mathbb{R}$ , is equivalent to

$$\int_A f d\mu = \int_A \alpha d\mu = \int 1_A \alpha d\mu = \alpha \cdot \int 1_A d\mu = \alpha \cdot \mu(A), \quad \forall A \in \mathcal{A}. \quad (3.46)$$

Furthermore,  $f \stackrel{\mu}{=} 0$  is equivalent to

$$\int_A f d\mu = 0, \quad \forall A \in \mathcal{A}, \quad (3.47)$$

using the convention  $0 \cdot \infty = 0$ , if necessary.

An immediate implication of Equation (3.46) for  $A = \Omega$  is

$$f \stackrel{\mu}{=} \alpha, \quad \alpha \in \mathbb{R} \quad \Rightarrow \quad \int f d\mu = \alpha \mu(\Omega). \quad (3.48)$$

For  $\alpha = 0$ , this yields

$$f \stackrel{\mu}{=} 0 \quad \Rightarrow \quad \int f d\mu = 0. \quad (3.49)$$

◁

**Remark 3.51 (Almost Everywhere)** The notion of  $\mu$ -equivalence of  $f$  and  $g$  is an example of a property that holds for all  $\omega \in \Omega \setminus A$  with  $\mu(A) = 0$ . We also say that such a property holds  $\mu$ -almost everywhere ( $\mu$ -a.e.). Another example is the property

$$f(\omega) \leq g(\omega), \quad \forall \omega \in \Omega \setminus A \text{ and } \mu(A) = 0,$$

which is denoted by  $f \stackrel{\mu}{\leq} g$ .

◁

The proposition of Lemma 3.47 analogously holds for the relations  $\stackrel{\mu}{\leq}$  and  $\stackrel{\mu}{<}$ . The following theorem generalizes Lemma 3.26.

**Theorem 3.52 (Monotonicity)**

Let  $f, g: (\Omega, \mathcal{A}, \mu) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  be measurable functions.

(i) If  $f$  and  $g$  are quasi- $\mu$ -integrable, then

$$f \stackrel{\mu}{\leq} g \quad \Rightarrow \quad \int f d\mu \leq \int g d\mu. \quad (\text{monotonicity}) \quad (3.50)$$

(ii) If  $\mu(\Omega) > 0$  and  $f, g$  are  $\mu$ -integrable, then

$$f \stackrel{\mu}{<} g \quad \Rightarrow \quad \int f d\mu < \int g d\mu. \quad (\text{strict monotonicity}) \quad (3.51)$$

(Proof p. 117)

### 3.2.2 Integral With Respect to a Weighted Sum of Measures

In Example 1.62 we already noted that a weighted sum of measures with nonnegative weights is again a measure. As a special case, if  $\mu$  is a measure on  $(\Omega, \mathcal{A})$  and  $\alpha$  is a nonnegative number, then  $\alpha \cdot \mu$  is a measure on  $(\Omega, \mathcal{A})$  as well. Furthermore, if  $f$  is  $\mu$ -integrable, then

$$\int f d(\alpha\mu) = \int \alpha f d\mu = \alpha \int f d\mu \quad (3.52)$$

(see Exercise 3-10). This is generalized in the following theorem.

**Theorem 3.53 (Integral With Respect to a Weighted Sum of Measures)**

If  $f : (\Omega, \mathcal{A}) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  is measurable and nonnegative,  $\mu_1, \mu_2, \dots$  are measures on  $(\Omega, \mathcal{A})$ , and  $\alpha_1, \alpha_2, \dots \in \mathbb{R}$  are nonnegative, then

$$\int f d\left(\sum_{i=1}^{\infty} \alpha_i \mu_i\right) = \sum_{i=1}^{\infty} \alpha_i \int f d\mu_i. \quad (3.53)$$

For a proof, see Equation (3.52) and Bauer (2001, Example 3 on page 61).

If we consider a *finite* weighted sum of measures, the assumption that  $f$  is nonnegative can be replaced by integrability of  $f$ . In the following theorem we consider a weighted sum of two measures. In Remark 3.55 we extend the result to a finite weighted sum of measures.

**Theorem 3.54 (Integral With Respect to a Weighted Sum of two Measures)**

Let  $\mu_1, \mu_2$  be measures on  $(\Omega, \mathcal{A})$ . If  $f : (\Omega, \mathcal{A}, \mu) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  is integrable with respect to  $\mu_1$  and  $\mu_2$ , and  $0 \leq \alpha_1, \alpha_2 \in \mathbb{R}$ , then  $f$  is integrable with respect to  $\alpha_1 \mu_1 + \alpha_2 \mu_2$ , and

$$\int f d(\alpha_1 \mu_1 + \alpha_2 \mu_2) = \alpha_1 \int f d\mu_1 + \alpha_2 \int f d\mu_2. \quad (3.54)$$

For a proof, see Equation (3.52) and Bauer (2001, Example 5 on page 67).

**Remark 3.55 (Integral With Respect to a Finite Weighted Sum of Measures)**

By induction, Theorem 3.54 yields, for nonnegative  $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ ,

$$\int f d\left(\sum_{i=1}^n \alpha_i \mu_i\right) = \sum_{i=1}^n \alpha_i \int f d\mu_i, \quad (3.55)$$

provided that  $f$  is integrable with respect to all measures  $\mu_1, \dots, \mu_n$ . ◁

**Example 3.56 (Integral With Respect to the Weighted Sum of Dirac Measures)**

Let  $(\Omega, \mathcal{A})$  be a measurable space and, for  $i \in \mathbb{N}$ , let  $\omega_i \in \Omega$ ,  $\alpha_i \in \mathbb{R}$ ,  $\alpha_i \geq 0$ , and  $\delta_{\omega_i}$  denote the Dirac measure at  $\omega_i$ . Then

$$\mu = \sum_{i=1}^{\infty} \alpha_i \delta_{\omega_i} \quad (3.56)$$

defined by  $\mu(A) = \sum_{i=1}^{\infty} \alpha_i \delta_{\omega_i}(A)$ , for all  $A \in \mathcal{A}$ , is a measure on  $(\Omega, \mathcal{A})$  (see Example 1.62). For any nonnegative measurable function  $f : \Omega \rightarrow \mathbb{R}$  we obtain

$$\int f d\mu = \int f d \sum_{i=1}^{\infty} \alpha_i \delta_{\omega_i} \quad [(3.56)]$$

$$= \sum_{i=1}^{\infty} \alpha_i \int f d\delta_{\omega_i} \quad [(3.53)] \quad (3.57)$$

$$= \sum_{i=1}^{\infty} \alpha_i f(\omega_i). \quad [(3.23)]$$

For  $\mu(A) = \sum_{i=1}^n \alpha_i \delta_{\omega_i}(A)$ , for all  $A \in \mathcal{A}$ , Equation (3.57) with  $\alpha_i = 0$  for  $i > n$ , yields

$$\int f d\mu = \sum_{i=1}^n \alpha_i f(\omega_i). \quad (3.58)$$

Hence, the integral of a nonnegative measurable function  $f$  with respect to a finite or countable weighted sum of Dirac measures with nonnegative weights is a weighted sum of values of  $f$ .  $\triangleleft$

### 3.2.3 Integral With Respect to an Image Measure

The next theorem is relevant whenever we consider the integral of a composition  $g \circ f$  of a mapping  $f$  with a numerical function  $g$  [see Eq. (2.25)] or the integral with respect to the image measure  $\mu_f$  of  $\mu$  under  $f$  [see Def. 2.80].

#### Theorem 3.57 (Transformation Theorem)

Let  $f: (\Omega, \mathcal{A}, \mu) \rightarrow (\Omega', \mathcal{A}')$  and  $g: (\Omega', \mathcal{A}') \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}}')$  be measurable.

(i) If  $g$  is nonnegative or integrable with respect to  $\mu_f$ , then

$$\int g d\mu_f = \int g \circ f d\mu. \quad (3.59)$$

(ii)  $g$  is integrable with respect to  $\mu_f$  if and only if  $g \circ f$  is  $\mu$ -integrable.

For a proof, see Bauer (2001, Corollary 19.2.1, p. 110).

If  $f: (\Omega, \mathcal{A}, \mu) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  is a numerical measurable function and we replace  $g$  by the identity function  $id: (\bar{\mathbb{R}}, \bar{\mathcal{B}}) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$ , then Theorem 3.57 implies the following corollary.

#### Corollary 3.58 (An Implication of the Transformation Theorem)

If  $f: (\Omega, \mathcal{A}, \mu) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  is nonnegative or  $\mu$ -integrable, then

$$\int id d\mu_f = \int_{\bar{\mathbb{R}}} id d\mu_f = \int f d\mu = \int_{\Omega} f d\mu. \quad (3.60)$$

Using the alternative notation of an integral introduced in Remark 3.32, Equation (3.59) can also be written

$$\int_{\overline{\mathbb{R}}} g(x) \mu_f(dx) = \int_{\Omega} g[f(\omega)] \mu(d\omega). \quad (3.61)$$

Correspondingly, Equation (3.60) can also be written

$$\int_{\overline{\mathbb{R}}} x \mu_f(dx) = \int_{\Omega} f(\omega) \mu(d\omega). \quad (3.62)$$

In Definition 3.10 we considered the case in which  $f = \sum_{i=1}^n \alpha_i 1_{A_i}$  is a nonnegative step function and defined its integral by  $\int f d\mu = \sum_{i=1}^n \alpha_i \mu(A_i)$ , presuming that  $A_1, \dots, A_n$  are pairwise disjoint. Now we consider a measurable function  $f$  with a finite number of values, which can be 0, positive, or negative.

**Corollary 3.59 (Integral of a Function With a Finite Number of Values)**

If  $(\Omega, \mathcal{A}, \mu)$  is a measure space and  $f = \sum_{i=1}^n \alpha_i 1_{A_i}$  with pairwise different  $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ ,  $\alpha_i \neq 0$ , and pairwise disjoint  $A_1, \dots, A_n \in \mathcal{A}$ , then  $f$  is  $\mu$ -integrable if and only if  $\mu(A_i) < \infty$  for all  $i = 1, \dots, n$ . If  $f$  is  $\mu$ -integrable, then

$$\int f d\mu = \sum_{i=1}^n \alpha_i \mu(A_i) = \sum_{i=1}^n \alpha_i \mu_f(\{\alpha_i\}). \quad (3.63)$$

(Proof p. 118)

### 3.2.4 Convergence Theorems

The next two theorems deal with convergence of integrals. In the first one, we assume that  $f_1, f_2, \dots$  is an increasing sequence of measurable functions that converge to  $f$ .

**Theorem 3.60 (Monotone Convergence; B. Levi)**

Let the functions  $f_n: (\Omega, \mathcal{A}, \mu) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  be measurable, for all  $n \in \mathbb{N}$ .

- (i) If the sequence  $f_1, f_2, \dots$  is increasing with  $\lim_{n \rightarrow \infty} f_n = f$  and the functions  $f_n$  are nonnegative for all  $n \in \mathbb{N}$  or  $\mu$ -integrable for all  $n \in \mathbb{N}$ , then

$$\int f d\mu = \lim_{i \rightarrow \infty} \int f_n d\mu. \quad (3.64)$$

- (ii) If the functions  $f_i$  are nonnegative for all  $i \in \mathbb{N}$ , then

$$\int \left( \sum_{i=1}^{\infty} f_i \right) d\mu = \sum_{i=1}^{\infty} \int f_i d\mu. \quad (3.65)$$

The integrals on both sides are finite or  $+\infty$ .

For a proof of (i), assuming nonnegativity, see Bauer (2001, Theorem 11.4, p. 59). For a proof of (i), assuming integrability, see Klenke (2013, Theorem 4.20, p. 93). For a proof of (ii), see Bauer (2001, Corollary 11.5, p. 60).

Note that, by definition of an ‘infinite sum’ (see Box 0.1), Equation (3.65) is equivalent to

$$\int \left( \lim_{n \rightarrow \infty} \sum_{i=1}^n f_i \right) d\mu = \lim_{n \rightarrow \infty} \sum_{i=1}^n \int f_i d\mu. \quad (3.66)$$

In the next theorem, we replace the assumption that  $f_1, f_2, \dots$  is increasing by the assumption that there is a  $\mu$ -integrable function  $g$  dominating the absolute value functions of all  $f_n$ .

**Theorem 3.61 (Dominated Convergence; Lebesgue Convergence Theorem)**

If  $g, f_n: (\Omega, \mathcal{A}, \mu) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ ,  $n \in \mathbb{N}$ , are  $\mu$ -integrable and there is a measurable function  $f: (\Omega, \mathcal{A}, \mu) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  with  $\lim_{n \rightarrow \infty} f_n = f$ , and  $|f_n| \leq g$  for all  $n \in \mathbb{N}$ , then

$$\int f d\mu = \lim_{n \rightarrow \infty} \int f_n d\mu, \quad (3.67)$$

and this integral is finite.

For a proof, see Bauer (2001, Theorem 15.6, p. 83).

### 3.3 Lebesgue and Riemann Integral

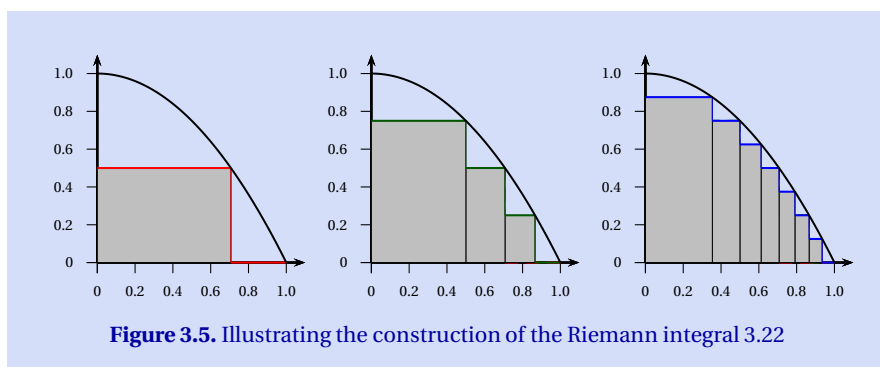
The Lebesgue measures  $\lambda_n$  on  $(\mathbb{R}^n, \mathcal{B}_n)$ ,  $n = 1, 2, 3$ , represent *length*, *area*, and *volume*, respectively. As the examples illustrated by Figure 3.4 shows, the integral of the  $(\mathcal{B}, \mathcal{B})$ -measurable function  $f: \mathbb{R} \rightarrow \mathbb{R}$  with respect to the Lebesgue measure  $\lambda = \lambda_1$ , i. e., the Lebesgue integral, yields the difference between the *areas* marked by + and the areas marked by –, respectively.

It is useful to know conditions under which the Lebesgue integral and the Riemann integrals are identical, because a lot of tools are available for Riemann integration (see, e. g., Ellis & Gulick, 2006). The following theorem is proved in Klenke (2013, Theorem 4.23, p. 96), who also provides a brief definition of the Riemann integral and Riemann integrability.

**Theorem 3.62 (Lebesgue Integral and Riemann Integral)**

Let  $\lambda$  denote the Lebesgue measure on  $(\mathbb{R}, \mathcal{B})$  and let  $[a, b]$ ,  $a, b \in \mathbb{R}$ ,  $a < b$ , be a closed interval. If  $f: [a, b] \rightarrow \mathbb{R}$  is Riemann integrable on  $[a, b]$ , then  $f$  is  $\lambda$ -integrable, and

$$\int_a^b f(x) dx = \int_{[a,b]} f d\lambda = \int_{[a,b]} f(x) \lambda(dx) = \int 1_{[a,b]} \cdot f d\lambda, \quad (3.68)$$



where  $\int_a^b f(x) dx$  denotes the Riemann integral from  $a$  to  $b$ .

Note that

$$\int_a^b f(x) dx = F(x) \Big|_a^b := F(b) - F(a), \quad (3.69)$$

where  $F$  is an antiderivative of  $f$ .

**Remark 3.63 (Lebesgue vs. Riemann Integral)** If we want to define the integral of a measurable function  $f: (\Omega, \mathcal{A}) \rightarrow (\mathbb{R}, \mathcal{B})$ , where the set  $\Omega$  is not necessarily a subset of  $\mathbb{R}$ , then this means that the traditional Riemann integral cannot be used. The *Riemann integral* is constructed by partitioning the *domain* of  $f$ , the set  $\mathbb{R}$  of real numbers into small intervals and adding the area of the rectangles on these intervals in order to approximate the area under the function  $f: \mathbb{R} \rightarrow \mathbb{R}$  (see Fig. 3.5). If  $\Omega \not\subset \mathbb{R}$ , then this idea does not work any more. Instead, the *Lebesgue integral* is constructed by partitioning the *codomain* of  $f$ , which is the set  $\mathbb{R}$  of real numbers into small intervals (see Fig. 3.3). This is also possible if  $\Omega \not\subset \mathbb{R}$ , and in this aspect, the Lebesgue integral is more general than the Riemann integral.  $\triangleleft$

Note, however, that even if the domain of  $f$  is a subset of the set of real numbers, there are functions for which the Lebesgue integral exists and the Riemann integral does not exist (see, e. g., Klenke, 2013, Example 4.24, p. 97). Also note that there are functions that are Riemann integrable on a half-open or unbounded interval but not Lebesgue integrable (see, e. g., Klenke, 2013, Remark 4.25, p. 97).

**Example 3.64 (Using the Riemann Integral)** As a simple application of Theorem 3.62 consider the function  $f$  defined by  $f(x) = 10 - x^2$  on a closed interval  $[a, b]$ . Because  $f$  is a continuous function, it is Riemann integrable. Hence, we can apply Equation (3.68). For  $a = -5$  and  $b = 5$ , this equation yields

$$\int_{[-5,5]} f d\lambda = \int_{-5}^5 f(x) dx = 16.\bar{6}$$

(see Exercise 3-11). This integral is the difference between the areas marked by + and the areas marked by – in Figure 3.4.  $\triangleleft$

### 3.4 Density

A density  $f$  can be interpreted as a weighting function of the values of the original measure  $\mu$ . If we consider a measure  $\mu$  on a measurable space  $(\Omega, \mathcal{A})$  with  $\mu(\{x\}) > 0$  for all  $x \in \Omega$ , then this means that the values  $\mu(\{x\})$  of the singletons  $\{x\}$  are multiplied by a nonnegative number  $f(x)$ . If we consider the Lebesgue measure  $\lambda$  on  $\mathbb{R}$ , then, intuitively speaking, any infinitesimal interval  $dx$  gets a weight  $f(x)$ . Using such a density, a new measure  $\nu$  on  $\mathcal{A}$  is introduced, where  $\nu(A)$  is the integral of  $f$  over  $A$  with respect to  $\mu$ . The most important examples are densities with respect to the Lebesgue measure (see Example 3.69).

#### Theorem 3.65 (Measure With Density)

Let  $f: (\Omega, \mathcal{A}, \mu) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  be a nonnegative measurable function. Then the function  $\nu: \mathcal{A} \rightarrow \bar{\mathbb{R}}$  defined by

$$\nu(A) = \int_A f d\mu, \quad \forall A \in \mathcal{A}, \quad (3.70)$$

is a measure, called the measure with density  $f$  with respect to  $\mu$ . It is denoted by  $f \circ \mu$ , i. e.,  $f \circ \mu := \nu$ .

For a proof, see Bauer (2001, Theorem 17.1, p. 96).

The notation  $f \circ \mu$  has been adopted from Elstrodt (2007, p. 127). Using this notation, Equation (3.70) can also be written

$$f \circ \mu(A) = \int_A f d\mu, \quad \forall A \in \mathcal{A}. \quad (3.71)$$

#### Definition 3.66 (Density)

Let  $\nu$  be a measure on  $(\Omega, \mathcal{A})$ . If  $f: (\Omega, \mathcal{A}, \mu) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  is a nonnegative measurable function satisfying Equation (3.70), then it is called a density of  $\nu$  with respect to  $\mu$ .

The following theorem generalizes Equation (3.52).

**Theorem 3.67 (Integral With Respect to a Measure With Density)**

Let  $(\Omega, \mathcal{A}, \mu)$  be a measure space and  $f: (\Omega, \mathcal{A}) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  a nonnegative measurable function. Furthermore, let  $f \circ \mu: \mathcal{A} \rightarrow \overline{\mathbb{R}}$  be the measure with density  $f$  with respect to  $\mu$  and let  $g: (\Omega, \mathcal{A}) \rightarrow (\mathbb{R}, \mathcal{B})$  be measurable.

(i) If  $g$  is nonnegative, then

$$\int g \, d f \circ \mu = \int g \cdot f \, d \mu. \quad (3.72)$$

(ii)  $g$  is integrable with respect to  $f \circ \mu$  if and only if  $g \cdot f$  is  $\mu$ -integrable.

(iii) If  $g$  is integrable with respect to  $f \circ \mu$ , then Equation (3.72) holds.

For a proof, see Bauer (2001, Theorem 17.3, p. 96, 97).

In the following theorem we summarize some necessary and sufficient conditions for  $\mu$ -equivalence of measurable functions on a measure space.

**Theorem 3.68 (Necessary and Sufficient Condition of  $\mu$ -Equivalence)**

Let  $f, g: (\Omega, \mathcal{A}, \mu) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  be measurable functions, let  $\mathcal{E} \subset \mathcal{A}$ , and consider:

(a)  $f \stackrel{\mu}{=} g$ .

(b)  $\int_A f \, d \mu = \int_A g \, d \mu, \quad \forall A \in \mathcal{A}.$

(c)  $f \circ \mu = g \circ \mu.$

(d)  $\int_A f \, d \mu = \int_A g \, d \mu, \quad \forall A \in \mathcal{E}.$

Then:

(i) (a)  $\Rightarrow$  (b), if  $f$  and  $g$  are quasi- $\mu$ -integrable.

(ii) (a)  $\Leftrightarrow$  (b), if  $f, g$  are  $\mu$ -integrable.

(iii) (a)  $\Leftrightarrow$  (c), if  $f, g$  are  $\mu$ -integrable and nonnegative.

(iv) (a)  $\Leftrightarrow$  (b)  $\Leftrightarrow$  (c)  $\Leftrightarrow$  (d), if  $f, g$  are  $\mu$ -integrable, nonnegative, and  $\mathcal{E} \subset \mathcal{A}$  is  $\cap$ -stable with  $\sigma(\mathcal{E}) = \mathcal{A}$ .

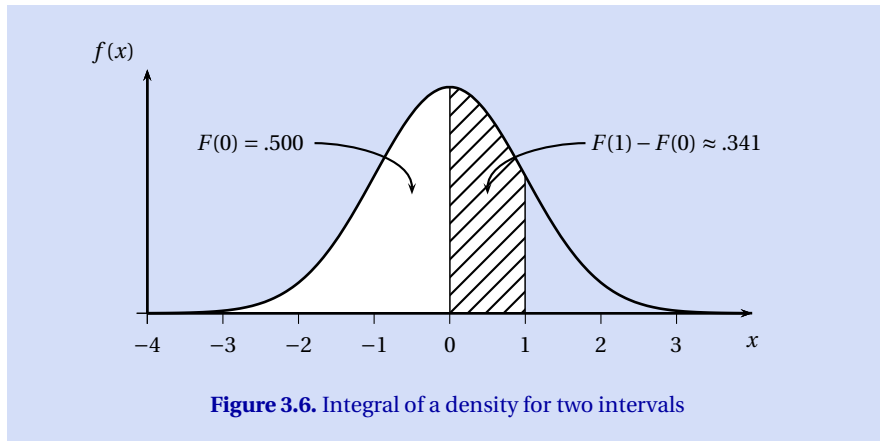
(Proof p. 119)

**Example 3.69 (A Density of the Normal Distribution)** As a special case of Equation (3.70) consider

$$\nu(A) = \int_A f \, d \lambda, \quad \forall A \in \mathcal{A}, \quad (3.73)$$

with

$$f(x) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left(\frac{-x^2}{2}\right), \quad x \in \mathbb{R}. \quad (3.74)$$



**Figure 3.6.** Integral of a density for two intervals

In this case, the measure  $\nu = f \circ \lambda$  is a probability measure and it is called the *standard normal distribution*. For an interval  $[a, b]$ , Theorem 3.67 yields

$$\nu([a, b]) = \int 1_{[a, b]} d\nu = \int 1_{[a, b]} d(f \circ \lambda) = \int 1_{[a, b]} f d\lambda = \int_a^b f(x) dx, \quad (3.75)$$

because  $f$  is Riemann  $\mu$ -integrable (see Th. 3.62). According to this equation, the value  $\nu([a, b])$  of the interval  $[a, b]$  can be represented as the area between the density and the  $x$ -axis above  $[a, b]$ . Figure 3.6 illustrates this fact for the interval  $[0, 1]$ . In this figure,

$$F(\alpha) = \int_{-\infty}^{\alpha} f(x) dx, \quad \alpha \in \mathbb{R}, \quad (3.76)$$

denotes the corresponding distribution function (see Def. 5.81), which is a special antiderivative of  $f$  [see Eq. (3.69)].  $\triangleleft$

### 3.5 Absolute Continuity and the Radon-Nikodym Theorem

Let  $\mu$  and  $\nu$  be measures on a measurable space  $(\Omega, \mathcal{A})$ . A necessary and sufficient condition for the existence of a density of  $\nu$  with respect to  $\mu$  is formulated in the Radon-Nikodym Theorem (see Th. 3.72), which is used not only for densities but also introducing conditional expectations (see ch. 10). The following definition prepares this theorem.

#### **Definition 3.70 (Absolute Continuity)**

Let  $\mu$  and  $\nu$  be measures on a measurable space  $(\Omega, \mathcal{A})$ .

- (i) The measure  $\nu$  is called *absolutely continuous with respect to  $\mu$* , denoted  $\nu \ll_{\mathcal{A}} \mu$ , if

$$\forall A \in \mathcal{A}: \mu(A) = 0 \Rightarrow \nu(A) = 0. \quad (3.77)$$

(ii) The measures  $\mu$  and  $\nu$  are called *null-set equivalent*, denoted  $\nu \approx \mu$ , if  $\nu \ll_{\mathcal{A}} \mu$  and  $\mu \ll_{\mathcal{A}} \nu$ , i. e., if

$$\forall A \in \mathcal{A}: \mu(A) = 0 \Leftrightarrow \nu(A) = 0. \quad (3.78)$$

If there is ambiguity about the measurable space, we use the terms *absolutely continuous on  $(\Omega, \mathcal{A})$*  and *null-set equivalent on  $(\Omega, \mathcal{A})$* .

**Remark 3.71 (An Implication)** If there is a density  $f$  of  $\nu$  with respect to  $\mu$ , then  $\nu \ll_{\mathcal{A}} \mu$ . This is a straightforward implication of Lemma 3.45 and (3.72) (see Exercise 3-12).  $\triangleleft$

Vice versa, if  $\nu \ll_{\mathcal{A}} \mu$ , then, according to the following theorem, there is a density  $f$  of  $\nu$  with respect to  $\mu$ , provided that  $\mu$  and  $\nu$  are  $\sigma$ -finite (see Definition 1.63).

**Theorem 3.72 (Radon-Nikodym)**

Let  $\mu$  and  $\nu$  be  $\sigma$ -finite measures on a measurable space  $(\Omega, \mathcal{A})$ .

- (i) Then  $\nu$  has a measurable density with respect to  $\mu$  if and only if  $\nu \ll_{\mathcal{A}} \mu$ . This density is denoted by  $\frac{d\nu}{d\mu}$  and called *Radon-Nikodym derivative*.
- (ii) If  $\nu \ll_{\mathcal{A}} \mu$ , then  $\frac{d\nu}{d\mu}$  is real-valued  $\mu$ -almost everywhere.

For a proof, see Klenke (2013, Corollary 7.34, p. 158) or Bauer (2001, Theorem 17.10, p. 102 and Theorem 17.11, p. 104).

**Remark 3.73 ( $\mu$ -Equivalence of Densities)** Note that, for  $\sigma$ -finite measures, all densities of  $\nu$  with respect to  $\mu$  are pairwise  $\mu$ -equivalent (for a proof see Bauer, 2001, Theorem 17.11, p. 104). If  $\nu$  is a finite measure, which is equivalent to  $\mu$ -integrability of the density  $\frac{d\nu}{d\mu}$ , then Theorem 3.68 (iii) implies  $\mu$ -equivalence of all densities of  $\nu$  with respect to  $\mu$ . The premise that  $\nu$  is a finite holds in particular if  $\nu$  is a probability measure.  $\triangleleft$

**Remark 3.74 (An Implication of the Radon-Nikodym Theorem)** Theorem 3.72 implies for  $\sigma$ -finite measures  $\nu$  and  $\mu$ : If  $\nu$  and  $\mu$  are null-set equivalent, then  $\frac{d\nu}{d\mu}$  and  $\frac{d\mu}{d\nu}$  both exist.  $\triangleleft$

The Radon-Nikodym theorem is used to prove the existence of the conditional expectation [see the proof of Theorem 10.9 (Bauer, 1996, Theorem 15.1, p. 111)]. The following corollary immediately follows from Theorem 3.67 and 3.72.

**Corollary 3.75 (An Implication of the Radon-Nikodym Theorem)**

Let  $\mu$  and  $\nu$  be  $\sigma$ -finite measures on a measurable space  $(\Omega, \mathcal{A})$ , and suppose  $\nu \ll_{\mathcal{A}} \mu$ . Furthermore, let  $g: (\Omega, \mathcal{A}) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  be a measurable function.

(i) If  $g$  is nonnegative or  $\nu$ -integrable, then

$$\int g d\nu = \int g \cdot \frac{d\nu}{d\mu} d\mu. \quad (3.79)$$

(ii)  $g$  is  $\nu$ -integrable if and only if  $g \cdot \frac{d\nu}{d\mu}$  is  $\mu$ -integrable.

**3.6 Integral With Respect to a Product Measure**

The following theorem shows that integration with respect to a product measure can be decomposed into a two-fold iterated integration where the order of integration is arbitrary.

**Theorem 3.76 (Fubini)**

Let  $(\Omega_i, \mathcal{A}_i, \mu_i)$ ,  $i = 1, 2$ , be  $\sigma$ -finite measure spaces and let  $f: \Omega_1 \times \Omega_2 \rightarrow \overline{\mathbb{R}}$  be  $(\mathcal{A}_1 \otimes \mathcal{A}_2, \overline{\mathcal{B}})$ -measurable. Furthermore, let  $f_i: \Omega_i \rightarrow \overline{\mathbb{R}}$ ,  $i = 1, 2$ , be defined by

$$f_1(\omega_1) := \int f(\omega_1, \omega_2) \mu_2(d\omega_2) \quad \text{and} \quad f_2(\omega_2) := \int f(\omega_1, \omega_2) \mu_1(d\omega_1).$$

If  $f$  is nonnegative or integrable with respect to the product measure  $\mu_1 \otimes \mu_2$ , then the functions  $f_i$  are  $(\mathcal{A}_i, \overline{\mathcal{B}})$ -measurable,  $i = 1, 2$ . Furthermore,

$$\begin{aligned} \int_{\Omega_1 \times \Omega_2} f d(\mu_1 \otimes \mu_2) &= \int_{\Omega_1 \times \Omega_2} f(\omega_1, \omega_2) \mu_1 \otimes \mu_2 [d(\omega_1, \omega_2)] \\ &= \int_{\Omega_1} \left( \int_{\Omega_2} f(\omega_1, \omega_2) \mu_2(d\omega_2) \right) \mu_1(d\omega_1) \\ &= \int_{\Omega_2} \left( \int_{\Omega_1} f(\omega_1, \omega_2) \mu_1(d\omega_1) \right) \mu_2(d\omega_2). \end{aligned} \quad (3.80)$$

For a proof, see Klenke (2013, Th. 14.16, p. 278). If  $f = 1_C$  for  $C \in \mathcal{A}_1 \otimes \mathcal{A}_2$ , then this theorem and (3.9) immediately yield the following corollary:

**Corollary 3.77 (Indicators)**

Let  $(\Omega_i, \mathcal{A}_i, \mu_i)$ ,  $i = 1, 2$ , be  $\sigma$ -finite measure spaces and let  $f: \Omega_1 \times \Omega_2 \rightarrow \overline{\mathbb{R}}$  be  $(\mathcal{A}_1 \otimes \mathcal{A}_2, \overline{\mathcal{B}})$ -measurable. Furthermore, let  $C \in \mathcal{A}_1 \otimes \mathcal{A}_2$  and define

$$\begin{aligned}
& \forall \omega_1 \in \Omega_1: C_{\omega_1} := \{\omega_2 \in \Omega_2: (\omega_1, \omega_2) \in C\} \\
& \text{and} \\
& \forall \omega_2 \in \Omega_2: C_{\omega_2} := \{\omega_1 \in \Omega_1: (\omega_1, \omega_2) \in C\}. \\
& \text{Then} \\
& \mu_1 \otimes \mu_2(C) = \int \mu_2(C_{\omega_1}) \mu_1(d\omega_1) = \int \mu_1(C_{\omega_2}) \mu_2(d\omega_2). \quad (3.81) \\
& \hspace{15em} (\text{Proof p. 120})
\end{aligned}$$

**Remark 3.78 (A Special Case)** Choosing  $C = A_1 \times A_2$ , Equation (3.81) yields

$$\begin{aligned}
\mu_1 \otimes \mu_2(A_1 \times A_2) &= \int \mu_2(A_2) \cdot \mathbf{1}_{A_1}(\omega_1) \mu_1(d\omega_1) = \mu_2(A_2) \cdot \int \mathbf{1}_{A_1}(\omega_1) \mu_1(d\omega_1) \\
&= \mu_1(A_1) \cdot \mu_2(A_2),
\end{aligned}$$

which is consistent with Equation (1.50).  $\triangleleft$

### 3.7 Proofs

#### *Proof of Lemma 3.33*

(i) If  $f$  is measurable and  $A \in \mathcal{A}$ , then  $\mathbf{1}_A$  is measurable as well (see Th. 2.57 and Example 2.12). Suppose that  $f$  is quasi- $\mu$ -integrable, i. e., suppose that  $\int f^+ d\mu$  or  $\int f^- d\mu$  are finite. Because

$$(\mathbf{1}_A f)^+ = \mathbf{1}_A f^+ \quad \text{and} \quad (\mathbf{1}_A f)^- = \mathbf{1}_A f^-$$

as well as

$$0 \leq \mathbf{1}_A f^+ \leq f^+ \quad \text{and} \quad 0 \leq \mathbf{1}_A f^- \leq f^-,$$

monotonicity of the integral of nonnegative measurable functions (Lemma 3.26) yields

$$0 \leq \int (\mathbf{1}_A f)^+ d\mu = \int \mathbf{1}_A f^+ d\mu \leq \int f^+ d\mu$$

and

$$0 \leq \int (\mathbf{1}_A f)^- d\mu = \int \mathbf{1}_A f^- d\mu \leq \int f^- d\mu,$$

which implies that  $\int (\mathbf{1}_A f)^+ d\mu$  or  $\int (\mathbf{1}_A f)^- d\mu$  is finite. Hence  $f$  is quasi- $\mu$ -integrable.

(ii) If  $f$  is  $\mu$ -integrable, then  $\int f^+ d\mu < \infty$  and  $\int f^- d\mu < \infty$ . Just like in the proof of (i) this implies  $\int (\mathbf{1}_A f)^+ d\mu < \infty$  and  $\int (\mathbf{1}_A f)^- d\mu < \infty$ . Hence,  $\mathbf{1}_A f$  is  $\mu$ -integrable.

#### *Proof of Theorem 3.36*

(i) Step 1: Let  $\alpha \geq 0$ ,  $f$  be a nonnegative step function, and  $f = \sum_{i=1}^n \alpha_i \mathbf{1}_{A_i}$  a normal representation (see Rem. 3.8). Then, according to (3.4),

$$\begin{aligned}
\int \alpha f d\mu &= \int \alpha \sum_{i=1}^n \alpha_i 1_{A_i} d\mu && \text{[Rem. 3.8]} \\
&= \int \sum_{i=1}^n (\alpha \alpha_i) 1_{A_i} d\mu \\
&= \sum_{i=1}^n (\alpha \alpha_i) \mu(A_i) && \text{[(3.4)]} \\
&= \alpha \sum_{i=1}^n \alpha_i \mu(A_i) \\
&= \alpha \int f d\mu. && \text{[(3.4)]}
\end{aligned}$$

Step 2: Let  $\alpha \geq 0$ ,  $f$  be a nonnegative measurable function, and  $f_1 \leq f_2 \leq \dots$  an increasing sequence of nonnegative step functions with  $\lim_{n \rightarrow \infty} f_n = f$  (see Th. 3.19). Then, according to Equation (3.22),

$$\begin{aligned}
\int \alpha f d\mu &= \int \alpha \lim_{n \rightarrow \infty} f_n d\mu \\
&= \int \lim_{n \rightarrow \infty} (\alpha f_n) d\mu \\
&= \lim_{n \rightarrow \infty} \int \alpha f_n d\mu && \text{[(3.22)]} \\
&= \lim_{n \rightarrow \infty} \alpha \int f_n d\mu && \text{[Step 1]} \\
&= \alpha \lim_{n \rightarrow \infty} \int f_n d\mu \\
&= \alpha \int f d\mu. && \text{[(3.22)]}
\end{aligned}$$

Step 3: Assume that  $\alpha \geq 0$  and that  $f$  is quasi- $\mu$ -integrable. Because  $\alpha f = \alpha(f^+ - f^-) = \alpha f^+ - \alpha f^-$ ,

$$\begin{aligned}
\int \alpha f d\mu &= \int \alpha f^+ d\mu - \int \alpha f^- d\mu && \text{[(3.27)]} \\
&= \alpha \int f^+ d\mu - \alpha \int f^- d\mu && \text{[Step 2]} \\
&= \alpha \left( \int f^+ d\mu - \int f^- d\mu \right) \\
&= \alpha \int f d\mu. && \text{[(3.27)]}
\end{aligned}$$

This proves Equation (3.32) for  $\alpha \geq 0$ . For  $\alpha < 0$ , note that

$$(\alpha f)^+ = -\alpha f^- \quad \text{and} \quad (\alpha f)^- = -\alpha f^+. \quad (3.82)$$

Therefore,

$$\int \alpha f d\mu = \int (\alpha f)^+ d\mu - \int (\alpha f)^- d\mu \quad \text{[(3.27)]}$$

$$\begin{aligned}
&= \int (-\alpha) f^- d\mu - \int (-\alpha) f^+ d\mu && [(3.82)] \\
&= -\alpha \int f^- d\mu - (-\alpha) \int f^+ d\mu && [-\alpha > 0, \text{ first part of Step 3}] \\
&= \alpha \left( \int f^+ d\mu - \int f^- d\mu \right) \\
&= \alpha \int f d\mu. && [(3.27)]
\end{aligned}$$

This shows that  $\int \alpha f d\mu = \alpha \int f d\mu$  holds for all  $\alpha \in \mathbb{R}$ , all quasi- $\mu$ -integrable  $f$ , and therefore also for all integrable  $f$ . This also implies that  $\alpha f$  is quasi- $\mu$ -integrable or  $\mu$ -integrable if  $f$  is quasi- $\mu$ -integrable or  $\mu$ -integrable, respectively.

(ii) Step 1: Let  $f$  and  $g$  be nonnegative step functions and let  $f = \sum_{i=1}^n \alpha_i \mathbf{1}_{A_i}$ ,  $g = \sum_{j=1}^m \beta_j \mathbf{1}_{B_j}$  be normal representations (see Rem. 3.8) with  $\bigcup_{i=1}^n A_i = \Omega$  and  $\bigcup_{j=1}^m B_j = \Omega$ . (Note that these latter requirements can always be fulfilled using  $A_n := \Omega \setminus \bigcup_{i=1}^{n-1} A_i$  and  $\alpha_n := 0$ , if  $f = \sum_{i=1}^{n-1} \alpha_i \mathbf{1}_{A_i}$  is already a normal representation.) Then  $f + g = \sum_{i=1}^n \alpha_i \mathbf{1}_{A_i} + \sum_{j=1}^m \beta_j \mathbf{1}_{B_j}$  is again a nonnegative step function (see Def. 3.9) and

$$f + g = \sum_{i=1}^n \sum_{j=1}^m \gamma_{ij} \mathbf{1}_{C_{ij}}$$

is a normal representation, where  $C_{ij} = A_i \cap B_j$  and  $\gamma_{ij} = \alpha_i + \beta_j$ . Note that some of these sets  $C_{ij}$  may be empty. Now,

$$\begin{aligned}
\int f + g d\mu &= \sum_{i=1}^n \sum_{j=1}^m \gamma_{ij} \mu(C_{ij}) && [(3.4)] \\
&= \sum_{i=1}^n \sum_{j=1}^m (\alpha_i + \beta_j) \mu(A_i \cap B_j) \\
&= \sum_{i=1}^n \sum_{j=1}^m \alpha_i \mu(A_i \cap B_j) + \sum_{i=1}^n \sum_{j=1}^m \beta_j \mu(A_i \cap B_j) \\
&= \sum_{i=1}^n \alpha_i \sum_{j=1}^m \mu(A_i \cap B_j) + \sum_{j=1}^m \beta_j \sum_{i=1}^n \mu(A_i \cap B_j) \\
&= \sum_{i=1}^n \alpha_i \mu(A_i) + \sum_{j=1}^m \beta_j \mu(B_j) && [\text{Rem. 1.47}] \\
&= \int f d\mu + \int g d\mu. && [(3.4)]
\end{aligned}$$

Step 2: Let  $f, g$  be nonnegative measure functions and  $f_1 \leq f_2 \leq \dots$ ,  $g_1 \leq g_2 \leq \dots$  increasing sequences of nonnegative step functions with  $\lim_{n \rightarrow \infty} f_n = f$  and  $\lim_{n \rightarrow \infty} g_n = g$ , respectively (see Th. 3.19). Then  $f_1 + g_1 \leq f_2 + g_2 \leq \dots$  is an increasing sequence of nonnegative step functions with  $\lim_{n \rightarrow \infty} (f_n + g_n) = \lim_{n \rightarrow \infty} f_n + \lim_{n \rightarrow \infty} g_n = f + g$ . Then

$$\begin{aligned}
\int f + g \, d\mu &= \int \lim_{n \rightarrow \infty} (f_n + g_n) \, d\mu \\
&= \lim_{n \rightarrow \infty} \int (f_n + g_n) \, d\mu && [(3.22)] \\
&= \lim_{n \rightarrow \infty} \left( \int f_n \, d\mu + \int g_n \, d\mu \right) && [\text{Step 1}] \\
&= \lim_{n \rightarrow \infty} \int f_n \, d\mu + \lim_{n \rightarrow \infty} \int g_n \, d\mu \\
&= \int f \, d\mu + \int g \, d\mu. && [(3.22)]
\end{aligned}$$

Step 3: Assume that  $f$  is quasi- $\mu$ -integrable and  $g$  is  $\mu$ -integrable. Then

$$f + g = f^+ - f^- + g^+ - g^-,$$

$$f + g = (f + g)^+ - (f + g)^-.$$

This implies

$$(f + g)^+ - (f + g)^- = f^+ - f^- + g^+ - g^-,$$

which is equivalent to

$$(f + g)^+ + f^- + g^- = (f + g)^- + f^+ + g^+.$$

Applying the result of Step 2 yields

$$\begin{aligned}
&\int (f + g)^+ \, d\mu + \int f^- \, d\mu + \int g^- \, d\mu \\
&= \int (f + g)^- \, d\mu + \int f^+ \, d\mu + \int g^+ \, d\mu.
\end{aligned} \tag{3.83}$$

If  $g$  is  $\mu$ -integrable, then  $\int g^+ \, d\mu$  and  $\int g^- \, d\mu$  are finite, and if  $f$  is quasi- $\mu$ -integrable, then at most one of  $\int f^+ \, d\mu$  and  $\int f^- \, d\mu$  is infinite, the other one is finite. Furthermore,  $(f + g)^+ \leq f^+ + g^+$  and  $(f + g)^- \leq f^- + g^-$ . Hence, Lemma 3.26 implies

$$\int (f + g)^+ \, d\mu \leq \int f^+ + g^+ \, d\mu = \int f^+ \, d\mu + \int g^+ \, d\mu \tag{3.84}$$

and

$$\int (f + g)^- \, d\mu \leq \int f^- + g^- \, d\mu = \int f^- \, d\mu + \int g^- \, d\mu.$$

Therefore, at most one of the integrals  $\int (f + g)^+ \, d\mu$  and  $\int (f + g)^- \, d\mu$  is infinite and this implies that  $f + g$  is quasi- $\mu$ -integrable. If  $\int (f + g)^+ \, d\mu = \infty$ , then

$$\int (f + g) \, d\mu = \int (f + g)^+ \, d\mu - \int (f + g)^- \, d\mu = \infty,$$

and, according to (3.84),  $\int f^+ d\mu = \infty$ . This implies

$$\int f d\mu + \int g d\mu = \int f^+ d\mu - \int f^- d\mu + \int g d\mu = \infty.$$

Analogously, if  $\int (f+g)^- d\mu = \infty$ , then

$$\int (f+g) d\mu = -\infty = \int f d\mu + \int g d\mu.$$

If both,  $\int (f+g)^+ d\mu$  and  $\int (f+g)^- d\mu$  are finite, then (3.83) is equivalent to

$$\int (f+g)^+ d\mu - \int (f+g)^- d\mu = \int f^+ d\mu - \int f^- d\mu + \int g^+ d\mu - \int g^- d\mu,$$

which in turn is equivalent to

$$\int (f+g) d\mu = \int f d\mu + \int g d\mu.$$

### ***Proof of Corollary 3.38***

Because  $|f| = f^+ + f^-$ , this proposition immediately follows from the definition of integrability (see Def. 3.28) and linearity of the integral [see Eq. (3.34)].

### ***Proof of Lemma 3.40***

If  $f$  is  $\mathcal{C}$ -measurable, then  $f^+$  and  $f^-$  are  $\mathcal{C}$ -measurable as well (see Th. 2.66). Furthermore,  $f^+$  and  $f^-$  can be represented as limits of increasing sequences of nonnegative step functions on  $(\Omega, \mathcal{C})$  [see Th. 3.19 (i)]. Hence, according to Equations (3.22) and (3.4), the values of the integrals  $\int f^+ d\mu$  and  $\int f^- d\mu$  only depend on the values  $\mu(A)$ ,  $A \in \mathcal{C}$ . Therefore, if  $\mu$  and  $\nu$  are identical on  $\mathcal{C}$ , then  $\int f d\mu = \int f d\nu$ , for all  $\mathcal{C}$ -measurable functions that are nonnegative or  $\mu$ -integrable.

### ***Proof of Lemma 3.41***

Define

$$A_+ := \{\omega \in \Omega: f(\omega) = \infty\} \in \mathcal{A} \quad \text{and} \quad A_- := \{\omega \in \Omega: f(\omega) = -\infty\} \in \mathcal{A}.$$

If  $\mu(A_+) > 0$ , then define the increasing sequence  $g_n: \Omega \rightarrow [0, \infty)$ ,  $n \in \mathbb{N}$ , by  $g_n = n \cdot 1_{A_+}$ . Because  $1_{A_+} \cdot f^+ = \lim_{n \rightarrow \infty} g_n$ ,

$$\begin{aligned}
\int f^+ d\mu &= \int 1_{A_+} \cdot f^+ d\mu + \int 1_{\Omega \setminus A_+} \cdot f^+ d\mu && [(3.36)] \\
&= \lim_{n \rightarrow \infty} \int g_n d\mu + \int 1_{\Omega \setminus A_+} \cdot f^+ d\mu && [\text{Def. 3.24}] \\
&= \lim_{n \rightarrow \infty} \int n \cdot 1_{A_+} d\mu + \int 1_{\Omega \setminus A_+} \cdot f^+ d\mu \\
&= \lim_{n \rightarrow \infty} n \cdot \mu(A_+) + \int 1_{\Omega \setminus A_+} \cdot f^+ d\mu && [(3.4)] \\
&= \infty.
\end{aligned}$$

Analogously we can prove that  $\int f^- d\mu = \infty$ , if  $\mu(A_-) > 0$  replacing  $f^+$  by  $f^-$  and  $A_+$  by  $A_-$ . Therefore, if  $f$  is quasi- $\mu$ -integrable and  $\int f^+ d\mu = \infty$ , then  $\int f^- d\mu$  is finite (see Def. 3.28) and  $\int f d\mu = \int f^+ d\mu - \int f^- d\mu = \infty$ . This proves (3.38). Analogously, if  $f$  is quasi- $\mu$ -integrable and  $\int f^- d\mu = \infty$ , then  $\int f^+ d\mu$  is finite (see Def. 3.28) and  $\int f d\mu = \int f^+ d\mu - \int f^- d\mu = -\infty$ , which proves (3.39). Finally, if  $\mu(A_+ \cup A_-) > 0$  and hence,  $\mu(A_+) > 0$  or  $\mu(A_-) > 0$ , then, according to (3.38) and (3.39),  $\int f d\mu$  is not defined or not finite. Thus, by contraposition, if  $f$  is  $\mu$ -integrable, then  $\mu(A_+ \cup A_-) = 0$ , i. e.,  $f$  is real-valued,  $\mu$ -almost everywhere.

#### **Proof of Lemma 3.44**

If  $f(\omega) > 0$ , for all  $\omega \in A$ , then  $1_A \cdot f: (\Omega, \mathcal{A}, \mu) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  is a nonnegative measurable function (see Th. 2.57). Hence,  $\int 1_A \cdot f d\mu \geq 0$  (see Defs. 3.24 and 3.10). Because  $\mu(\{\omega \in \Omega: (1_A \cdot f(\omega) > 0)\}) = \mu(A)$ , the assumption  $\mu(A) > 0$  implies that  $1_A \cdot f \stackrel{\mu}{=} 0$  does *not hold*. Therefore, according to Equation (3.40),  $\int 1_A \cdot f d\mu \neq 0$ , and we can conclude:  $\int 1_A \cdot f d\mu > 0$ .

#### **Proof of Lemma 3.45**

If  $f: (\Omega, \mathcal{A}, \mu) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  is measurable and  $A \in \mathcal{A}$  with  $\mu(A) = 0$ , then  $1_A \cdot f$  is measurable (see Th. 2.57) and  $1_A \cdot f \stackrel{\mu}{=} 0$  (see Exercise 3-9). This implies  $(1_A \cdot f)^+ \stackrel{\mu}{=} 0$  and  $(1_A \cdot f)^- \stackrel{\mu}{=} 0$ . Now Equation (3.40) yields  $\int (1_A \cdot f)^+ d\mu = 0$  and  $\int (1_A \cdot f)^- d\mu = 0$ . Hence,  $\int 1_A \cdot f d\mu$  exists (see Def. 3.28) and Equation (3.27) implies

$$\int 1_A \cdot f d\mu = \int (1_A \cdot f)^+ d\mu - \int (1_A \cdot f)^- d\mu = 0.$$

#### **Proof of Lemma 3.47**

Define  $A := \{\omega \in \Omega: f(\omega) \neq g(\omega)\}$  and note that  $A \in \mathcal{A}$  [see Rem. 2.67 (c)]. Therefore,  $f \stackrel{\mu}{=} g$  implies  $\mu(A) = 0$ . Hence,

$$\begin{aligned}
\int f d\mu &= \int_{\Omega \setminus A} f d\mu && [(3.43)] \\
&= \int_{\Omega \setminus A} g d\mu && [\text{Def. of } A] \\
&= \int g d\mu. && [(3.43)]
\end{aligned}$$

**Proof of Theorem 3.48**

(a)

$f, g$  are  $\mu$ -integrable and  $f \stackrel{\mu}{=} g$

$$\Rightarrow \forall A \in \mathcal{A}: 1_A \cdot f \stackrel{\mu}{=} 1_A \cdot g \text{ and } 1_A \cdot f, 1_A \cdot g \text{ are } \mu\text{-integrable} \quad [(2.33), (3.29)]$$

$$\Rightarrow \forall A \in \mathcal{A}: \int_A f d\mu = \int_A g d\mu. \quad [(3.44)]$$

(b) If  $f, g$  are  $\mu$ -integrable, then  $f, g$  are real-valued  $\mu$ -a.e. (see Lemma 3.41). Hence, for

$$B := \{\omega \in \Omega: f(\omega) \in \{-\infty, \infty\}\} \cup \{\omega \in \Omega: g(\omega) \in \{-\infty, \infty\}\},$$

$B \in \mathcal{A}$  and  $\mu(B) = 0$ . Now define

$$A_{>} := \{\omega \in \Omega: f(\omega) > g(\omega)\} \quad \text{and} \quad A_{<} := \{\omega \in \Omega: f(\omega) < g(\omega)\}.$$

According to Remark 2.67 (b) and (a),  $A_{>} \in \mathcal{A}$  and  $A_{<} \in \mathcal{A}$ . Then

$$\begin{aligned}
\forall A \in \mathcal{A}: \int_A f d\mu &= \int_A g d\mu \\
\Rightarrow \forall A \in \mathcal{A}: \int_{A \cap B^c} f d\mu &= \int_{A \cap B^c} g d\mu && [(3.43)] \\
\Rightarrow \forall A \in \mathcal{A}: \int_{A \cap B^c} (f - g) d\mu &= 0 && [f, g \text{ } \mu\text{-integrable, (3.34)}] \\
\Rightarrow \int_{A_{>} \cap B^c} (f - g) d\mu &= 0 \quad \text{and} \quad \int_{A_{<} \cap B^c} (f - g) d\mu = 0 && [A_{>}, A_{<} \in \mathcal{A}] \\
\Rightarrow \mu(A_{>} \cap B^c) &= 0 \text{ and } \mu(A_{<} \cap B^c) = 0 && [(3.41)] \\
\Rightarrow \mu(A_{>} \cup A_{<}) &= \mu((A_{>} \cap B^c) \cup (A_{>} \cap B) \cup (A_{<} \cap B^c) \cup (A_{<} \cap B)) = 0 && [\text{Box 1.1 (ii)}] \\
\Rightarrow f &\stackrel{\mu}{=} g. && [\text{Def. 2.68}]
\end{aligned}$$

**Proof of Theorem 3.52**

(i) Define  $A := \{\omega \in \Omega: f(\omega) > g(\omega)\}$ . According to Remark 2.67 (b),  $A \in \mathcal{A}$ . Therefore,  $f \stackrel{\mu}{\leq} g$  implies  $\mu(A) = 0$ . Furthermore, define

$$A_{-\infty} := \{\omega \in \Omega \setminus A: f(\omega) = -\infty\} \quad \text{and} \quad A_{\infty} := \{\omega \in \Omega \setminus A: f(\omega) = g(\omega) = \infty\}.$$

If  $\mu(A_{-\infty}) > 0$ , then  $\int f d\mu = -\infty$  [see (3.39)]. Therefore,  $\int f d\mu \leq \int g d\mu$ . If  $\mu(A_{\infty}) > 0$ , then  $\int f d\mu = \int g d\mu = \infty$  [see (3.38)], and therefore  $\int f d\mu \leq \int g d\mu$ .

If  $\mu(A_{-\infty}) = \mu(A_{\infty}) = 0$ , then  $f, g$  are real-valued  $\mu$ -a.e. Now define  $B := (\Omega \setminus A) \setminus (A_{\infty} \cup A_{-\infty})$ , which implies  $\mu(\Omega \setminus B) = 0$ , and

$$\forall \omega \in B: f(\omega), g(\omega) \text{ are finite, } f(\omega) \leq g(\omega).$$

If  $\int (1_B f)^- d\mu = \infty$ , then  $\int_B f d\mu = -\infty$ . Hence,  $\int f d\mu \leq \int g d\mu$ . If  $\int (1_B f)^+ d\mu = \infty$ , then  $1_B f \leq 1_B g$  implies  $(1_B f)^+ \leq (1_B g)^+$  and  $\int (1_B g)^+ d\mu = \infty$  [see Eq. (3.24)]. Therefore,  $\int f d\mu = \int g d\mu = \infty$ , which implies  $\int f d\mu \leq \int g d\mu$ .

Now, if all  $(1_B f)^+, (1_B f)^-, (1_B g)^+, (1_B g)^-$  are  $\mu$ -integrable, then

$$\begin{aligned} \int g d\mu &= \int_B g d\mu && [(3.43)] \\ &= \int_B (f + g - f) d\mu && [1_B f, 1_B g \text{ are real-valued}] \\ &= \int_B f d\mu + \int_B (g - f) d\mu && [(3.34)] \\ &\geq \int_B f d\mu && [1_B(g - f) \geq 0, (3.24)] \\ &= \int f d\mu. && [(3.43)] \end{aligned}$$

(ii) Define

$$B := \{\omega \in \Omega: f(\omega) \in \{-\infty, \infty\}\} \cup \{\omega \in \Omega: g(\omega) \in \{-\infty, \infty\}\}.$$

If  $f, g$  are  $\mu$ -integrable, then Lemma 3.41 implies  $\mu(B) = 0$ . Furthermore, define  $A := \{\omega \in \Omega \setminus B: f(\omega) \geq g(\omega)\}$ . Then  $f \leq_\mu g$  implies  $\mu(A) = 0$ . Hence,  $\mu(A \cup B) = 0$  [see Box 1.1 (xi)]. Now,

$$\begin{aligned} \int g d\mu &= \int_{\Omega \setminus (A \cup B)} g d\mu && [(3.43)] \\ &= \int_{\Omega \setminus (A \cup B)} (f + g - f) d\mu && [1_{\Omega \setminus (A \cup B)} f, 1_{\Omega \setminus (A \cup B)} g \text{ are real-valued}] \\ &= \int_{\Omega \setminus (A \cup B)} f d\mu + \int_{\Omega \setminus (A \cup B)} (g - f) d\mu && [(3.34)] \\ &> \int_{\Omega \setminus (A \cup B)} f d\mu. && [1_{\Omega \setminus (A \cup B)}(g - f) > 0, \mu(\Omega \setminus (A \cup B)) > 0, \text{Lem. 3.44}] \end{aligned}$$

### **Proof of Corollary 3.59**

Assume that  $f: (\Omega, \mathcal{A}) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  is measurable with a finite number of *positive* values  $\alpha_1, \dots, \alpha_m > 0$  and a finite number of *negative* values  $\alpha_{m+1}, \dots, \alpha_n < 0$ . By

convention, if  $m = n$ , then  $\sum_{i=m+1}^n \alpha_i 1_{A_i} = 0$ , and if  $m = 0$ , then  $\sum_{i=1}^m \alpha_i 1_{A_i} = 0$ . Then

$$f = \sum_{i=1}^n \alpha_i 1_{A_i} = \sum_{i=1}^m \alpha_i 1_{A_i} + \sum_{i=m+1}^n \alpha_i 1_{A_i}$$

and  $f^+ = \sum_{i=1}^m \alpha_i 1_{A_i}$  as well as  $f^- = -\sum_{i=m+1}^n \alpha_i 1_{A_i} = \sum_{i=m+1}^n -\alpha_i 1_{A_i}$ . Therefore,

$$\int f^+ d\mu = \sum_{i=1}^m \alpha_i \mu(A_i) \quad \text{and} \quad \int f^- d\mu = \sum_{i=m+1}^n -\alpha_i \mu(A_i),$$

and  $\int f^+ d\mu$  as well as  $\int f^- d\mu$  are finite if and only if  $\mu(A_i) < \infty$ , for all  $i = 1, \dots, n$ . Now,  $\mu(A_i) < \infty$ , for all  $i = 1, \dots, n$ , implies

$$\begin{aligned} \int f d\mu &= \int f^+ d\mu - \int f^- d\mu && \text{[Def. 3.28]} \\ &= \sum_{i=1}^m \alpha_i \mu(A_i) + \sum_{i=m+1}^n \alpha_i \mu(A_i) \\ &= \sum_{i=1}^n \alpha_i \mu(A_i) \\ &= \sum_{i=1}^n \alpha_i \mu_f(\{\alpha_i\}) && \text{[Def. 2.80, (2.10)].} \end{aligned}$$

Note that, in the last equation, we used the assumption that the  $\alpha_1, \dots, \alpha_n$  are pairwise different.

### ***Proof of Theorem 3.68***

- (i) This is the proposition of Lemma 3.47.
- (ii) This proposition is Theorem 3.48.
- (iii) If  $f, g$  are  $\mu$ -integrable and nonnegative, then it suffices to show: (b)  $\Leftrightarrow$  (c) [see (ii)]. Now

$$\begin{aligned} \int_A f d\mu &= \int_A g d\mu, \quad \forall A \in \mathcal{A} \\ \Leftrightarrow \int 1_A f d\mu &= \int 1_A g d\mu, \quad \forall A \in \mathcal{A} && \text{[Def. 3.35]} \\ \Leftrightarrow \int 1_A df \circ \mu &= \int 1_A dg \circ \mu, \quad \forall A \in \mathcal{A} && \text{[Th. 3.67 (i)]} \\ \Leftrightarrow f \circ \mu(A) &= g \circ \mu(A), \quad \forall A \in \mathcal{A} && \text{[(3.9)]} \\ \Leftrightarrow f \circ \mu &= g \circ \mu. \end{aligned}$$

- (iv) If  $f, g$  are  $\mu$ -integrable and nonnegative, then the equivalence of (a), (b), and (c) follows from (ii) and (iii). Hence, it suffices to show: (c)  $\Leftrightarrow$  (d). Because  $\mu$ -integrability of  $f$  and  $g$  implies that  $f \circ \mu$  and  $g \circ \mu$  are finite measures, applying Theorem 1.72 completes the proof.

**Proof of Corollary 3.77**

Note that

$$\forall (\omega_1, \omega_2) \in \Omega_1 \times \Omega_2: 1_C(\omega_1, \omega_2) = 1_{C_{\omega_1}}(\omega_2) = 1_{C_{\omega_2}}(\omega_1). \quad (3.85)$$

Now,

$$\mu_1 \otimes \mu_2(C) = \int 1_C d(\mu_1 \otimes \mu_2) \quad [(3.9)]$$

$$= \int \int 1_C(\omega_1, \omega_2) \mu_2(d\omega_2) \mu_1(d\omega_1) \quad [(3.80)]$$

$$= \int \int 1_{C_{\omega_1}}(\omega_2) \mu_2(d\omega_2) \mu_1(d\omega_1) \quad [(3.85)]$$

$$= \int \mu_2(C_{\omega_1}) \mu_1(d\omega_1). \quad [(3.9)]$$

The proof of the second equation is analog.

**3.8 Exercises**

▷ **Exercise 3-1** Construct a representation of the identity function on  $\Omega = \{1, \dots, n\}$  as a weighted sum of indicators of elements of  $\mathcal{A}$ , where  $n \in \mathbb{N}$  and  $(\Omega, \mathcal{A})$  with  $\mathcal{A} = \mathcal{P}(\Omega)$ .

▷ **Exercise 3-2** Prove that, for every nonnegative step function, there exists a normal representation (see Rem. 3.8).

▷ **Exercise 3-3** Consider the measure space  $(\Omega, \mathcal{A}, \lambda)$ , where  $\lambda$  is the Lebesgue measure. Show that the number  $\sum_{i=1}^n \alpha_i \lambda(A_i)$  assigned to the nonnegative step function  $f$  defined in Example 3.7 is identical for the four specified representations of  $f$ , two of which are nonnormal representations.

▷ **Exercise 3-4** Let  $(\Omega, \mathcal{A})$  be a measurable space and let  $A \in \mathcal{A}$ . Show that if  $f = \sum_{i=1}^n \alpha_i 1_{A_i}$  is a normal representation of a nonnegative step function, then the product  $1_A f$  of the indicator  $1_A$  and  $f$  is also a normal representation of a nonnegative step function, and  $1_A f = \sum_{i=1}^n \alpha_i 1_{A \cap A_i}$ .

▷ **Exercise 3-5** Prove Equation (3.8).

▷ **Exercise 3-6** Compute the integral of the identity mapping  $id: \Omega \rightarrow \Omega$  with respect to the counting measure  $\mu_{\#}$  on  $\mathcal{P}(\Omega)$ , where  $\Omega = \{1, \dots, n\}$ . Then look at it for  $n = 5$ .

▷ **Exercise 3-7** Compute the integrals  $\int f_1 d\lambda$  and  $\int f_2 d\lambda$  of the functions  $f_1$  and  $f_2$  defined in Example 3.17.

▷ **Exercise 3-8** Prove the propositions of Example 3.39.

▷ **Exercise 3-9** Let  $f: (\Omega, \mathcal{A}, \mu) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  be measurable and  $A \in \mathcal{A}$ . Show that  $1_A f \stackrel{\mu}{=} 0$  if  $\mu(A) = 0$ .

▷ **Exercise 3-10** Prove Equation (3.52).

▷ **Exercise 3-11** Compute the integral of the function  $f(x) = 10 - x^2$  considered in Example 3.64.

▷ **Exercise 3-12** Prove the proposition of Remark 3.71.

## Solutions

▷ **Solution 3-1** The identity function  $id: \Omega \rightarrow \Omega$  on  $\Omega = \{1, \dots, n\}$  is defined by

$$id(i) = i, \quad \forall i \in \Omega.$$

According to Example 2.9 it is  $(\mathcal{A}, \mathcal{A}_0)$ -measurable for all  $\sigma$ -algebras  $\mathcal{A}_0 \subset \mathcal{A}$ . Now consider the set  $\{1, \dots, n\}$  of values of  $id$  and the partition  $\{\{1\}, \dots, \{n\}\}$  of  $\Omega$ . Then

$$id = \sum_{i=1}^n i \cdot 1_{\{i\}} = 1 \cdot 1_{\{1\}} + \dots + n \cdot 1_{\{n\}}.$$

▷ **Solution 3-2** Let  $f: (\Omega, \mathcal{A}) \rightarrow (\mathbb{R}, \mathcal{B})$  be a nonnegative step function, with  $f = \sum_{i=1}^n \alpha_i 1_{A_i}$ . Define, for all nonempty  $J \subset \{1, \dots, n\}$ ,

$$B_J := \left( \bigcap_{i \in J} A_i \right) \cap \left( \bigcap_{i \notin J} A_i^c \right).$$

These are  $2^n - 1$  sets, where several of them may be empty, and all are pairwise disjoint. Then

$$f = \sum_{J: B_J \neq \emptyset} \left( \sum_{i \in J} \alpha_i \right) \cdot 1_{B_J}$$

is a normal representation of  $f$ .

▷ **Solution 3-3** We compute the sum for all four representations of  $f$ . The first one is:

$$\begin{aligned} \sum_{i=1}^4 \alpha_i \lambda(A_i) &= 2 \cdot (1-0) + 5 \cdot (2-1) + 4 \cdot (3-2) + 1 \cdot (4-3) \\ &= 2 \cdot 1 + 5 \cdot 1 + 4 \cdot 1 + 1 \cdot 1 = 12. \end{aligned}$$

The sum for the second representation of  $f$  is:

$$\begin{aligned} \sum_{i=1}^5 \gamma_i \lambda(C_i) &= 2 \cdot (.5-0) + 2 \cdot (1-.5) + 5 \cdot (2-1) + 4 \cdot (3-2) + 1 \cdot (4-3) \\ &= 2 \cdot .5 + 2 \cdot .5 + 5 \cdot 1 + 4 \cdot 1 + 1 \cdot 1 = 12. \end{aligned}$$

The sum for the third representation of  $f$ , which is nonnormal, is:

$$\begin{aligned} \sum_{i=1}^3 \beta_i \lambda(B_i) &= 2 \cdot (2-0) + 3 \cdot (3-1) + 1 \cdot (4-2) \\ &= 2 \cdot 2 + 3 \cdot 2 + 1 \cdot 2 = 12. \end{aligned}$$

The sum for the fourth representation of  $f$ , which is also nonnormal, is:

$$\begin{aligned} \sum_{i=1}^4 \delta_i \lambda(D_i) &= 1 \cdot (4-0) + 1 \cdot (3-0) + 2 \cdot (3-1) + 1 \cdot (2-1) \\ &= 1 \cdot 4 + 1 \cdot 3 + 2 \cdot 2 + 1 \cdot 1 = 12. \end{aligned}$$

Obviously, all four sums are identical.

▷ **Solution 3-4** Let  $f = \sum_{i=1}^n \alpha_i 1_{A_i}$ , where  $A_1, \dots, A_n \in \mathcal{A}$ . This implies  $A \cap A_1, \dots, A \cap A_n \in \mathcal{A}$ , for  $A \in \mathcal{A}$ . Therefore, and because of  $1_A \cdot f = \sum_{i=1}^n \alpha_i 1_{A \cap A_i}$ , the function  $1_A \cdot f$  is a non-negative step function. If  $f = \sum_{i=1}^n \alpha_i 1_{A_i}$  is a normal representation, then  $A_i \cap A_j = \emptyset$  for  $i \neq j$ , which implies  $(A \cap A_i) \cap (A \cap A_j) = A \cap (A_i \cap A_j) = A \cap \emptyset = \emptyset$ , for  $i \neq j$ . Therefore,  $1_A \cdot f = \sum_{i=1}^n \alpha_i 1_{A \cap A_i}$  is a normal representation as well.

▷ **Solution 3-5**  $\int_A \alpha d\mu = \int \alpha 1_A d\mu = \alpha \mu(A)$  [see Eq. (3.4)].

▷ **Solution 3-6** Consider the elements  $\{1\}, \dots, \{n\}$  of  $\mathcal{A} = \mathcal{P}(\Omega)$  and  $id = \sum_{\omega=1}^n \omega \cdot 1_{\{\omega\}}$ . According to Definition 3.10,

$$\begin{aligned} \int id d\mu_{\#} &= \sum_{\omega=1}^n \omega \cdot \mu_{\#}(\{\omega\}) \\ &= 1 \cdot \mu_{\#}(\{1\}) + 2 \cdot \mu_{\#}(\{2\}) + \dots + n \cdot \mu_{\#}(\{n\}) \\ &= \sum_{i=1}^n i = \frac{n(n+1)}{2} \end{aligned} \quad (3.86)$$

is the integral of  $id$  over  $\Omega$  with respect to the measure  $\mu_{\#}$ . Hence, in this example, the integral  $\int id d\mu_{\#}$  is the sum over all elements in  $\Omega$ . For  $n = 5$  this formula yields  $\int id d\mu_{\#} = 15$ .

▷ **Solution 3-7** In Example 3.17 we considered the measure space  $(\mathbb{R}, \mathcal{B}, \lambda)$ , where  $\lambda$  is the Lebesgue (or length) measure on  $\mathcal{B}$ . Remember, the Lebesgue measure satisfies

$$\lambda(]a, b]) = \lambda([a, b]) = b - a,$$

for  $a < b$  [see Eq. (1.54)]. We also considered  $f_1 = \alpha_1 1_{A_1}$  with  $A_1 = [0, (1 - .50)^{1/2}]$  and  $\alpha_1 = .50$ . Hence,  $f_1 = .50 \cdot 1_{A_1}$ . Therefore,

$$\int f_1 d\lambda = \alpha_1 \cdot \lambda(A_1) = .50 \cdot \lambda(A_1) = .50 \cdot (1 - .50)^{1/2} = .50 \cdot .50^{1/2} \approx .3536.$$

This is the area shaded in the left part of Figure 3.3.

Similarly, in Example 3.17 we also considered  $f_2 = \sum_{i=1}^3 \beta_i 1_{B_i}$  with the three intervals

$$B_1 = [0, (1 - .75)^{1/2}], \quad B_2 = ](1 - .75)^{1/2}, (1 - .50)^{1/2}], \quad B_3 = ](1 - .50)^{1/2}, (1 - .25)^{1/2}].$$

Again note that  $B_1, B_2, B_3$  is a sequence of elements of  $\mathcal{A}$ . Furthermore,  $f_2 = \sum_{i=1}^3 \beta_i 1_{B_i}$  with  $\beta_1 = .75$ ,  $\beta_2 = .50$ , and  $\beta_3 = .25$ . Hence, the integral of  $f_2 = \sum_{i=1}^3 \beta_i 1_{B_i}$  with respect to  $\lambda$  is

$$\begin{aligned} \int f_2 d\lambda &= \sum_{i=1}^3 \beta_i \cdot \lambda(B_i) = .75 \cdot \lambda(B_1) + .50 \cdot \lambda(B_2) + .25 \cdot \lambda(B_3) \\ &= .75 \cdot [(1 - .75)^{1/2}] + .50 \cdot [(1 - .50)^{1/2} - (1 - .75)^{1/2}] + .25 \cdot [(1 - .25)^{1/2} - (1 - .50)^{1/2}] \\ &\approx .75 \cdot .50 + .50 \cdot .2071 + .25 \cdot .1589 \approx 0.3750 + .1036 + .0397 = .5183. \end{aligned}$$

This is the area shaded in the middle part of Figure 3.3. The integral of  $f_3 = \sum_{i=1}^7 \gamma_i 1_{C_i}$  can be computed correspondingly. It is the area shaded in the right part of Figure 3.3.

▷ **Solution 3-8** If  $f: (\Omega, \mathcal{A}, \mu) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  is  $\mu$ -integrable and  $A, B \subset \mathcal{A}$ , then  $1_{A \cup B} f$  is  $\mu$ -integrable (see Lemma 3.33) and

$$\int_{A \cup B} f \, d\mu = \int 1_{A \cup B} f \, d\mu \quad [(3.30)]$$

$$= \int (1_A + 1_B - 1_{A \cap B}) f \, d\mu \quad [(1.34)]$$

$$= \int (1_A f + 1_B f - 1_{A \cap B} f) \, d\mu$$

$$= \int 1_A f \, d\mu + \int 1_B f \, d\mu - \int 1_{A \cap B} f \, d\mu \quad [(3.34)]$$

$$= \int_A f \, d\mu + \int_B f \, d\mu - \int_{A \cap B} f \, d\mu. \quad [(3.30)]$$

If  $A \cap B = \emptyset$  and  $f: (\Omega, \mathcal{A}, \mu) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  is quasi- $\mu$ -integrable, then  $1_{A \cup B} \cdot f = 1_A \cdot f + 1_B \cdot f$ , and these three functions are quasi- $\mu$ -integrable (see Lem. 3.33). If  $\int f^+ \, d\mu$  is finite, then  $\int 1_{A \cup B} \cdot f^+ \, d\mu$ ,  $\int 1_A \cdot f^+ \, d\mu$ , and  $\int 1_B \cdot f^+ \, d\mu$  are finite as well (see Lem. 3.33). If  $\int f^+ \, d\mu$  is infinite, then quasi- $\mu$ -integrability of  $f$  implies that  $\int f^- \, d\mu$  is finite and, according to Lemma 3.33, also the integrals  $\int 1_{A \cup B} \cdot f^- \, d\mu$ ,  $\int 1_A \cdot f^- \, d\mu$ , and  $\int 1_B \cdot f^- \, d\mu$ . Hence, in both cases

$$\int 1_{A \cup B} \cdot f \, d\mu = \int 1_{A \cup B} \cdot f^+ \, d\mu - \int 1_{A \cup B} \cdot f^- \, d\mu \quad [(3.27)]$$

$$= \int 1_A \cdot f^+ \, d\mu + \int 1_B \cdot f^+ \, d\mu - \left( \int 1_A \cdot f^- \, d\mu + \int 1_B \cdot f^- \, d\mu \right) \quad [(3.34)]$$

$$= \left( \int 1_A \cdot f^+ \, d\mu - \int 1_A \cdot f^- \, d\mu \right) + \left( \int 1_B \cdot f^+ \, d\mu - \int 1_B \cdot f^- \, d\mu \right)$$

$$= \int 1_A \cdot f \, d\mu + \int 1_B \cdot f \, d\mu. \quad [(3.27)]$$

▷ **Solution 3-9**

$$1_A(\omega) \cdot f(\omega) = \begin{cases} 0, & \text{if } \omega \notin A \\ f(\omega), & \text{if } \omega \in A. \end{cases}$$

Therefore,  $\{\omega \in \Omega: 1_A(\omega) \cdot f(\omega) \neq 0\} \subset A$ . Hence,  $\mu(\{\omega \in \Omega: 1_A(\omega) \cdot f(\omega) \neq 0\}) \leq \mu(A) = 0$ .

▷ **Solution 3-10** Let  $\alpha \geq 0$  and remember that the measure  $\alpha\mu$  on  $(\Omega, \mathcal{A})$  is defined by  $(\alpha\mu)(A) = \alpha\mu(A)$  for all  $A \in \mathcal{A}$ . The proof is conducted in three steps: (a) for a nonnegative step function, (b) for a nonnegative numerical measurable function, and (c) for an  $\mu$ -integrable numerical function (see Rem. 3.30).

(a) If  $f$  is a nonnegative step function and  $f = \sum_{i=1}^n \alpha_i 1_{A_i}$  a normal representation, then

$$\int f \, d(\alpha\mu) = \sum_{i=1}^n \alpha_i (\alpha\mu)(A_i) = \alpha \sum_{i=1}^n \alpha_i \mu(A_i) = \alpha \int f \, d\mu.$$

(b) If  $f$  is a nonnegative numerical measurable function and  $f_i, i \in \mathbb{N}$ , is an increasing sequence of nonnegative step functions with  $\lim_{i \rightarrow \infty} f_i = f$ , then

$$\begin{aligned} \int f \, d(\alpha\mu) &= \lim_{i \rightarrow \infty} \int f_i \, d(\alpha\mu) \\ &= \lim_{i \rightarrow \infty} \alpha \int f_i \, d\mu && [(a)] \\ &= \alpha \lim_{i \rightarrow \infty} \int f_i \, d\mu = \alpha \int f \, d\mu. \end{aligned}$$

(c) If  $f$  an  $\mu$ -integrable numerical function, then

$$\begin{aligned}
\int f d(\alpha\mu) &= \int f^+ d(\alpha\mu) - \int f^- d(\alpha\mu) \\
&= \alpha \int f^+ d\mu - \alpha \int f^- d\mu \quad [(b)] \\
&= \alpha \int f d\mu.
\end{aligned}$$

▷ **Solution 3-11** Because the derivative  $g'(x)$  of a function  $g(x) = \alpha + \beta x + \gamma x^n$ ,  $\alpha, \beta, \gamma \in \mathbb{R}$ ,  $n \in \mathbb{N}$ , is  $g'(x) = \beta + \gamma n x^{n-1}$ , the indefinite integral of  $f(x)$  is

$$\int f(x) dx = F(x) = 10x - \frac{x^3}{3} + c, \quad c \in \mathbb{R},$$

and therefore,

$$\int_a^b f(x) dx = F(x) \Big|_a^b := F(b) - F(a) = \left(10x - \frac{x^3}{3}\right) \Big|_a^b.$$

For  $a = -5$  and  $b = 5$ , this equation yields

$$\int_{-5}^5 f(x) dx = \left(10x - \frac{x^3}{3}\right) \Big|_{-5}^5 = \left(50 - \frac{125}{3}\right) - \left(-50 + \frac{125}{3}\right) = 100 - \frac{250}{3} = 16.\bar{6}.$$

Hence,

$$\int_{[-5,5]} f d\lambda = \int_{-5}^5 f(x) dx = 16.\bar{6}.$$

▷ **Solution 3-12** Let  $A \in \mathcal{A}$  with  $\mu(A) = 0$  and let  $f$  be a density  $f$  of  $\nu$  with respect to  $\mu$ , i. e.,  $\nu = f \circ \mu$ . Then

$$\begin{aligned}
\nu(A) &= \int 1_A d\nu = \int 1_A d f \circ \mu && [(3.8), \nu = f \circ \mu] \\
&= \int 1_A \cdot f d\mu = 0. && [(3.72), (3.42)]
\end{aligned}$$

Hence,  $\nu \ll_{\mathcal{A}} \mu$ .

**Part II**  
**Probability, Random Variable and its**  
**Distribution**



## Chapter 4

# Probability Measure

In chapter 1 we introduced the concept of a *measure*, treated various examples of measures, and some of their properties. In this chapter, we turn to a special class of examples, called *probability measures*. We start with the definition of a probability measure, then turn to conditional probabilities and the most important theorems related to conditional probability: the *multiplication rule*, the *theorem of total probability*, and *Bayes' Theorem*. Furthermore, we introduce the concept of a *conditional-probability measure*. Next, we define *independence of events* and of *independence of sets of events* with respect to a probability measure. A section on *conditional independence given an event* concludes this chapter.

### 4.1 Probability Measure and Probability Space

Now we introduce the concept of a *probability measure* as defined by Kolmogorov (1933/1977) (for the English version of this book see Kolmogorov, 1956). As we shall see, a probability measure is a special finite measure that is standardized.

#### 4.1.1 Definition

**Definition 4.1 (Probability Measure)**

Let  $(\Omega, \mathcal{A})$  be a measurable space. Then the function  $P: \mathcal{A} \rightarrow [0, 1]$  is called a *probability measure* on  $(\Omega, \mathcal{A})$ , if the following conditions hold:

- (a)  $P(\Omega) = 1$  (*standardization*).
- (b)  $P(A) \geq 0, \forall A \in \mathcal{A}$  (*nonnegativity*).
- (c)  $A_1, A_2, \dots \in \mathcal{A}$  are pairwise disjoint  $\Rightarrow P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$  ( *$\sigma$ -additivity*).

**Remark 4.2 (Probability and Probability Space)** Let  $P$  be a probability measure on  $(\Omega, \mathcal{A})$ . Then the triple  $(\Omega, \mathcal{A}, P)$  is called a *probability space* and a value  $P(A)$  of  $P$  is called the *probability* of  $A$ .  $\triangleleft$

**Remark 4.3 (Elementary Event and Event)** Let  $(\Omega, \mathcal{A}, P)$  be a probability space. Then  $A \in \mathcal{A}$  is called an *event*, and a singleton  $\{\omega\}, \omega \in \Omega$ , is called an *elemen-*

*tary event*, if  $\{\omega\} \in \mathcal{A}$ . Note the distinction between an *outcome*  $\omega \in \Omega$  and an *elementary event*  $\{\omega\} \in \mathcal{A}$  (see Exercise 4-1). Also note that the term *event* is only used in the context of a probability space  $(\Omega, \mathcal{A}, P)$ . Otherwise,  $A \in \mathcal{A}$  is called a *measurable set*.  $\triangleleft$

#### 4.1.2 Formal and Substantive Meaning of Probabilistic Terms

We distinguish between the mathematical or formal meaning of probabilistic terms and the meaning of these terms if used in an application of probability theory to a concrete real-world phenomenon. Often such a real-world phenomenon is a random experiment such as flipping a coin. In such a case, the set  $\Omega$  represents the set of possible outcomes of the random experiment considered and, in this sense,  $\Omega$  has a real-world interpretation. In other words, in this case  $\Omega$  is not only an abstract set anymore. This *real-world meaning* or *substantive meaning* is additional to their mathematical or *formal* meaning, namely being elements of the (abstract) set  $\Omega$ .

The terms ‘probability measure’, ‘probability of an event’ etc. hint at an important area of application of probability theory: real-world phenomena called random experiments. However, formally speaking, a probability measure is simply a label for a measure on a measurable space  $(\Omega, \mathcal{A})$  satisfying  $P(\Omega) = 1$ . If  $\Omega$  is not just an abstract set but represents a concrete random experiment, then the probability of an event  $A \in \mathcal{A}$  corresponds to the common language meaning of the term probability.

**Remark 4.4 (No Time Order Between Events)** The intuitive concept of an event often implies that events are ordered with respect to time. That is, in common language an ‘event’ is prior, simultaneous, or posterior to another ‘event’. In contrast, events as defined in probability theory are not necessarily ordered with respect to time. However, time order between events and sets of events can be introduced with respect to a filtration (see Def. 4.17, Fig. 4.1 and Example 4.19).  $\triangleleft$

**Remark 4.5 (A Priori Perspective)** If we apply probability theory to real-world phenomena, then we consider random experiments from the *a priori perspective*. Hence, the possible outcomes of a random experiment and events are considered *before* they happen. Even if an event already happened, we do *as if* it did not happen when we talk about its probability (see also Rem. 4.13).  $\triangleleft$

#### 4.1.3 Properties of a Probability Measure

Comparing conditions (a) to (c) of the definition of a probability measure to the conditions defining a measure (see Def. 1.43) shows that (b) and (c) are identical; only condition (a) differs. However,  $P(\Omega) = 1$  implies  $P(\emptyset) = 0$ , because  $\sigma$ -additivity of  $P$  yields

$$P(\Omega) = P(\Omega \cup \emptyset \cup \emptyset \cup \dots) = P(\Omega) + \sum_{i=1}^{\infty} P(\emptyset).$$

**Box 4.1 Rules of Computation for Probabilities**

Let  $(\Omega, \mathcal{A}, P)$  be a probability space.

If  $A_1, A_2, \dots \in \mathcal{A}$  are pairwise disjoint, then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i) \quad (\sigma\text{-additivity}) \quad (\text{i})$$

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i), \quad \forall n \in \mathbf{N}. \quad (\text{finite additivity}) \quad (\text{ii})$$

If  $A, B \in \mathcal{A}$ , then:

$$P(A) = P(A \cap B) + P(A \setminus B) \quad (\text{iii})$$

$$P(A^c) = 1 - P(A) \quad (\text{iv})$$

$$P(A) \leq P(B), \quad \text{if } A \subset B \quad (\text{monotonicity}) \quad (\text{v})$$

$$P(A \setminus B) = P(A) - P(A \cap B) \quad (\text{vi})$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (\text{vii})$$

$$P(A) = 1 \Rightarrow P(A \cap B) = P(B) \quad (\text{viii})$$

$$P(A) = 0 \Rightarrow P(A \cup B) = P(B). \quad (\text{ix})$$

Let  $A \in \mathcal{A}$  and let  $\Omega_0 \subset \Omega$  be finite or countable with  $P(\Omega_0) = 1$ .

If, for all  $\omega \in \Omega_0$ ,  $\{\omega\} \in \mathcal{A}$ , then

$$P(A) = \sum_{\omega \in A \cap \Omega_0} P(\{\omega\}). \quad (\text{x})$$

If  $A_1, A_2, \dots \in \mathcal{A}$ , then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} P(A_i). \quad (\sigma\text{-subadditivity}) \quad (\text{xi})$$

Therefore,  $\sum_{i=1}^{\infty} P(\emptyset) = P(\Omega) - P(\Omega) = 0$  and this yields  $P(\emptyset) = 0$ . This proves the following corollary:

**Corollary 4.6 (A Probability Measure is a Measure)**

A measure  $P$  on  $(\Omega, \mathcal{A})$  is a probability measure on  $\mathcal{A}$  if and only if  $P(\Omega) = 1$ .

A direct implication of this corollary is that all rules of computation for a measure (see Box 1.1) also hold for a probability measure. For convenience, these

rules are explicitly formulated for probability measures in Box 4.1 using the additional property  $P(\Omega) = 1$ .

**Remark 4.7 (Distribution vs. Probability Measure)** A probability measure on  $(\Omega, \mathcal{A})$  is also called a *distribution on  $(\Omega, \mathcal{A})$* . Although this term is preferably used in the context of a random variable (see Def. 5.3), the term ‘distribution’ is well-defined without referring to a random variable.  $\triangleleft$

#### 4.1.4 Examples

**Example 4.8 (Continuous Uniform Distribution)** Let  $\mathcal{B}_2$  denote the Borel  $\sigma$ -algebra on  $\mathbb{R}^2$  and consider a probability space  $(\Omega, \mathcal{A}, P)$ , where  $\Omega \in \mathcal{B}_2$ ,  $\mathcal{A} = \mathcal{B}_2|_{\Omega} := \{\Omega \cap A : A \in \mathcal{B}_2\}$  is the trace of  $\mathcal{B}_2$  in  $\Omega$  (see Example 1.10). Furthermore, let  $\lambda_2$  denote the Lebesgue measure on  $(\mathbb{R}^2, \mathcal{B}_2)$ , assume  $0 < \lambda_2(\Omega) < \infty$ , and define

$$P(A) = \frac{\lambda_2(A)}{\lambda_2(\Omega)}, \quad \forall A \in \mathcal{A}. \quad (4.1)$$

Then  $P$  is the *continuous uniform distribution* over  $\Omega$ . The relative size of the set  $A \in \mathcal{A}$  represents the probability  $P(A)$ , and Figure 1.1 (p. 4) can be used to illustrate some of its properties, e. g., Rules (iii) to (ix) of Box 4.1. This example will be generalized in section 8.2.1.  $\triangleleft$

**Example 4.9 (Joe and Ann With Random Assignment – continued)** In Example 1.9, we specified the set

$$\Omega = \{(Joe, no, -), (Joe, no, +), \dots, (Ann, yes, +)\},$$

which is also presented in the first column of Table 4.1. In this table we also specify the probability measure  $P$  on  $\mathcal{A} = \mathcal{P}(\Omega)$  by the probabilities of the eight elementary events  $\{\omega\} \in \Omega$ . Except for the empty set, which has probability  $P(\emptyset) = 0$ , all  $2^8 = 256$  elements of  $\mathcal{A}$  are either one of the eight elementary events  $\{(Joe, no, -)\}, \{(Joe, no, +)\}, \dots, \{(Ann, yes, +)\}$  or a union of some of these elementary events. Note that elementary events are always pairwise *disjoint*, i. e.,  $\{\omega_i\} \cap \{\omega_j\} = \emptyset$ , if  $\omega_i \neq \omega_j$ . Therefore, the probabilities of their unions can easily be computed using finite additivity of the probability measure [see Rule (ii) of Box 4.1]. In order to illustrate this point, consider the event that *Joe is drawn*,

$$A = \{(Joe, no, -), (Joe, no, +), (Joe, yes, -), (Joe, yes, +)\},$$

and the event that *the drawn person is successful*,

$$C = \{(Joe, no, +), (Joe, yes, +)\}, \{(Ann, no, +), (Ann, yes, +)\}.$$

The event  $A$  has the probability

$$\begin{aligned} P(A) &= P[\{(Joe, no, -)\}] + P[\{(Joe, no, +)\}] + P[\{(Joe, yes, -)\}] + P[\{(Joe, yes, +)\}] \\ &= .09 + .21 + .04 + .16 = .50. \end{aligned}$$

**Table 4.1.** Joe and Ann With Random Assignment: Probability Measures

Elements of $\Omega$			Probability measures			
Unit	Treatment	Success	$P_{\{\omega\}}$	$P^B_{\{\omega\}}$	$P^A_{\{\omega\}}$	$P^{A^c}_{\{\omega\}}$
(Joe, no, -)			.09	0	.18	0
(Joe, no, +)			.21	0	.42	0
(Joe, yes, -)			.04	.10	.08	0
(Joe, yes, +)			.16	.40	.32	0
(Ann, no, -)			.24	0	0	.48
(Ann, no, +)			.06	0	0	.12
(Ann, yes, -)			.12	.30	0	.24
(Ann, yes, +)			.08	.20	0	.16

Note.  $P, P^B, P^A,$  and  $P^{A^c}$  are probability measures on  $(\Omega, \mathcal{A})$ .

Similarly, the event  $C$  has the probability

$$P(C) = P\{(Joe, no, +)\} + P\{(Joe, yes, +)\} + P\{(Ann, no, +)\} + P\{(Ann, yes, +)\} \\ = .21 + .16 + .06 + .08 = .51,$$

and the event *Joe is drawn and is successful*,  $A \cap C = \{(Joe, no, +), (Joe, yes, +)\}$ , has the probability

$$P(A \cap C) = P\{(Joe, no, +)\} + P\{(Joe, yes, +)\} = .21 + .16 = .37.$$

The probability measures specified in the last three columns of Table 4.1 will be treated in Examples 4.34 and 4.35. ◁

**Remark 4.10 (Other Examples)** In section 8.1 probability measures on the measurable space  $(\mathbb{N}_0, \mathcal{P}(\mathbb{N}_0))$  are considered, e. g., the *binomial distribution* (see Def. 8.7), the *Poisson distribution* (see Def. 8.14), and the *geometric distribution* (see Def. 8.20). In all these examples, a probability measure on  $(\mathbb{N}_0, \mathcal{P}(\mathbb{N}_0))$  is uniquely defined, if the probabilities of the elementary events  $\{x\}$  are determined for all  $x \in \mathbb{N}_0$  [see Box 4.1 (x) for  $\Omega_0 = \mathbb{N}_0$ ].

The example of the Poisson distribution shows that, even for the countably infinite set  $\mathbb{N}_0$ , there are probability measures  $\mathcal{P}_\lambda$  on  $(\mathbb{N}_0, \mathcal{P}(\mathbb{N}_0))$  with  $\mathcal{P}_\lambda(\{x\}) > 0$ , for all  $x \in \mathbb{N}_0$ , and

$$\mathcal{P}_\lambda(\mathbb{N}_0) = \sum_{x \in \mathbb{N}_0} \mathcal{P}_\lambda(\{x\}) = \sum_{x=0}^{\infty} \mathcal{P}_\lambda(\{x\}) = 1.$$

For another example see Exercise 4-2. ◁

**Example 4.11 (Finite Mixture of Probability Measures)** In Example 1.62 we already noted that the weighted sum of measures on a measurable space  $(\Omega, \mathcal{A})$  is again a measure on  $(\Omega, \mathcal{A})$ . With an additional assumption this also applies to probability measures. More precisely, if  $P_1, \dots, P_n$  are probability measures on  $(\Omega, \mathcal{A})$ ,  $\alpha_i \geq 0$ ,  $i = 1, \dots, n$ , and we additionally assume  $\sum_{i=1}^n \alpha_i = 1$ , then  $\sum_{i=1}^n \alpha_i P_i$  is again a probability measure on  $(\Omega, \mathcal{A})$ . It is called a *finite mixture* of  $P_1, \dots, P_n$ . Such a finite mixture of probability measures will be illustrated by Example 4.35 using conditional-probability measures.  $\triangleleft$

## 4.2 Conditional Probability

Conditional probabilities can be used to describe *dependencies* between two events  $A, B \in \mathcal{A}$  with respect to a probability measure  $P$  on  $\mathcal{A}$ . In section 4.2.7 we will also use this concept in order to introduce the concept of a *conditional-probability measure*.

### 4.2.1 Definition

**Definition 4.12 (Conditional Probability)**

Let  $(\Omega, \mathcal{A}, P)$  be a probability space, let  $A, B \in \mathcal{A}$ , and let  $P(B) > 0$ . Then

$$P(A|B) := \frac{P(A \cap B)}{P(B)} \quad (4.2)$$

is called the *conditional probability of A given B with respect to P*.

**Remark 4.13 (A Priori Perspective)** The conditional probability  $P(A|B)$  is the probability of the event  $A$  if it is known that the event  $B$  occurred. In order to compute  $P(A|B)$  according to Equation (4.2), we need the (unconditional) probability  $P(B)$ . Using  $P(B)$  we do as if  $B$  did not yet happen (see Rem. 4.5). The fact that  $B$  occurred is reflected by  $P(B|B) = 1$ .  $\triangleleft$

**Remark 4.14 (Continuous Uniform Distribution – continued)** In Example 4.8 we defined the continuous uniform distribution on  $(\Omega, \mathcal{A})$  by Equation (4.1). Using the area of the ellipses presented in Figure 1.1 (p. 4), the conditional probability  $P(A|B)$  corresponds to the area of the intersection  $A \cap B$  divided by the area of  $B$ .  $\triangleleft$

**Example 4.15 (Flipping a Coin two Times)** Consider the random experiment of flipping a coin two times, the measurable space  $(\Omega, \mathcal{A})$  of which is the same as in subsection 2.2.2 and the probability measure has been specified by Equation 2.43. The conditional probability  $P(B|A)$  that we flip *heads* in the second flip ( $B$ ) given that we flip *heads* in the first flip ( $A$ ) is  $1/2$ , which is equal to the *unconditional* probability  $P(B)$  of flipping *heads* in the second flip. In such a case the

two events  $A$  and  $B$  are *independent* (see section 4.3). Note that the conditional probability  $P(A|B)$  that we flip *heads* in the first flip ( $A$ ) given that we flip *heads* in the second flip ( $B$ ) is also equal to the unconditional probability  $P(A)$  of flipping *heads* in the first flip. This example shows that we may condition on events that occur later in time and that a conditional probability does not necessarily describe a causal dependence. Note, however, that conditional probabilities *can* be used to describe causal dependencies, provided that additional assumptions hold (see Examples 4.16 and 4.36).

As another example consider the event *flipping at least one heads* ( $A$ ) and the event *no heads are flipped in the first flip* ( $B$ ). In this case

$$P(A|B) = \frac{1}{2} \neq P(A) = \frac{3}{4},$$

and the two events are not independent (see section 4.3). ◁

**Example 4.16 (Joe and Ann With Random Assignment – continued)** Consider again Table 4.1 (p. 131), define  $\Omega_U = \{Joe, Ann\}$  and  $\Omega_X = \{yes, no\}$ , and let

$$C = \Omega_U \times \Omega_X \times \{+\} = \{(Joe, no, +), (Joe, yes, +), (Ann, no, +), (Ann, yes, +)\}$$

be the event that the *drawn person is successful*. Furthermore, let

$$B = \Omega_U \times \{yes\} \times \Omega_Y = \{(Joe, yes, -), (Joe, yes, +), (Ann, yes, -), (Ann, yes, +)\}$$

denote the event that the *drawn person is treated*. Then Equation (4.2) yields:

$$P(C|B) = \frac{P(C \cap B)}{P(B)} = \frac{P(\Omega_U \times \{yes\} \times \{+\})}{P(\Omega_U \times \{yes\} \times \Omega_Y)} = \frac{.16 + .08}{.04 + .16 + .12 + .08} = .60.$$

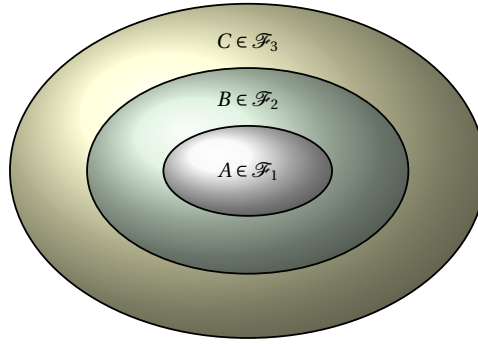
Conditioning on the event  $B^c$  that the *drawn person is not treated* yields

$$P(C|B^c) = \frac{P(C \cap B^c)}{P(B^c)} = \frac{P(\Omega_U \times \{no\} \times \{+\})}{P(\Omega_U \times \{no\} \times \Omega_Y)} = \frac{.21 + .06}{.09 + .21 + .24 + .06} = .45.$$

In this example, the difference  $P(C|B) - P(C|B^c) = .60 - .45$  *can* be used to evaluate the average effect of the treatment. This will be substantiated in more detail in Example 4.36. ◁

### 4.2.2 Filtration and Time Order Between Events and Sets of Events

As mentioned in Remark 4.4, the definition of an event in probability theory does not presume that there is a time order between events and set of events. However, in many applications of probability theory such a time order is important. In Example 4.16, for instance, it is crucial that the event  $C$  is *posterior* to the event  $B$ . Such a time order is formalized in the theory of stochastic processes (see, e. g., Bauer, 1996; Klenke, 2013) and in the theory of causal effects (see, e. g., Steyer, Mayer, & Fiege, 2014; Mayer, Thoemmes, Rose, Steyer, & West, 2014).



**Figure 4.1.** A filtration with  $T = \{1, 2, 3\}$

**Definition 4.17 (Filtration)**

Let  $(\Omega, \mathcal{A})$  be a measurable space and  $T \subset \mathbb{R}$ . A family  $(\mathcal{F}_t, t \in T)$  of sub- $\sigma$ -algebras  $\mathcal{F}_t$  of  $\mathcal{A}$  is called a filtration in  $\mathcal{A}$ , if  $\mathcal{F}_s \subset \mathcal{F}_t$  for all  $s, t \in T$  with  $s \leq t$ .

Referring to such a filtration, time order between events can be introduced as follows.

**Remark 4.18 (Event  $A$  is Prior, Simultaneous, Posterior to Event  $B$ )**

Let  $(\Omega, \mathcal{A}, P)$  be a probability space,  $(\mathcal{F}_t, t \in T)$  a filtration in  $\mathcal{A}$ , and  $A, B \in \mathcal{A}$ .

- (i) The event  $A$  is called *prior to  $B$*  (and  $B$  posterior to  $A$ ) in  $(\mathcal{F}_t, t \in T)$ , if there is an  $s \in T$  such that  $A \in \mathcal{F}_s, B \notin \mathcal{F}_s$ , and a  $t \in T, t > s$ , such that  $B \in \mathcal{F}_t$ .
- (ii) The event  $A$  is called *simultaneous to  $B$*  in  $(\mathcal{F}_t, t \in T)$ , if there is a  $t \in T$  such that  $A, B \in \mathcal{F}_t$  and no  $s \in T, s < t$ , such that  $A \in \mathcal{F}_s$  or  $B \in \mathcal{F}_s$ .

◁

**Example 4.19 (Joe and Ann With Random Assignment – continued)** In the random experiment described by Table 4.1, the event  $A$  that Joe is drawn (see Example 4.9) is prior to the event

$$B = \{(Joe, yes, +), (Joe, yes, -), (Ann, yes, +), (Ann, yes, -)\}$$

that the drawn person is treated, which itself is prior to the event  $C$  that the drawn person is successful. This time order in the real-world can formally be represented by the following filtration:

$$\mathcal{F}_1 := \sigma(\{A\}), \quad \mathcal{F}_2 := \sigma(\{A, B\}), \quad \mathcal{F}_3 := \sigma(\{A, B, C\}) = \mathcal{P}(\Omega),$$

using the concept of a  $\sigma$ -algebra generated by a set system (see Def. 1.13). With respect to the filtration  $(\mathcal{F}_t, t \in T)$ ,  $T = \{1, 2, 3\}$ , the event  $A$  is prior to  $B$ , because  $A \in \mathcal{F}_1, B \notin \mathcal{F}_1$ , but  $B \in \mathcal{F}_2$  (see Fig. 4.1, Rem. 4.18, and Exercise 4-3). ◁

**Remark 4.20 (Formal and Substantive Meaning of Time Order)** As noted in section 4.1.2, we distinguish between the mathematical or formal meaning of a probabilistic term on one side and the meaning of these terms if used in an application of probability theory to a concrete real-world phenomenon on the other side. This also applies to the terms *prior*, *simultaneous*, and *posterior* with respect to a filtration. In applications in which the elements of the set  $T$  represents time points, these terms do not only have a formal meaning that is specified by their mathematical definition, but also a substantive meaning: In the *real world*, an event  $A$  that is prior to  $B$  with respect to  $(\mathcal{F}_t, t \in T)$ , occurs before the event  $B$ . In other applications, the term *prior* may not express the actual substantive meaning. Of course, this applies to the terms *simultaneous* and *posterior* as well.  $\triangleleft$

### 4.2.3 Multiplication Rule

Now we treat some theorems involving conditional probabilities. The first one shows how the probability  $P(A_1 \cap \dots \cap A_n)$  can be factorized into a product of an unconditional probability and conditional probabilities.

**Remark 4.21 (Multiplication Rule for Two and for Three Events)** For two events  $A_1$  and  $A_2$ , the multiplication rule is

$$P(A_1 \cap A_2) = P(A_1) \cdot P(A_2 | A_1), \quad (4.3)$$

provided that  $P(A_1) > 0$ . This equation directly follows from the definition of the conditional probability  $P(A_2 | A_1)$  [see Eq. (4.2)]. For three events  $A_1$ ,  $A_2$ , and  $A_3$ , the multiplication rule is

$$P(A_1 \cap A_2 \cap A_3) = P(A_1) \cdot P(A_2 | A_1) \cdot P(A_3 | A_1 \cap A_2), \quad (4.4)$$

provided that  $P(A_1 \cap A_2) > 0$ . This equation follows from the definition of the conditional probability

$$P(A_3 | A_1 \cap A_2) = \frac{P(A_1 \cap A_2 \cap A_3)}{P(A_1 \cap A_2)}, \quad (4.5)$$

inserting Equation (4.3) for  $P(A_1 \cap A_2)$ , and solving the resulting equation for  $P(A_1 \cap A_2 \cap A_3)$ .  $\triangleleft$

For  $n$  events  $A_1, \dots, A_n$ , the multiplication rule is formulated in the following theorem.

#### Theorem 4.22 (Multiplication Rule)

Let  $(\Omega, \mathcal{A}, P)$  be a probability space and  $A_1, \dots, A_n \in \mathcal{A}$ , where  $2 \leq n \in \mathbb{N}$ . If  $P(\bigcap_{i=1}^{n-1} A_i) > 0$ , then

$$P\left(\bigcap_{i=1}^n A_i\right) = P(A_1) \cdot \prod_{j=2}^n P\left(A_j \mid \bigcap_{i=1}^{j-1} A_i\right). \quad (4.6)$$

(Proof p. 148)

#### 4.2.4 Examples

**Example 4.23 (Joe and Ann With Random Assignment – continued)** Consider again the Example presented in Table 4.1 (p. 131) and let

$$A = \{(Joe, no, -), (Joe, no, +), (Joe, yes, -), (Joe, yes, +)\}$$

denote the event that Joe is drawn,

$$B = \{(Joe, yes, -), (Joe, yes, +), (Ann, yes, -), (Ann, yes, +)\}$$

the event that the drawn person is treated, and

$$C = \{(Joe, no, +), (Joe, yes, +), (Ann, no, +), (Ann, yes, +)\}$$

the event that there is success, irrespective of the drawn person and treatment received. Then

$$A \cap B \cap C = \{(Joe, yes, +)\}$$

is the event that Joe is drawn, receives the treatment, and is successful. According to Equation (4.4), the probability of this event can be computed by

$$\begin{aligned} P(A \cap B \cap C) &= P(A) \cdot P(B|A) \cdot P(C|A \cap B) \\ &= (.09 + .21 + .04 + .16) \cdot \frac{.04 + .16}{.09 + .21 + .04 + .16} \cdot \frac{.16}{.04 + .16} \\ &= .50 \cdot .40 \cdot .80 = .16 \end{aligned}$$

(see Exercise 4-4). Of course, Equation (4.4) can also be applied to the other seven sets  $A \cap B \cap C^c$  to  $A^c \cap B^c \cap C^c$  in Figure 4.2. In this example,  $P(A \cap B \cap C) = P(\{(Joe, yes, +)\}) = .16$  is the probability of an elementary event (see Table 4.1).  $\triangleleft$

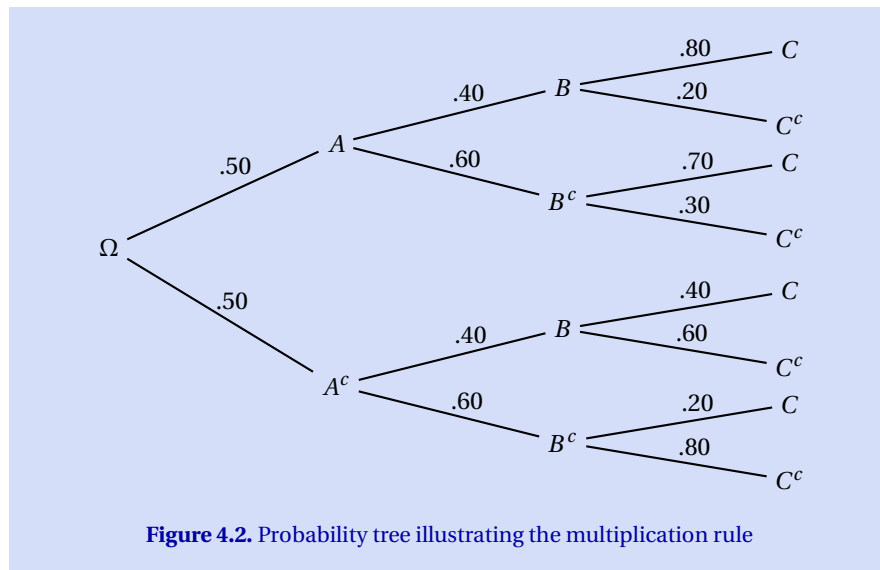
**Example 4.24 (Drawing Three Balls)** Consider drawing three balls without replacement from an urn containing two white balls and four black balls. Furthermore, let us consider the three events  $A_i$  to *draw a black ball at time  $i$* , where  $i = 1, 2, 3$ . According to Theorem 4.22, the probability of drawing three black balls is

$$P(A_1 \cap A_2 \cap A_3) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1 \cap A_2),$$

where  $P(A_1) = 4/6$ ,  $P(A_2|A_1) = 3/5$ , and  $P(A_3|A_1 \cap A_2) = 2/4$ . Hence,

$$P(A_1 \cap A_2 \cap A_3) = \frac{4}{6} \cdot \frac{3}{5} \cdot \frac{2}{4} = \frac{24}{120} = \frac{1}{5}.$$

$\triangleleft$



#### 4.2.5 Theorem of Total Probability

In our next theorem, called the *theorem of total probability*, we show how the probability of an event  $B \subset A_1 \cup \dots \cup A_n$  can additively be decomposed into the products  $P(B|A_i) \cdot P(A_i)$  of conditional and unconditional probabilities. In this theorem we assume that the events  $A_1, \dots, A_n$  are *pairwise disjoint*, i. e., we assume  $A_i \cap A_j = \emptyset$ , for all  $i, j = 1, \dots, n$ , with  $i \neq j$ .

##### Theorem 4.25 (Theorem of Total Probability)

Let  $(\Omega, \mathcal{A}, P)$  be a probability space and  $B \in \mathcal{A}$ .

(i) If

(a)  $A_1, \dots, A_n \in \mathcal{A}$  are pairwise disjoint,

(b)  $B \subset \bigcup_{i=1}^n A_i$ ,

then

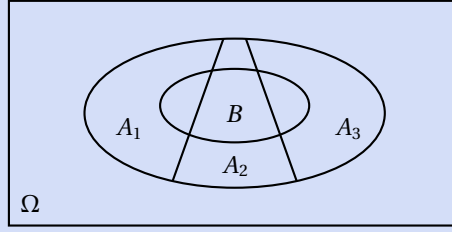
$$P(B) = \sum_{i=1}^n P(B \cap A_i). \quad (4.7)$$

(ii) If (a) and (b) of (i) hold as well as

(c)  $P(A_i) > 0, \quad \forall i = 1, \dots, n$ ,

then

$$P(B) = \sum_{i=1}^n P(B|A_i) \cdot P(A_i). \quad (4.8)$$



**Figure 4.3.** Venn diagram illustrating a partition of a set

(iii) If

- (a)  $A_1, A_2, \dots \in \mathcal{A}$  are pairwise disjoint,  
 (b)  $B \subset \bigcup_{i=1}^{\infty} A_i$ ,

then

$$P(B) = \sum_{i=1}^{\infty} P(B \cap A_i). \quad (4.9)$$

(iv) If (a) and (b) of (iii) hold as well as

- (c)  $P(A_i) > 0, \quad \forall i = 1, 2, \dots$ ,

then

$$P(B) = \sum_{i=1}^{\infty} P(B | A_i) \cdot P(A_i). \quad (4.10)$$

(Proof p. 148)

Equation (4.7) can be illustrated by Figure 4.3. If we assume that  $(\Omega, \mathcal{A}, P)$  is the probability space specified in Example 4.8, then the figure visualizes that  $P(B) = P(B \cap A_1) + P(B \cap A_2) + P(B \cap A_3)$ . The crucial points are:

- (a) If the events  $A_1, \dots, A_n$  are pairwise disjoint, then  $B \cap A_1, \dots, B \cap A_n$  are pairwise disjoint as well.  
 (b) The probability measure  $P$  is additive.

#### 4.2.6 Bayes' Theorem

Our next theorem, called *Bayes' theorem*, reveals how the conditional probabilities  $P(B | A_i)$  are related to the conditional probabilities  $P(A_i | B)$ . Using the definitions of the conditional probabilities  $P(A_i | B)$  and  $P(B | A_i)$  yields

$$P(A_i | B) = \frac{P(B | A_i) \cdot P(A_i)}{P(B)}. \quad (4.11)$$

Inserting Equation (4.8) for  $P(B)$  then proves the following theorem.

**Theorem 4.26 (Bayes' Theorem)**

Let  $(\Omega, \mathcal{A}, P)$  be a probability space,  $B \in \mathcal{A}$ , and  $P(B) > 0$ . Under the assumptions (a) to (c) of Theorem 4.25 (i) and (ii),

$$P(A_i | B) = \frac{P(B | A_i) \cdot P(A_i)}{\sum_{j=1}^n P(B | A_j) \cdot P(A_j)}, \quad \forall i = 1, \dots, n. \quad (4.12)$$

Analogously, under the assumptions (a) to (c) of Theorem 4.25 (iii) and (iv),

$$P(A_i | B) = \frac{P(B | A_i) \cdot P(A_i)}{\sum_{j=1}^{\infty} P(B | A_j) \cdot P(A_j)}, \quad \forall i \in \mathbb{N}. \quad (4.13)$$

**Example 4.27 (Joe and Ann With Random Assignment – continued)** Let

$$A = \{(Joe, no, -), (Joe, no, +), (Joe, yes, -), (Joe, yes, +)\}$$

denote the event that Joe is drawn,

$$A^c = \{(Ann, no, -), (Ann, no, +), (Ann, yes, -), (Ann, yes, +)\}$$

the event that Ann is drawn, and

$$B = \{(Joe, yes, -), (Joe, yes, +), (Ann, yes, -), (Ann, yes, +)\}$$

the event that the *drawn person is treated*. Then

$$\begin{aligned} P(A|B) &= \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|A^c) \cdot P(A^c)} \\ &= \frac{.40 \cdot .50}{.40 \cdot .50 + .40 \cdot .50} = .50 \end{aligned}$$

is the conditional probability that *Joe is drawn* given that the *drawn person is treated* (see Table 4.1, p. 131). The corresponding probability that *Ann is drawn* given that the *drawn person is treated* is identical in this example, i. e.,  $P(A^c | B) = .50$ . Hence, given treatment, each person has the same probability to be drawn. This is the *sampling perspective* of a randomized experiment supplementing the *assignment perspective*, according to which the treatment probability is the same for each person, i. e.,  $P(B | A) = P(B | A^c) = .40$  (see again Table 4.1).  $\triangleleft$

### 4.2.7 Conditional-Probability Measure

Just like probabilities, conditional probabilities of events  $A \in \mathcal{A}$  given  $B$  are values of a probability measure.

**Theorem 4.28 (Conditional-Probability Measure)**

Let  $(\Omega, \mathcal{A}, P)$  be a probability space. If  $B \in \mathcal{A}$  and  $P(B) > 0$ , then the function  $P^B: \mathcal{A} \rightarrow [0, 1]$  defined by

$$P^B(A) = P(A|B), \quad \forall A \in \mathcal{A}, \quad (4.14)$$

is a probability measure on  $(\Omega, \mathcal{A})$ .

(Proof p. 148)

According to this theorem, for each  $B \in \mathcal{A}$  with  $P(B) > 0$ , the triple  $(\Omega, \mathcal{A}, P^B)$  is a probability space.

**Definition 4.29 (Conditional-Probability Measure)**

Let  $(\Omega, \mathcal{A}, P)$  be a probability space, let  $B \in \mathcal{A}$ , and let  $P(B) > 0$ . Then the function  $P^B$  defined by (4.14) is called the  $B$ -conditional-probability measure on  $(\Omega, \mathcal{A})$ .

In the following lemma we consider the relationship between conditional probabilities with respect to the measures  $P^B$  and  $P$ .

**Lemma 4.30 (Conditional Probabilities With Respect to  $P^B$ )**

Let  $(\Omega, \mathcal{A}, P)$  be a probability space. If  $A, B, C \in \mathcal{A}$  and  $P(B \cap C) > 0$ , then

$$P^B(A|C) = P(A|B \cap C). \quad (4.15)$$

(Proof p. 149)

**Remark 4.31 (Total Conditional Probability)** Suppose  $A, B, C \in \mathcal{A}$ ,  $P(B \cap C) > 0$ , and  $P(B \cap C^c) > 0$ . This implies  $P(B) > 0$  and  $P^B(C) = P(C|B) = P(C \cap B) / P(B) > 0$ . Applying Equation (4.8) to the measure  $P^B$  then yields

$$P^B(A) = P^B(A|C) \cdot P^B(C) + P^B(A|C^c) \cdot P^B(C^c), \quad (4.16)$$

and Equations (4.14) and (4.15) imply

$$P(A|B) = P(A|B \cap C) \cdot P(C|B) + P(A|B \cap C^c) \cdot P(C^c|B). \quad (4.17)$$

◁

According to the following lemma,  $P^B$  is *absolutely continuous with respect to*  $P$ , i. e.

$$\forall A \in \mathcal{A}: P(A) = 0 \Rightarrow P^B(A) = 0. \quad (4.18)$$

This is denoted by  $P^B \ll_{\mathcal{A}} P$  [see Def. 3.70 (i)]. In contrast,  $P \ll_{\mathcal{A}} P^B$  does not always hold.

**Lemma 4.32 (Absolute Continuity of the Conditional-Probability Measure)**

Let  $(\Omega, \mathcal{A}, P)$  be a probability space,  $B \in \mathcal{A}$ , and  $P(B) > 0$ . Then

- (i)  $P^B \ll_{\mathcal{A}} P$ , and  
(ii)  $P^B = \left( \frac{1}{P(B)} \cdot 1_B \right) \odot P$ .

(Proof p. 149)

**Remark 4.33 ( $P^B$  is a Measure With Density)** Proposition (ii) of Lemma 4.32 implies that  $P^B$  is a measure with density  $1_B/P(B)$  with respect to  $P$ . The following equations show how  $P^B(A)$  can be written as an integral in various ways:

$$\begin{aligned} \forall A \in \mathcal{A}: P^B(A) &= \int_A dP^B && [(3.8)] \\ &= \int 1_A d\left(\frac{1}{P(B)} \cdot 1_B \odot P\right) && [(3.30), \text{Lem. 4.32 (ii)}] \\ &= \int 1_A \cdot 1_B \frac{1}{P(B)} dP && [(3.72)] \\ &= \frac{1}{P(B)} \cdot \int 1_{A \cap B} dP. && [(1.33), (3.32)] \end{aligned} \quad (4.19)$$

Note that, according to Theorem 3.72 (i), the density  $1_B/P(B)$  can be written as a Radon-Nikodym derivate of  $P^B$  with respect to  $P$ , i. e.,

$$\frac{1}{P(B)} \cdot 1_B = \frac{dP^B}{dP}. \quad (4.20)$$

◁

**Example 4.34 (Joe and Ann With Random Assignment – continued)** Consider the example presented in Table 4.1 (p. 131). We specify the  $B$ -conditional-probability measure  $P^B: \mathcal{A} \rightarrow [0, 1]$  for the event that the *drawn person is treated*, i. e., for

$$B = \{(Joe, yes, -), (Joe, yes, +), (Ann, yes, -), (Ann, yes, +)\}.$$

For the first two elementary events,  $P^B(\{(Joe, no, +)\}) = P^B(\{(Joe, no, -)\}) = 0$ , because the intersections  $\{(Joe, no, -)\} \cap B$  and  $\{(Joe, no, +)\} \cap B$  are empty. For the next two elementary events, the  $B$ -conditional probabilities are

$$P^B(\{(Joe, yes, -)\}) = \frac{P(\{(Joe, yes, -)\} \cap B)}{P(B)} = \frac{.04}{.40} = .10$$

and

$$P^B(\{(Joe, yes, +)\}) = \frac{P(\{(Joe, yes, +)\} \cap B)}{P(B)} = \frac{.16}{.40} = .40.$$

For the next two elementary events,  $P^B(\{(Ann, no, -)\}) = P^B(\{(Ann, no, +)\}) = 0$ , because the intersections  $\{(Ann, no, -)\} \cap B$  and  $\{(Ann, no, +)\} \cap B$  are again empty. Finally, for the last two elementary events, the  $B$ -conditional probabilities are

$$P^B(\{(Ann, yes, -)\}) = \frac{P(\{(Ann, yes, -)\} \cap B)}{P(B)} = \frac{.12}{.40} = .30$$

and

$$P^B(\{(Ann, yes, +)\}) = \frac{P(\{(Ann, yes, +)\} \cap B)}{P(B)} = \frac{.08}{.40} = .20.$$

These probability are summarized in the third column of Table 4.1. Except for  $\emptyset$ , all other events are unions of these elementary events. Because the elementary events are *disjoint*, the probabilities of their unions can easily be computed using finite additivity of the probability measure [see Rule (ii) of Box 4.1 and Exercise 4-5].  $\triangleleft$

**Example 4.35 (Joe and Ann With Random Assignment – continued)** Two other conditional-probability measures on  $(\Omega, \mathcal{A})$  are  $P^A$  and  $P^{A^c}$ , where  $A$  is the event

$$A = \{(Joe, no, -), (Joe, no, +), (Joe, yes, -), (Joe, yes, +)\}$$

that Joe is sampled and  $A^c$  the event

$$A^c = \{(Ann, no, -), (Ann, no, +), (Ann, yes, -), (Ann, yes, +)\}$$

that Ann is sampled. The values of these conditional-probability measures are presented in the last two columns of Table 4.1. These measures can also be used to illustrate a *mixture of two probability measures*. As is easily seen

$$P = .50 \cdot P^A + .50 \cdot P^{A^c},$$

i. e., the measure  $P$  is a mixture of the two conditional-probability measures  $P^A$  and  $P^{A^c}$  (see Examples 4.11 and 1.62).  $\triangleleft$

**Example 4.36 (Joe and Ann With Random Assignment– continued)** In Example 4.16 we computed the two conditional probabilities  $P(C|B) = .60$  and  $P(C|B^c) = .45$  of success given treatment and no treatment, respectively. These are conditional probabilities with respect to the measure  $P$ . Let us now consider the *individual treatment effects* of Joe and of Ann. These individual effects can be computed using the  $P^A$ - and  $P^{A^c}$ -conditional probability measures, respectively. For Joe the individual treatment effect is

$$\begin{aligned} P^A(C|B) - P^A(C|B^c) &= \frac{P^A(\Omega_U \times \{yes\} \times \{+\})}{P^A(\Omega_U \times \{yes\} \times \Omega_Y)} - \frac{P^A(\Omega_U \times \{no\} \times \{+\})}{P^A(\Omega_U \times \{no\} \times \Omega_Y)} \\ &= \frac{.32 + 0}{.08 + .32 + 0 + 0} - \frac{.32 + 0}{.18 + .42 + 0 + 0} \\ &= .80 - .70 = .10, \end{aligned}$$

and for Ann it is

$$\begin{aligned}
P^{A^c}(C|B) - P^{A^c}(C|B^c) &= \frac{P^{A^c}(\Omega_U \times \{\text{yes}\} \times \{+\})}{P^{A^c}(\Omega_U \times \{\text{yes}\} \times \Omega_Y)} - \frac{P^{A^c}(\Omega_U \times \{\text{no}\} \times \{+\})}{P^{A^c}(\Omega_U \times \{\text{no}\} \times \Omega_Y)} \\
&= \frac{.16 + 0}{.24 + .16 + 0 + 0} - \frac{.12 + 0}{.48 + .12 + 0 + 0} \\
&= .40 - .20 = .20.
\end{aligned}$$

Hence, the treatment effect  $P(C|B) - P(C|B^c) = .15$  (see Example 4.16), is just the weighted average  $.50 \cdot .10 + .50 \cdot .20 = .15$  of the two individual treatment effects, where the weights are .50 for Joe and for Ann (see Example 4.35). Note that this property does not always hold [see Table 11.2 (p. 341) and Example 11.28].  $\triangleleft$

## 4.3 Independence

### 4.3.1 Independence of Events

Independence of two events  $A$  and  $B$  means that the conditional and unconditional probabilities are the same, i. e.,  $P(A|B) = P(A)$  and  $P(B|A) = P(B)$ . This definition presupposes that  $P(A), P(B) > 0$ , because otherwise the two conditional probabilities are not defined. The following definition does not rest on this requirement and extends the concept of independence to more than two events.

#### Definition 4.37 (Independence of Events)

Let  $(\Omega, \mathcal{A}, P)$  be a probability space.

(i) Two events  $A, B \in \mathcal{A}$  are called  $P$ -independent, denoted  $A \perp\!\!\!\perp_P B$ , if

$$P(A \cap B) = P(A) \cdot P(B). \quad (4.21)$$

(ii) Let  $I$  be a nonempty set and let  $A_i \in \mathcal{A}$ ,  $i \in I$ . Then  $(A_i, i \in I)$  is called a family of  $P$ -independent events, denoted  $\perp\!\!\!\perp_P (A_i, i \in I)$ , if

$$P\left(\bigcap_{i \in I_0} A_i\right) = \prod_{i \in I_0} P(A_i), \quad \forall \text{ finite } I_0 \subset I. \quad (4.22)$$

**Remark 4.38 (Pairwise and Triple-Wise Independence)** For  $n$  events  $A_1, \dots, A_n$ ,  $P$ -independence will also be denoted by

$$\perp\!\!\!\perp_P A_1, \dots, A_n.$$

For three events, for instance, it means that

$$P(A_i \cap A_j) = P(A_i) \cdot P(A_j), \quad i \neq j, \quad i, j = 1, 2, 3, \quad (4.23)$$

(pairwise  $P$ -independence) and

$$P(A_1 \cap A_2 \cap A_3) = P(A_1) \cdot P(A_2) \cdot P(A_3) \quad (4.24)$$

(triple-wise  $P$ -independence) hold.

Note that pairwise  $P$ -independence of more than two events does not imply  $P$ -independence of these events. Furthermore, triple-wise  $P$ -independence, for instance, does not imply pairwise  $P$ -independence. For more propositions on independence of events see Box 4.2 (p. 147).  $\triangleleft$

**Remark 4.39 (Independence of any event  $A$  with  $\Omega$  and  $\emptyset$ )** For any probability space  $(\Omega, \mathcal{A}, P)$ ,

$$\forall A \in \mathcal{A}: A \perp\!\!\!\perp_P \Omega \quad \text{and} \quad A \perp\!\!\!\perp_P \emptyset. \quad (4.25)$$

(see Exercise 4-7).  $\triangleleft$

### 4.3.2 Independence of Set Systems

Now we extend the concept of  $P$ -independence to *set systems*, i. e., to sets of events, and illustrate independence by an example.

#### Definition 4.40 (Family of Independent Set Systems)

Let  $(\Omega, \mathcal{A}, P)$  be a probability space and  $\mathcal{E}_i \subset \mathcal{A}$ ,  $i \in I \neq \emptyset$ . Then  $(\mathcal{E}_i, i \in I)$  is called a family of  $P$ -independent set systems, denoted  $\perp\!\!\!\perp_P (\mathcal{E}_i, i \in I)$ , if  $\perp\!\!\!\perp_P (A_i, i \in I)$  holds for all families  $(A_i, i \in I)$  with  $A_i \in \mathcal{E}_i$ ,  $i \in I$ . If  $I = \{1, 2\}$  we also use the notation  $\mathcal{E}_1 \perp\!\!\!\perp_P \mathcal{E}_2$  instead of  $\perp\!\!\!\perp_P (\mathcal{E}_i, i \in I)$ .

**Remark 4.41 (Independence of an Event and a Set System)** Let  $(\Omega, \mathcal{A}, P)$  be a probability space. An event  $A \in \mathcal{A}$  and a set system  $\mathcal{E} \subset \mathcal{A}$  are called  $P$ -independent, denoted  $A \perp\!\!\!\perp_P \mathcal{E}$ , if  $\{A\} \perp\!\!\!\perp_P \mathcal{E}$ .  $\triangleleft$

**Remark 4.42 (Independence of  $\sigma$ -Algebras)** Note that  $\sigma$ -algebras are special set systems referred to in the definition above. Hence, a family  $(\mathcal{A}_i, i \in I)$  of sub- $\sigma$ -algebras of  $\mathcal{A}$  can be  $P$ -independent as well. This fact will be used introducing the concept of  $P$ -independence of random variables (see section 5.4).  $\triangleleft$

**Example 4.43 (Joe and Ann With Random Assignment – continued)** Suppose  $A = \{\text{Joe}\} \times \Omega_X \times \Omega_Y$  denotes the event that Joe is sampled and  $B = \Omega_U \times \{\text{yes}\} \times \Omega_Y$  the event that the person sampled is treated. Then  $A$  and  $B$  are independent, because

$$P(A \cap B) = P(\{\text{Joe}\} \times \{\text{yes}\} \times \Omega_Y) = .04 + .16 = .20$$

and

$$\begin{aligned} P(A) \cdot P(B) &= P(\{\text{Joe}\} \times \Omega_X \times \Omega_Y) \cdot P(\Omega_U \times \{\text{yes}\} \times \Omega_Y) \\ &= (.09 + .21 + .04 + .16) \cdot (.04 + .16 + .12 + .08) \\ &= .50 \cdot .40 = .20. \end{aligned}$$

Hence,  $P(A \cap B) = P(A) \cdot P(B)$ . This implies that the  $\sigma$ -algebras  $\{A, A^c, \Omega, \emptyset\}$  and  $\{B, B^c, \Omega, \emptyset\}$  are independent as well [see Box 4.2 (iii)]. In fact, this is a special

case of the following theorem, because the set systems  $\mathcal{E}_1 := \{A\}$  and  $\mathcal{E}_2 := \{B\}$  are  $\cap$ -stable (see Def. 1.36) and  $\sigma(\mathcal{E}_1) = \{A, A^c, \Omega, \emptyset\}$  and  $\sigma(\mathcal{E}_2) = \{B, B^c, \Omega, \emptyset\}$  are the  $\sigma$ -algebras generated by  $\mathcal{E}_1$  and  $\mathcal{E}_2$ , respectively (see Def. 1.13 and Example 1.17).  $\triangleleft$

According to the following theorem, it is sufficient to check  $P$ -independence of a family of  $\cap$ -stable generating system in order to check  $P$ -independence of a family of  $\sigma$ -algebras. In this theorem,  $(\sigma(\mathcal{E}_i), i \in I)$  denotes the family of  $\sigma$ -algebras generated by the set systems  $\mathcal{E}_i, i \in I$ .

**Theorem 4.44 ( $\cap$ -Stable Set Systems and Independence)**

If  $(\Omega, \mathcal{A}, P)$  is a probability space and  $\mathcal{E}_i \subset \mathcal{A}, i \in I$ , are  $\cap$ -stable, then

$$\prod_P (\mathcal{E}_i, i \in I) \Rightarrow \prod_P (\sigma(\mathcal{E}_i), i \in I). \quad (4.26)$$

For a proof see Georgii (2008, Theorem 3.19, p. 65).

## 4.4 Conditional Independence Given an Event

Now we extend the concept of independence of events and of sets of events introducing *conditional independence of events and of sets of events given an event*.

### 4.4.1 Conditional Independence of Events Given an Event

**Definition 4.45 (Conditional Independence of Two Events)**

Let  $(\Omega, \mathcal{A}, P)$  be a probability space,  $A, B, C \in \mathcal{A}$ , and  $P(B) > 0$ . Then the events  $A$  and  $C$  are called  $B$ -conditionally  $P$ -independent, denoted  $A \perp\!\!\!\perp_P C | B$ , if

$$P(A \cap C | B) = P(A | B) \cdot P(C | B). \quad (4.27)$$

**Remark 4.46 (A Condition Equivalent to Conditional Independence)** Suppose  $P(B \cap C) > 0$ . Then Equation (4.27) is equivalent to

$$P(A | B \cap C) = P(A | B) \quad (4.28)$$

[see Box 4.2 (xii)]. Exchanging  $A$  and  $C$  immediately yields: If  $P(A \cap B) > 0$  then Equation (4.27) is equivalent to

$$P(C | A \cap B) = P(C | B). \quad (4.29)$$

$\triangleleft$

**Remark 4.47 (Independence of Events With Respect to  $P^B$ )** Using the conditional-probability measure  $P^B$  defined by (4.14) we can rewrite Equation (4.27):

$$P^B(A \cap C) = P^B(A) \cdot P^B(C). \quad (4.30)$$

This equation shows that  $B$ -conditional  $P$ -independence of  $A$  and  $C$  is equivalent to  $P^B$ -independence of  $A$  and  $C$ , which will also be denoted by  $A \perp\!\!\!\perp_{P^B} C$ .  $\triangleleft$

**Remark 4.48 (Independence and Conditional Independence)** Assume that  $B \in \mathcal{A}$  with  $P(B) > 0$ . Then  $P$ -independence of  $A$  and  $C$  neither implies nor is implied by  $B$ -conditional  $P$ -independence of  $A$  and  $C$  (see Exercise 4-8). However,  $P$ -independence of  $A, B$ , and  $C$  *does* imply  $B$ -conditional  $P$ -independence of  $A$  and  $C$  [see Box 4.2 (x)]. For more propositions on conditional independence of events see Box 4.2 (p. 147), which is proved in Exercise 4-9.  $\triangleleft$

#### 4.4.2 Conditional Independence of Set Systems Given an Event

Now we extend the concept of conditional  $P$ -independence to *set systems*. In Remark 4.47 we already noted that  $B$ -conditional  $P$ -independence of two events  $A$  and  $C$  is equivalent to  $P^B$ -independence of  $A$  and  $C$ . Correspondingly,  $B$ -conditional  $P$ -independence of a family  $(\mathcal{E}_i, i \in I)$  of events is defined as  $P^B$ -independence of  $(\mathcal{E}_i, i \in I)$ .

##### Definition 4.49 (Family of Conditionally Independent Set Systems)

Let  $(\Omega, \mathcal{A}, P)$  be a probability space,  $B \in \mathcal{A}$  with  $P(B) > 0$ , and  $\mathcal{E}_i \subset \mathcal{A}$ ,  $i \in I$ . Then  $(\mathcal{E}_i, i \in I)$  is called a *family of  $B$ -conditionally  $P$ -independent set systems*, denoted  $\perp\!\!\!\perp_{P^B}(\mathcal{E}_i, i \in I) | B$ , if  $\perp\!\!\!\perp_{P^B}(\mathcal{E}_i, i \in I)$ .

**Remark 4.50 (Conditional Independence of  $\sigma$ -Algebras)** Again,  $\sigma$ -algebras can be such set systems referred to in the definition above. Hence, a family  $(\mathcal{A}_i, i \in I)$  of sub- $\sigma$ -algebras of  $\mathcal{A}$  can be  $B$ -conditionally  $P$ -independent as well.  $\triangleleft$

**Remark 4.51 (Independence of Set Systems With Respect to  $P^B$ )** According to Theorem 4.44, under the assumptions of Definition 4.49,  $\cap$ -stability of the set systems  $\mathcal{E}_i$ ,  $i \in I$ , implies

$$\perp\!\!\!\perp_{P^B}(\mathcal{E}_i, i \in I) \Rightarrow \perp\!\!\!\perp_{P^B}(\sigma(\mathcal{E}_i), i \in I). \quad (4.31)$$

$\triangleleft$

Together with Definition 4.49, this remark immediately implies the following corollary.

**Box 4.2 Independence and Conditional Independence of Events**

Let  $(\Omega, \mathcal{A}, P)$  be a probability space and  $A, B, C \in \mathcal{A}$ . Then:

$$A \perp\!\!\!\perp_p B \Leftrightarrow P(A \cap B) = P(A) \cdot P(B) \quad (\text{i})$$

$$A \perp\!\!\!\perp_p B \Leftrightarrow A^c \perp\!\!\!\perp_p B \quad (\text{ii})$$

$$A \perp\!\!\!\perp_p B \Leftrightarrow \sigma(\{A\}) \perp\!\!\!\perp_p \sigma(\{B\}). \quad (\text{iii})$$

$$\begin{aligned} \perp\!\!\!\perp_p A, B, C \Leftrightarrow & P(A \cap B) = P(A) \cdot P(B), \\ & P(A \cap C) = P(A) \cdot P(C), \\ & P(B \cap C) = P(B) \cdot P(C), \\ & P(A \cap B \cap C) = P(A) \cdot P(B) \cdot P(C). \end{aligned} \quad (\text{iv})$$

$$\perp\!\!\!\perp_p A, B, C \Rightarrow A \perp\!\!\!\perp_p B, A \perp\!\!\!\perp_p C, B \perp\!\!\!\perp_p C. \quad (\text{v})$$

If  $P(B) > 0$ , then

$$A \perp\!\!\!\perp_p B \Leftrightarrow P(A|B) = P(A) \quad (\text{vi})$$

$$A \perp\!\!\!\perp_p C|B \Leftrightarrow P(A \cap C|B) = P(A|B) \cdot P(C|B) \quad (\text{vii})$$

$$A \perp\!\!\!\perp_p C|B \Leftrightarrow A \perp\!\!\!\perp_{p^B} C \quad (\text{viii})$$

$$A \perp\!\!\!\perp_p C|B \Leftrightarrow A \perp\!\!\!\perp_p C^c|B \quad (\text{ix})$$

$$\perp\!\!\!\perp_p A, B, C \Rightarrow A \perp\!\!\!\perp_p C|B. \quad (\text{x})$$

If  $P(B), P(B^c) > 0$ , then

$$A \perp\!\!\!\perp_p B \Leftrightarrow P(A|B) = P(A|B^c). \quad (\text{xi})$$

If  $P(B \cap C) > 0$ , then

$$A \perp\!\!\!\perp_p C|B \Leftrightarrow P(A|B \cap C) = P(A|B). \quad (\text{xii})$$

If  $P(B \cap C^c) > 0$ , then

$$A \perp\!\!\!\perp_p C|B \Leftrightarrow P(A|B \cap C^c) = P(A|B). \quad (\text{xiii})$$

If  $P(B \cap C), P(B \cap C^c) > 0$ , then

$$A \perp\!\!\!\perp_p C|B \Leftrightarrow P(A|B \cap C) = P(A|B \cap C^c) \quad (\text{xiv})$$

$$B \perp\!\!\!\perp_p C \Rightarrow P(A|B) = P(A|B \cap C) \cdot P(C) + P(A|B \cap C^c) \cdot P(C^c) \quad (\text{xv})$$

$$A \perp\!\!\!\perp_p C|B \Rightarrow P(A|B) = P(A|B \cap C) \cdot P(C) + P(A|B \cap C^c) \cdot P(C^c). \quad (\text{xvi})$$

**Corollary 4.52 ( $\cap$ -Stable Set Systems and Conditional Independence)**

If  $(\Omega, \mathcal{A}, P)$  is a probability space,  $B \in \mathcal{A}$  with  $P(B) > 0$ , and  $(\mathcal{E}_i, i \in I)$  is a family of  $\cap$ -stable set systems  $\mathcal{E}_i \subset \mathcal{A}$ , then

$$\prod_P (\mathcal{E}_i, i \in I) | B \Rightarrow \prod_P (\sigma(\mathcal{E}_i), i \in I) | B. \quad (4.32)$$

**4.5 Proofs****Proof of Theorem 4.22**

In Remark 4.21 we have already shown that Equation (4.6) holds for  $n = 2$  [see Box 4.1 v]. Hence, for an induction over  $n$  it suffices to show that (4.6) holds for  $A_1, \dots, A_n$  if it holds for  $A_1, \dots, A_{n-1}$ . Note that  $P(\bigcap_{i=1}^{n-1} A_i) > 0$  implies  $P(\bigcap_{i=1}^{j-1} A_i) > 0$  for  $2 \leq j \leq n$ . Hence,

$$\begin{aligned} P\left(\bigcap_{i=1}^n A_i\right) &= P\left(\bigcap_{i=1}^{n-1} A_i \cap A_n\right) \\ &= P\left(\bigcap_{i=1}^{n-1} A_i\right) \cdot P\left(A_n \mid \bigcap_{i=1}^{n-1} A_i\right) && [(4.6)] \\ &= P(A_1) \cdot \left[ \prod_{j=2}^{n-1} P\left(A_j \mid \bigcap_{i=1}^{j-1} A_i\right) \right] \cdot P\left(A_n \mid \bigcap_{i=1}^{n-1} A_i\right) && [(4.6), \text{ ass. of induction}] \\ &= P(A_1) \cdot \prod_{j=2}^n P\left(A_j \mid \bigcap_{i=1}^{j-1} A_i\right). \end{aligned}$$

**Proof of Theorem 4.25**

(i) This equation immediately follows from (1.29).

(ii) If  $P(A_i) > 0$ , then  $P(B \cap A_i) = P(B|A_i) \cdot P(A_i)$  [see Eq. (4.2)]. Hence, (4.7) immediately implies

$$P(B) = \sum_{i=1}^n P(B|A_i) \cdot P(A_i).$$

(iii) This proposition immediately follows from (1.30).

(iv) This proposition immediately follows from (iii) inserting  $P(B \cap A_i) = P(B|A_i) \cdot P(A_i)$  [see Eq. (4.2)].

**Proof of Theorem 4.28**

4.1 (a)

$$\begin{aligned}
P^B(\Omega) &= \frac{P(B \cap \Omega)}{P(B)} && [(4.2)] \\
&= \frac{P(B)}{P(B)} && [B \subset \Omega] \\
&= 1.
\end{aligned}$$

4.1 (b) We assume  $P(B) > 0$ . Therefore,  $P(A \cap B) \geq 0$ , for all  $A \in \mathcal{A}$ , implies that  $P^B(A) = P(A \cap B)/P(B) \geq 0$ , for all  $A \in \mathcal{A}$ .

4.1 (c) If  $A_1, A_2, \dots$  are pairwise disjoint, then  $A_1 \cap B, A_2 \cap B, \dots$  are pairwise disjoint. Therefore,

$$\begin{aligned}
P^B\left(\bigcup_{i=1}^{\infty} A_i\right) &= \frac{P\left[\left(\bigcup_{i=1}^{\infty} A_i\right) \cap B\right]}{P(B)} && [(4.2)] \\
&= \frac{P\left[\bigcup_{i=1}^{\infty} (A_i \cap B)\right]}{P(B)} \\
&= \frac{\sum_{i=1}^{\infty} P(A_i \cap B)}{P(B)} && [\text{Def. 4.1 (c)}] \\
&= \sum_{i=1}^{\infty} P^B(A_i). && [(4.2)]
\end{aligned}$$

### ***Proof of Lemma 4.30***

$$\begin{aligned}
P(A|B \cap C) &= \frac{P(A \cap B \cap C)}{P(B \cap C)} && [(4.2)] \\
&= \frac{P(A \cap C|B) \cdot P(B)}{P(C|B) \cdot P(B)} && [(4.2)] \\
&= \frac{P^B(A \cap C)}{P^B(C)} && [(4.14)] \\
&= P^B(A|C). && [(4.2)]
\end{aligned}$$

### ***Proof of Lemma 4.32***

(i) For all  $A \in \mathcal{A}$ ,

$$\begin{aligned}
P(A) = 0 &\Rightarrow P(A \cap B) = 0 && [\text{Box 4.1 (v)}] \\
&\Rightarrow \frac{P(A \cap B)}{P(B)} = 0 && [P(B) > 0] \\
&\Rightarrow P(A|B) = 0 && [(4.2)] \\
&\Rightarrow P^B(A) = 0. && [(4.14)]
\end{aligned}$$

Hence,  $P^B \ll_{\mathcal{A}} P$  [see Def. 3.70].

(ii) For all  $A \in \mathcal{A}$ ,

$$\begin{aligned}
P^B(A) &= \frac{P(A \cap B)}{P(B)} && [(4.14), (4.2)] \\
&= \frac{1}{P(B)} \int 1_{A \cap B} dP && [(3.9)] \\
&= \frac{1}{P(B)} \int 1_A \cdot 1_B dP && [(1.33)] \\
&= \int_A \frac{1}{P(B)} \cdot 1_B dP. && [(3.30), (3.32)]
\end{aligned}$$

According to Theorem 3.65, this means  $P^B = \left( \frac{1}{P(B)} \cdot 1_B \right) \circledast P$ .

## 4.6 Exercises

▷ **Exercise 4-1** Consider flipping a coin  $n$  times and the event  $A_1 = \text{flipping heads at the first flip}$ . Specify the set  $\Omega$  of possible outcomes of this random experiment and the set  $A_1$  as a subset of  $\Omega$ . How many elements has  $\Omega$ ? How many elements has the event  $A_1$ ?

▷ **Exercise 4-2** Draw the interval  $[0, 1]$ , cut it in two halves, cut the right-hand piece in two halves, cut the remaining part in two halves, etc. In this way you can visualize the sequence  $1/2, 1/4, 1/8, \dots$  by lengths of intervals. This sequence can also be written:  $1/2^i, i \in \mathbb{N}$ . Note that all terms  $1/2^i$  of this sequence are positive, i. e.,  $1/2^i > 0$  for all  $i \in \mathbb{N}$ . Determine

$$\sum_{i=1}^{\infty} \frac{1}{2^i} = \lim_{n \rightarrow \infty} \sum_{i=1}^n \frac{1}{2^i}.$$

▷ **Exercise 4-3** Consider Example 4.19 and list all elements of the  $\sigma$ -algebras  $\mathcal{F}_1$  and  $\mathcal{F}_2$  referring explicitly to the elements of  $\Omega$ .

▷ **Exercise 4-4** Compute the probabilities  $P(A)$  and  $P(C | A \cap B)$  of the events defined in Example 4.23.

▷ **Exercise 4-5** Compute the  $B$ -conditional probability of the event  $\{(Ann, \text{yes}, -), (Ann, \text{yes}, +)\}$ , i. e., the event that *Ann is sampled and treated*. Use the results already obtained in Example 4.34.

▷ **Exercise 4-6** In Example 4.16 we computed the conditional probabilities  $P(C | B) = .60$  and  $P(C | B^c) = .45$ , where  $C$  is the event that the *drawn person is successful* and  $B$  is the event that the *drawn person is treated*. What is the conclusion the effect of the treatment if we compare these two conditional probabilities to each other?

▷ **Exercise 4-7** Prove the proposition of Remark 4.39.

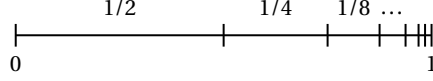
▷ **Exercise 4-8** Show by examples that, for  $B \in \mathcal{A}$  with  $P(B) > 0$ ,  $P$ -independence of  $A$  and  $C$  neither implies nor is implied by  $B$ -conditional  $P$ -independence of  $A$  and  $C$ .

▷ **Exercise 4-9** Prove the propositions of Box 4.2.

## Solutions

▷ **Solution 4-1** The set of possible outcomes is  $\Omega = \{h, t\}^n = \{h, t\} \times \dots \times \{h, t\}$  ( $n$ -times). The event *flipping heads at time 1* is  $A_1 = \{h\} \times \{h, t\}^{n-1}$ . The set  $\Omega$  has  $2^n$  elements and  $A_1$  has  $2^n/2 = 2^{n-1}$  elements.

▷ **Solution 4-2** The picture of this interval is



and this illustrates that  $\sum_{i=1}^{\infty} \frac{1}{2^i} = \lim_{n \rightarrow \infty} \sum_{i=1}^n \frac{1}{2^i} = 1$ .

▷ **Solution 4-3**

$$\begin{aligned} \mathcal{F}_1 &= \sigma(\{A\}) = \{A, A^c, \Omega, \emptyset\} \\ &= \left\{ \{ (Joe, no, -), (Joe, no, +), (Joe, yes, -), (Joe, yes, +) \}, \right. \\ &\quad \left. \{ (Ann, no, -), (Ann, no, +), (Ann, yes, -), (Ann, yes, +) \}, \Omega, \emptyset \right\}. \end{aligned}$$

$$\begin{aligned} \mathcal{F}_2 &= \sigma(\{A, B\}) \\ &= \{A, A^c, B, B^c, (A \cap B) \cup (A^c \cap B^c), (A \cap B^c) \cup (A^c \cap B), \\ &\quad A \cap B, A^c \cap B, A \cap B^c, A^c \cap B^c, \\ &\quad A \cup B, A^c \cup B, A \cup B^c, A^c \cup B^c, \Omega, \emptyset\} \end{aligned}$$

where

$$\begin{aligned} A &= \{ (Joe, yes, -), (Joe, yes, +), (Joe, no, -), (Joe, no, +) \}, \\ A^c &= \{ (Ann, yes, -), (Ann, yes, +), (Ann, no, -), (Ann, no, +) \}, \\ B &= \{ (Joe, yes, -), (Joe, yes, +), (Ann, yes, -), (Ann, yes, +) \}, \\ B^c &= \{ (Joe, no, -), (Joe, no, +), (Ann, no, -), (Ann, no, +) \}, \\ (A \cap B) \cup (A^c \cap B^c) &= \{ (Joe, yes, -), (Joe, yes, +), (Ann, no, -), (Ann, no, +) \}, \\ (A \cap B^c) \cup (A^c \cap B) &= \{ (Joe, no, -), (Joe, no, +), (Ann, yes, -), (Ann, yes, +) \}, \end{aligned}$$

$$\begin{aligned} A \cap B &= \{ (Joe, yes, -), (Joe, yes, +) \}, \\ A^c \cap B &= \{ (Ann, yes, -), (Ann, yes, +) \}, \\ A \cap B^c &= \{ (Joe, no, -), (Joe, no, +) \}, \\ A^c \cap B^c &= \{ (Ann, no, -), (Ann, no, +) \}, \end{aligned}$$

$$\begin{aligned} A \cup B &= \{ (Joe, yes, -), (Joe, yes, +), (Joe, no, -), (Joe, no, +), (Ann, yes, -), (Ann, yes, +) \}, \\ A^c \cup B &= \{ (Ann, yes, -), (Ann, yes, +), (Ann, no, -), (Ann, no, +), (Joe, yes, -), (Joe, yes, +) \}, \\ A \cup B^c &= \{ (Joe, yes, -), (Joe, yes, +), (Joe, no, -), (Joe, no, +), (Ann, no, -), (Ann, no, +) \}, \\ A^c \cup B^c &= \{ (Ann, yes, -), (Ann, yes, +), (Ann, no, -), (Ann, no, +), (Joe, no, -), (Joe, no, +) \}. \end{aligned}$$

▷ **Solution 4-4** Because the four events  $\{(Joe, no, -)\}, \dots, \{(Joe, yes, +)\}$  are pairwise disjoint, we can simply add their probabilities. Hence,  $P(A) = .09 + .21 + .04 + .16 = .5$  (see the second column of Table 4.1, p. 131). In order to compute  $P(C|A \cap B)$  note that  $A \cap B = \{(Joe, yes, -), (Joe, yes, +)\}$  is the event that *Joe is drawn and treated*. Again, because the two elementary events involved are disjoint,  $P(A \cap B) = .04 + .16 = 0.2$ . Furthermore,  $A \cap B \cap C = \{(Joe, yes, +)\}$  is the event that *Joe is drawn, treated, and successful*. Its probability is  $P(A \cap B \cap C) = P(\{(Joe, yes, +)\}) = .16$ . Hence,

$$P(C|A \cap B) = \frac{P(A \cap B \cap C)}{P(A \cap B)} = \frac{.16}{.04 + .16} = .8.$$

▷ **Solution 4-5** In Example 4.34 we already computed the two  $B$ -conditional probabilities  $P^B(\{(Ann, yes, -)\}) = .30$  and  $P^B(\{(Ann, yes, +)\}) = .20$ . Because these elementary events are *disjoint*, the probabilities of their union can easily be computed using the additivity property of the probability measure  $P^B$ . Hence,  $P^B(\{(Ann, yes, -), (Ann, yes, +)\}) = .30 + .20 = .50$ .

▷ **Solution 4-6** Although this question and the concepts needed for an answer are beyond the scope of this book, the difference  $P(C|B) - P(C|B^c) = .60 - .45 = .15$  is the *average total treatment effect* (see Steyer et al., 2014). It is the average of the two *individual total treatment effects* of Joe and of Ann. For Joe this individual treatment effect is  $.80 - .70 = .10$ , (probability of *success* given *Joe* and *treatment* – probability of *success* given *Joe* and no *treatment*) whereas it is  $.40 - .20 = 0$  for Ann.

▷ **Solution 4-7** Let  $(\Omega, \mathcal{A}, P)$  be a probability space. Then for all  $A \in \mathcal{A}$ :  
 $P(\Omega \cap A) = P(A) = 1 \cdot P(A) = P(\Omega) \cdot P(A)$  and  $P(\emptyset \cap A) = P(\emptyset) = 0 = 0 \cdot P(A) = P(\emptyset) \cdot P(A)$ .

▷ **Solution 4-8** Consider Example 2.2.2, let  $A = \{(h, t), (h, h)\}$  denote the event to flip heads with the first coin,  $B = \{(t, t), (h, h)\}$  the event to flip tails or to flip heads with both coins, and  $C = \{(t, h), (h, h)\}$  the event to flip heads with the second coin. All three events have the same probability  $P(A) = P(B) = P(C) = .5$ . Now,

$$P(A \cap C) = P(\{(h, h)\}) = .25 = .5 \cdot .5 = P(A) \cdot P(C)$$

and

$$P(B \cap C) = P(\{(h, h)\}) = .25 = .5 \cdot .5 = P(B) \cdot P(C).$$

Hence,  $A$  and  $C$  as well as  $B$  and  $C$  are  $P$ -independent, which implies  $P(A|B) = .5$  and  $P(C|B) = .5$ . However,

$$\begin{aligned} P(A \cap C|B) &= \frac{P(A \cap C \cap B)}{P(B)} = \frac{.25}{.5} = .5 \\ &\neq .25 = .5 \cdot .5 = P(A|B) \cdot P(C|B), \end{aligned}$$

which shows that  $A$  and  $C$  are not  $B$ -conditionally  $P$ -independent.

Now we present an example in which  $A$  and  $C$  are  $B$ -conditionally  $P$ -independent but not (unconditionally)  $P$ -independent. Consider flipping three coins. This random experiment is represented by the probability space  $(\Omega, \mathcal{A}, P)$ , where  $\Omega = \{h, t\}^3$ ,  $\mathcal{A} = \mathcal{P}(\Omega)$ , and  $P: \mathcal{A} \rightarrow [0, 1]$ , satisfying  $P(\{\omega\}) = .125$  for all  $\omega \in \Omega$ . Furthermore, let  $A = \{(t, t, t), (t, t, h)\}$  denote the event to flip tails with the first two coins,  $B = \{(t, t, t), (t, t, h), (t, h, t), (t, h, h)\}$  the event to flip tails with the first coin, and  $C = \{(t, t, h), (t, h, h)\}$  the event to flip tails with the first coin and heads with the third coin. The two events  $A$  and  $C$  have the same probability  $P(A) = P(C) = .25$  and  $P(B) = .5$ . Because

$$P(A \cap C) = P(\{(t, t, h)\}) = .125 \neq .25 \cdot .25 = P(A) \cdot P(C),$$

$A$  and  $C$  are *not*  $P$ -independent. Further,  $P(A \cap B) = P(\{(t, t, t), (t, t, h)\}) = .25$ ,  $P(C \cap B) = P(\{(t, t, h), (t, h, h)\}) = .25$ , and

$$\begin{aligned} P(A \cap C | B) &= \frac{P(A \cap B \cap C)}{P(B)} = \frac{.125}{.5} = .25 \\ &= .5 \cdot .5 = \frac{P(A \cap B)}{P(B)} \cdot \frac{P(B \cap C)}{P(B)} = P(A | B) \cdot P(C | B). \end{aligned}$$

This shows that  $A$  and  $C$  are  $B$ -conditionally  $P$ -independent.

▷ **Solution 4-9** (i) This is the definition of  $A \perp\!\!\!\perp_P B$ .

$$\begin{aligned} \text{(ii)} \quad P(A^c \cap B) &= P(B \setminus A) \\ &= P(B) - P(A \cap B) && \text{[Box 4.1 (vi)]} \\ &= P(B) - P(A) \cdot P(B) && [A \perp\!\!\!\perp_P B] \\ &= [1 - P(A)] \cdot P(B) \\ &= P(A^c) \cdot P(B) && \text{[Box 4.1 (iv)],} \end{aligned}$$

which is  $A^c \perp\!\!\!\perp_P B$ .

(iii) We have to show that  $A \perp\!\!\!\perp_P B$  implies

$$P(A_i \cap B_j) = P(A_i) \cdot P(B_j), \quad \forall A_i \in \{A, A^c, \Omega, \emptyset\} \text{ and } \forall B_j \in \{B, B^c, \Omega, \emptyset\}.$$

Whenever  $A_i$  or  $B_j$  is  $\Omega$  or  $\emptyset$  this equation holds [see (4.25)]. Furthermore,  $P(A \cap B) = P(A) \cdot P(B)$  is equivalent to our premise  $A \perp\!\!\!\perp_P B$ , and  $P(A^c \cap B) = P(A^c) \cdot P(B)$  is proposition (ii). The corresponding argument holds for  $P(A \cap B^c) = P(A) \cdot P(B^c)$  and  $P(A^c \cap B^c) = P(A^c) \cdot P(B^c)$ , exchanging the roles of  $A$  and  $B$ .

(iv) This is the definition of  $\perp\!\!\!\perp_P A, B, C$ .

(v) This proposition immediately follows from (iv) and (i).

(vi) We assume  $P(B) > 0$ . Then

$$\begin{aligned} A \perp\!\!\!\perp_P B &\Leftrightarrow P(A \cap B) = P(A) \cdot P(B) && \text{[(i)]} \\ &\Leftrightarrow \frac{P(A \cap B)}{P(B)} = P(A) \\ &\Leftrightarrow P(A | B) = P(A). && \text{[(4.2)]} \end{aligned}$$

(vii) This is the definition of  $A \perp\!\!\!\perp_P C | B$ .

(viii) We assume  $P(B) > 0$ . Therefore,

$$\begin{aligned} A \perp\!\!\!\perp_P C | B &\Leftrightarrow P(A \cap C | B) = P(A | B) \cdot P(C | B) && \text{[(vii)]} \\ &\Leftrightarrow P^B(A \cap C) = P^B(A) \cdot P^B(C) && \text{[(4.14)]} \\ &\Leftrightarrow A \perp\!\!\!\perp_{P^B} C && \text{[(i)]} \end{aligned}$$

(ix) We assume  $P(B) > 0$ . Therefore,

$$\begin{aligned} A \perp\!\!\!\perp_P C | B &\Leftrightarrow A \perp\!\!\!\perp_{P^B} C && \text{[(viii)]} \\ &\Leftrightarrow A \perp\!\!\!\perp_{P^B} C^c. && \text{[(ii)]} \end{aligned}$$

(x) We assume  $P(B) > 0$ . Then

$$\begin{aligned}
P(A \cap C | B) &= \frac{P(A \cap B \cap C)}{P(B)} && [(4.2)] \\
&= \frac{P(A) \cdot P(B) \cdot P(C)}{P(B)} && [A \perp_p B, C, \text{ (iv)}] \\
&= P(A|B) \cdot P(C|B). && [(v), (vi)]
\end{aligned}$$

(xi) We assume  $P(B), P(B^c) > 0$ . Then

$$\begin{aligned}
P(A|B) = P(A|B^c) &\Leftrightarrow \frac{P(A \cap B)}{P(B)} = \frac{P(A \cap B^c)}{1 - P(B)} && [(4.2), \text{Box 4.1 (iv)}] \\
&\Leftrightarrow P(A \cap B) \cdot [1 - P(B)] = P(A \cap B^c) \cdot P(B) \\
&\Leftrightarrow P(A \cap B) = [P(A \cap B) + P(A \cap B^c)] \cdot P(B) \\
&\Leftrightarrow P(A \cap B) = P(A) \cdot P(B) && [(4.7)] \\
&\Leftrightarrow A \perp_p B. && [(i)]
\end{aligned}$$

(xii) We assume  $P(B \cap C) > 0$ . This implies  $P(B) > 0$  and

$$\begin{aligned}
A \perp_p C | B &\Leftrightarrow P(A \cap C | B) = P(A|B) \cdot P(C|B) && [(vii)] \\
&\Leftrightarrow \frac{P(A \cap B \cap C)}{P(B)} = \frac{P(A \cap B)}{P(B)} \cdot \frac{P(B \cap C)}{P(B)} && [(4.2)] \\
&\Leftrightarrow \frac{P(A \cap B \cap C)}{P(B \cap C)} = \frac{P(A \cap B)}{P(B)} \\
&\Leftrightarrow P(A|B \cap C) = P(A|B). && [(4.2)]
\end{aligned}$$

(xiii) We assume  $P(B \cap C^c) > 0$ . This implies  $P(B) > 0$  and

$$\begin{aligned}
A \perp_p C | B &\Leftrightarrow A \perp_{P^B} C && [(viii)] \\
&\Leftrightarrow A \perp_{P^B} C^c && [(ii)] \\
&\Leftrightarrow P^B(A|C^c) = P^B(A) && [(vi)] \\
&\Leftrightarrow P(A|B \cap C^c) = P(A|B) && [(4.15), (4.14)]
\end{aligned}$$

(xiv) We assume  $P(B \cap C), P(B \cap C^c) > 0$ .

$$\begin{aligned}
P(A|B \cap C) = P(A|B \cap C^c) &\Leftrightarrow P^B(A|C) = P^B(A|C^c) && [(4.15)] \\
&\Leftrightarrow A \perp_{P^B} C && [(xi)] \\
&\Leftrightarrow A \perp_p C | B. && [(viii)]
\end{aligned}$$

(xv) We assume  $P(B \cap C), P(B \cap C^c) > 0$ .

$$\begin{aligned}
B \perp_p C &\Rightarrow P(C|B) = P(C), P(C^c|B) = P(C^c) && [(vi), (ii)] \\
&\Rightarrow P(A|B) = P(A|B \cap C) \cdot P(C) + P(A|B \cap C^c) \cdot P(C^c). && [(4.17)]
\end{aligned}$$

(xvi) We assume  $P(B \cap C), P(B \cap C^c) > 0$  and  $A \perp_p C | B$ .

$$\begin{aligned}
P(A|B) &= P(A|B \cap C) && [A \perp_p C | B, (xii)] \\
&= P(A|B \cap C) \cdot [P(C) + P(C^c)] && [\text{Box 4.1 (iv)}] \\
&= P(A|B \cap C) \cdot P(C) + P(A|B \cap C) \cdot P(C^c) \\
&= P(A|B \cap C) \cdot P(C) + P(A|B \cap C^c) \cdot P(C^c). && [(xiv)]
\end{aligned}$$

## Chapter 5

# Random Variable, Distribution, Density, and Distribution Function

In chapter 4 we translated the concepts *measure* and *measure space* to probability theory introducing the notions *probability measure* and *probability space*. In this chapter we define a *random variable* as a measurable mapping and its *distribution* as the image measure of a measurable mapping with respect to a probability measure (see ch. 2). The distribution of a random variable contains the comprehensive information about its properties. It informs us about the probability of each event that can be represented by this random variable. Expectation, variance and other moments of a random variable are determined by its distribution (see ch. 6). For a multivariate random variable, the (joint) distribution also contains the information about the dependencies between its components. It also determines the conditional expectations (see ch. 10). In this chapter, we apply the concept of independence of families of events in order to introduce *independence of random variables* and *families of random variables*. Finally, the last sections of this chapter are devoted to the concept of a *probability function*, and, for a real-valued random variable, the notions of a *distribution function* and a *probability density*, which are very useful for describing a distribution, for calculations (see, e. g., ch. 6), and for providing instructive illustrations of the underlying distributions (see ch. 8).

### 5.1 Random Variable and its Distribution

In section 2.6 we introduced the notation

$$f: (\Omega, \mathcal{A}, \mu) \rightarrow (\Omega', \mathcal{A}'),$$

which expresses that  $f: \Omega \rightarrow \Omega'$  is an  $(\mathcal{A}, \mathcal{A}')$ -measurable mapping and that  $\mu$  is a measure on the measurable space  $(\Omega, \mathcal{A})$ . If  $\mu$  is a probability measure, then a measurable mapping is also called a *random variable* and its image measure  $\mu_f$  is also called its *distribution*. This change of terms goes along with a change in notation. Instead of  $f$ ,  $g$ , and  $h$ , we preferably use letters such as  $X$ ,  $Y$ , and  $Z$ .

#### Definition 5.1 (Random Variable)

If  $(\Omega, \mathcal{A}, P)$  is a probability space and  $X: (\Omega, \mathcal{A}) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  a measurable mapping, i. e., if  $X: \Omega \rightarrow \Omega'_X$  satisfies

$$X^{-1}(A') \in \mathcal{A}, \quad \forall A' \in \mathcal{A}'_X, \quad (5.1)$$

then  $X$  is called a random variable on  $(\Omega, \mathcal{A}, P)$  with values in  $(\Omega'_X, \mathcal{A}'_X)$ . If  $(\Omega'_X, \mathcal{A}'_X) = (\mathbb{R}, \mathcal{B})$ , then  $X$  is called real-valued, and if  $(\Omega'_X, \mathcal{A}'_X) = (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ , then  $X$  is called numerical.

**Remark 5.2 (Measurability of Inverse Images)** Equation (5.1) implies that all inverse images

$$X^{-1}(A') := \{\omega \in \Omega: X(\omega) \in A'\}, \quad A' \in \mathcal{A}'_X,$$

are elements of the  $\sigma$ -algebra  $\mathcal{A}$  on  $\Omega$ . Because the measure  $P: \mathcal{A} \rightarrow [0, 1]$  assigns a probability to *all* elements of  $\mathcal{A}$ , the probabilities  $P[X^{-1}(A')]$  of these inverse images are determined by  $P$  (see Exercises 5-1 and 5-2).  $\triangleleft$

**Definition 5.3 (Distribution of a Random Variable)**

Suppose that  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  is a random variable. Then the function  $P_X: \mathcal{A}'_X \rightarrow [0, 1]$  defined by

$$P_X(A') = P[X^{-1}(A')], \quad \forall A' \in \mathcal{A}'_X, \quad (5.2)$$

is called the distribution of  $X$  (with respect to  $P$ ).

**Remark 5.4 (Notation  $P(X \in A')$  and  $P(X=x)$ )** If  $A' \in \mathcal{A}'_X$ , we use the notation

$$P(X \in A') := P[X^{-1}(A')] \quad (5.3)$$

for the probability of the event  $\{X \in A'\} = X^{-1}(A')$ , i. e., the event that  $X$  takes on a value in the subset  $A'$  of  $\Omega'_X$ . If  $\{x\} \in \mathcal{A}'_X$ ,

$$P(X=x) := P[X^{-1}(\{x\})] \quad (5.4)$$

for the probability of the event  $\{X=x\} = X^{-1}(\{x\}) = \{\omega \in \Omega: X(\omega) = x\}$ . If we write  $P(X=x)$ , then we always assume  $\{x\} \in \mathcal{A}'_X$ , even if not mentioned explicitly.  $\triangleleft$

**Remark 5.5 (A New Probability Space)** Definition 5.1 implies that *every* random variable  $X$  on a probability space  $(\Omega, \mathcal{A}, P)$  has a distribution  $P_X$ . Furthermore,  $P_X: \mathcal{A}'_X \rightarrow [0, 1]$  is also a measure, the *image measure of  $P$  under  $X$*  (see Th. 2.79 and Def. 2.80). Because  $P_X(\Omega'_X) = P(\Omega) = 1$ , we can conclude that  $P_X$  is a probability measure, and  $(\Omega'_X, \mathcal{A}'_X, P_X)$  is also a probability space. Therefore, we use the notation

$$X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X, P_X)$$

expressing that

- (a)  $X: \Omega \rightarrow \Omega'_X$  is a random variable on the probability space  $(\Omega, \mathcal{A}, P)$ ,
- (b)  $\mathcal{A}'_X$  is a  $\sigma$ -algebra on  $\Omega'_X$ , and

(c)  $P_X$  is the distribution of  $X$ .

◁

**Definition 5.6 (Identically Distributed Random Variables)**

Let  $X: (\Omega^{(1)}, \mathcal{A}^{(1)}, P^{(1)}) \rightarrow (\Omega', \mathcal{A}')$  and  $Y: (\Omega^{(2)}, \mathcal{A}^{(2)}, P^{(2)}) \rightarrow (\Omega', \mathcal{A}')$  be random variables. If  $P_X = P_Y$ , then we say that  $X$  and  $Y$  are identically distributed.

Note that, oftentimes,  $(\Omega^{(1)}, \mathcal{A}^{(1)}, P^{(1)}) = (\Omega^{(2)}, \mathcal{A}^{(2)}, P^{(2)})$ . Now we consider the distribution of a composition  $g(X)$  of a random variable  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  and a measurable function  $g: (\Omega'_X, \mathcal{A}'_X) \rightarrow (\Omega', \mathcal{A}')$ . According to Remark 5.5, the mapping  $g$  is a random variable on the probability space  $(\Omega'_X, \mathcal{A}'_X, P_X)$ . Furthermore, according to the following lemma,  $g(X)$  is a random variable on  $(\Omega, \mathcal{A}, P)$  and the distribution of  $g(X)$  is the image measure of  $P_X$  under  $g$ . The notation of this image measure is  $(P_X)_g$ .

**Lemma 5.7 (Distribution of a Composition)**

Let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  be a random variable and  $g: (\Omega'_X, \mathcal{A}'_X) \rightarrow (\Omega', \mathcal{A}')$  a measurable function. Then the composition  $g(X): (\Omega, \mathcal{A}, P) \rightarrow (\Omega', \mathcal{A}')$  is a random variable and

$$(P_X)_g = P_{g(X)}. \quad (5.5)$$

(Proof p. 188)

**Example 5.8 (Indicator (Variable) of an Event)** If  $(\Omega, \mathcal{A}, P)$  is a probability space and  $A \in \mathcal{A}$ , then the mapping  $1_A: (\Omega, \mathcal{A}, P) \rightarrow (\{0, 1\}, \mathcal{P}(\{0, 1\}))$  is a random variable. It is called the *indicator (variable) of  $A$* . The distribution of  $1_A$  is

$$P_{1_A}(\{0\}) = P(A^c), \quad P_{1_A}(\{1\}) = P(A), \quad P_{1_A}(\{0, 1\}) = P(\Omega) = 1, \quad P_{1_A}(\emptyset) = P(\emptyset) = 0.$$

If we consider the same event  $A$  and the measurable space  $(\mathbb{R}, \mathcal{B})$ , then we can also write  $1_A: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  in order to express that  $1_A$  is also  $(\mathcal{A}, \mathcal{B})$ -measurable. Note, however, that now the distribution of  $1_A$  is a probability measure on  $(\mathbb{R}, \mathcal{B})$ , and for all  $B \in \mathcal{B}$ ,

$$\begin{aligned} P_{1_A}(B) &= P[1_A^{-1}(B)] = P[\{\omega \in \Omega: 1_A(\omega) \in B\}] && [(5.2), (2.2)] \\ &= \begin{cases} P(\emptyset) = 0, & \text{if } 0 \notin B, 1 \notin B, \\ P(A), & \text{if } 0 \notin B, 1 \in B, \\ P(A^c), & \text{if } 0 \in B, 1 \notin B, \\ P(\Omega) = 1, & \text{if } \{0, 1\} \subset B. \end{cases} \end{aligned}$$

◁

**Example 5.9 (Indicator of an Inverse Image)** If  $(\Omega, \mathcal{A}, P)$  is a probability space,  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  a random variable, and  $A' \in \mathcal{A}'_X$ , then  $1_{X^{-1}(A')}: (\Omega, \mathcal{A}, P) \rightarrow (\{0, 1\}, \mathcal{P}(\{0, 1\}))$  is a random variable on  $(\Omega, \mathcal{A}, P)$  and

$$1_{X \in A'} := 1_{X^{-1}(A')} = 1_{A'}(X) = 1_{A'} \circ X \quad (5.6)$$

(see Exercise 5-3). The distribution of  $1_{X \in A'}$  is

$$\begin{aligned} P_{1_{X \in A'}}(\{0\}) &= P(X \notin A'), & P_{1_{X \in A'}}(\{1\}) &= P(X \in A'), \\ P_{1_{X \in A'}}(\{0, 1\}) &= P(\Omega) = 1, & P_{1_{X \in A'}}(\emptyset) &= P(\emptyset) = 0. \end{aligned}$$

◁

**Example 5.10 (Dichotomous Random Variable)** Let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  be a random variable on  $(\Omega, \mathcal{A}, P)$ . Then  $X$  is called *dichotomous with values  $x_1$  and  $x_2$*  if  $\{x_1\}, \{x_2\} \in \mathcal{A}'_X$ ,  $P(X \in \{x_1, x_2\}) = 1$  and  $0 < P(X=x_1) < 1$ . If  $X$  is dichotomous with values 0 and 1, then  $X \stackrel{p}{=} 1_{X=1}$ .

◁

**Example 5.11 (Flipping two Coins – continued)** In Examples 2.2.2 and 2.83, we considered flipping two coins and defined  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{P}(\Omega'_X))$ , a random variable representing with its values the *number of flipping heads*. Its possible values are 0, 1, or 2. Hence, we can choose  $\Omega'_X := \{0, 1, 2\}$  and

$$\begin{aligned} P(X=0) &= P_X(\{0\}) = P[X^{-1}(\{0\})] = P[\{(t, t)\}] = \frac{1}{4}, \\ P(X=1) &= P_X(\{1\}) = P[X^{-1}(\{1\})] = P[\{(h, t), (t, h)\}] = \frac{1}{2}, \\ P(X=2) &= P_X(\{2\}) = P[X^{-1}(\{2\})] = P[\{(h, h)\}] = \frac{1}{4} \end{aligned}$$

are the probabilities assigned to the singletons  $\{0\}$ ,  $\{1\}$ , and  $\{2\}$ , whereas

$$\begin{aligned} P(X \in \{0, 1\}) &= P_X(\{0, 1\}) = P[X^{-1}(\{0, 1\})] = P[\{(t, t), (h, t), (t, h)\}] = \frac{3}{4}, \\ P(X \in \{0, 2\}) &= P_X(\{0, 2\}) = P[X^{-1}(\{0, 2\})] = P[\{(t, t), (h, h)\}] = \frac{2}{4}, \\ P(X \in \{1, 2\}) &= P_X(\{1, 2\}) = P[X^{-1}(\{1, 2\})] = P[\{(h, t), (t, h), (h, h)\}] = \frac{3}{4} \end{aligned}$$

are the probabilities assigned to the sets  $\{0, 1\}$ ,  $\{0, 2\}$ , and  $\{1, 2\}$ , which consist of two elements of  $\Omega'_X$ . Finally,  $P_X(\Omega'_X) = P[X^{-1}(\Omega'_X)] = P(\Omega) = 1$  and  $P_X(\emptyset) = P[X^{-1}(\emptyset)] = P(\emptyset) = 0$ .

◁

**Example 5.12 (Tom, Jim, and Kate)** Now we consider an example that is similar to the experiment with Joe and Ann. However, the set of persons is now  $\Omega_U := \{Tom, Jim, Kate\}$ , and we consider three treatments, the elements of the set  $\Omega_X := \{Con, BTh, PTh\}$ , where *Con* could be *no treatment*. The random experiment consists of: drawing a unit  $u$  from the set  $\Omega_U$ , assigning it to one of the three

**Table 5.1.** Tom, Jim, and Kate

Elements of $\Omega$			Random variables			
Unit	Treatment	Success	Probabilities of elementary events $P(\{\omega\})$	Person variable $U$	Treatment variable $X$	Outcome variable $Y$
<i>(Tom, Con, -)</i>			10/99	<i>Tom</i>	0	0
<i>(Tom, Con, +)</i>			10/99	<i>Tom</i>	0	1
<i>(Tom, BTh, -)</i>			2/99	<i>Tom</i>	1	0
<i>(Tom, BTh, +)</i>			6/99	<i>Tom</i>	1	1
<i>(Tom, PTh, -)</i>			1/99	<i>Tom</i>	2	0
<i>(Tom, PTh, +)</i>			4/99	<i>Tom</i>	2	1
<i>(Jim, Con, -)</i>			5/99	<i>Jim</i>	0	0
<i>(Jim, Con, +)</i>			15/99	<i>Jim</i>	0	1
<i>(Jim, BTh, -)</i>			3/99	<i>Jim</i>	1	0
<i>(Jim, BTh, +)</i>			5/99	<i>Jim</i>	1	1
<i>(Jim, PTh, -)</i>			2/99	<i>Jim</i>	2	0
<i>(Jim, PTh, +)</i>			3/99	<i>Jim</i>	2	1
<i>(Kate, Con, -)</i>			12/99	<i>Kate</i>	0	0
<i>(Kate, Con, +)</i>			8/99	<i>Kate</i>	0	1
<i>(Kate, BTh, -)</i>			5/99	<i>Kate</i>	1	0
<i>(Kate, BTh, +)</i>			3/99	<i>Kate</i>	1	1
<i>(Kate, PTh, -)</i>			4/99	<i>Kate</i>	2	0
<i>(Kate, PTh, +)</i>			1/99	<i>Kate</i>	2	1

treatment conditions *Con*, *BTh*, or *PTh* and observing whether (+) or not (-) a success criterion is reached. Hence, the set of possible outcomes of this random experiment is

$$\Omega := \Omega_U \times \Omega_X \times \Omega_Y = \{ (Tom, Con, -), (Tom, Con, +), \dots, (Kate, PTh, +) \}.$$

It consists of the  $3 \cdot 3 \cdot 2 = 18$  triples  $(u, \omega_X, \omega_Y)$  listed in the first column of Table 5.1. As the set of possible events  $\mathcal{A}$  we consider the power set  $\mathcal{P}(\Omega)$ . This set has  $2^{18} = 262144$  elements, where 18 is the number of elements of  $\Omega$ . The probabilities of the 18 elementary events  $\{\omega\}$ ,  $\omega \in \Omega$ , are displayed in the second column of the table. With these specifications, the probabilities  $P(A)$  of all  $2^{18}$  elements  $A \in \mathcal{A}$  are determined [see Rule (x) of Box 4.1]. Hence, the probability space  $(\Omega, \mathcal{A}, P)$  is completely specified.

Table 5.1 also displays the values of the three random variables  $U: (\Omega, \mathcal{A}, P) \rightarrow [\Omega_U, \mathcal{P}(\Omega_U)]$ ,  $X: (\Omega, \mathcal{A}, P) \rightarrow [\Omega'_X, \mathcal{P}(\Omega'_X)]$ , and  $Y: (\Omega, \mathcal{A}, P) \rightarrow [\Omega'_Y, \mathcal{P}(\Omega'_Y)]$ , where

$\Omega'_X := \{0, 1, 2\}$  and  $\Omega'_Y := \{0, 1\}$ . For the singletons  $\{x\}$ ,  $x \in \Omega'_X$ , the values  $P_X(\{x\}) = P[X^{-1}(\{x\})]$  of the distribution of  $X$  are

$$P_X(\{0\}) = 60/99, \quad P_X(\{1\}) = 24/99, \quad P_X(\{2\}) = 15/99,$$

for the sets that consist of two elements of  $\Omega'_X$ , they are

$$P_X(\{0, 1\}) = 84/99, \quad P_X(\{0, 2\}) = 75/99, \quad P_X(\{1, 2\}) = 39/99,$$

and for  $\Omega'_X$  and  $\emptyset$ , they are  $P_X(\Omega'_X) = 1$  and  $P_X(\emptyset) = 0$ .

For the singletons  $\{u\}$ ,  $u \in \Omega_U$ , the values  $P_U(\{u\}) = P[U^{-1}(\{u\})]$  of the distribution of  $U$  are

$$P_U(\{Tom\}) = P_U(\{Jim\}) = P_U(\{Kate\}) = 1/3,$$

for the sets that consist of two elements of  $\Omega_U$  they are

$$P_U(\{Tom, Jim\}) = P_U(\{Tom, Kate\}) = P_U(\{Jim, Kate\}) = 2/3,$$

and for  $\Omega_U$  and  $\emptyset$ , they are  $P_U(\Omega_U) = 1$  and  $P_U(\emptyset) = 0$ . ◁

### Time Order Between Random Variables

In Example 5.12 and also in the examples with Joe and Ann, there is a time order between the random variables involved. Obviously, the person variable  $U$  represents events that are *prior* to the events represented by the treatment variable  $X$  and to the events represented by the outcome variable  $Y$ . In the following definition we extend the definitions introduced in section 4.2.2 to random variables.

#### Definition 5.13 (*X is Prior, Simultaneous, Posterior to Y*)

Let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$ ,  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Y, \mathcal{A}'_Y)$  be random variables and  $(\mathcal{F}_t, t \in T)$  a filtration in  $\mathcal{A}$ . Then:

- (i)  $X$  is called *prior to Y* (and  $Y$  *posterior to X*) in  $(\mathcal{F}_t, t \in T)$ , if there is an  $s \in T$  such that  $\sigma(X) \subset \mathcal{F}_s$ ,  $\sigma(Y) \not\subset \mathcal{F}_s$ , and a  $t \in T$ ,  $t > s$ , such that  $\sigma(Y) \subset \mathcal{F}_t$ .
- (ii)  $X$  is called *simultaneous to Y* in  $(\mathcal{F}_t, t \in T)$ , if there is a  $t \in T$  such that  $\sigma(X), \sigma(Y) \subset \mathcal{F}_t$  and no  $s \in T$ ,  $s < t$ , such that  $\sigma(X) \subset \mathcal{F}_s$  or  $\sigma(Y) \subset \mathcal{F}_s$ .

**Remark 5.14 (Filtration Generated by a Family of Random Variables)** Definition 5.13 does not presume that all pairs of random variables can be ordered and compared to each other with respect to a filtration  $(\mathcal{F}_t, t \in T)$ . However, given a family  $(X_i, i \in I)$  of random variables where  $I \subset \mathbb{R}$ , we can define a filtration  $(\mathcal{F}_i, i \in I)$  by  $\mathcal{F}_i := \sigma(X_j, j \leq i)$  for all  $i \in I$ . This filtration is called the *filtration generated by*  $(X_i, i \in I)$  (see Klenke, 2013, p. 191). In this filtration  $X_i$  is prior to  $X_j$  if and only if  $i < j$ , where  $i, j \in I$ . ◁

**Example 5.15 (Flipping a Coin Twice)** If we consider the random experiment of flipping the same coin twice, the probability space  $(\Omega, \mathcal{A}, P)$  is identical to the one specified in section 2.2.2 and Example 2.83, where we considered the random experiment of flipping two coins. Hence, the set of possible outcomes is

$$\Omega = \{(h, h), (h, t), (t, h), (t, t)\}.$$

The possible outcome  $(t, h)$  now represents to obtain tails in the first flip and heads in the second flip. Now, for  $i = 1, 2$ , define the random variables  $X_i: \Omega \rightarrow \{0, 1\}$  by

$$X_i[(a_1, a_2)] := \begin{cases} 1, & \text{if } a_i = h, \\ 0, & \text{if } a_i = t, \end{cases} \quad \forall (a_1, a_2) \in \Omega. \quad (5.7)$$

Hence, the value 1 of  $X_i$  indicates that the outcome of the  $i$ th flip is heads. If we define the filtration  $(\mathcal{F}_1, \mathcal{F}_2)$  by  $\mathcal{F}_1 := \sigma(X_1)$  and  $\mathcal{F}_2 := \sigma(X_1, X_2)$ , then  $X_1$  is prior to  $X_2$  in the filtration  $(\mathcal{F}_1, \mathcal{F}_2)$ . Hence, this filtration serves to introduce time order between the first and the second flip of the coin. It can be shown, e. g., that  $X_1$  is also prior to  $X_1 \cdot X_2$  in  $(\mathcal{F}_1, \mathcal{F}_2)$  (see Exercise 5-4).  $\triangleleft$

**Example 5.16 (Joe and Ann With Random Assignment – continued)** In Example 5.37, the random variable  $U$  is prior to  $X$  in the filtration  $(\mathcal{F}_t, t \in \{1, 2, 3\})$  specified in Example 4.19, because  $\sigma(U) \subset \mathcal{F}_1$ ,  $\sigma(X) \not\subset \mathcal{F}_1$ , and  $\sigma(X) \subset \mathcal{F}_2$ . Analogously it can be shown that, in this example,  $X$  is prior to  $Y$ .  $\triangleleft$

## 5.2 Equivalence of Two Random Variables With Respect to a Probability Measure

### 5.2.1 Identical and $P$ -Equivalent Random Variables

Let  $X, Y: (\Omega, \mathcal{A}, P) \rightarrow (\Omega', \mathcal{A}')$  be two random variables. Then  $X$  and  $Y$  are called *identical* if

$$\forall \omega \in \Omega: X(\omega) = Y(\omega). \quad (5.8)$$

**Remark 5.17 ( $P$ -Equivalent Random Variables)** Let  $X, Y: (\Omega, \mathcal{A}, P) \rightarrow (\Omega', \mathcal{A}')$  be two random variables. Then  $X$  and  $Y$  are *almost surely identical* with respect to  $P$  or  *$P$ -equivalent*, denoted  $X \stackrel{P}{=} Y$ , if

$$\exists A \in \mathcal{A}: (\forall \omega \in \Omega \setminus A: X(\omega) = Y(\omega) \quad \text{and} \quad P(A) = 0) \quad (5.9)$$

(see Def. 2.68). Another notation for  $X \stackrel{P}{=} Y$  is  $X(\omega) \stackrel{P\text{-a.a.}}{=} Y(\omega)$ , which is a shortcut for

$$X(\omega) = Y(\omega), \quad \text{for } P\text{-a.a. } \omega \in \Omega, \quad (5.10)$$

meaning that the values of  $X$  and  $Y$  are identical for  $P$ -almost all  $\omega \in \Omega$  (see Rem. 2.70).  $\triangleleft$

**Remark 5.18 (Singleton With a Positive Probability)** If  $X \stackrel{P}{=} Y$  or, equivalently, if  $X(\omega) \stackrel{P-a.a.}{=} Y(\omega)$ , and  $\{\omega^*\} \in \mathcal{A}$ , with  $P(\{\omega^*\}) > 0$ , then  $X(\omega^*) = Y(\omega^*)$  [see Rem. 2.71].  $\triangleleft$

**Example 5.19 (Indicator of a Null Set)** Let  $(\Omega, \mathcal{A}, P)$  be a probability space and  $A \in \mathcal{A}$ . If  $P(A) = 0$ , then

$$1_A \stackrel{P}{=} 0 \quad \text{and} \quad 1_{A^c} = 1 - 1_A \stackrel{P}{=} 1 \quad (5.11)$$

(see Example 5.8).  $\triangleleft$

**Remark 5.20 (Q-Equivalence)** Note that the definition of equivalence of two random variables  $X$  and  $Y$  with respect to a probability measure only presumes that  $X$  and  $Y$  are measurable with respect to a  $\sigma$ -algebra on  $\Omega$  and that the measure considered is a probability measure on this  $\sigma$ -algebra. Hence, we can consider the equivalence of  $X$  and  $Y$  with respect to different probability measures, say  $P$  and  $Q$  and study their relationship.  $\triangleleft$

In the following lemma we consider the relationship between  $P$ -equivalence and  $Q$ -equivalence, presuming  $Q \ll_{\mathcal{A}} P$  (absolute continuity), i. e., presuming

$$\forall C \in \mathcal{C}: P(C) = 0 \Rightarrow Q(C) = 0$$

(see Def. 3.70).

**Lemma 5.21 ( $P$ -Equivalence and  $Q$ -Equivalence)**

Let  $X, Y: (\Omega, \mathcal{A}, P) \rightarrow (\Omega', \mathcal{A}')$  be random variables. If  $Q$  is a probability measure on  $(\Omega, \mathcal{A})$  such that  $Q \ll_{\mathcal{A}} P$ , then  $X \stackrel{P}{=} Y$  implies  $X \stackrel{Q}{=} Y$ .

(Proof p. 188)

According to Lemma 4.32 (i),  $P^B \ll_{\mathcal{A}} P$ , provided that  $B \in \mathcal{A}$  is an event for which  $P(B) > 0$ . Hence, Lemma 5.21 immediately implies the following corollary:

**Corollary 5.22 ( $P$ -Equivalence Implies  $P^B$ -Equivalence)**

Let  $X, Y: (\Omega, \mathcal{A}, P) \rightarrow (\Omega', \mathcal{A}')$  be two random variables and let  $B \in \mathcal{A}$  with  $P(B) > 0$ . Then  $X \stackrel{P}{=} Y$  implies  $X \stackrel{P^B}{=} Y$ .

**Example 5.23 ( $P^B$ -Equivalence Does not Imply  $P$ -Equivalence)** Consider the set  $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$  with the  $\sigma$ -algebra  $\mathcal{A} = \mathcal{P}(\Omega)$ , and the set  $\Omega' = \{a, b, c, d\}$ , with the  $\sigma$ -algebra  $\mathcal{A}' = \mathcal{P}(\Omega')$ . Furthermore, let  $P: \mathcal{A} \rightarrow [0, 1]$  satisfy  $P(\{\omega_1\}) = .25$ ,  $P(\{\omega_2\}) = .25$ ,  $P(\{\omega_3\}) = 0$ , and  $P(\{\omega_4\}) = .50$ . Finally, define  $X, Y: (\Omega, \mathcal{A}, P) \rightarrow (\Omega', \mathcal{A}')$  by

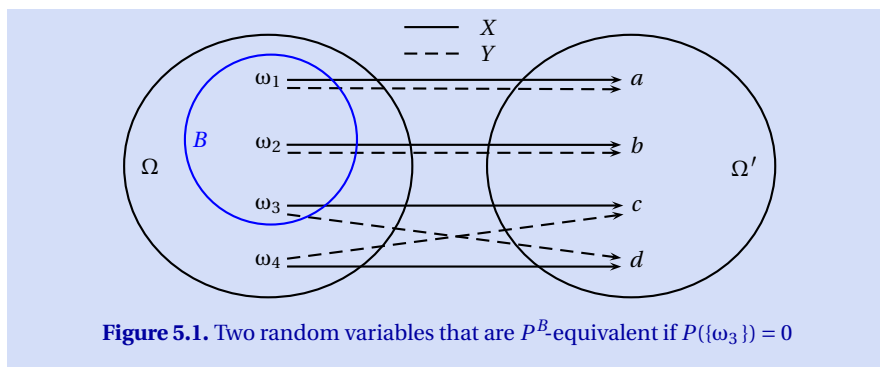


Figure 5.1. Two random variables that are  $P^B$ -equivalent if  $P(\{\omega_3\}) = 0$

$$X(\omega) = \begin{cases} a, & \text{if } \omega = \omega_1 \\ b, & \text{if } \omega = \omega_2 \\ c, & \text{if } \omega = \omega_3 \\ d, & \text{if } \omega = \omega_4 \end{cases} \quad Y(\omega) = \begin{cases} a, & \text{if } \omega = \omega_1 \\ b, & \text{if } \omega = \omega_2 \\ d, & \text{if } \omega = \omega_3 \\ c, & \text{if } \omega = \omega_4 \end{cases}$$

(see Fig. 5.1). If  $B = \{\omega_1, \omega_2, \omega_3\}$ , then  $X \stackrel{P^B}{=} Y$ , but neither  $X=Y$  nor  $X \stackrel{P}{=} Y$  (see Exercise 5-5). Therefore, equivalence with respect to  $P^B$  does not imply equivalence with respect to  $P$ .  $\triangleleft$

Theorem 2.85 on the equivalence of image measures immediately implies the following corollary on the equivalence of the distributions of two  $P$ -equivalent random variables:

**Corollary 5.24 ( $P$ -Equivalence Implies Equal Distributions)**

Let  $X, Y: (\Omega, \mathcal{A}, P) \rightarrow (\Omega', \mathcal{A}')$  be random variables with distributions  $P_X$  and  $P_Y$ , respectively. If  $X \stackrel{P}{=} Y$ , then  $P_X = P_Y$ .

In other words, if  $X$  and  $Y$  are  $P$ -equivalent, then they are identically distributed. Note, however, that identical distributions of  $X$  and  $Y$  do not imply that  $X$  and  $Y$  are  $P$ -equivalent.

In chapter 6 we shall see that Corollary 5.24 implies that also the expectations, variances, and other moments of  $X$  and  $Y$  are identical if  $X, Y: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  are  $P$ -equivalent numerical random variables, provided that the expectations, variances, and other moments of  $X$  and  $Y$  exist.

The following corollary is an immediate implication of Theorem 2.86.

**Corollary 5.25 ( $P$ -Equivalence of Compositions)**

Let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  be a random variable with distribution  $P_X$  and let  $g, g^*: (\Omega'_X, \mathcal{A}'_X) \rightarrow (\mathbb{R}, \mathcal{B})$  be measurable functions. Then:

$$(i) \quad g(X) \stackrel{P}{=} g^*(X) \Leftrightarrow g \stackrel{P_X}{=} g^*.$$

$$(ii) \quad g(X) \stackrel{P}{<} g^*(X) \Leftrightarrow g \stackrel{P_X}{<} g^*.$$

$$(iii) \quad g(X) \stackrel{P}{\leq} g^*(X) \Leftrightarrow g \stackrel{P_X}{\leq} g^*.$$

(Proof p. 188)

**Remark 5.26 (Alternative Notations)** Note that

$$g \stackrel{P_X}{=} g^* \Leftrightarrow g(x) = g^*(x), \quad \text{for } P_X\text{-a.a. } x \in \Omega'_X, \quad (5.12)$$

$$g \stackrel{P_X}{<} g^* \Leftrightarrow g(x) < g^*(x), \quad \text{for } P_X\text{-a.a. } x \in \Omega'_X, \quad (5.13)$$

$$g \stackrel{P_X}{\leq} g^* \Leftrightarrow g(x) \leq g^*(x), \quad \text{for } P_X\text{-a.a. } x \in \Omega'_X. \quad (5.14)$$

&lt;

### 5.2.2 $P$ -Equivalence, $P^B$ -Equivalence, and Absolute Continuity

Now we consider the relationship between equivalence of two random variables  $X, Y: (\Omega, \mathcal{A}, P) \rightarrow (\Omega', \mathcal{A}')$  with respect to  $P$  and  $P^B$ , and absolute continuity of  $P_X$  with respect to  $P_X^B$ , the distribution of  $X$  with respect to the conditional-probability measure  $P^B$ . Remember, for  $B \in \mathcal{A}$  and  $P(B) > 0$ , we defined the  $B$ -conditional probability measure  $P^B$  (see Def. 4.29). Referring to such a measure,  $X \stackrel{P^B}{=} Y$  means

$$\exists A \in \mathcal{A}: \left( \forall \omega \in \Omega \setminus A: X(\omega) = Y(\omega) \quad \text{and} \quad P^B(A) = 0 \right), \quad (5.15)$$

[see (5.9)]. If  $B$  denotes the event  $\{X=x\} := \{\omega \in \Omega: X(\omega)=x\}$ , then we define  $P^{X=x} := P^B$  and call it the  $(X=x)$ -conditional probability measure on  $(\Omega, \mathcal{A})$ .

#### Lemma 5.27 (An Implication of Absolute Continuity)

Let  $X, Y: (\Omega, \mathcal{A}, P) \rightarrow (\Omega', \mathcal{A}')$  be random variables that are measurable with respect to the  $\sigma$ -algebra  $\mathcal{C} \subset \mathcal{A}$ , and assume  $\{X \neq Y\} \in \mathcal{C}$ . Furthermore, let  $B \in \mathcal{A}$  with  $P(B) > 0$ . If  $X \stackrel{P^B}{=} Y$  and  $P \ll_{\mathcal{C}} P^B$ , then  $X \stackrel{P}{=} Y$ .

(Proof p. 189)

**Example 5.28 (No Treatment For Joe)** Consider Table 5.2. In this example we define the set  $\Omega_U = \{Joe, Jim, Ann\}$  and

$$\mathcal{A}_U = \mathcal{P}(\Omega_U) = \{\{Joe\}, \{Jim\}, \{Ann\}, \{Joe, Jim\}, \{Joe, Ann\}, \{Jim, Ann\}, \Omega_U, \emptyset\}.$$

Using these sets, not only  $U: (\Omega, \mathcal{A}, P) \rightarrow (\Omega_U, \mathcal{A}_U)$  is a random variable, but also  $U^*: (\Omega, \mathcal{A}, P) \rightarrow (\Omega_U, \mathcal{A}_U)$  defined in Table 5.2. Now the distribution of  $U$  is specified by  $P_U(\{Joe\}) = .5$ ,  $P_U(\{Jim\}) = 0$ , and  $P_U(\{Ann\}) = .5$ . The probabilities of the

other five elements of  $\mathcal{A}_U$  are obtained using Rule (x) of Box 4.1. Furthermore,  $P_{U^*} = P_U$ .

Considering the measure  $P^{X=0}$  we find  $U \underset{P^{X=0}}{=} U^*$ , because

$$P^{X=0}(\{U \neq U^*\}) = P^{X=0}(\{(Joe, yes, +), (Joe, yes, -)\}) = 0.$$

Furthermore, there are only two sets  $A \in \sigma(U)$  with  $P^{X=0}(A) = 0$ . These are the sets  $U^{-1}(\{Jim\}) = \{(Joe, yes, +), (Joe, yes, -)\}$  and  $\emptyset$ , and for these sets we find  $P(U^{-1}(\{Jim\})) = P(\emptyset) = 0$ . Hence,  $P \ll_{\sigma(U)} P^{X=0}$ , and according to Lemma 5.27 this implies  $U \underset{P}{=} U^*$ . In fact, we find

$$P(\{U \neq U^*\}) = P(\{(Joe, yes, +), (Joe, yes, -)\}) = 0.$$

◁

### Lemma 5.29 (Absolute Continuity)

Let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  be a random variable and  $B \in \mathcal{A}$  with  $P(B) > 0$ . Then

$$P \ll_{\sigma(X)} P^B \Leftrightarrow P_X \ll_{\mathcal{A}'_X} P_X^B. \quad (5.16)$$

(Proof p. 189)

**Example 5.30 (No Treatment For Joe – continued)** In Example 5.28, we already found

$$P \ll_{\sigma(U)} P^{X=0}.$$

There are only two sets  $A' \in \mathcal{A}_U$  with  $P_U^{X=0}(A') = 0$ , the sets  $\{Jim\}$  and  $\emptyset$ , and for these sets we find  $P_U(\{Jim\}) = P_U(\emptyset) = 0$ . Hence, in this example,

$$P_U \ll_{\mathcal{A}_U} P_U^{X=0}$$

holds as well.

◁

Lemmas 5.29 and 4.32 immediately imply the following corollary.

### Corollary 5.31 (Null-Set Equivalence)

Let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  be a random variable and  $B \in \mathcal{A}$  with  $P(B) > 0$ . Then  $P$  and  $P^B$  are null-set equivalent on  $(\Omega, \sigma(X))$  if and only if  $P_X$  and  $P_X^B$  are null-set equivalent on  $(\Omega'_X, \mathcal{A}'_X)$ .

According to Lemma 4.32, absolute continuity of  $P_X^B$  with respect to  $P_X$  always holds. In other words,  $P_X^B \ll_{\mathcal{A}'_X} P_X$ , which is equivalent to

$$\forall A' \in \mathcal{A}'_X: P_X(A') = 0 \Rightarrow P_X^B(A') = 0, \quad (5.17)$$

always holds. In contrast,  $P_X \ll_{\mathcal{A}'_X} P_X^B$  is *not necessarily true*.

**Table 5.2.** No Treatment for Joe

Outcomes $\omega$			Observables						
Unit	Treatment	Success							
			$P(\{\omega\})$	Person variable $U$	Person variable $U^*$	Treatment variable $X$	Outcome variable $Y$		
							$P^{X=0}(\{\omega\})$ (rounded)	$P^{X=1}(\{\omega\})$	
(Joe, no, -)			.152	Joe	Joe	0	0	.245	0
(Joe, no, +)			.348	Joe	Joe	0	1	.561	0
(Joe, yes, -)			0	Joe	Jim	1	0	0	0
(Joe, yes, +)			0	Joe	Jim	1	1	0	0
(Ann, no, -)			.096	Ann	Ann	0	0	.155	0
(Ann, no, +)			.024	Ann	Ann	0	1	.039	0
(Ann, yes, -)			.228	Ann	Ann	1	0	0	.60
(Ann, yes, +)			.152	Ann	Ann	1	1	0	.40

**Example 5.32 (No Treatment For Joe – continued)** Table 5.2 displays an example illustrating absolute continuity of  $P_U$  with respect to  $P_U^B$  for a discrete random variable  $U$ . Consider the event  $B = \{X=1\} = \{\omega \in \Omega: X(\omega) = 1\}$ . Using this notation,  $P_U$  is not absolutely continuous with respect to  $P_U^{X=1}$ , i. e.,  $P_U \ll_{\mathcal{A}_U} P_U^{X=1}$  does *not hold*. In contrast,  $P_U \ll_{\mathcal{A}_U} P_U^{X=0}$  does *hold*.

In this example, the eight elements of  $\Omega$  are listed in the first column of the table. Furthermore, we choose  $\mathcal{A} = \mathcal{P}(\Omega)$  and the probability measure on  $(\Omega, \mathcal{A})$  is specified by the probabilities of the singletons  $\{\omega\}$  specified in the second column of the table [see Box 4.1 (x)]. The random variables  $U: (\Omega, \mathcal{A}, P) \rightarrow (\Omega_U, \mathcal{A}_U)$ , with  $\Omega_U = \{Joe, Ann\}$ ,  $\mathcal{A}_U = \mathcal{P}(\Omega_U)$ , and  $X, Y: (\Omega, \mathcal{A}, P) \rightarrow (\Omega', \mathcal{P}(\Omega'))$  with  $\Omega' = \{0, 1\}$ , are specified in columns 3, 5, and 6 of Table 5.2. (The random variable  $U^*$  has been used in Example 5.28.) Note that the distribution of  $U$  is:

$$\begin{aligned}
 P_U(\{Joe\}) &= P(\{(Joe, no, -), (Joe, no, +), (Joe, yes, +), (Joe, yes, -)\}) = .5, \\
 P_U(\{Ann\}) &= P(\{(Ann, no, -), (Ann, no, +), (Ann, yes, +), (Ann, yes, -)\}) = .5, \\
 P_U(\Omega_U) &= 1, \quad P_U(\emptyset) = 0.
 \end{aligned}$$

Now, we compute the  $(X=1)$ -conditional probabilities of the elementary events:

$$\begin{aligned}
 P^{X=1}(\{\omega_1\}) &= P^{X=1}(\{(Joe, no, -)\}) = \frac{P(\{(Joe, no, -)\} \cap \{X=1\})}{P(X=1)} \\
 &= 0 / (.228 + .152) = 0,
 \end{aligned}$$

and the same result is obtained for  $\omega_2$  to  $\omega_6$ . In contrast,

$$\begin{aligned} P^{X=1}(\{\omega_7\}) &= P^{X=1}(\{(Ann, yes, -)\}) = \frac{P(\{(Ann, yes, -)\} \cap \{X=1\})}{P(X=1)} \\ &= .228/ (.228 + .152) = .60, \end{aligned}$$

and

$$\begin{aligned} P^{X=1}(\{\omega_8\}) &= P^{X=1}(\{(Ann, yes, +)\}) = \frac{P(\{(Ann, yes, +)\} \cap \{X=1\})}{P(X=1)} \\ &= .152/ (.228 + .152) = .40. \end{aligned}$$

These results are displayed in the last column of Table 5.2 and the last but one column shows the probabilities  $P^{X=0}(\{\omega\})$  of the singletons with respect to  $P^{X=0}$ .

Now consider the element  $\{Joe\} \in \mathcal{A}_U$ . Inspecting the last and the second columns of Table 5.2 shows that

$$P_U^{X=1}(\{Joe\}) = 0 \quad \text{and} \quad P_U(\{Joe\}) = .5.$$

According to Definition 3.70 (i), this implies that  $P_U \ll_{\mathcal{A}_U} P_U^{X=1}$  does *not hold*. In contrast, none of the four elements  $A' \in \mathcal{A}_U$  satisfies

$$P_U^{X=0}(A') = 0 \quad \text{and} \quad P_U(A') \neq 0.$$

Therefore, in this example,  $P_U \ll_{\mathcal{A}_U} P_U^{X=0}$  *does hold*. ◁

### 5.3 Multivariate Random Variable

Univariate random variables take on their values in sets such as  $\Omega' \subset \bar{\mathbb{R}}$ ,  $\Omega' = \{male, female\}$ , or  $\Omega' = \{low, medium, high\}$ , whereas bivariate random variables take on their values in sets such as  $\Omega' \subset \bar{\mathbb{R}}^2$  or

$$\Omega' = \{male, female\} \times \{low, medium, high\}.$$

The values of bivariate random variables are pairs such as (5,8) or (male, low). The values of  $n$ -variate random variables are  $n$ -tuples. If  $X$  takes on values such as *male* or (male, low), then we call  $X$  *qualitative*. If  $X$  takes on values in a subset of  $\mathbb{R}^n$ ,  $n \in \mathbb{N}$ , we call it  *$n$ -variate real-valued*. If  $X$  takes on values in a subset of  $\bar{\mathbb{R}}^n$ ,  $n \in \mathbb{N}$ , we call it  *$n$ -variate numerical*.

**Remark 5.33 (Joint and Marginal Distributions)** Definition 5.1 also applies to an  $n$ -variate random variable  $X$ , i. e., to a random variable

$$X = (X_1, \dots, X_n): (\Omega, \mathcal{A}, P) \rightarrow \left( \prod_{i=1}^n \Omega'_i, \bigotimes_{i=1}^n \mathcal{A}'_i \right) \quad (5.18)$$

that consists of  $n$  univariate random variables  $X_i: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_i, \mathcal{A}'_i)$ . Hence,

$$X(\omega) = [X_1(\omega), \dots, X_n(\omega)], \quad \forall \omega \in \Omega. \quad (5.19)$$

The distribution  $P_X = P_{X_1, \dots, X_n}$  of  $X$  is also called the *joint distribution* of the random variables  $X_i, i = 1, \dots, n$ .

Because  $\pi_i(X_1, \dots, X_n) = X_i$ ,

$$P_{X_i} = P_{\pi_i(X_1, \dots, X_n)}, \quad i = 1, \dots, n, \quad (5.20)$$

[see Eq. (2.20) defining the projection  $\pi_i$ ]. In this context,  $P_{X_i}$  is called the (one-dimensional) *marginal distribution of  $X_i$* . Equation (5.20) shows that the joint distribution uniquely determines all marginal distributions, but not vice versa! More specifically, for  $i = 1, \dots, n$ ,

$$P_{X_i}(A'_i) = P_{X_1, \dots, X_n}(\Omega'_1 \times \dots \times \Omega'_{i-1} \times A'_i \times \Omega'_{i+1} \times \dots \times \Omega'_n), \quad \forall A'_i \in \mathcal{A}'_i. \quad (5.21)$$

Analogously, we may also describe the marginal distribution of  $(X_{i_1}, \dots, X_{i_m})$ , where  $\{i_1, \dots, i_m\} \subset \{1, \dots, n\}$ .  $\triangleleft$

**Remark 5.34 (Joint Distribution vs. Other Quantities)** The joint distribution of a multivariate random variable contains the essential information about the random variables  $X_1, \dots, X_n$ . All other quantities such as expectations  $E(X_i)$ , variances  $\text{Var}(X_i)$ , covariances  $\text{Cov}(X_i, X_j)$ , or conditional expectations such as  $E(X_1 | X_2, \dots, X_n)$ , which are introduced in succeeding chapters, are determined by the joint distribution, and usually they contain less information. Nevertheless, these other quantities often reveal certain properties of a multivariate random variable more clearly than the joint distribution.  $\triangleleft$

**Example 5.35 (Flipping two Coins – continued)** In Examples 2.2.2 and 2.83, we considered the random experiment of flipping two coins and defined a random variable  $X$  representing with its values the number of flipping heads. Additional to  $X$  we may also define the random variables  $X_1, X_2: (\Omega, \mathcal{A}, P) \rightarrow (\{0, 1\}, \mathcal{P}(\{0, 1\}))$  by

$$X_1(\omega) = \begin{cases} 1, & \text{if } \omega \in \{(h, t), (h, h)\} \\ 0, & \text{if } \omega \in \{(t, h), (t, t)\} \end{cases} \quad (5.22)$$

and

$$X_2(\omega) = \begin{cases} 1, & \text{if } \omega \in \{(t, h), (h, h)\} \\ 0, & \text{if } \omega \in \{(h, t), (t, t)\}. \end{cases} \quad (5.23)$$

They indicate with their value 1 if *heads* are flipped at the first and second flip, respectively. Obviously,  $X = X_1 + X_2$ . Furthermore,

$$(X_1, X_2): (\Omega, \mathcal{A}, P) \rightarrow (\{0, 1\} \times \{0, 1\}, \mathcal{P}(\{0, 1\}) \otimes \mathcal{P}(\{0, 1\}))$$

is a two-dimensional random variable with values  $(0, 0)$ ,  $(0, 1)$ ,  $(1, 0)$ , and  $(1, 1)$ . The joint distribution  $P_{X_1, X_2}$  is uniquely defined by

$$P_{X_1, X_2}(\{(i, j)\}) = \frac{1}{4}, \quad \forall i, j = 0, 1.$$

The marginal distribution of  $X_1$  is

$$P_{X_1}(\{i\}) = P_{X_1, X_2}(\{(i, 0)\}) + P_{X_1, X_2}(\{(i, 1)\}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}, \quad i = 0, 1,$$

$P_{X_1}(\{0, 1\}) = 1$ , and  $P_{X_1}(\emptyset) = 0$ . Obviously,  $P_{X_1}$  is completely determined by the joint distribution  $P_{X_1, X_2}$ , and the same applies to the marginal distribution  $P_{X_2}$ .  $\triangleleft$

**Example 5.36 (Tom, Jim, and Kate – continued)** The second column of Table 5.1 also displays the probabilities  $P_{U, X, Y}(\{(u, x, y)\}) = P(\{\omega\})$  of the three-dimensional random variable  $(U, X, Y)$  that maps the elements  $\omega \in \Omega$  onto the set

$$\Omega' := \{Tom, Jim, Kate\} \times \{0, 1, 2\} \times \{0, 1\}$$

on which we consider the  $\sigma$ -algebra

$$\mathcal{A}' := \mathcal{P}(\Omega') = \mathcal{P}(\{Tom, Jim, Kate\}) \otimes \mathcal{P}(\{0, 1, 2\}) \otimes \mathcal{P}(\{0, 1\}).$$

The probabilities  $P_{U, X, Y}(\{(u, x, y)\})$ ,  $(u, x, y) \in \Omega'$ , uniquely determine the joint distribution  $P_{U, X, Y}$  as well as the one-dimensional marginal distributions  $P_U$ ,  $P_X$ ,  $P_Y$ , and the two-dimensional marginal distributions  $P_{U, X}$ ,  $P_{U, Y}$ , and  $P_{X, Y}$ .  $\triangleleft$

For another example of a joint distribution, which refers to Example 2.34 (p. 53) see Exercise 5-6.

## 5.4 Independence of Random Variables

The concepts of independence of events and of set systems, i. e., of sets of events, which have been introduced in Definition 4.40, can be used to define *stochastic independence of random variables*. Remember that

$$\sigma(X) := X^{-1}(\mathcal{A}') := \{X^{-1}(A') : A' \in \mathcal{A}'\}$$

is a  $\sigma$ -algebra on  $\Omega$ , called the  $\sigma$ -algebra generated by  $X$  (see Def. 2.26). Hence, we can define the random variables  $X_1: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_1, \mathcal{A}'_1)$  and  $X_2: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_2, \mathcal{A}'_2)$  to be  $P$ -independent if  $X_1^{-1}(\mathcal{A}'_1)$  and  $X_2^{-1}(\mathcal{A}'_2)$  are  $P$ -independent. In other words,  $X_1$  and  $X_2$  are defined to be  $P$ -independent, if and only if

$$P(A \cap B) = P(A) \cdot P(B), \quad \forall (A, B) \in X_1^{-1}(\mathcal{A}'_1) \times X_2^{-1}(\mathcal{A}'_2). \quad (5.24)$$

Using the notation introduced in Remark 5.4 and

$$P(X_1 \in A', X_2 \in B') := P(\{X_1 \in A'\} \cap \{X_2 \in B'\}), \quad (5.25)$$

this equation is equivalent to

$$P(X_1 \in A', X_2 \in B') = P(X_1 \in A') \cdot P(X_2 \in B'), \quad \forall (A', B') \in \mathcal{A}'_1 \times \mathcal{A}'_2. \quad (5.26)$$

Independence of the random variables  $X_1$  and  $X_2$  with respect to  $P$  is denoted by  $X_1 \perp\!\!\!\perp_P X_2$ .

**Example 5.37 (Joe and Ann With Random Assignment – continued)** In Table 2.2 we presented the random experiment of drawing a person from a set of persons,  $\Omega_U = \{Joe, Ann\}$ , performing a randomized assignment of the drawn person to one of two treatment conditions represented by the elements of the set  $\Omega_X = \{yes, no\}$ , and observing success or failure, represented by the elements of the set  $\Omega_Y = \{-, +\}$ . Hence, the set of possible outcomes of this random experiment is

$$\Omega = \Omega_U \times \Omega_X \times \Omega_Y,$$

which consists of the eight triples listed in the first column of Table 2.2. In that table we defined the three random variables

$$U: (\Omega, \mathcal{P}(\Omega)) \rightarrow (\Omega_U, \mathcal{P}(\Omega_U)) \quad \text{and} \quad X, Y: (\Omega, \mathcal{P}(\Omega)) \rightarrow (\Omega', \mathcal{A}'),$$

where  $\Omega' = \{0, 1\}$  and  $\mathcal{A}' = \{\{0\}, \{1\}, \Omega', \emptyset\}$ . In order to check if Equation (5.26) actually holds, we choose the two sets  $\{Joe\} \in \mathcal{P}(\Omega_U)$  and  $\{0\} \in \mathcal{A}'$  and compare the probability

$$P(X=0, U=Joe) = P(\{(Joe, no, -), (Joe, no, +)\}) = .3$$

(see the first two rows of Table 2.2) to the product of the two probabilities

$$P(X=0) = P(\{(Joe, no, -), (Joe, no, +), (Ann, no, -), (Ann, no, +)\}) = .6$$

and

$$P(U=Joe) = P(\{(Joe, no, -), (Joe, no, +), (Joe, no, -), (Joe, no, +)\}) = .5.$$

Obviously, Equation (5.26) holds for the pair  $(\{0\}, \{Joe\}) \in \mathcal{A}' \times \mathcal{P}(\Omega_U)$ . Repeating the corresponding comparisons for all pairs of  $\mathcal{A}' \times \mathcal{P}(\Omega_U)$  shows that Equation (5.26) actually holds in this example (see also Exercise 5-7).  $\triangleleft$

**Remark 5.38 (A Methodological Note on Random Assignment)** In random experiments such as the one presented in Example 5.37, with a randomized assignment of the drawn person to one of several treatment conditions, we create independence of  $X$  and the person variable  $U$ . According to Equation 5.26, this implies that we create independence of  $X$  and all  $U$ -measurable mappings  $f(U)$ , because  $\sigma[f(U)] \subset \sigma(U)$  (see Cor. 2.53). More generally, randomized assignment of an observational unit (such as a person) creates independence of  $X$  and all pretreatment variables.  $\triangleleft$

Using definition 4.40, the following definition extends the concept of independence of two random variables to a family of random variables. This includes a finite sequence of random variables  $X_i, i \in I := \{1, \dots, n\}$ , an infinite sequence of random variables  $X_i, i \in I := \{1, 2, \dots\}$ , and a family  $(X_i, i \in I)$  of random variables in which the index set  $I$  may be *any* set, including, e. g.,  $I \subset \mathbb{R}$ .

**Definition 5.39 (Family of Independent Random Variables)**

A family  $(X_i, i \in I)$  of random variables  $X_i: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_i, \mathcal{A}'_i)$  is called a family of  $P$ -independent random variables, denoted  $\perp\!\!\!\perp_P (X_i, i \in I)$ , if  $(X_i^{-1}(\mathcal{A}'_i), i \in I)$  is a family of  $P$ -independent  $\sigma$ -algebras.

**Remark 5.40 (Independence of Three Random Variables)** Hence, three random variables  $X_1, X_2, X_3$  are *independent*, denoted  $\perp\!\!\!\perp_P X_1, X_2, X_3$ , if and only if

$$\begin{aligned} P(X_1 \in A', X_2 \in B', X_3 \in C') &= P(X_1 \in A') \cdot P(X_2 \in B') \cdot P(X_3 \in C'), \\ \forall (A', B', C') \in \mathcal{A}'_1 \times \mathcal{A}'_2 \times \mathcal{A}'_3 \end{aligned} \quad (5.27)$$

(see Rem. 4.38). Note that pairwise independence of  $X_1, X_2, X_3$  follows from choosing  $A' = \Omega'_1$ ,  $B' = \Omega'_2$ , or  $C' = \Omega'_3$ , respectively.  $\triangleleft$

**Remark 5.41 (Independence of  $n$  Random Variables)** Correspondingly, the random variables  $X_1, \dots, X_n$  are *independent*, denoted  $\perp\!\!\!\perp_P X_1, \dots, X_n$ , if and only if

$$\begin{aligned} P(X_1 \in A'_1, \dots, X_n \in A'_n) &= P(X_1 \in A'_1) \cdot \dots \cdot P(X_n \in A'_n), \\ \forall (A'_1, \dots, A'_n) \in \mathcal{A}'_1 \times \dots \times \mathcal{A}'_n. \end{aligned} \quad (5.28)$$

$\triangleleft$

**Remark 5.42 (Sample)** Oftentimes, we assume that  $X_1, \dots, X_n$  is a sequence of independent and identically distributed (abbreviated i. i. d.) random variables (see, e. g., chs. 6 and 8). In statistics, a sequence  $X_1, \dots, X_n$  of i. i. d. random variables is called a *random sample*.

An important example of i. i. d. random variables is treated in the section on Bernoulli trials (see section 8.1.2).  $\triangleleft$

**Remark 5.43 (Independence With Respect to a Probability Measure)** If there is no ambiguity we also use the term *independence* of events, sets of events, random variables, and sets of random variables. Note that, if  $Q$  is another probability measure on  $(\Omega, \mathcal{A})$ , then events, sets of events, and random variables can be  $P$ -independent although they are not  $Q$ -independent.  $\triangleleft$

**Remark 5.44 (A Random Variable and a Set System)** Independence of a set system and a random variable is defined in the same way. A set system  $\mathcal{E} \subset \mathcal{A}$  and a random variable  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  are called *independent*, denoted  $\mathcal{E} \perp\!\!\!\perp_P X$ , if  $\mathcal{E}$  and  $\sigma(X)$  are independent. Of course,  $\mathcal{E}$  can also be a sub- $\sigma$ -algebra of  $\mathcal{A}$ .  $\triangleleft$

**Remark 5.45 (A Random Variable and a Family of Random Variables)** Independence of a random variable  $X$  and a family  $(Y_i, i \in I)$  of random variables, denoted  $X \perp\!\!\!\perp_P (Y_i, i \in I)$ , is defined by  $X \perp\!\!\!\perp_P \sigma(Y_i, i \in I)$  [see Eq. (2.18)]. Note that  $X \perp\!\!\!\perp_P (Y_i, i \in I)$  implies  $X \perp\!\!\!\perp_P \sigma(Y_i)$ , for all  $i \in I$ .  $\triangleleft$

**Remark 5.46 (Equivalent Propositions)** Let  $(\Omega, \mathcal{A}, P)$  be a probability space,  $B \in \mathcal{A}$ , and  $\mathcal{C} \subset \mathcal{A}$ . Then the following propositions are equivalent to each other:

$$1_B \perp\!\!\!\perp_P \mathcal{C}, \quad \sigma(\{B\}) \perp\!\!\!\perp_P \mathcal{C}, \quad \{B\} \perp\!\!\!\perp_P \mathcal{C}, \quad B \perp\!\!\!\perp_P \mathcal{C}$$

(see Rem. 4.41 and Exercise 5-8). ◁

In Corollary 5.24 we noted that  $P$ -equivalent random variables have identical distributions. According to the following lemma this also has implications for independence of random variables.

**Lemma 5.47 ( $P$ -Equivalence and Independence)**

Let  $X_i: (\Omega, \mathcal{A}, P) \rightarrow (\Omega', \mathcal{A}')$ ,  $i = 1, 2$ , and  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\Omega_Y', \mathcal{A}_Y')$  be random variables. Then

$$(X_1 \stackrel{P}{\sim} X_2 \wedge X_1 \perp\!\!\!\perp_P Y) \Rightarrow X_2 \perp\!\!\!\perp_P Y. \quad (5.29)$$

(Proof p. 189)

Now we consider the probability measure  $P$  and the  $B$ -conditional-probability measure  $P^B$  on  $(\Omega, \mathcal{A})$  (see Def. 4.29). In Lemma 4.32 we have shown that  $P^B$  is absolutely continuous (see Def. 3.70) with respect to  $P$  on  $(\Omega, \mathcal{A})$ . In the following lemma we show that  $P$  is absolutely continuous with respect to  $P^B$  on  $(\Omega, \mathcal{C})$ ,  $\mathcal{C} \subset \mathcal{A}$ , provided that  $B$  and  $\mathcal{C}$  are independent.

**Lemma 5.48 (Independence and Absolute Continuity)**

Let  $(\Omega, \mathcal{A}, P)$  be a probability space, let  $\mathcal{C} \subset \mathcal{A}$  be a  $\sigma$ -algebra, and  $B \in \mathcal{A}$  with  $P(B) > 0$ . Then  $1_B \perp\!\!\!\perp_P \mathcal{C}$  implies  $P \ll_{\mathcal{C}} P^B$ .

(Proof p. 189)

In the following lemma,  $P_{X_1} \otimes \dots \otimes P_{X_n}$  denotes the product measure of the marginal distributions (see Def. 1.67 and Rem. 5.33).

**Lemma 5.49 (Independence and Product Measure)**

Let  $X = (X_1, \dots, X_n)$  be an  $n$ -variate random variable as specified in (5.18). Then

$$\perp\!\!\!\perp_P X_1, \dots, X_n \Leftrightarrow P_{X_1, \dots, X_n} = P_{X_1} \otimes \dots \otimes P_{X_n}. \quad (5.30)$$

(Proof p. 190)

**Example 5.50 (Tom, Jim, and Kate – continued)** In example 5.12 we considered the random variables  $X$  and  $U$ , which have been constructed such that they are

independent. All  $8 \cdot 8 = 64$  pairs  $(A, B)$  of elements  $A \in X^{-1}(\mathcal{A}'_X)$  and  $B \in U^{-1}(\mathcal{A}'_U)$  satisfy  $P(A \cap B) = P(A) \cdot P(B)$ . Let us consider, e. g.,  $A_1 := X^{-1}(\{0\})$  and  $B_1 := U^{-1}(\{Tom\})$ ,  $B_2 := U^{-1}(\{Jim\})$ , and  $B_3 := U^{-1}(\{Kate\})$ . Then

$$P(A_1 \cap B_j) = \frac{20}{99}, \quad j = 1, 2, 3$$

and

$$P(A_1) \cdot P(B_j) = \frac{60}{99} \cdot \frac{1}{3} = \frac{20}{99}, \quad j = 1, 2, 3.$$

Similarly, considering the events  $A_2 := X^{-1}(\{1\})$  and  $B_j$ ,

$$P(A_2 \cap B_j) = \frac{8}{99}, \quad j = 1, 2, 3$$

and

$$P(A_2) \cdot P(B_j) = \frac{24}{99} \cdot \frac{1}{3} = \frac{8}{99}, \quad j = 1, 2, 3.$$

Finally, considering the events  $A_3 := X^{-1}(\{2\})$  and  $B_j$ , yields

$$P(A_3 \cap B_j) = \frac{5}{99}, \quad j = 1, 2, 3$$

and

$$P(A_3) \cdot P(B_j) = \frac{15}{99} \cdot \frac{1}{3} = \frac{5}{99}, \quad j = 1, 2, 3.$$

Because  $\emptyset$  and all sets  $A \in \mathcal{A}$  are independent, this implies that independence holds for all pairs  $(A, B) \in \{A_1, A_2, A_3, \emptyset\} \times \{B_1, B_2, B_3, \emptyset\}$ . Furthermore, because

- (a)  $\mathcal{E}_1 := \{A_1, A_2, A_3, \emptyset\}$  and  $\mathcal{E}_2 := \{B_1, B_2, B_3, \emptyset\}$  are  $\cap$ -stable set systems on  $\mathcal{A}$ ,
- (b)  $\sigma(\mathcal{E}_1) = X^{-1}(\mathcal{A}'_X)$  and  $\sigma(\mathcal{E}_2) = U^{-1}(\mathcal{A}'_U)$ ,

we can conclude that  $P(A \cap B) = P(A) \cdot P(B)$  holds for *all* elements  $A \in X^{-1}(\mathcal{A}'_X)$  and  $B \in U^{-1}(\mathcal{A}'_U)$  (see Th. 4.44). Therefore, according to Equation (5.24),  $X$  and  $U$  are independent.  $\triangleleft$

**Lemma 5.51 (Independence of a Constant and a Set of Events)**

Let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  be a random variable and  $\mathcal{C} \subset \mathcal{A}$ . If  $X \stackrel{P}{=} \alpha$  and  $\{\alpha\} \in \mathcal{A}'_X$ , then  $X$  and  $\mathcal{C}$  are independent.

(Proof p. 190)

Now we consider mappings of independent random variables. If two random variables  $X_1$  and  $X_2$  are independent and  $f_i: (\Omega'_i, \mathcal{A}'_i) \rightarrow (\Omega''_i, \mathcal{A}''_i)$ ,  $i = 1, 2$ , are measurable mappings, then the two random variables  $f_1(X_1)$  and  $f_2(X_2)$  are independent as well. More generally, if  $f_i: (\Omega'_i, \mathcal{A}'_i) \rightarrow (\Omega''_i, \mathcal{A}''_i)$ ,  $i = 1, \dots, n$ , is a sequence of measurable mappings, then

$$\perp\!\!\!\perp_P f_1(X_1), \dots, f_n(X_n),$$

i. e., then  $f_1(X_1), \dots, f_n(X_n)$  is a sequence of independent random variables on  $(\Omega, \mathcal{A}, P)$ , provided that  $X_1, \dots, X_n$  are independent. In the following theorem we generalize this proposition.

**Theorem 5.52 (Mappings of Families of Independent Random Variables)**

Let  $X_i: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_i, \mathcal{A}'_i)$ ,  $i = 1, \dots, n$ , be random variables and, for  $m \in \mathbb{N}$ , let  $I_1 = \{1, \dots, i_1\}$ ,  $I_2 = \{i_1 + 1, \dots, i_2\}, \dots, I_m = \{i_{m-1} + 1, \dots, n\}$ . Furthermore, let

$$f_j: \left( \prod_{i \in I_j} \Omega'_i, \bigotimes_{i \in I_j} \mathcal{A}'_i \right) \rightarrow (\Omega''_j, \mathcal{A}''_j), \quad j = 1, \dots, m,$$

be measurable mappings. If  $X_1, \dots, X_n$  are independent, then

$$f_1(X_1, \dots, X_{i_1}), f_2(X_{i_1+1}, \dots, X_{i_2}), \dots, f_m(X_{i_{m-1}+1}, \dots, X_n)$$

are independent.

For a generalization and a proof see Bauer (1996, Theorem 9.6, p. 63).

**Example 5.53 (Sums of Independent Random Variables)** Let  $X_1, \dots, X_{2n}$ ,  $n \in \mathbb{N}$ , be independent real-valued random variables, then the  $n$  random variables

$$X_1 + X_2, X_3 + X_4, \dots, X_{2n-1} + X_{2n}$$

are independent as well (see Th. 2.57).  $\triangleleft$

**Example 5.54 (Tom, Jim, and Kate – continued)** In Example 5.50 we showed that the random variables  $X$  and  $U$  are independent. Now we consider the mappings  $f: \{0, 1, 2\} \rightarrow \{0, 1\}$  and  $g: \Omega_U \rightarrow \{male, female\}$  defined by

$$f(x) = \begin{cases} 0, & \text{if } x = 0 \\ 1, & \text{if } x = 1 \text{ or } x = 2 \end{cases}$$

and

$$g(u) = \begin{cases} male, & \text{if } u = Tom \text{ or } u = Jim \\ female, & \text{if } u = Kate, \end{cases}$$

respectively. According to Theorem 5.52 the mappings  $f(X)$  (control vs. any of the two treatments) and  $g(U)$  (*sex*) are independent as well (see Exercise 5-9).  $\triangleleft$

**Remark 5.55 (Conditional Independence of Random Variables Given an Event)**

In chapter 4 we also considered *conditional* independence of events and families of events *given an event*  $B$ . If, for random variables  $X_1, \dots, X_n$  (or, more generally, families of random variables) we consider the set systems  $\sigma(X_1), \dots, \sigma(X_n)$ , then we can use Definition 4.49 in order to define conditional independence of  $X_1, \dots, X_n$  given an event  $B$ , presuming  $P(B) > 0$ . According to Definitions 4.49 and 5.39, conditional independence given  $B$  is equivalent to independence with respect to the probability measure  $P^B$ . In chapter 16 we generalize this concept and study it in more detail.  $\triangleleft$

## 5.5 Probability Function of a Discrete Random Variable

The distribution of a discrete random variable can be described by its *probability function* that is now introduced. Remember, if  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  is a random variable, then the distribution  $P_X$  of  $X$  is a probability measure on  $(\Omega'_X, \mathcal{A}'_X)$ .

### Definition 5.56 (Discrete Random Variable and its Probability Function)

Let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  be a random variable and assume that  $\Omega'_0 \subset \Omega'_X$  is finite or countable with  $P_X(\Omega'_0) = 1$  and  $\{x\} \in \mathcal{A}'_X$  for all  $x \in \Omega'_0$ . Then  $X$  and its distribution  $P_X$  are called *discrete*, and the function  $p_X: \Omega'_X \rightarrow [0, 1]$  defined by

$$p_X(x) = \begin{cases} P_X(\{x\}), & \text{if } x \in \Omega'_0, \\ 0, & \text{if } x \in \Omega'_X \setminus \Omega'_0, \end{cases} \quad (5.31)$$

is called the *probability function* of  $X$ .

**Remark 5.57 (Notation)** Note that  $P(X=x) = p_X(x)$ , using the notation introduced in Remark 5.4.  $\triangleleft$

**Remark 5.58 (Probability Function vs. Distribution)** The *distribution*  $P_X$  is defined for *every* random variable, whereas the *probability function*  $p_X$  only applies to *discrete* random variables. While  $P_X$  assigns probabilities to *subsets* of the codomain  $\Omega'_X$  of  $X$ , the probability function  $p_X$  assigns a probability to each *element*  $x$  in  $\Omega'_X$ . Note that  $p_X$  is a real-valued random variable on the probability space  $(\Omega'_X, \mathcal{A}'_X, P_X)$ .  $\triangleleft$

### Remark 5.59 (The Probability Function Uniquely Determines the Distribution)

Note that  $\sigma$ -additivity of the probability measure  $P_X$  implies that  $P_X$  is uniquely determined by the probability function  $p_X$  [see Rule (x) in Box 4.1]. Vice versa, according to Definition 5.56,  $P_X$  defines  $p_X$ . Hence, if  $X, Y: (\Omega, \mathcal{A}, P) \rightarrow (\Omega', \mathcal{A}')$  are discrete random variables, then

$$p_X = p_Y \Leftrightarrow P_X = P_Y. \quad (5.32)$$

$\triangleleft$

**Remark 5.60 (Probability Function of a Discrete Distribution)** Note that (5.32) allows us to use the term *probability function of a discrete distribution* instead of *probability function of a discrete random variable*.  $\triangleleft$

### Lemma 5.61 (Characterizations of a Discrete Random Variable)

Let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  be a random variable.

- (i) Then  $X$  is discrete if and only if there is a finite or countable  $\Omega'_0 \subset \Omega'_X$  such that  $\{x\} \in \mathcal{A}'_X$  for all  $x \in \Omega'_0$ ,  $P_X(\Omega'_0) = 1$ , and

$$1_{X=x_0} \stackrel{p}{=} 1 - \sum_{x \in \Omega'_0 \setminus \{x_0\}} 1_{X=x}, \quad \forall x_0 \in \Omega'_0. \quad (5.33)$$

(ii) Now assume that  $X: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  is real-valued. Then  $X$  is discrete if and only if there is a finite or countable  $\Omega' \subset \mathbb{R}$  such that

$$X \stackrel{p}{=} \sum_{x \in \Omega'} x \cdot 1_{X=x}. \quad (5.34)$$

(Proof p. 191)

**Remark 5.62 (A Caveat)** Note that Equation (5.33) is equivalent to  $P(\Omega'_0) = 1$ . In proposition (ii), we can choose  $\Omega'$  such that  $0 \notin \Omega'$  even if  $P(X=0) > 0$ . In this case, the set  $\Omega'_0$  referred to in (i) can be chosen such that  $\Omega'_0 := \Omega' \cup \{0\}$ .  $\triangleleft$

**Corollary 5.63 (Discrete Real-Valued Random Variable)**

Assume that  $X: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  is a real-valued random variable. Then  $X$  is discrete if and only if the following two conditions hold:

(a)  $\Omega'_> := \{x \in \mathbb{R} : P(X=x) > 0\}$  is finite or countable.

(b)  $X \stackrel{p}{=} \sum_{x \in \Omega'_>} x \cdot 1_{X=x}$ .

(Proof p. 191)

Condition (b) may equivalently be written

$$X \stackrel{p}{=} \sum_{\substack{x \in \mathbb{R} \\ P(X=x) > 0}} x \cdot 1_{X=x}. \quad (5.35)$$

**Example 5.64 (Flipping two Coins – continued)** Consider again Example 5.35 and let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  denote the number of flipping heads. If we assume that  $P(\{\omega\}) = \frac{1}{4}$  for all  $\omega \in \Omega$ , then

$$p_X(0) = P_X(\{0\}) = \frac{1}{4} \quad p_X(1) = P_X(\{1\}) = \frac{1}{2} \quad p_X(2) = P_X(\{2\}) = \frac{1}{4}$$

are the values of the probability function  $p_X$  of  $X$ . They are the probabilities of the events that  $X$  takes on the value 0, 1, and 2, respectively. For simplicity, we also denote these probabilities by  $P(X=0)$ ,  $P(X=1)$ , and  $P(X=2)$ . In this example, we may choose different measurable spaces  $(\Omega'_X, \mathcal{A}'_X)$ . If we choose  $(\Omega'_X, \mathcal{A}'_X) = (\{0, 1, 2\}, \mathcal{P}(\{0, 1, 2\}))$ , then  $\Omega'_0 = \Omega'_X$  (see Def. 5.56). If we choose  $(\Omega'_X, \mathcal{A}'_X) = (\mathbb{R}, \mathcal{B})$ , then  $\Omega'_0 = \{0, 1, 2\}$  and  $\mathcal{A}'_X|_{\Omega'_0} = \mathcal{B}|_{\Omega'_0} = \mathcal{P}(\Omega'_0)$  (see Remark 1.29). According to Equation (5.34),

$$X = 0 \cdot 1_{X=0} + 1 \cdot 1_{X=1} + 2 \cdot 1_{X=2}.$$

This example is a special case of a random variable with a binomial distribution. The general case is treated in Definition 8.7. Other examples of a discrete

random variable and their probability function are random variables that have a *Poisson distribution* or a *geometric distribution*. In both cases, the random variable considered takes on an infinite and countable number of values, each of which has a probability greater than 0. These examples are treated in chapter 8 (see Defs. 8.14 and 8.20).  $\triangleleft$

**Example 5.65 (Flipping two Coins – continued)** In Example 5.35 we introduced the random variables  $X_1$  and  $X_2$ , which indicate if we flip *heads* in the first and second trial, respectively. The probability function of the bivariate random variable  $X = (X_1, X_2)$  is

$$p_{X_1, X_2}(x_1, x_2) = \frac{1}{4}, \quad \forall (x_1, x_2) \in \{0, 1\}^2.$$

$\triangleleft$

**Lemma 5.66 (Probability Function of a Marginal Distribution)**

Consider a multivariate random variable  $X = (X_1, \dots, X_n)$  as specified in (5.18) and assume that there is a finite or countable set  $\Omega'_0 \subset \prod_{i=1}^n \Omega'_i$  with  $P(\Omega'_0) = 1$  and  $\{x\} \in \otimes_{i=1}^n \mathcal{A}'_i$  for all  $x \in \Omega'_0$ . Furthermore, for all  $x_i \in \Omega'_i$ , define

$$\Omega'_{0, x_i} := \{(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) : (x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n) \in \Omega'_0\}.$$

Then, for all  $x_i \in \Omega'_i$ ,

$$p_{X_i}(x_i) = \sum_{(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) \in \Omega'_{0, x_i}} p_X(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n), \quad (5.36)$$

where  $p_{X_i}$  denotes the probability function of  $X_i$ ,  $i = 1, \dots, n$ , which is also called the *marginal probability function* of  $X_i$ .

(Proof p. 192)

Now we turn to a condition that is equivalent to independence of discrete random variables.

**Remark 5.67 (Support Sets of Discrete Random Variables)** Under the assumptions of Lemma 5.66 we define the support sets of the distributions  $P_{X_i}$ , i. e., the sets

$$\Omega'_{0, i} := \{x_i \in \Omega'_i : p_{X_i}(x_i) > 0\}, \quad i = 1, \dots, n. \quad (5.37)$$

Obviously,  $\Omega'_{0, i}$  is finite or countable for all  $i = 1, \dots, n$ . Hence,  $\Omega'_{sn} := \prod_{i=1}^n \Omega'_{0, i}$  is finite or countable as well. Furthermore,  $P(X \in \Omega'_{sn}) = 1$ , because, for  $(x_1, \dots, x_n) \in \Omega'_0 \setminus \Omega'_{sn}$ , there is at least one  $i$  such that  $X_i \notin \Omega'_{0, i}$  and therefore  $p_{X_i} = P(X_i = x_i) = 0$ , which implies  $P(X_1 = x_1, \dots, X_i = x_i, \dots, X_n = x_n) = 0$ .  $\triangleleft$

**Lemma 5.68 (A Condition Equivalent to Independence)**

Let  $X$  be a multivariate random variable as specified in (5.18) and assume that there is a finite or countable set  $\Omega'_0 \subset \prod_{i=1}^n \Omega'_i$  with  $P(\Omega'_0) = 1$  and  $\{x\} \in \mathcal{A}'_X$ , for all  $x \in \Omega'_0$ . Furthermore, let  $p_X, p_{X_1}, \dots, p_{X_n}$  denote the probability functions of  $X, X_1, \dots, X_n$ , respectively, and let  $\Omega'_{0,1}, \dots, \Omega'_{0,n}$  be the sets defined in (5.37). Then  $X_1, \dots, X_n$  are independent if and only if

$$p_X(x_1, \dots, x_n) = p_{X_1}(x_1) \cdot \dots \cdot p_{X_n}(x_n), \quad \forall (x_1, \dots, x_n) \in \prod_{i=1}^n \Omega'_{0,i}. \quad (5.38)$$

(Proof p. 192)

Note that, in Lemmas 5.66 and 5.68, the set  $\Omega'_0$  is not necessarily a Cartesian product. We only require that it is a *subset* of a Cartesian product.

In the following section we shall see that a probability function is a special probability density (see Th. 5.77).

**5.6 Probability Density With Respect to a Measure**

Some probability measures can also be described by a density with respect to the Lebesgue measure on  $(\mathbb{R}, \mathcal{B})$  or the counting measure on  $(\Omega, \mathcal{P}(\Omega))$ , where  $\Omega$  is a finite or countable set (see Th. 3.65). Such a density is useful for explicit numerical calculations and comparing distributions to each other. We start by translating some concepts and results of chapter 3 to probability measures.

**5.6.1 General Concepts and Properties**

According to Theorem 3.65 and Definition 3.66, a nonnegative measurable function  $f: (\Omega, \mathcal{A}, \mu) \rightarrow (\overline{\mathbb{R}}, \mathcal{B})$  is called a *density of  $\nu$  with respect to  $\mu$* , if

$$\nu(A) = \int_A f \, d\mu, \quad \forall A \in \mathcal{A}. \quad (5.39)$$

The function  $\nu: \mathcal{A} \rightarrow \overline{\mathbb{R}}$  defined by (5.39) is a measure, which is also denoted by  $f \circ \mu$ . Hence,  $f \circ \mu(A) = \int_A f \, d\mu, \forall A \in \mathcal{A}$ . Theorem 3.65 implies the following corollary.

**Corollary 5.69 (Probability Measure With Density)**

Let  $(\Omega, \mathcal{A}, \mu)$  be a measure space. If  $f: (\Omega, \mathcal{A}, \mu) \rightarrow (\overline{\mathbb{R}}, \mathcal{B})$  is a nonnegative  $\mu$ -integrable function with  $\int f \, d\mu = 1$ , then  $P := f \circ \mu$  is a probability measure on  $(\Omega, \mathcal{A})$ .

Corollary 5.69 justifies the following definition.

**Definition 5.70 (Probability Density)**

Let  $(\Omega, \mathcal{A}, P)$  be a probability space and  $\mu$  a measure on  $(\Omega, \mathcal{A})$ . If  $P = f \circ \mu$ , then  $f$  is called a probability density of  $P$  with respect to  $\mu$ .

**Remark 5.71 (Probability Density of a Random Variable)** Consider the random variable  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  and let  $\mu$  be a measure on  $(\Omega'_X, \mathcal{A}'_X)$ . If  $P_X = f_X \circ \mu$ , then  $f_X$  is also referred to as a probability density of  $X$  with respect to  $\mu$ .  $\triangleleft$

Applying Equation (5.39) yields the following corollary.

**Corollary 5.72 (Characterizing the Probability Measure by a Density)**

Let  $f: (\Omega, \mathcal{A}, \mu) \rightarrow (\mathbb{R}, \mathcal{B})$  be  $\mu$ -integrable and nonnegative and  $P$  a probability measure on  $(\Omega, \mathcal{A})$ . Then  $f$  is a (probability) density of  $P$  with respect to  $\mu$  if and only if it satisfies

$$P(A) = \int_A f \, d\mu, \quad \forall A \in \mathcal{A}. \quad (5.40)$$

Theorem 3.68 (a) and (c) imply the following corollary.

**Corollary 5.73 (Probability Densities are  $\mu$ -Equivalent)**

Let  $\mu$  and  $P$  be measures on the measurable space  $(\Omega, \mathcal{A})$ , where  $P$  is a probability measure. If  $f, f^*: (\Omega, \mathcal{A}, \mu) \rightarrow (\mathbb{R}, \mathcal{B})$  are probability densities of  $P$  with respect to  $\mu$ , then  $f \stackrel{\mu}{=} f^*$ .

Now we translate the Radon-Nikodym Theorem (see Th. 3.72), which yields the following corollary.

**Corollary 5.74 (An Implication of the Radon-Nikodym Theorem)**

Let  $(\Omega, \mathcal{A}, \mu)$  be a measure space. If  $\mu$  is  $\sigma$ -finite,  $P$  is a probability measure on  $(\Omega, \mathcal{A})$ , and  $P \ll \mu$ , then there is a probability density  $f$  of  $P$  with respect to  $\mu$  (also called a Radon-Nikodym derivative), i. e.,

$$f = \frac{dP}{d\mu}. \quad (5.41)$$

**Example 5.75 (Conditional-Probability Measure)** In Lemma 4.32 (ii) we showed that  $1_B/P(B)$  is a density of  $P^B$  with respect to  $P$ , and according to Remark 5.71 it is a probability density of  $P^B$ .  $\triangleleft$

### 5.6.2 Density of a Discrete Random Variable

As a special case, we consider a discrete random variable (see section 5.5).

**Remark 5.76 (A Sum of Dirac Measures)** Let the assumptions of Definition 5.56 hold, i. e., let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  be a random variable and suppose  $\Omega'_0 \subset \Omega'_X$  is finite or countable with  $P_X(\Omega'_0) = 1$  and  $\{x\} \in \mathcal{A}'_X$  for all  $x \in \Omega'_0$ . Furthermore, define

$$\mu := \sum_{x \in \Omega'_0} \delta_x, \quad (5.42)$$

the sum of Dirac measures at  $x$ ,  $x \in \Omega'_0$ . According to Example 1.57,  $\mu$  is a measure on  $(\Omega'_X, \mathcal{A}'_X)$ , which is  $\sigma$ -finite.  $\triangleleft$

The following theorem asserts that the probability function  $p_X$  is a density of  $P_X$  with respect to  $\mu$ .

#### Theorem 5.77 (The Probability Function is a Density)

Let the assumptions of Definition 5.56 be satisfied and let  $\mu$  be defined by Equation (5.42). Then:

- (i)  $P_X \ll_{\mathcal{A}'_X} \mu$ .
- (ii) The probability function  $p_X$  is a density of  $P_X$  with respect to  $\mu$ , i. e.,

$$p_X = \frac{dP_X}{d\mu} \quad (5.43)$$

and, for all  $A' \in \mathcal{A}'_X$ ,

$$P_X(A') = \int_{A'} p_X d\mu \quad (5.44)$$

$$= \sum_{x \in \Omega'_0} 1_{A'}(x) \cdot p_X(x) \quad (5.45)$$

$$= \sum_{x \in A'} p_X(x). \quad (5.46)$$

(Proof p. 193)

Hence, each probability  $P_X(A')$ ,  $A' \in \mathcal{A}'_X$ , can be computed from the probability function  $p_X$ .

### 5.6.3 Density of a Bivariate Random Variable

Now we consider bivariate random variables. However, extending the following notation and propositions to general multivariate random variables is straightforward.

**Lemma 5.78 (Absolute Continuity of Marginal Distributions)**

Let  $(X, Y): (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X \times \Omega'_Y, \mathcal{A}'_X \otimes \mathcal{A}'_Y)$  be a random variable and suppose that  $(\Omega'_X, \mathcal{A}'_X, \mu)$  and  $(\Omega'_Y, \mathcal{A}'_Y, \nu)$  are  $\sigma$ -finite measure spaces. Then

$$P_{X,Y} \ll_{\mathcal{A}'_X \otimes \mathcal{A}'_Y} \mu \otimes \nu \Rightarrow P_X \ll_{\mathcal{A}'_X} \mu \text{ and } P_Y \ll_{\mathcal{A}'_Y} \nu.$$

(Proof p. 193)

Let

$$f_{X,Y} := \frac{dP_{X,Y}}{d\mu \otimes \nu}, \quad f_X := \frac{dP_X}{d\mu}, \quad f_Y := \frac{dP_Y}{d\nu}$$

denote Radon-Nikodym derivatives (see Th. 3.72). In the following lemma, we use the notation ' $\stackrel{\mu}{\equiv}$ ' introduced in Remark 2.70.

**Lemma 5.79 (Marginal Densities)**

Let  $(X, Y): (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X \times \Omega'_Y, \mathcal{A}'_X \otimes \mathcal{A}'_Y)$  be a random variable,  $(\Omega'_X, \mathcal{A}'_X, \mu)$ ,  $(\Omega'_Y, \mathcal{A}'_Y, \nu)$  be measure spaces,  $\mu, \nu$  be  $\sigma$ -finite, and assume  $P_{X,Y} \ll_{\mathcal{A}'_X \otimes \mathcal{A}'_Y} \mu \otimes \nu$ .

Then

$$f_X(x) \stackrel{\mu}{\equiv} \int f_{X,Y}(x, y) \nu(dy), \quad (5.47)$$

and

$$f_Y(y) \stackrel{\nu}{\equiv} \int f_{X,Y}(x, y) \mu(dx). \quad (5.48)$$

The functions  $f_X$  and  $f_Y$  are also called *marginal densities* of  $X$  and  $Y$ , respectively.

(Proof p. 194)

**Remark 5.80 (Marginal and Joint Density)** Suppose that  $X$  and  $Y$  are real-valued. Then for the Lebesgue measure  $\mu = \nu = \lambda$  and a Riemann integrable density  $f_{X,Y}$ , Equations (5.47) and (5.48) yield

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy \quad (5.49)$$

and

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx, \quad (5.50)$$

respectively (see Th. 3.62).  $\triangleleft$ 

Note that, for a discrete random variable  $X = (X_1, \dots, X_n)$ , the probability function  $p_X$  is a density of  $P_X$  with respect to the measure  $\mu$  specified in Equation (5.42), and the marginal probability functions  $p_{X_i}$ ,  $i = 1, \dots, n$ , are marginal densities. In section 5.7.4 we consider multivariate densities with respect to the Lebesgue measure  $\mu = \nu = \lambda$ .

## 5.7 Uni- or Multivariate Real-Valued Random Variable

The remaining sections of this chapter show how to describe distributions of *real-valued* uni- and multivariate random variables.

### 5.7.1 Distribution Function of a Univariate Real-Valued Random Variable

If we consider a univariate real-valued random variable  $X$ , then the *distribution function*  $F_X$  assigns to each  $x \in \mathbb{R}$  the probability  $P(X \leq x)$  of the event  $\{X \leq x\} = \{\omega \in \Omega: X(\omega) \leq x\}$  that  $X$  takes on a value *smaller or equal* than  $x$ . As we shall see, the distribution function uniquely determines the distribution  $P_X$ .

#### Definition 5.81 (Distribution Function)

Let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B}, P_X)$  denote a real-valued random variable. Then the (cumulative) distribution function  $F_X: \mathbb{R} \rightarrow [0, 1]$  of  $X$  is defined by:

$$F_X(x) := P_X(]-\infty, x]) = P(X \leq x), \quad \forall x \in \mathbb{R}. \quad (5.51)$$

**Remark 5.82 (Probabilities of Intervals)** This definition implies that we can compute the probability  $P(a < X \leq b)$  of  $X$  taking a value in the interval  $]a, b]$  by

$$P(a < X \leq b) = F_X(b) - F_X(a), \quad \text{if } a < b, \quad (5.52)$$

because

$$P(a < X \leq b) = P_X(]-\infty, b] \setminus ]-\infty, a]) = P_X(]-\infty, b]) - P_X(]-\infty, a])$$

[see Rule (vi), Box 4.1]. ◁

**Remark 5.83 (The Distribution Function Determines the Distribution)** Every random variable  $X$  has a distribution  $P_X$ . Therefore, the distribution function  $F_X$  exists for all real-valued random variables. The distribution function uniquely determines the distribution  $P_X$  of a real-valued random variable, because a finite measure on  $(\Omega, \mathcal{A})$  is already uniquely specified by its values on a  $\cap$ -stable generating system of  $\mathcal{A}$  (see Th. 1.72) and the set system  $\{]-\infty, x]: x \in \mathbb{R}\}$  is a  $\cap$ -stable generating system of  $\mathcal{B}$ , the Borel  $\sigma$ -algebra on  $\mathbb{R}$  [see Eq. (1.19)]. Hence,  $P_X$  uniquely determines  $F_X$ , which implies the following theorem. ◁

#### Theorem 5.84 (Uniqueness)

Let  $P_X, P_Y$  denote the distributions and  $F_X, F_Y$  the distribution functions of two real-valued random variables  $X, Y: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$ . Then

$$F_X = F_Y \Leftrightarrow P_X = P_Y. \quad (5.53)$$

This theorem facilitates describing distributions and calculations involving distributions considerably, because the distribution function is defined on the set  $\mathbb{R}$  of real numbers, whereas  $P_X$  is defined on a much more complex domain, the Borel  $\sigma$ -algebra  $\mathcal{B}$ .

**Example 5.85 (Flipping two Coins – continued)** In Example 5.11, we considered flipping two coins and specified the distribution  $P_X$  of  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{P}(\Omega'_X))$ , representing with its values  $x$  the number of flipping heads. The distribution  $P_X$  assigns a probability to all  $2^3 = 8$  subsets of  $\Omega'_X := \{0, 1, 2\}$ . Because  $\{0, 1, 2\} \subset \mathbb{R}$ , the random variable  $X$  is also a random variable  $X: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  with values in  $\mathbb{R}$ . In this case,  $F_X$  is a step function defined by

$$F_X(x) = P_X([-\infty, x]) = P(X \leq x) = \begin{cases} 0, & \text{if } x < 0 \\ 1/4, & \text{if } 0 \leq x < 1 \\ 3/4, & \text{if } 1 \leq x < 2 \\ 1, & \text{if } x \geq 2. \end{cases}$$

According to Theorem 5.84 the distribution  $P_X$  is uniquely defined by these four values. In other words, if we know these four values, then we know the probabilities  $P_X(A')$  for all elements  $A'$  of the Borel  $\sigma$ -algebra  $\mathcal{B}$  (see Exercise 5-10).  $\triangleleft$

Now we turn to the *quantile function*, which, in some cases, is the inverse of the distribution function. Sometimes this function is also called the *pseudo-inverse* of  $F_X$ . It assigns to each  $p \in [0, 1]$  the smallest real number  $x$  for which  $P(X \leq x) = F_X(x) \geq p$ .

**Definition 5.86 (Quantile Function)**

Let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  be a real-valued random variable with distribution function  $F_X$ . Then the quantile function  $Q_X: [0, 1] \rightarrow \bar{\mathbb{R}}$  of  $X$  is defined by:

$$\begin{aligned} Q_X(p) &= \inf \{x \in \mathbb{R} : F_X(x) \geq p\}, \quad \forall p \in ]0, 1[, \\ Q_X(0) &= \begin{cases} \inf \{x \in \mathbb{R} : F_X(x) > 0\}, & \text{if } \exists x \in \mathbb{R} \text{ with } F_X(x) = 0 \\ -\infty, & \text{if } F_X(x) > 0, \forall x \in \mathbb{R}, \end{cases} \quad (5.54) \\ Q_X(1) &= \begin{cases} \inf \{x \in \mathbb{R} : F_X(x) = 1\}, & \text{if } \exists x \in \mathbb{R} \text{ with } F_X(x) = 1 \\ \infty, & \text{if } F_X(x) < 1, \forall x \in \mathbb{R}. \end{cases} \end{aligned}$$

A value  $Q_X(p)$ ,  $p \in [0, 1]$ , is called the *quantile of  $p$  with respect to  $F_X$* .

**Remark 5.87 (Inverse Function of  $F_X$ )** If  $F_X$  is continuous and strictly monotone, i. e., if  $x_1 < x_2$  implies  $F_X(x_1) < F_X(x_2)$ , then

$$Q_X(p) = F_X^{-1}(p), \quad \forall p \in ]0, 1[. \quad (5.55)$$

where  $F_X^{-1}$  denotes the inverse function of  $F_X$ .  $\triangleleft$

**Example 5.88 (Flipping two Coins – continued)** In Example 5.85, we specified the distribution function of  $X := \text{number of flipping heads}$  for the random experiment of flipping two coins. The corresponding quantile function takes on the three values

$$Q_X(p) = \begin{cases} 0, & \text{if } 0 \leq p \leq 1/4 \\ 1, & \text{if } 1/4 < p \leq 3/4 \\ 2, & \text{if } 3/4 < p \leq 1. \end{cases}$$

&lt;

### 5.7.2 Distribution Function of a Multivariate Real-Valued Random Variable

Now we extend the concept of a distribution function to the multivariate case. In the following definition we use the notation introduced in Equation (5.25).

#### Definition 5.89 (Joint Distribution Function)

Let  $(X_1, \dots, X_n): (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}^n, \mathcal{B}_n)$  be a random variable. Its distribution function  $F_{X_1, \dots, X_n}: \mathbb{R}^n \rightarrow [0, 1]$  is defined by

$$F_{X_1, \dots, X_n}(x_1, \dots, x_n) := P(X_1 \leq x_1, \dots, X_n \leq x_n), \quad \forall (x_1, \dots, x_n) \in \mathbb{R}^n. \quad (5.56)$$

It is also called the joint distribution function of  $X_1, \dots, X_n$ .

**Example 5.90 (Flipping two Coins – continued)** In example 5.35, we considered flipping two coins and defined the random variables  $X_1$  and  $X_2$  indicating whether or not we flip *heads* at first and second flip, respectively. In this example, the bivariate distribution function  $F_{X_1, X_2}$  takes on the values

$$F_{X_1, X_2}(x_1, x_2) = \begin{cases} 0 & \text{if } x_1 < 0 \text{ or } x_2 < 0, \\ 1/4 & \text{if } 0 \leq x_1 < 1, 0 \leq x_2 < 1, \\ 2/4 & \text{if } x_1 \geq 1, 0 \leq x_2 < 1, \\ 2/4 & \text{if } 0 \leq x_1 < 1, x_2 \geq 1, \\ 1 & \text{if } x_1 \geq 1, x_2 \geq 1. \end{cases}$$

&lt;

Just like in Theorem 5.84 we can prove uniqueness, using a  $\cap$ -stable generating system for  $\mathcal{B}_n$ , now referring to Equation (1.21).

#### Theorem 5.91 (Uniqueness)

Let  $P_X, P_Y$  denote the distributions and  $F_X, F_Y$  the distribution functions of two  $n$ -variate real-valued random variables  $X, Y: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}^n, \mathcal{B}_n)$ . Then

$$F_X = F_Y \Leftrightarrow P_X = P_Y. \quad (5.57)$$

As an implication of Equation (5.21) with  $A_i' = ] - \infty, X_i]$  and  $\Omega_j' = \mathbb{R}$ ,  $i \neq j$ , we obtain the next corollary. In the special case of a bivariate real-valued random variable  $(X_1, X_2)$ , this corollary asserts that the value of the *marginal distribution function of  $X_1$*  for the argument  $x_1$ , i. e.,  $\lim_{x_2 \rightarrow \infty} F_{X_1, X_2}(x_1, x_2) = F_{X_1}(x_1)$ . In this corollary,

$$\lim_{\substack{x_j \rightarrow \infty \\ j \neq i}} F_{X_1, \dots, X_n}(x_1, \dots, x_n), \quad (5.58)$$

denotes the limit of the distribution function of  $(X_1, \dots, X_n)$  for  $x_j \rightarrow \infty$ , for all  $j = 1, \dots, i-1, i+1, \dots, n$ . This limit is the value of the *marginal distribution function of  $X_i$*  for the argument  $x_i$ , and the corollary asserts that this limit is identical to the value  $F_{X_i}(x_i)$  of the distribution function of  $X_i$  for the argument  $X_i$ .

**Corollary 5.92 (Joint and Marginal Distribution Function)**

Let  $(X_1, \dots, X_n): (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}^n, \mathcal{B}_n)$  be a random variable. Then

$$F_{X_i}(x_i) = \lim_{\substack{x_j \rightarrow \infty \\ j \neq i}} F_{X_1, \dots, X_n}(x_1, \dots, x_n), \quad \forall x_i \in \mathbb{R}. \quad (5.59)$$

(Proof p. 194)

The next corollary shows how independence of the random variables  $X_1, \dots, X_n$  can be formulated in terms of their distribution functions.

**Corollary 5.93 (Independence and Joint Distribution Function)**

Let  $(X_1, \dots, X_n): (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}^n, \mathcal{B}_n)$  be a random variable. Then the following two propositions are equivalent to each other:

- (a)  $X_1, \dots, X_n$  are independent
- (b)  $F_{X_1, \dots, X_n}(x_1, \dots, x_n) = F_{X_1}(x_1) \cdot \dots \cdot F_{X_n}(x_n)$ ,  $\forall (x_1, \dots, x_n) \in \mathbb{R}^n$ .

(Proof p. 194)

**5.7.3 Density of a Continuous Univariate Real-Valued Random Variable**

As a special case, we consider a random variable  $X: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  for which there is a nonnegative measurable function  $f_X: (\mathbb{R}, \mathcal{B}, P_X) \rightarrow (\mathbb{R}, \mathcal{B})$  such that

$$P_X(B) = \int_B f_X d\lambda, \quad \forall B \in \mathcal{B}, \quad (5.60)$$

where  $\lambda$  denotes the Lebesgue measure on  $(\mathbb{R}, \mathcal{B})$ . According to Theorem 3.68 (ii), this equation is equivalent to  $P_X = f_X \circ \lambda$ .

The following definition is a special case of Definition 5.70.

**Definition 5.94 (Continuous Random Variable and its Density)**

Let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  be a real-valued random variable with distribution  $P_X$ . We call  $X$  continuous, if there is a nonnegative function  $f_X: \mathbb{R} \rightarrow \mathbb{R}$  that is integrable with respect to the Lebesgue measure  $\lambda$  and

$$P_X = f_X \circ \lambda. \quad (5.61)$$

A function  $f_X$  satisfying (5.61) is called a (probability) density of  $X$ .

Note that Equation (5.61) is equivalent to

$$F_X(x) = \int_{]-\infty, x]} f_X d\lambda, \quad \forall x \in \mathbb{R}, \quad (5.62)$$

because

$$F_X(x) = P_X(]-\infty, x]) = \int 1_{]-\infty, x]} dP_X = \int_{]-\infty, x]} f_X d\lambda, \quad \forall x \in \mathbb{R}, \quad (5.63)$$

(see Th. 3.67). Furthermore, Theorem 3.62 immediately implies the following corollary:

**Corollary 5.95 (Riemann Integral of the Density)**

If  $f_X$  is a density of the random variable  $X: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  and  $f_X$  is Riemann integrable, then

$$F_X(x) = \int_{-\infty}^x f_X(t) dt, \quad \forall x \in \mathbb{R}. \quad (5.64)$$

**Remark 5.96 (Interpretation of Densities)** Note that the term  $f_X(t)$  in Equation (5.64) is not a probability, instead it is a value of the density for  $t \in \mathbb{R}$ . However, the probability  $P(a < X \leq b)$  that  $X$  takes on a value in the interval  $]a, b]$  can be computed using Equation (5.52) and the density  $f_X$ , provided that it exists and that it is Riemann integrable:

$$P(a < X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(x) dx, \quad \text{if } a < b. \quad (5.65)$$

This probability can be represented as the area between the density and the  $x$ -axis above the interval  $[a, b]$  (see Fig. 3.6).  $\triangleleft$

**Remark 5.97 (Continuity of  $X$  Implies  $P(X = x) = 0$ )** Consider a continuous random variable  $X: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$ . Definition 5.94 and Remark 3.71 imply that  $P_X \ll \lambda$ . Because  $\lambda(\{x\}) = 0$  [see Eq. (1.53)] we can conclude  $P(X=x) = 0$ , for all  $x \in \mathbb{R}$ . Hence, additivity of  $P$  yields, for all  $a, b \in \mathbb{R}$ ,  $a < b$ ,

$$P(a < X \leq b) = P(a \leq X \leq b) = P(a \leq X < b) = P(a < X < b), \quad (5.66)$$

provided that  $X$  is continuous.  $\triangleleft$

**Example 5.98 (Continuous Random Variables and Their Densities)** In section 8.2 we present some examples of continuous random variables and their densities, such as the densities of normal distributions, central  $\chi^2$ -distributions, central  $t$ -distributions, and central  $F$ -distributions.  $\triangleleft$

### 5.7.4 Density of a Continuous Multivariate Real-Valued Random Variable

**Remark 5.99 (Multivariate Case)** Let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}^n, \mathcal{B}_n)$  be a multivariate random variable with distribution  $P_X$ . If  $f_X: \mathbb{R}^n \rightarrow \mathbb{R}$  is nonnegative and integrable with respect to the Lebesgue measure  $\lambda_n$  on  $(\mathbb{R}^n, \mathcal{B}_n)$ , and  $P_X = f_X \circ \lambda_n$ , then  $X$  is continuous with probability density  $f_X$ , and

$$F_X(x_1, \dots, x_n) = \int_B f_X d\lambda_n, \quad \forall (x_1, \dots, x_n) \in \mathbb{R}^n, \quad (5.67)$$

where  $B := \times_{i=1}^n ]-\infty, x_i]$ . If  $f_X$  is Riemann integrable, then

$$F_X(x_1, \dots, x_n) = \int_{-\infty}^{x_n} \dots \int_{-\infty}^{x_1} f_X(t_1, \dots, t_n) dt_1 \dots dt_n, \quad \forall (x_1, \dots, x_n) \in \mathbb{R}^n. \quad (5.68)$$

More generally, for any  $B \in \mathcal{B}_n$ ,

$$P_X(B) = P(X \in B) = \int 1_B \cdot f_X d\lambda_n. \quad (5.69)$$

The probability  $P_X(B)$  can be represented as the  $(n+1)$ -dimensional *volume* between the joint density and the  $(x_1, \dots, x_n)$ -hyperplane above  $B$  (see Fig. 8.8).  $\triangleleft$

In analogy to Corollary 5.93, independence of continuous real-valued random variables can also be formulated in terms of probability densities, using the marginal densities  $f_{X_1}, \dots, f_{X_n}$  (see Lemma 5.79).

#### Corollary 5.100 (Independence and Probability Densities)

Let  $(X_1, \dots, X_n): (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}^n, \mathcal{B}_n)$ ,  $n > 1$ , be a random variable and suppose that all random variables  $X_i$ ,  $i = 1, \dots, n$ , have a density  $f_{X_i}$  with respect to the Lebesgue measure  $\lambda$  on  $(\mathbb{R}, \mathcal{B})$ . Then  $X_1, \dots, X_n$  are independent if and only if

$$f_{X_1, \dots, X_n}(x_1, \dots, x_n) := f_{X_1}(x_1) \cdot \dots \cdot f_{X_n}(x_n), \quad \forall (x_1, \dots, x_n) \in \mathbb{R}^n, \quad (5.70)$$

is a joint density of  $(X_1, \dots, X_n)$  with respect to  $\lambda_n$ .

(Proof p. 195)

**Remark 5.101 (Independence, Densities, and Riemann Integrals)** If all densities  $f_{X_i}$ ,  $i = 1, \dots, n$ , are Riemann integrable, then  $f_{X_1, \dots, X_n}$  in (5.70) is Riemann integrable as well (see, e. g., Ellis & Gulick, 2006).  $\triangleleft$

**Remark 5.102 (Other Random Variables)** In this section we considered univariate and multivariate real-valued random variables  $X$ . All these random variables have a distribution  $P_X$  and also a (cumulative) distribution function  $F_X$ . If  $X$  is discrete, then its distribution can be described by the probability function, and the distribution function  $F_X$  is a step function (see Example 5.90 and Fig. 8.1). If  $X$  is continuous, then its distribution can be specified by a density with respect to the Lebesgue measure. The distribution functions of continuous random variables do not have discontinuities or jumps (see Rem. 5.97 and Fig. 8.7 for an example). Note, however, that there are random variables that are neither discrete nor continuous. Their distribution functions have jumps but are not step functions.  $\triangleleft$

## 5.8 Proofs

### *Proof of Lemma 5.7*

According to Theorem 2.49, the composition  $g(X)$  of  $X$  and  $g$  is  $(\mathcal{A}, \mathcal{A}')$ -measurable. Therefore, it is a random variable on  $(\Omega, \mathcal{A}, P)$ . For all  $A \in \mathcal{A}$ ,

$$\begin{aligned}
 P_{g(X)}(A) &= P\{[g(X)]^{-1}(A)\} && [(5.2)] \\
 &= P\{\{\omega \in \Omega: g[X(\omega)] \in A\}\} && [\text{Def. 2.1}] \\
 &= P\{\{\omega \in \Omega: X(\omega) \in g^{-1}(A)\}\} && [\text{Def. 2.1}] \\
 &= P\{X^{-1}[g^{-1}(A)]\} && [\text{Def. 2.1}] \\
 &= P_X[g^{-1}(A)] && [(5.2)] \\
 &= (P_X)_g(A). && [(5.2)]
 \end{aligned}$$

### *Proof of Lemma 5.21*

$$\begin{aligned}
 X \stackrel{P}{=} Y &\Rightarrow \exists A \in \mathcal{A}: (\forall \omega \in \Omega \setminus A: X(\omega) = Y(\omega) \text{ and } P(A) = 0) && [(5.9)] \\
 &\Rightarrow \exists A \in \mathcal{A}: (\forall \omega \in \Omega \setminus A: X(\omega) = Y(\omega) \text{ and } Q(A) = 0) && [Q \ll_{\mathcal{A}} P] \\
 &\Rightarrow X \stackrel{Q}{=} Y. && [(5.9)]
 \end{aligned}$$

### *Proof of Corollary 5.25*

Proposition (i) is a special case of Theorem 2.86. Analogously to the proof of Theorem 2.86, we can prove Propositions (ii) and (iii) using Remark 2.67.

**Proof of Lemma 5.27**

Let  $A := \{\omega \in \Omega: X(\omega) \neq Y(\omega)\} \in \mathcal{C}$ . Hence, if  $X \stackrel{P^B}{=} Y$ , then the conjunction of  $P^B(A) = 0$  and  $P \ll_{\mathcal{C}} P^B$  implies  $P(A) = 0$ .

**Proof of Lemma 5.29**

$P \ll_{\sigma(X)} P^B \Rightarrow P_X \ll_{\mathcal{A}'_X} P_X^B$ : If  $P \ll_{\sigma(X)} P^B$ , then, for all  $C' \in \mathcal{A}'_X$ :

$$\begin{aligned} P_X^B(C') = 0 &\Rightarrow P^B[X^{-1}(C')] = 0 && [(5.2)] \\ &\Rightarrow P[X^{-1}(C')] = 0 && [P \ll_{\sigma(X)} P^B] \\ &\Rightarrow P_X(C') = 0. && [(5.2)] \end{aligned}$$

$P_X \ll_{\mathcal{A}'_X} P_X^B \Rightarrow P \ll_{\sigma(X)} P^B$ : If  $P_X \ll_{\mathcal{A}'_X} P_X^B$  and  $C \in \sigma(X)$ , then there is a  $C' \in \mathcal{A}'_X$  such that

$$C = X^{-1}(C').$$

Hence,

$$\begin{aligned} P^B(C) = 0 &\Rightarrow P^B[X^{-1}(C')] = 0 \\ &\Rightarrow P_X^B(C') = 0 && [(5.2)] \\ &\Rightarrow P_X(C') = 0 && [P_X \ll_{\mathcal{A}'_X} P_X^B] \\ &\Rightarrow P[X^{-1}(C')] = 0 && [(5.2)] \\ &\Rightarrow P(C) = 0. \end{aligned}$$

**Proof of Lemma 5.47**

If  $X_1 \stackrel{P}{=} X_2$ , then

$$\begin{aligned} X_1 &\perp\!\!\!\perp_P Y \\ &\Rightarrow \forall (A', B') \in \mathcal{A}' \times \mathcal{A}'_Y: \\ &\quad P[X_1^{-1}(A') \cap Y^{-1}(B')] = P[X_1^{-1}(A')] \cdot P[Y^{-1}(B')] && [(5.24)] \\ &\Rightarrow \forall (A', B') \in \mathcal{A}' \times \mathcal{A}'_Y: \\ &\quad P[X_2^{-1}(A') \cap Y^{-1}(B')] = P[X_2^{-1}(A')] \cdot P[Y^{-1}(B')] \quad [X_1 \stackrel{P}{=} X_2, \text{Def. 5.3, Cor. 5.24}] \\ &\Rightarrow X_2 \perp\!\!\!\perp_P Y. && [(5.24)] \end{aligned}$$

**Proof of Lemma 5.48**

According to Remark 5.46,  $1_B \perp\!\!\!\perp_P \mathcal{C} \Leftrightarrow B \perp\!\!\!\perp_P \mathcal{C}$ . Hence, for all  $C \in \mathcal{C}$ ,

$$\begin{aligned}
P^B(C) = 0 &\Rightarrow P(B \cap C) = 0 \\
&\Rightarrow P(B) \cdot P(C) = 0 && [B \perp\!\!\!\perp_p C] \\
&\Rightarrow P(C) = 0. && [P(B) > 0]
\end{aligned}$$

### **Proof of Lemma 5.49**

$\Rightarrow$  If  $X_1, \dots, X_n$  are independent, then, for all  $A'_i \in \mathcal{A}'_i$ ,  $i = 1, \dots, n$ ,

$$\begin{aligned}
P_{X_1, \dots, X_n}(A'_1 \times \dots \times A'_n) &= P[X_1^{-1}(A'_1) \cap \dots \cap X_n^{-1}(A'_n)] \\
&= P[X_1^{-1}(A'_1)] \cdot \dots \cdot P[X_n^{-1}(A'_n)] \\
&= P_{X_1}(A'_1) \cdot \dots \cdot P_{X_n}(A'_n) \\
&= (P_{X_1} \otimes \dots \otimes P_{X_n})(A'_1 \times \dots \times A'_n).
\end{aligned}$$

Hence, according to the definition of the product measure (see Def. 1.67),

$$P_{X_1, \dots, X_n} = P_{X_1} \otimes \dots \otimes P_{X_n}.$$

$\Leftarrow$  If the right-hand side of (5.30) holds, then, for all  $A'_i \in \mathcal{A}'_i$ ,  $i = 1, \dots, n$ ,

$$\begin{aligned}
P[X_1^{-1}(A'_1) \cap \dots \cap X_n^{-1}(A'_n)] &= P_{X_1, \dots, X_n}(A'_1 \times \dots \times A'_n) \\
&= P_{X_1}(A'_1) \cdot \dots \cdot P_{X_n}(A'_n) \\
&= P[X_1^{-1}(A'_1)] \cdot \dots \cdot P[X_n^{-1}(A'_n)].
\end{aligned}$$

According to Equation (5.28), this implies independence of  $X_1, \dots, X_n$ .

### **Proof of Lemma 5.51**

Assume that  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  is a random variable and that there is an  $\alpha \in \Omega'_X$  and  $\{\alpha\} \in \mathcal{A}'_X$  such that  $X \stackrel{p}{=} \alpha$ . If  $A := \{\omega \in \Omega: X(\omega) \neq \alpha\}$ , then  $A = \Omega \setminus X^{-1}(\{\alpha\}) \in \mathcal{A}$ ,  $P(A) = 0$  and  $P(A^c) = 1$ . This implies, for all  $A' \in \mathcal{A}'_X$ ,

$$\begin{aligned}
P[X^{-1}(A')] &= P[X^{-1}(A') \cap A^c] && [\text{Box 4.1 (viii)}] \\
&= P[\{\omega \in \Omega: X(\omega) \in A', X(\omega) = \alpha\}] \\
&= \begin{cases} P(\emptyset), & \text{if } \alpha \notin A' \\ P(A^c), & \text{if } \alpha \in A' \end{cases} \\
&= \begin{cases} 0, & \text{if } \alpha \notin A' \\ 1, & \text{if } \alpha \in A'. \end{cases} && (5.71)
\end{aligned}$$

This implies, for all  $A' \in \mathcal{A}'_X$  and all  $C \in \mathcal{C}$ ,

$$\begin{aligned}
P[X^{-1}(A') \cap C] &= \begin{cases} 0, & \text{if } \alpha \notin A', \\ P(C), & \text{if } \alpha \in A', \end{cases} && [\text{Box 4.1 (v), (viii)}] \\
&= P[X^{-1}(A')] \cdot P(C). && [(5.71)]
\end{aligned}$$

**Proof of Lemma 5.61**

(i) ( $\Rightarrow$ ) If  $X$  is discrete, then there is a finite or countable  $\Omega'_0 \subset \Omega'_X$  with  $\{x\} \in \mathcal{A}'_X$  for all  $x \in \Omega'_0$  and  $P_X(\Omega'_0) = 1$ . This implies

$$P\left(1 \neq \sum_{x \in \Omega'_0} 1_{X=x}\right) = P(X \notin \Omega'_0) = 0,$$

and therefore

$$1 \stackrel{P}{=} \sum_{x \in \Omega'_0} 1_{X=x}.$$

( $\Leftarrow$ ) If Equation (5.33) holds and  $\Omega'_0 \subset \Omega'_X$  is finite or countable with  $\{x\} \in \mathcal{A}'_X$  for all  $x \in \Omega'_0$ , then  $1 \stackrel{P}{=} \sum_{x \in \Omega'_0} 1_{X=x}$  and therefore

$$1 = P\left(1 = \sum_{x \in \Omega'_0} 1_{X=x}\right) = P(X \in \Omega'_0) = P_X(\Omega'_0).$$

(ii) If  $X$  is real-valued, then  $X$  is discrete if and only if there is a finite or countable  $\Omega' \subset \mathbb{R}$  such that

$$\begin{aligned} X &= X \cdot 1 \\ &\stackrel{P}{=} X \cdot \sum_{x \in \Omega'} 1_{X=x} && \text{[(i)]} \\ &= \sum_{x \in \Omega'} X \cdot 1_{X=x} \\ &= \sum_{x \in \Omega'} x \cdot 1_{X=x}. && [1_{X=x}(\omega) = 0 \text{ if } X(\omega) \neq x, \text{ (5.6)}] \end{aligned}$$

**Proof of Corollary 5.63**

$\Rightarrow$ . If  $X$  is discrete and  $\Omega'_0 \subset \Omega'_X$  (see Def. 5.56), i. e., if  $\Omega'_0$  is finite or countable and  $P_X(\Omega'_0) = 1$ , then  $\Omega'_> \subset \Omega'_0$ , because  $P_X(\Omega'_X \setminus \Omega'_0) = 0$ . Hence,  $\Omega'_>$  is finite or countable as well, which proves (a). Furthermore, for finite or countable sets  $\Omega'_0, \Omega'_>$ :

$$P_X(\Omega'_0 \setminus \Omega'_>) = \sum_{x \in \Omega'_0 \setminus \Omega'_>} P(X=x) = 0. \quad \text{[Def. of } \Omega'_> \text{]}$$

Hence,

$$P_X(\Omega'_X \setminus \Omega'_>) = P_X(\Omega'_X \setminus \Omega'_0) + P_X(\Omega'_0 \setminus \Omega'_>) = 0$$

and, according to (5.11),

$$1_{X \in (\Omega'_X \setminus \Omega'_>)} \stackrel{P}{=} 0. \quad (5.72)$$

Therefore,

$$X = 1_{X \in \Omega'_>} \cdot X + 1_{X \in (\Omega'_X \setminus \Omega'_>)} \cdot X \quad [1 = 1_{X \in \Omega'_>} + 1_{X \in (\Omega'_X \setminus \Omega'_>)}]$$

$$\begin{aligned}
& \stackrel{p}{=} X \cdot \sum_{x \in \Omega'_>} 1_{X=x} && [(5.72), \Omega'_> \text{ is finite or countable}] \\
& = \sum_{x \in \Omega'_>} X \cdot 1_{X=x} \\
& = \sum_{x \in \Omega'_>} x \cdot 1_{X=x}. && [X \cdot 1_{X=x} = x \cdot 1_{X=x}]
\end{aligned}$$

←. This is an immediate implication of Lemma 5.61 (ii).

### **Proof of Lemma 5.66**

If  $x_i \in \Omega'_i$  and  $\Omega'_{0,x_i} = \emptyset$ , then  $p_{X_i}(x_i) = P(X_i = x_i) = 0$ . If  $\Omega'_{0,x_i} \neq \emptyset$ , then it is finite or countable, because  $\Omega'_0$  is finite or countable. Then, using  $P_X(\Omega'_0) = 1$ ,

$$\begin{aligned}
& p_{X_i}(x_i) \\
& = P(X_i = x_i) \\
& = P(X_1 \in \Omega'_1, \dots, X_{i-1} \in \Omega'_{i-1}, X_i = x_i, X_{i+1} \in \Omega'_{i+1}, \dots, X_n \in \Omega'_n) && [\text{Box 4.1 (viii)}] \\
& = P((X_1, \dots, X_n) \in (\Omega'_1 \times \dots \times \Omega'_{i-1} \times \{x_i\} \times \Omega'_{i+1} \times \dots \times \Omega'_n)) \\
& = P((X_1, \dots, X_n) \in ((\Omega'_1 \times \dots \times \Omega'_{i-1} \times \{x_i\} \times \Omega'_{i+1} \times \dots \times \Omega'_n) \cap \Omega'_0)) && [\text{Box 4.1 (viii)}] \\
& = \sum_{(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) \in \Omega'_{0,x_i}} P(X_1 = x_1, \dots, X_{i-1} = x_{i-1}, X_i = x_i, X_{i+1} = x_{i+1}, \dots, X_n = x_n) && [\text{Box 4.1 (x)}] \\
& = \sum_{(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) \in \Omega'_{0,x_i}} p_X(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n). && [(5.31)]
\end{aligned}$$

### **Proof of Lemma 5.68**

If  $X = (X_1, \dots, X_n)$  and  $X_1, \dots, X_n$  are independent, then, for all  $(x_1, \dots, x_n) \in \prod_{i=1}^n \Omega'_{0,i}$ ,

$$p_X(x_1, \dots, x_n) = P((X_1, \dots, X_n) = (x_1, \dots, x_n)) \quad [(5.31)]$$

$$= P(X_1 = x_1, \dots, X_n = x_n)$$

$$= P(X_1 = x_1) \cdot \dots \cdot P(X_n = x_n) \quad [(5.28)]$$

$$= p_{X_1}(x_1) \cdot \dots \cdot p_{X_n}(x_n). \quad [(5.31)]$$

For all  $A'_1 \in \mathcal{A}'_1, \dots, A'_n \in \mathcal{A}'_n$ ,

$$(A'_1 \times \dots \times A'_n) \cap (\Omega'_{0,1} \times \dots \times \Omega'_{0,n}) = (A'_1 \cap \Omega'_{0,1}) \times \dots \times (A'_n \cap \Omega'_{0,n}). \quad (5.73)$$

Now assume that Equation (5.38) holds and define  $\Omega'_{sn} := \prod_{i=1}^n \Omega'_{0,i}$ . Then,

$$\begin{aligned}
& P(X_1 \in A'_1, \dots, X_n \in A'_n) \\
& = P[(X_1, \dots, X_n) \in (A'_1 \times \dots \times A'_n)] \\
& = P[(X_1, \dots, X_n) \in (A'_1 \times \dots \times A'_n) \cap \Omega'_{sn}] && [P_X(\Omega'_{sn}) = 1, \text{Box 4.1, (viii)}]
\end{aligned}$$

$$= \sum_{(x_1, \dots, x_n) \in (A'_1 \times \dots \times A'_n) \cap \Omega'_{sn}} P[(X_1, \dots, X_n) = (x_1, \dots, x_n)] \quad [\text{Box 4.1 (x)}]$$

$$= \sum_{(x_1, \dots, x_n) \in (A'_1 \times \dots \times A'_n) \cap \Omega'_{sn}} p_X(x_1, \dots, x_n) \quad [(5.36)]$$

$$= \sum_{(x_1, \dots, x_n) \in (A'_1 \times \dots \times A'_n) \cap \Omega'_{sn}} p_{X_1}(x_1) \cdot \dots \cdot p_{X_n}(x_n) \quad [(5.38)]$$

$$= \left( \sum_{x_1 \in A'_1 \cap \Omega'_{0,1}} p_{X_1}(x_1) \right) \cdot \dots \cdot \left( \sum_{x_n \in A'_n \cap \Omega'_{0,n}} p_{X_n}(x_n) \right) \quad [(5.73)]$$

$$= P(X_1 \in A'_1) \cdot \dots \cdot P(X_n \in A'_n), \quad [P_{X_i}(\Omega'_{0,i}) = 1, \text{ Box 4.1 (viii), (x)}]$$

which implies independence of  $X_1, \dots, X_n$  [see (5.28)].

### **Proof of Theorem 5.77**

For  $A' \in \mathcal{A}'_X$ ,

$$P_X(A') = \sum_{x \in A' \cap \Omega'_0} P_X(\{x\}) \quad [P_X(\Omega'_0) = 1, \text{ Box 4.1 (viii), (x)}]$$

$$= \sum_{x \in A' \cap \Omega'_0} p_X(x) \quad [(5.31)]$$

$$= \sum_{x \in \Omega'_0} p_X(x) \cdot 1_{A'}(x)$$

$$= \sum_{x \in \Omega'_0} \int p_X \cdot 1_{A'} d\delta_x \quad [(3.23)]$$

$$= \int 1_{A'} \cdot p_X d\left( \sum_{x \in \Omega'_0} \delta_x \right) \quad [(3.53)]$$

$$= \int_{A'} p_X d\left( \sum_{x \in \Omega'_0} \delta_x \right) \quad [(3.30)]$$

$$= \int_{A'} p_X d\mu. \quad [(5.42)]$$

Hence, according to Definition 5.70, the probability function  $p_X$  is a density of  $P_X$  with respect to the measure  $\mu$  on the measurable space  $(\Omega'_X, \mathcal{A}'_X)$  and  $p_X = \frac{dP}{d\mu}$  (see Th. 3.72).

### **Proof of Lemma 5.78**

Let  $A' \in \mathcal{A}'_X$  with  $\mu(A') = 0$ . Then

$$\mu \otimes \nu(A' \times \Omega'_Y) = \mu(A') \cdot \nu(\Omega'_Y) = 0 \cdot \nu(\Omega'_Y) = 0,$$

and this holds even if  $\nu(\Omega'_Y) = \infty$ . Together with  $P_{X,Y} \ll_{\mathcal{A}'_X \otimes \mathcal{A}'_Y} \mu \otimes \nu$  this implies

$$P_X(A') = P(X \in A', Y \in \Omega'_Y) = P_{X,Y}(A' \times \Omega'_Y) = 0.$$

Therefore,  $P_X \ll_{\mathcal{A}'_X} \mu$ . The proof for  $P_Y$  is analogous.

**Proof of Lemma 5.79**

For all  $A' \in \mathcal{A}'_X$ ,

$$\begin{aligned}
& \int \mathbf{1}_{A'}(x) \left( \int f_{X,Y}(x,y) \nu(dy) \right) \mu(dx) \\
&= \int \mathbf{1}_{A'}(x) f_{X,Y}(x,y) \mu \otimes \nu [d(x,y)] \quad [\text{Th. 3.76}] \\
&= \int \mathbf{1}_{A'}(x) P_{X,Y}[d(x,y)] \quad [\text{Th. 3.67}] \\
&= P_{X,Y}(A' \times \Omega'_Y) \quad [(3.30), (3.8)] \\
&= P_X(A'). \quad [(5.21)]
\end{aligned}$$

Theorem 3.65 then implies the lemma. The proof for  $f_Y$  is analogous.

**Proof of Corollary 5.92**

Note that

$$\mathbb{R}^{n-1} = \bigcup_{m=1}^{\infty} ]-\infty, m]^{n-1}. \quad (5.74)$$

For all  $X_i \in \mathbb{R}$ ,

$$\begin{aligned}
F_{X_i}(x_i) &= P(X_i \leq x_i) \\
&= P_{X_1, \dots, X_n}(\mathbb{R} \times \dots \times \mathbb{R} \times ]-\infty, x_i] \times \mathbb{R} \times \dots \times \mathbb{R}) \quad [(5.21)] \\
&= P(X_1 \in \mathbb{R}, \dots, X_{i-1} \in \mathbb{R}, X_i \in ]-\infty, x_i], X_{i+1} \in \mathbb{R}, \dots, X_n \in \mathbb{R}) \quad [(5.2)] \\
&= P\left((X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n) \in \mathbb{R}^{n-1}, X_i \in ]-\infty, x_i]\right) \\
&= P\left(X_j \in \bigcup_{m_j=1}^{\infty} ]-\infty, m_j], j \in \{1, \dots, i-1, i+1, \dots, n\}, X_i \in ]-\infty, x_i]\right) \quad [(5.74)] \\
&= \lim_{m_j \rightarrow \infty, j \neq i} P\left(X_j \in ]-\infty, m_j], j \in \{1, \dots, i-1, i+1, \dots, n\}, X_i \in ]-\infty, x_i]\right) \\
&\quad [\text{Th. 1.68 (i)}] \\
&= \lim_{m_j \rightarrow \infty, j \neq i} F_{X_1, \dots, X_n}(m_1, \dots, m_{i-1}, x_i, m_{i+1}, \dots, m_n) \quad [(5.56)] \\
&= \lim_{x_j \rightarrow \infty, j \neq i} F_{X_1, \dots, X_n}(x_1, \dots, x_n).
\end{aligned}$$

The limits exist, because  $F_{X_1, \dots, X_n}$  is monotone in all coordinates.

**Proof of Corollary 5.93**

(a)  $\Rightarrow$  (b) For all  $(x_1, \dots, x_n) \in \mathbb{R}^n$ ,

$$\begin{aligned}
F_{X_1, \dots, X_n}(x_1, \dots, x_n) &= P_{X_1, \dots, X_n}([-\infty, x_1] \times \dots \times ]-\infty, x_n]) \quad [(5.56)] \\
&= P_{X_1} \otimes \dots \otimes P_{X_n}([-\infty, x_1] \times \dots \times ]-\infty, x_n]) \quad [\text{Lemma 5.49 and (a)}] \\
&= P_{X_1}([-\infty, x_1]) \cdot \dots \cdot P_{X_n}([-\infty, x_n]) \quad [(1.50)] \\
&= F_{X_1}(x_1) \cdot \dots \cdot F_{X_n}(x_n). \quad [\text{Def. 5.81}]
\end{aligned}$$

(b)  $\Rightarrow$  (a) For all  $(x_1, \dots, x_n) \in \mathbb{R}^n$ ,

$$\begin{aligned}
 & P_{X_1, \dots, X_n}([-\infty, x_1] \times \dots \times ]-\infty, x_n]) \\
 &= F_{X_1, \dots, X_n}(x_1, \dots, x_n) && [(5.56)] \\
 &= F_{X_1}(x_1) \cdot \dots \cdot F_{X_n}(x_n) && [(b)] \\
 &= P_{X_1}([-\infty, x_1]) \cdot \dots \cdot P_{X_n}([-\infty, x_n]) && [\text{Def. 5.81}] \\
 &= P_{X_1} \otimes \dots \otimes P_{X_n}([-\infty, x_1] \times \dots \times ]-\infty, x_n]). && [(1.50)]
 \end{aligned}$$

Because  $\{]-\infty, x_1] \times \dots \times ]-\infty, x_n] : (x_1, \dots, x_n) \in \mathbb{R}^n\}$  is a  $\cap$ -stable generating system of  $\mathcal{B}_n$  (see (1.21)), Theorem 1.72 yields  $P_{X_1, \dots, X_n} = P_{X_1} \otimes \dots \otimes P_{X_n}$ . Applying Lemma 5.49 then completes the proof.

### **Proof of Corollary 5.100**

(i)  $(X_1, \dots, X_n)$  are independent  $\Rightarrow f_{X_1, \dots, X_n}$  in (5.70) is a joint density.

Because  $f_{X_1, \dots, X_n}$  defined in (5.70) is nonnegative and integrable with respect to  $\lambda_n$ , we can conclude that  $f_{X_1, \dots, X_n} \odot \lambda_n$  defines a finite measure on  $(\mathbb{R}^n, \mathcal{B}_n)$  (see Th. 3.65). Furthermore, if  $X_1, \dots, X_n$  are independent and  $f_{X_i}$  is a density of  $X_i$  for all  $i = 1, \dots, n$ , then, for all  $(x_1, \dots, x_n) \in \mathbb{R}^n$  and  $B = \prod_{i=1}^n ]-\infty, x_i]$ ,

$$\begin{aligned}
 & \int_B f_{X_1, \dots, X_n}(t_1, \dots, t_n) \lambda_n[d(t_1, \dots, t_n)] \\
 &= \int_B f_{X_1}(t_1) \cdot \dots \cdot f_{X_n}(t_n) \lambda_n[d(t_1, \dots, t_n)] && [(5.70)] \\
 &= \prod_{i=1}^n \int_{]-\infty, x_i]} f_{X_i}(t_i) \lambda(dt_i) && [\text{Th. 3.76}] \\
 &= \prod_{i=1}^n F_{X_i}(x_i) && [(5.62)] \\
 &= F_{X_1, \dots, X_n}(x_1, \dots, x_n) && [\text{Cor. 5.93}] \\
 &= P_{X_1, \dots, X_n}(B).
 \end{aligned}$$

This shows that  $f_{X_1, \dots, X_n} \odot \lambda_n(B) = P_{X_1, \dots, X_n}(B)$ , and this implies that  $f_{X_1, \dots, X_n}$  defined in (5.70) is a density of  $(X_1, \dots, X_n)$  with respect to  $\lambda_n$  (see Def. 5.70).

(ii)  $f_{X_1, \dots, X_n}$  in (5.70) is a density of  $(X_1, \dots, X_n) \Rightarrow X_1, \dots, X_n$  are independent. If Equation (5.70) holds and  $B = \prod_{i=1}^n ]-\infty, x_i]$ , then for all  $(x_1, \dots, x_n) \in \mathbb{R}^n$

$$\begin{aligned}
 & F_{X_1, \dots, X_n}(x_1, \dots, x_n) \\
 &= \int_B f_{X_1, \dots, X_n}(t_1, \dots, t_n) \lambda_n[d(t_1, \dots, t_n)] && [(5.67)] \\
 &= \int_{]-\infty, x_n]} \dots \int_{]-\infty, x_1]} f_{X_1}(t_1) \cdot \dots \cdot f_{X_n}(t_n) \lambda(dt_1) \dots \lambda(dt_n) && [\text{Th. 3.76}] \\
 &= \prod_{i=1}^n \int_{]-\infty, x_i]} f_{X_i}(t_i) \lambda(dt_i) \\
 &= \prod_{i=1}^n F_{X_i}(x_i). && [(5.62)]
 \end{aligned}$$

Now, Corollary 5.93 implies that  $X_1, \dots, X_n$  are independent.

### 5.9 Exercises

▷ **Exercise 5-1** Consider the random variable  $X$  defined in Example 2.34. Which are the elements  $\omega$  in the inverse image  $X^{-1}(\{1\})$  and which are the probabilities of the events  $\{\omega\}$ ?

▷ **Exercise 5-2** Consider again the random variable  $X$  defined in Example 2.34. What are the values of the distribution of  $X$  and the distribution of  $U$ ?

▷ **Exercise 5-3** Consider Example 5.9. Show that  $1_{X \in A'}$  is a random variable on  $(\Omega, \mathcal{A}, P)$  and that  $1_{X \in A'} = 1_{A'}(X) = 1_{A'} \circ X$ .

▷ **Exercise 5-4** Consider Example 5.15 and show that  $X_1$  is prior to  $X_1 \cdot X_2$  in  $(\mathcal{F}_1, \mathcal{F}_2)$ .

▷ **Exercise 5-5** Show that  $X \stackrel{p_B}{=} Y$  for  $B = \{\omega_1, \omega_2, \omega_3\}$  and that  $X \stackrel{p}{=} Y$  does *not* hold in Example 5.23.

▷ **Exercise 5-6** Consider the random variable  $(U, X): \Omega \rightarrow \Omega_U \times \{0, 1\}$  defined in Example 2.34 (p. 53). Which are the elements  $\omega$  in the inverse image  $(U, X)^{-1}(\{(Joe, 1)\})$  and which are the probabilities to the events  $\{\omega\}$ ?

▷ **Exercise 5-7** Show that the random variables  $U$  and  $X$  presented in Table 2.2 (p. 53) are independent.

▷ **Exercise 5-8** Let  $(\Omega, \mathcal{A}, P)$  be a probability space,  $B \in \mathcal{B}$ , and  $\mathcal{C} \subset \mathcal{A}$  a  $\sigma$ -algebra. Prove that the following propositions are equivalent to each other:  $1_B \perp\!\!\!\perp_P \mathcal{C}$ ,  $\sigma(\{B\}) \perp\!\!\!\perp_P \mathcal{C}$ ,  $\{B\} \perp\!\!\!\perp_P \mathcal{C}$ ,  $B \perp\!\!\!\perp_P \mathcal{C}$ .

▷ **Exercise 5-9** In Example 5.50 we showed that  $X$  and  $U$  are independent and in Example 5.54 we defined the mappings  $f: \{0, 1, 2\} \rightarrow \{0, 1\}$  and  $g: \Omega_U \rightarrow \{male, female\}$ . Use Equation (5.24) to show that the mappings  $f(X)$  (control vs. any of the two treatments) and  $g(U)$  (sex) are independent as well.

▷ **Exercise 5-10** In Example 5.85 we specified the values of the distribution function  $F_X$  for the random variable *number of flipping heads*. Use these values to compute all eight values of the distribution  $P_X$ .

### Solutions

▷ **Solution 5-1** The inverse image of the set  $\{1\}$  under  $X$  is

$$X^{-1}(\{1\}) = \{(Joe, yes, -), (Joe, yes, +), (Ann, yes, -), (Ann, yes, +)\}.$$

The events  $\{\omega\}$ ,  $\omega \in X^{-1}(\{1\})$  have the probabilities  $P[\{(Joe, yes, -)\}] = .04$ ,  $P[\{(Joe, yes, +)\}] = .16$ ,  $P[\{(Ann, yes, -)\}] = .12$ , and  $P[\{(Ann, yes, +)\}] = .08$ .

▷ **Solution 5-2** The random variable  $X$  (the treatment variable) has the following distribution:

$$\begin{aligned} P_X(\{1\}) &= P[X^{-1}(\{1\})] = .40, & P_X(\{0\}) &= P[X^{-1}(\{0\})] = .60, \\ P_X(\Omega') &= P[X^{-1}(\Omega')] = 1, & P_X(\emptyset) &= P[X^{-1}(\emptyset)] = 0, \end{aligned}$$

whereas the distribution of the random variable  $U$  (the observational-unit variable), is:

$$\begin{aligned} P_U(\{Joe\}) &= P[U^{-1}(\{Joe\})] = .50 & P_U(\{Ann\}) &= P[U^{-1}(\{Ann\})] = .50 \\ P_U(\Omega_U) &= P[U^{-1}(\Omega_U)] = 1 & P_U(\emptyset) &= P[U^{-1}(\emptyset)] = 0. \end{aligned}$$

▷ **Solution 5-3** Let  $A' \in \mathcal{A}'_X$  and consider the indicator function  $1_{X \in A'}: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$ .  
Measurability: For all  $B \in \mathcal{B}$ ,

$$\begin{aligned} (1_{X \in A'})^{-1}(B) &= (1_{X^{-1}(A')})^{-1}(B) && [(5.6)] \\ &= \{\omega \in \Omega: 1_{X^{-1}(A')}(\omega) \in B\} && [(2.2)] \\ &= \begin{cases} \emptyset, & \text{if } 0 \notin B, 1 \notin B, \\ X^{-1}(A'), & \text{if } 0 \notin B, 1 \in B \\ \Omega \setminus X^{-1}(A'), & \text{if } 0 \in B, 1 \notin B \\ \Omega, & \text{if } \{0, 1\} \subset B, \end{cases} \end{aligned}$$

and all these sets are elements of  $\mathcal{A}$ , because  $X$  is assumed to be a random variable.

Furthermore, for all  $\omega \in \Omega$ ,

$$\begin{aligned} 1_{X \in A'}(\omega) &= 1_{X^{-1}(A')}(\omega) && [(5.6)] \\ &= \begin{cases} 1, & \text{if } \omega \in X^{-1}(A') \\ 0, & \text{if } \omega \notin X^{-1}(A') \end{cases} \\ &= \begin{cases} 1, & \text{if } X(\omega) \in A' \\ 0, & \text{if } X(\omega) \notin A' \end{cases} && [(2.2)] \\ &= 1_{A'}[X(\omega)] \\ &= 1_{A'} \circ X(\omega). && [(2.25)] \end{aligned}$$

▷ **Solution 5-4**

The filtration  $(\mathcal{F}_1, \mathcal{F}_2)$  is defined by  $\mathcal{F}_1 := \sigma(X_1)$  and  $\mathcal{F}_2 := \sigma(X_1, X_2)$ . Hence,

$$\sigma(X_1) \subset \mathcal{F}_1 = \{\Omega, \emptyset, \{(h, h), (h, t)\}, \{(t, h), (t, t)\}\}$$

In contrast,

$$\sigma(X_1 \cdot X_2) = \{\Omega, \emptyset, \{(h, h)\}, \{(h, t), (t, h), (t, t)\}\} \not\subset \mathcal{F}_1.$$

However,  $\sigma(X_1 \cdot X_2) \subset \mathcal{F}_2$ , because  $\mathcal{F}_2 = \mathcal{P}(\Omega)$ . Hence, according to Definition 5.13 (i),  $X_1$  is prior to  $X_1 \cdot X_2$  in the filtration  $(\mathcal{F}_1, \mathcal{F}_2)$ .

▷ **Solution 5-5** In order to prove  $X \stackrel{P^B}{=} Y$  for  $B = \{\omega_1, \omega_2, \omega_3\}$ , it is sufficient to show  $P^B(A) = 0$ , where  $A := \{\omega \in \Omega: X(\omega) \neq Y(\omega)\} \in \mathcal{A} = \mathcal{P}(\Omega)$ . In this example,  $A = \{\omega_3, \omega_4\}$ . Now  $P(\{\omega_3\}) = 0$ , which implies

$$P^B(\{\omega_3\}) = \frac{P(B \cap \{\omega_3\})}{P(B)} = \frac{P(\{\omega_3\})}{P(B)} = \frac{0}{P(B)} = 0.$$

Furthermore,

$$P^B(\{\omega_4\}) = \frac{P(B \cap \{\omega_4\})}{P(B)} = \frac{P(\emptyset)}{P(B)} = \frac{0}{P(B)} = 0.$$

Additivity of the measure  $P^B$  then implies  $P^B(A) = 0$ .

In order to prove that  $X \stackrel{P}{=} Y$  does not hold, it suffices to show  $P(A) \neq 0$ . Now  $P(\{\omega_3\}) = 0$  and  $P(\{\omega_4\}) = .50$ . Therefore,  $P(A) = P(\{\omega_4\}) = .50 > 0$ .

▷ **Solution 5-6** The inverse image of the set  $\{1\}$  under  $X$  is

$$X^{-1}(\{1\}) = \{(Joe, yes, -), (Joe, yes, +), (Ann, yes, -), (Ann, yes, +)\}.$$

The events  $\{\omega\}$ ,  $\omega \in X^{-1}(\{1\})$  have the probabilities  $P[\{(Joe, yes, -)\}] = .04$ ,  $P[\{(Joe, yes, +)\}] = .16$ ,  $P[\{(Ann, yes, -)\}] = .12$ , and  $P[\{(Ann, yes, +)\}] = .08$ .

▷ **Solution 5-7** In Example 4.43, we already showed that the events  $A := \{Joe\} \times \Omega_X \times \Omega_Y$  (that Joe is sampled) and  $B := \Omega_U \times \{yes\} \times \Omega_Y$  (that the person sampled is treated) are independent. According to Box 4.2 (iii), this implies that the  $\sigma$ -algebras  $\{A, A^c, \Omega, \emptyset\}$  and  $\{B, B^c, \Omega, \emptyset\}$  are independent as well. If  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega', \mathcal{A}')$  and  $U: (\Omega, \mathcal{A}, P) \rightarrow (\Omega_U, \mathcal{A}_U)$  are the random variables defined Table 2.2 (p. 53), then  $X^{-1}(\mathcal{A}') = \{A, A^c, \Omega, \emptyset\}$  and  $U^{-1}(\mathcal{A}_U) = \{B, B^c, \Omega, \emptyset\}$  are the  $\sigma$ -algebras generated by  $U$  and  $X$ , respectively. Hence, in this example,  $X$  and  $U$  are independent.

▷ **Solution 5-8** First of all,

$$\{B\} \perp\!\!\!\perp_P \mathcal{C} \Leftrightarrow B \perp\!\!\!\perp_P \mathcal{C} \quad [\text{Rem. 4.41}].$$

Furthermore,  $\sigma(1_B) = \{\Omega, \emptyset, B, B^c\}$  (see Example 2.31) and  $\sigma(\{B\}) = \{\Omega, \emptyset, B, B^c\}$ . Hence,  $\sigma(1_B) = \sigma(\{B\})$ , and this implies

$$1_B \perp\!\!\!\perp_P \mathcal{C} \Leftrightarrow \sigma(\{B\}) \perp\!\!\!\perp_P \mathcal{C} \quad [\text{Rem. 5.44}].$$

Finally,

$$\sigma(\{B\}) \perp\!\!\!\perp_P \mathcal{C} \Leftrightarrow \{B\} \perp\!\!\!\perp_P \mathcal{C} \quad [\text{Box 4.2 (ii)}].$$

▷ **Solution 5-9** Consider the two events  $A := f(X)^{-1}(\{0\})$  and  $B := g(U)^{-1}(\{male\})$ .

$$P(A \cap B) = \frac{40}{99}$$

and

$$P(A) \cdot P(B) = \frac{60}{99} \cdot \frac{2}{3} = \frac{40}{99}.$$

Because  $f(X)^{-1}(\mathcal{P}(\{0, 1\})) = \{A, A^c, \Omega, \emptyset\}$  and  $g(U)^{-1}(\mathcal{P}(\{male, female\})) = \{B, B^c, \Omega, \emptyset\}$ , this proves that  $f(X)$  and  $g(U)$  are independent (see Exercise 5-8).

▷ **Solution 5-10** We consider all elements  $B \in \mathcal{B}$  of the Borel  $\sigma$ -algebra on  $\mathbb{R}$  and assign the following values:

$$P_X(B) = \begin{cases} F_X(-1) = 0, & \text{if } 0, 1, 2 \notin B \\ F_X(0) = \frac{1}{4}, & \text{if } 0 \in B, 1, 2 \notin B \\ F_X(1) = \frac{3}{4}, & \text{if } 0, 1 \in B, 2 \notin B \\ F_X(2) = 1, & \text{if } 0, 1, 2 \in B \\ F_X(1) - F_X(0) = \frac{2}{4}, & \text{if } 1 \in B, 0, 2 \notin B \\ F_X(2) - F_X(0) = \frac{3}{4}, & \text{if } 1, 2 \in B, 0 \notin B \\ F_X(2) - F_X(1) = \frac{1}{4}, & \text{if } 2 \in B, 0, 1 \notin B \\ F_X(2) - F_X(1) + F_X(0) = \frac{1}{2}, & \text{if } 0, 2 \in B, 1 \notin B. \end{cases}$$

## Chapter 6

# Expectation, Variance, and Other Moments

In chapter 4 we introduced a probability measure as a special finite measure and in chapter 5 we defined a random variable as a measurable mapping on a probability space. In this chapter we will translate integration theory (see ch. 3) to probability theory introducing *expectations* of numerical random variables and other important concepts that are special expectations: *central* and *noncentral moments*, and *variances*. Even *covariances* and *correlations* are special expectations (see ch. 7). All these quantities describe important properties of random variables, although, in general, they do not determine the complete distribution.

### 6.1 Expectation

#### 6.1.1 Definition

Reading the following definition remember that a random variable  $Y$  is called *quasi-integrable* with respect to  $P$  if  $\int Y^+ dP$  or  $\int Y^- dP$  are finite, where  $Y^+$  and  $Y^-$  denote the positive and negative parts of  $Y$ , respectively (see Rem. 2.62 and Def. 3.28).

#### **Definition 6.1 (Expectation)**

Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  be a numerical random variable that is quasi-integrable with respect to  $P$ . Then we define

$$E(Y) := \int Y dP, \quad (6.1)$$

call it the *expectation* of  $Y$  (w.r.t.  $P$ ), and say that it *exists*. Instead of *expectation* w.r.t.  $P$  we also use the term  *$P$ -expectation*.

**Remark 6.2 (Existence of the Expectation)** Note that  $E(Y)$  can be infinite. Furthermore, if  $E(Y)$  exists we also say that  $Y$  is a random variable *with expectation*  $E(Y)$ . If  $Y$  is not quasi-integrable with respect to  $P$  and therefore also not  $P$ -integrable, then we say that the expectation of  $Y$  with respect to  $P$  does *not exist*.  $\triangleleft$

**Remark 6.3 (Notation and Synonymous Terms)** A synonym for expectation is *expectation value*. The reference to the measure  $P$  is usually omitted if the context is unambiguous. If we consider the expectation with respect to another probability measure on  $(\Omega, \mathcal{A})$ , e. g., the conditional-probability measure  $P^B$  (see Def. 4.29), then we adapt the notation correspondingly:

$$E^B(Y) := \int Y dP^B. \quad (6.2)$$

*Expectation with respect to  $P^B$*  is used synonymously to  *$P^B$ -expectation*.  $\triangleleft$

**Remark 6.4 (Random Variables With Finite Expectations)** Without substantial loss of generality, we will often assume that a random variable  $Y$  is *real-valued* if its expectation is finite. According to Remark 3.42, if the random variable  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  has a finite expectation, then there is random variable  $Y^*: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  such that  $Y \stackrel{P}{=} Y^*$ .  $\triangleleft$

### 6.1.2 Expectation of a Discrete Random Variable

In this section we use the notation  $P(Y=y_i) := P[Y^{-1}(\{y_i\})]$  introduced in Remark 5.4.

**Remark 6.5 (Random Variable With a Finite Number of Real Values)** Suppose  $y_1, \dots, y_n \in \mathbb{R}$ ,  $n \in \mathbb{N}$ , denote all (negative, 0, or positive) values of a real-valued random variable  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$ , let  $P_Y$  denote the distribution, and  $p_Y$  the probability function of  $Y$ . Then the expectation  $E(Y)$  exists, and

$$E(Y) = \sum_{i=1}^n y_i \cdot P(Y=y_i) = \sum_{i=1}^n y_i \cdot P_Y(\{y_i\}) = \sum_{i=1}^n y_i \cdot p_Y(y_i) \quad (6.3)$$

[see Cor. 3.59, (3.45), Eq. (5.34), and Def. 5.56]. Hence, if  $Y$  has only a finite number  $n$  of values, then its expectation is simply the sum of its values, each one weighted by its probability  $P(Y=y_i) = P_Y(\{y_i\}) = p_Y(y_i)$ .  $\triangleleft$

**Example 6.6 (Expectation of an Indicator)** If  $(\Omega, \mathcal{A}, P)$  is a probability space and  $1_A$  is the indicator of  $A \in \mathcal{A}$ , then

$$E(1_A) = 0 \cdot P(1_A=0) + 1 \cdot P(1_A=1) = P(1_A=1) = P(A) \quad (6.4)$$

is the expectation of  $1_A$  (see also Example 3.14). Considering the event  $\{Y=y\}$  and using the notation  $1_{Y=y} := 1_{\{Y=y\}}$ , this yields

$$E(1_{Y=y}) = P(Y=y). \quad (6.5)$$

$\triangleleft$

**Example 6.7 (Joe and Ann With Random Assignment – continued)** In Example 1.9 we defined the set

$$B = \{(Joe, yes, -), (Joe, yes, +), (Ann, yes, -), (Ann, yes, +)\} = \Omega_U \times \{yes\} \times \Omega_Y,$$

the event that the drawn person is treated, and the set

$$C = \{(Joe, no, +), (Joe, yes, +), (Ann, no, +), (Ann, yes, +)\} = \Omega_U \times \Omega_X \times \{+\},$$

the event that  $\{+\}$  (success) occurs, irrespective of which person is drawn and whether or not the person is treated. In Table 2.2 (p. 53) we assigned probabilities to each elementary event  $\{\omega\}$ ,  $\omega \in \Omega$  and defined  $X := 1_B$ , the treatment variable, as well as  $Y := 1_C$  the outcome variable. Applying Equation (6.4) to the indicator  $1_B$  yields:

$$\begin{aligned} E(X) &= E(1_B) = P(B) \\ &= P[\{(Joe, yes, -)\}] + P[\{(Joe, yes, +)\}] + P[\{(Ann, yes, -)\}] + P[\{(Ann, yes, +)\}] \\ &= .04 + .16 + .12 + .08 = .40. \end{aligned}$$

Similarly, for the indicator  $1_C$ , we obtain

$$\begin{aligned} E(Y) &= E(1_C) = P(C) \\ &= P[\{(Joe, no, +)\}] + P[\{(Joe, yes, +)\}] + P[\{(Ann, no, +)\}] + P[\{(Ann, yes, +)\}] \\ &= .21 + .16 + .06 + .08 = .51. \end{aligned}$$

&lt;

**Example 6.8 (Tossing a Dice – continued)** In Example 3.6 we considered the random variable  $X = \text{number of dots}$ . In this example we specified the probability space  $(\Omega, \mathcal{A}, P)$  by  $\Omega := \{\omega_1, \dots, \omega_6\}$ ,  $\mathcal{A} := \mathcal{P}(\Omega)$ , and  $P(\{\omega_1\}) = \dots = P(\{\omega_6\}) = 1/6$ . Because  $P(X=x_i) = P_X(\{i\}) = P(\{\omega_i\})$ ,  $i = 1, \dots, 6$ , Equation (6.3) yields

$$E(X) = \sum_{i=1}^6 i \cdot P(X=i) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \dots + 6 \cdot \frac{1}{6} = 3.5.$$

&lt;

**Remark 6.9 (Random Variable With a Countable Number of Real Values)** Let  $y_1, y_2, \dots \in \mathbb{R}$  denote the values of a real-valued random variable  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  and suppose that the expectation of  $Y$  exists. Then

$$E(Y) = \sum_{i=1}^{\infty} y_i \cdot P(Y=y_i) = \lim_{n \rightarrow \infty} \sum_{i=1}^n y_i \cdot P(Y=y_i) \quad (6.6)$$

[see (3.45), Eq. (5.34), and Th. 3.60]. Examples in which the expectation of a random variable is the ‘infinite sum’ of its values weighted by their probabilities are random variables with a Poisson distribution and with a geometric distribution (see Theorems 8.16 and 8.22). <

**Example 6.10 (A Discrete Random Variable Without Expectation)** Suppose that  $\Omega = \mathbb{N}_0 = \{0, 1, 2, \dots\}$  and consider the random variable  $Y: (\Omega, \mathcal{P}(\Omega), P) \rightarrow (\mathbb{R}, \mathcal{B})$  defined by

$$Y(i) = y_i = (-1)^i i! e, \quad \forall i \in \mathbb{N}_0, \quad (6.7)$$

with

$$P(Y=y_i) = \frac{1}{e} \cdot \frac{1}{i!}, \quad \forall i \in \mathbb{N}_0. \quad (6.8)$$

Note that  $e = \sum_{i=0}^{\infty} \frac{1}{i!}$ . Hence, dividing both sides by  $e$  yields  $\sum_{i=0}^{\infty} \frac{1}{i!e} = 1$ . Therefore, Equation (6.8) specifies a probability distribution. Now consider

$$\sum_{i=0}^n y_i \cdot P(Y=y_i) = \sum_{i=0}^n (-1)^i = \begin{cases} 1, & \text{if } n \text{ is even} \\ 0, & \text{if } n \text{ is odd.} \end{cases} \quad (6.9)$$

Obviously, in this example, the limit

$$\lim_{n \rightarrow \infty} \sum_{i=0}^n y_i \cdot P(Y=y_i)$$

occurring in Equation (6.6) does not exist. Therefore, according to Definitions 6.1 and 3.28  $E(Y)$  is not defined.  $\triangleleft$

### 6.1.3 Computing the Expectation Using a Density

According to the following theorem, the expectation of a continuous real-valued random variable can also be computed using its density  $f_Y$  with respect to the Lebesgue measure (see Def. 5.94) and the *Riemann integral*.

#### **Theorem 6.11 (Expectation of a Continuous Random Variable)**

Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  be a continuous random variable with expectation  $E(Y)$  and a density  $f_Y$  that is Riemann integrable. Then

$$E(Y) = \int_{-\infty}^{\infty} y f_Y(y) dy. \quad (6.10)$$

(Proof p. 212)

Examples of continuous random variables and their expectations are treated in chapter 8, section 8.2.

**Example 6.12 (A Continuous Random Variable Without Expectation)** Consider the continuous random variable  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  with density

$$f_Y(y) = \frac{1}{\pi} \cdot \frac{1}{1+y^2}, \quad \forall y \in \mathbb{R}, \quad (6.11)$$

and distribution function

$$F_Y(y) = \frac{1}{2} + \frac{1}{\pi} \cdot \arctan y, \quad \forall y \in \mathbb{R}. \quad (6.12)$$

A distribution  $P_Y$  with density (6.11) is called a *standard Cauchy distribution* or *central  $t$ -distribution* with 1 degree of freedom.

The integral of the positive part of  $y \cdot f_Y(y)$  is

$$\int_{-\infty}^{\infty} [y \cdot f_Y(y)]^+ dy = \int_0^{\infty} y \cdot \frac{1}{\pi} \cdot \frac{1}{1+y^2} dy = \frac{1}{2\pi} \ln(1+y^2) \Big|_0^{-\infty} = \infty,$$

and the integral of the negative part is

$$\int_{-\infty}^{\infty} [y \cdot f_Y(y)]^- dy = \int_{-\infty}^0 -y \cdot \frac{1}{\pi} \cdot \frac{1}{1+y^2} dy = -\frac{1}{2\pi} \ln(1+y^2) \Big|_{-\infty}^0 = \infty$$

[for the notation cf. Eq. (3.69)]. Hence,  $y \cdot f_Y(y)$  is not quasi- $P$ -integrable on  $\mathbb{R}$  (see Def. 3.28) and  $E(Y)$  does not exist.  $\triangleleft$

#### 6.1.4 Transformation Theorem

Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Y, \mathcal{A}'_Y)$  and  $g: (\Omega'_Y, \mathcal{A}'_Y, P_Y) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  be random variables. Then we denote the expectation of  $g$  with respect to the distribution  $P_Y$  by  $E_Y(g)$ . The following corollary immediately follows from Theorem 3.57, translating the measure theory terms to probability theory. This corollary is relevant whenever we consider the expectation of a composition  $g \circ Y = g(Y)$  of a numerical function  $g$  [see Eq. (2.25)] and a mapping  $Y$  or the expectation of  $g$  with respect to the distribution  $P_Y$  [see Eq. (5.2)].

##### Theorem 6.13 (Transformation Theorem)

Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Y, \mathcal{A}'_Y)$  be a random variable and  $g: (\Omega'_Y, \mathcal{A}'_Y) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  be measurable.

(i) If  $g$  is nonnegative or has a finite expectation  $E_Y(g)$ , then

$$E_Y(g) = \int g dP_Y = \int g(y) P_Y(dy) = \int g(Y) dP = E[g(Y)]. \quad (6.13)$$

(ii)  $E_Y(g)$  is finite if and only if  $E[g(Y)]$  is finite.

The virtue of Equation (6.13) is that we do not have to know the distribution of  $g(Y)$ . Instead, the distribution of  $Y$  suffices.

**Remark 6.14 (A Special Case)** If we consider the special case in which  $g$  is the identity function  $id: \bar{\mathbb{R}} \rightarrow \bar{\mathbb{R}}$ , defined by  $id(y) = y$ , for all  $y \in \bar{\mathbb{R}}$ , then  $id(Y) = Y$  and Equations (6.13) yield

$$E(Y) = \int Y dP = \int y P_Y(dy). \quad (6.14)$$

$\triangleleft$

**Remark 6.15 (Finite Number of Values)** If  $Y$  takes on only a finite number of different values  $y_1, \dots, y_n \in \mathbb{R}$ , then Equation (6.13) simplifies to

$$E[g(Y)] = E_Y(g) = \int g \, dP_Y = \sum_{i=1}^n g(y_i) \cdot P(Y=y_i), \quad (6.15)$$

where  $P(Y=y_i) = P_Y(\{y_i\}) = p_Y(y_i)$ ,  $i = 1, \dots, n$ , and  $p_Y$  denotes the probability function of  $Y$  (see Def. 5.56).  $\triangleleft$

**Remark 6.16 (Countable Number of Values)** If  $Y$  takes on a countable number of different values  $y_1, y_2, \dots \in \mathbb{R}$  and  $\sum_{i=1}^{\infty} g^+(y_i) \cdot P(Y=y_i) < \infty$  or  $\sum_{i=1}^{\infty} g^-(y_i) \cdot P(Y=y_i) < \infty$ , then

$$E[g(Y)] = E_Y(g) = \int g \, dP_Y = \sum_{i=1}^{\infty} g(y_i) \cdot P(Y=y_i). \quad (6.16)$$

Note that (6.16) also applies if  $g$  is nonnegative, because in this case  $g^- = 0$  holds for the negative part of  $g$ , which implies  $\sum_{i=1}^{\infty} g^-(y_i) = 0 < \infty$ .  $\triangleleft$

Equation (6.13) immediately implies the following corollary according to which the expectations of two random variables  $X$  and  $Y$  are identical if they have identical distributions, provided that the expectations exist (see also Remark 6.27).

**Corollary 6.17 (Identical Distributions Imply Identical Expectations)**

Let  $X, Y: (\Omega, \mathcal{A}, P) \rightarrow (\Omega', \mathcal{A}')$  be random variables and  $g: (\Omega', \mathcal{A}') \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  a measurable function that is nonnegative or with expectation  $E_X(g)$ . If  $P_X = P_Y$ , then  $E[g(X)] = E[g(Y)]$ .

This property allows us to use the term *expectation of a distribution* instead of expectation of a random variable.

In the following lemma we consider a bivariate random variable  $(X, Y)$  and a numerical function  $g(X)$ , i. e., a function that only depends on  $X$ . According to this lemma, the expectation of  $g$  with respect to the joint distribution  $P_{X,Y}$  is identical to the expectation of  $g$  with respect to the marginal distribution  $P_X$ .

**Lemma 6.18 (Expectation w.r.t. Joint and Marginal Distributions)**

Let  $(X, Y): (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X \times \Omega'_Y, \mathcal{A}'_X \otimes \mathcal{A}'_Y)$  be a bivariate random variable with joint distribution  $P_{X,Y}$  and let  $g: (\Omega'_X, \mathcal{A}'_X) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  be a measurable function that is nonnegative or with expectation  $E_X(g)$ . Then

$$E_{X,Y}(g) = E_X(g), \quad (6.17)$$

which is equivalent to

$$\int g(x) P_{X,Y} [d(x, y)] = \int g(x) P_X(dx). \quad (6.18)$$

(Proof p. 212)

**Example 6.19 (Flipping two Coins – continued)** Consider again the random variable  $X = \text{number of flipping heads}$  and the indicator  $1_H: \Omega \rightarrow \mathbb{R}$  of the event that *at least one heads is tossed*. In Example 2.47 we showed that  $1_H = g \circ X$ , where  $g: \mathbb{R} \rightarrow \mathbb{R}$  is defined by

$$g(x) = \begin{cases} 1 & \text{if } x \in \{1, 2\}, \\ 0, & \text{otherwise} \end{cases} \quad \text{for all } x \in \mathbb{R}.$$

According to (6.15),

$$E(1_H) = E[g(X)] = E_X(g) = 0 \cdot P(\{X=0\}) + 1 \cdot P(X \in \{1, 2\}) = P(X \in \{1, 2\}) = \frac{3}{4}.$$

&lt;

**Example 6.20 (Expectation of  $Y^2$ )** Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  be a real-valued random variable, let  $g: (\mathbb{R}, \mathcal{B}) \rightarrow (\mathbb{R}, \mathcal{B})$  be measurable and let  $g(Y) := Y^2$ . Then, according to Equation (6.13),

$$E(Y^2) = E[g(Y)] = E_Y(g) = \int g dP_Y = \int y^2 P_Y(dy).$$

[Note that  $Y^2$  is nonnegative and  $E(Y^2)$  can be infinite.] This equation shows that the expectation of  $Y^2$  solely depends on the distribution  $P_Y$  of  $Y$ , which illustrates Corollary 6.17. Using the integral  $\int g dP_Y$  is often the most convenient way of computing the expectation  $E(Y^2)$ . If  $Y$  takes on only a finite number of values  $y_1, \dots, y_n \in \mathbb{R}$ , then this equation simplifies to

$$E(Y^2) = \sum_{i=1}^n y_i^2 \cdot P_Y(\{y_i\}) = \sum_{i=1}^n y_i^2 \cdot P(Y=y_i). \quad (6.19)$$

These equations only involve the probabilities  $P(Y=y_i) = P_Y(\{y_i\})$ , not the probabilities  $P(Y^2=y^2)$ . <

**Example 6.21 (Multiplication With Indicators)** Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  be a numerical random variable with expectation  $E(Y)$ . If  $A \in \mathcal{A}$  and  $P(A) = 0$ , then  $(1_A Y) \stackrel{p}{=} 0$  and Rule (i) of Box 6.1 implies

$$E(1_A Y) = 0. \quad (6.20)$$

If  $C = A \cup B$ ,  $A \cap B = \emptyset$ , and  $A, B \in \mathcal{A}$ , then  $1_C Y = 1_A Y + 1_B Y$  and Rule (vi) of Box 6.1 implies

$$E(1_C Y) = E(1_A Y) + E(1_B Y). \quad (6.21)$$

&lt;

The following corollary shows how to compute the expectation of the composition  $g(Y)$  using the density of  $Y$ . As mentioned before, the virtue of Equation (6.22) is that we do not have to know the density of  $g(Y)$ , the density of  $Y$  suffices.

**Corollary 6.22 (Transformation Theorem, Continuous Random Variable)**

Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  be a continuous random variable with a Riemann integrable density  $f_Y$ . If  $g: (\mathbb{R}, \mathcal{B}) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  is a measurable function that is nonnegative or numerical with finite expectation  $E_Y(g) = \int g dP_Y$ , then

$$E[g(Y)] = E_Y(g) = \int_{-\infty}^{\infty} g(y) f_Y(y) dy. \quad (6.22)$$

(Proof p. 212)

**6.1.5 Rules of Computation**

Some rules of computation for expectations are gathered in Box 6.1 (for proofs see Exercise 6-1).

**Example 6.23 (Expectation of a Sample Mean)** Let  $Y_1, \dots, Y_n$  be a random sample. This means that the random variables  $Y_1, \dots, Y_n$  are i.i.d. (see Rem. 5.42). Furthermore, if the  $Y_1, \dots, Y_n$  are real-valued, define

$$\bar{Y} = \frac{1}{n} \cdot \sum_{i=1}^n Y_i, \quad (6.23)$$

the *arithmetic mean*, which in statistics is also called the *sample mean*. If  $Y_1$  is nonnegative or with finite expectation and

$$\mu_Y := E(Y_1) \quad (6.24)$$

denotes the identical expectations of the variables  $Y_1, \dots, Y_n$ , then

$$E(\bar{Y}) = \mu_Y \quad (6.25)$$

(see Exercise 6-4). ◁

Now we turn to a generalization of Rule (x) of Box 6.1.

**Theorem 6.24 (Expectation of the Product of Random Variables)**

Let  $Y_i: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$ ,  $i = 1, \dots, n$ , be real-valued random variables that are nonnegative or with finite expectations and assume that the  $Y_1, \dots, Y_n$  are independent. Then

$$E\left(\prod_{i=1}^n Y_i\right) = \prod_{i=1}^n E(Y_i). \quad (6.26)$$

If the expectations  $E(Y_i)$ ,  $i = 1, \dots, n$ , are finite, then  $E\left(\prod_{i=1}^n Y_i\right)$  is finite, too.

For a proof see, e. g., Bauer (1996, Theorem 8.1, p. 49). Later we will weaken the independence assumption [see Rem. 7.10 and Box 7.1 (i)]. However, if the variables  $Y_i$  are *not independent*, then Equation (6.26) does *not necessarily hold* (see Remark 7.10).

**Box 6.1 Rules of Computation for Expectations**

Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  be a random variable with expectation  $E(Y)$  and let  $\alpha \in \mathbb{R}$ . Then:

$$Y \stackrel{P}{=} \alpha \Rightarrow E(Y) = \alpha. \quad (\text{i})$$

$$E(\alpha + Y) = \alpha + E(Y). \quad (\text{ii})$$

$$E(\alpha \cdot Y) = \alpha \cdot E(Y). \quad (\text{iii})$$

Let  $A, B \in \mathcal{A}$ . Then

$$E(\mathbf{1}_A \cdot \mathbf{1}_B) = P(A \cap B). \quad (\text{iv})$$

$$E(\mathbf{1}_A \cdot Y) = 0, \quad \text{if } P(A) = 0. \quad (\text{v})$$

If  $Y_1, Y_2$  are nonnegative or real-valued with finite expectation, then

$$E(Y_1 + Y_2) = E(Y_1) + E(Y_2). \quad (\text{vi})$$

For  $i = 1, \dots, n$ , let  $Y_i: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  be random variables with finite expectations  $E(Y_i)$  and  $\alpha_i \in \mathbb{R}$ . Then

$$E\left(\sum_{i=1}^n \alpha_i \cdot Y_i\right) = \sum_{i=1}^n \alpha_i \cdot E(Y_i). \quad (\text{vii})$$

Let  $X, Y: (\Omega, \mathcal{A}, P) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  be random variables that are nonnegative or with finite expectations. Then:

$$X \stackrel{P}{=} Y \Rightarrow E(X) = E(Y). \quad (\text{viii})$$

$$X \stackrel{P}{=} Y \Leftrightarrow \forall A \in \mathcal{A}: E(\mathbf{1}_A X) = E(\mathbf{1}_A Y). \quad (\text{ix})$$

$$X \perp\!\!\!\perp_P Y \Rightarrow E(X \cdot Y) = E(X) \cdot E(Y). \quad (\text{x})$$

**6.2 Moments, Variance, and Standard Deviation**

The expectation  $E(Y)$  of a numerical random variable  $Y$  is also called the *first moment of  $Y$* , provided that this expectation exists, whereas the expectation  $E(Y^2)$  is called the *second moment of  $Y$*  (see Example 6.20). For second (and higher) moments we distinguish between *moments* and *central moments*.

**Remark 6.25 (Higher Moments of  $Y$ )** Analogously to Example 6.20, we may consider  $g(Y) = |Y^n|$  or  $g(Y) = Y^n$  with  $n \geq 1$ . Note that

- (i)  $E(|Y^n|) = E(|Y|^n)$  always exists, because  $|Y^n|$  is nonnegative (see Def. 6.1).
- (ii) If, for  $n \in \mathbb{N}$ , the expectation  $E(|Y|^n)$  is finite, then  $E(Y^n)$  exists and is finite (see Cor. 3.38).
- (iii) If, for  $n \in \mathbb{N}$ , the expectation  $E(|Y|^n)$  is finite, then  $E(Y^m)$  exists and is finite as well for all  $m$  with  $1 \leq m \leq n$  (see Exercise 6-2).
- (iv) If  $n$  is even, then the random variable  $Y^n$  is nonnegative and  $E(Y^n)$  exists. In contrast, if  $n$  is odd, then the expectation  $E(Y^n)$  does not necessarily exist.

◁

**Definition 6.26 (Moments)**

Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  be a numerical random variable and let  $n \in \mathbb{N}$ .

- (i) The expectation  $E(|Y|^n)$  is called the *nth absolute moment* of  $Y$ .
- (ii) If  $Y$  is nonnegative or such that  $E(Y^n)$  is finite, then  $E(Y^n)$  is called the *nth moment* of  $Y$ .
- (iii) If  $E(Y^n)$  is finite or if  $Y$  is nonnegative and the expectation  $E(Y)$  is finite, then we call  $E([Y - E(Y)]^n)$  the *nth central moment* of  $Y$ .

**Remark 6.27 (Moments Under  $P$ -Equivalence)** If the expectations of  $Y^n$  and therefore also of  $[Y - E(Y)]^n$  are finite, then they can be represented as expectations of functions  $g(Y)$  of  $Y$ , where  $g: (\overline{\mathbb{R}}, \overline{\mathcal{B}}) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  is a measurable function with finite expectation  $E_Y(g)$  [see Proposition (ii) of Th. 6.13]. Therefore, according to Corollary 6.17, all moments (central or noncentral) of a numerical random variable  $Y$  solely depend on its distribution  $P_Y$ . Hence, if two random variables  $Y_1$  and  $Y_2$  have the same distribution  $P_{Y_1} = P_{Y_2}$ , then they have the same moments. For instance, if  $Y_1 \stackrel{P}{=} Y_2$  and the expectations  $E(Y_1^2)$  and  $E(Y_2^2)$  are finite, then  $E(Y_1) = E(Y_2)$  and  $E(Y_1^2) = E(Y_2^2)$ . This allows us to use the terms (central) moments of a distribution instead of (central) moments of a random variable. ◁

Variance and standard deviation are the most important parameters describing the *variability* of a random variable. They are defined as follows:

**Definition 6.28 (Variance and Standard Deviation)**

Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  be a numerical random variable and assume that  $E(Y^2) < \infty$ . Then the *variance* of  $Y$  is defined by

$$\text{Var}(Y) := E([Y - E(Y)]^2), \quad (6.27)$$

and the *standard deviation* of  $Y$  by the positive square root of the variance, i. e.,

$$\text{SD}(Y) := \sqrt{\text{Var}(Y)}. \quad (6.28)$$

**Box 6.2 Rules of Computation for Variances**

Let  $X, Y: (\Omega, \mathcal{A}, P) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  be random variables with finite second moments and let  $\alpha \in \mathbb{R}$ . Then:

$$\text{Var}(Y) = E(Y^2) - E(Y)^2. \quad (\text{i})$$

$$\text{Var}(\alpha + Y) = \text{Var}(Y). \quad (\text{ii})$$

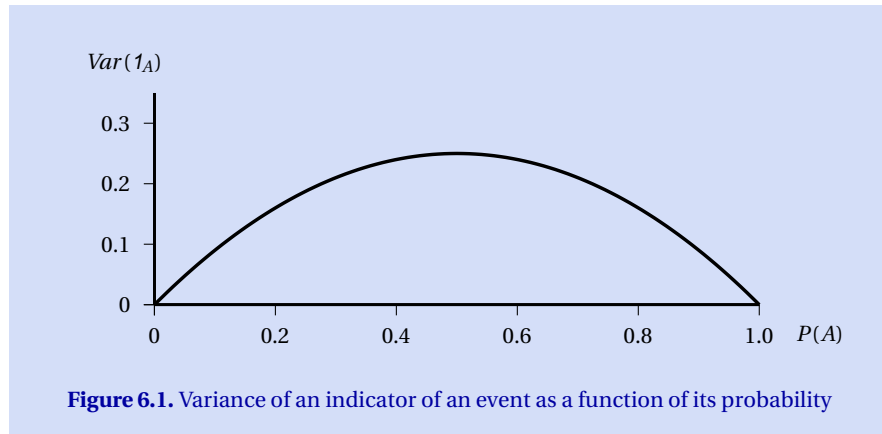
$$\text{Var}(\alpha \cdot Y) = \alpha^2 \cdot \text{Var}(Y). \quad (\text{iii})$$

$$\exists \alpha \in \mathbb{R} : Y \stackrel{P}{=} \alpha \Leftrightarrow \text{Var}(Y) = 0. \quad (\text{iv})$$

$$X \stackrel{P}{=} Y \Rightarrow \text{Var}(X) = \text{Var}(Y). \quad (\text{v})$$

For  $i = 1, \dots, n$ , let the random variables  $Y_i: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  be independent with finite second moments and  $\alpha_i \in \mathbb{R}$ . Then

$$\text{Var}\left(\sum_{i=1}^n \alpha_i \cdot Y_i\right) = \sum_{i=1}^n \alpha_i^2 \cdot \text{Var}(Y_i). \quad (\text{vi})$$



**Figure 6.1.** Variance of an indicator of an event as a function of its probability

According to this definition,  $\text{Var}(Y)$  is the expectation of the squared *mean centered* random variable  $Y - E(Y)$ . Hence, the variance of  $Y$  is the second central moment of  $Y$ . Note that variances and standard deviations are nonnegative. The variance of  $Y$  is also denoted by  $\sigma_Y^2$  and the standard deviation by  $\sigma_Y$ . Box 6.2 summarizes some important properties of variances (see Exercise 6-3).

**Example 6.29 (Location vs. Variability)** Consider two real-valued random variables  $X, Y: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  with  $P(X=-1) = P(X=1) = 0.5$  and  $P(Y=-10) = P(Y=10) = 0.5$ . Then  $E(X) = E(Y) = 0$  but  $\text{Var}(X) = 1 \neq \text{Var}(Y) = 100$ . In contrast, if  $P(X=-1) = P(X=1) = 0.5$  and  $P(Y=9) = P(Y=11) = 0.5$ , then  $E(X) = 0 \neq E(Y) = 10$  but  $\text{Var}(X) = \text{Var}(Y) = 1$ . This illustrates that the expectation describes

the ‘location’ of a random variable [see Rule (ii) of Box 6.1], while the variance is invariant with respect to translations [see Rule (ii) of Box 6.2]. In contrast, the variance describes the ‘variability’ of a random variable, whereas, in general, the expectation does not.  $\triangleleft$

**Example 6.30 (Variance of an Indicator)** Let  $(\Omega, \mathcal{A}, P)$  be a probability space and let  $1_A$  denote the indicator of  $A \in \mathcal{A}$ . Then

$$\begin{aligned} \text{Var}(1_A) &= E(1_A^2) - [E(1_A)]^2 && \text{[Box 6.2, (i)]} \\ &= E(1_A) - [E(1_A)]^2 && [1_A^2 = 1_A] \\ &= E(1_A) \cdot [1 - E(1_A)] \\ &= P(A) \cdot [1 - P(A)]. && \text{[(6.4)]} \end{aligned} \quad (6.29)$$

According to Equation (6.4), the expectation of the indicator  $1_A$  is  $P(A)$ , and Equation (6.29) shows that its variance is  $P(A) \cdot [1 - P(A)]$ . Obviously, the variance of an indicator variable does not contain any information additional to the expectation  $E(1_A) = P(A)$ . In fact, in this case,  $E(1_A)$  contains the full information about the distribution of  $1_A$ . This is not surprising because the distribution of  $1_A$  is completely determined by the single parameter  $P(A)$ . Unlike the expectation, the variance of  $1_A$  *does not* comprise the full information about the distribution of  $1_A$ . For example,  $\text{Var}(1_A) = .1 \cdot .9 = .09$  if  $P(A) = .1$  or  $P(A) = .9$ . The variance of  $1_A$  has its maximum for  $P(A) = 1/2$  and goes to 0 if  $P(A)$  approaches 0 or 1 (see Fig. 6.1).  $\triangleleft$

**Example 6.31 (Joe and Ann With Random Assignment– continued)** In the example presented in Table 2.2 (p. 53),  $X$  is an indicator variable. Hence, its variance is most easily computed as follows:

$$\text{Var}(X) = P(X=1) \cdot [1 - P(X=1)] = .40 \cdot .60 = .24.$$

Similarly, the variance of  $Y$  is obtained by

$$\text{Var}(Y) = P(Y=1) \cdot [1 - P(Y=1)] = .51 \cdot (1 - .51) = 0.2499.$$

$\triangleleft$

**Example 6.32 (Variance and Standard Error of the Sample Mean)** Let  $Y_1, \dots, Y_n$  be a sample (see Example 6.23), and  $\bar{Y}$  the sample mean [see Eq. (6.23)]. If  $E(Y_1^2) < \infty$ , and

$$\sigma_Y^2 := \text{Var}(Y_1) \quad (6.30)$$

denotes the identical variances of the  $Y_1, \dots, Y_n$ , then

$$\sigma_{\bar{Y}}^2 := \text{Var}(\bar{Y}) = \frac{\sigma_Y^2}{n} \quad (6.31)$$

(see Exercise 6-4). Hence,

$$\sigma_{\bar{Y}} := \sqrt{\sigma_{\bar{Y}}^2} = \text{SD}(\bar{Y}) = \frac{\sigma_Y}{\sqrt{n}}. \quad (6.32)$$

In statistics  $\text{SD}(\bar{Y})$  is also denoted  $\text{SE}(\bar{Y})$  and called the *standard error of the sample mean*.  $\triangleleft$

**Remark 6.33 (Z-Transformation)** Every real-valued random variable  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  with finite and positive variance  $\text{Var}(Y)$  can be transformed by

$$Z_Y := \frac{Y - E(Y)}{SD(Y)}. \quad (6.33)$$

Note that  $E(Z_Y) = 0$  and  $\text{Var}(Z_Y) = 1$  (see Exercise 6-5). This transformation is called the *Z-transformation*.

As a special case consider the mean  $\bar{Y}$  of a sample  $Y_1, \dots, Y_n$  with expectation  $\mu_Y := E(Y_1)$  and finite standard deviation  $\sigma_Y := SD(Y_1) > 0$  (see Example 6.32). Then the Z-transformation of  $\bar{Y}$  is

$$Z_{\bar{Y}} = \sqrt{n} \cdot \frac{\bar{Y} - \mu_Y}{\sigma_Y} \quad (6.34)$$

(see Exercise 6-6). The random variable  $Z_{\bar{Y}}$  will be used in the Central Limit Theorem (see Th. 8.34).  $\triangleleft$

**Remark 6.34 (The Expectation Minimizes the Mean Squared Error)** If  $E(Y^2) < \infty$ , then  $\alpha \in \mathbb{R}$  minimizes  $MSE(a) := E[(Y - a)^2]$  if and only if  $\alpha = E(Y)$ . This means that the expectation  $E(Y)$  that the expectation minimizes the mean squared error (see Exercise 6-7).  $\triangleleft$

**Definition 6.35 (Coefficient of Variation)**

Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  be a nonnegative numerical random variable. If  $E(Y^2)$  is finite and  $E(Y) \neq 0$ , then

$$CV(Y) := \frac{SD(Y)}{|E(Y)|} \quad (6.35)$$

is called the *coefficient of variation* of  $Y$ .

**Remark 6.36 (Properties of the Coefficient of Variation)** This coefficient is a non-negative dimensionless quantity that expresses the variability of  $Y$  in units of its expectation. Note that sometimes the coefficient of variation is defined as  $SD(Y)/E(Y)$ . If  $\alpha \neq 0$ , then  $CV(\alpha Y) = CV(Y)$  [see Box 6.2 (iii), (6.28), and 6.1 (iii)], i. e.,  $CV(Y)$  is invariant under multiplication with (nonzero) constant. In contrast, if  $\beta \neq -E(Y)$ , then  $CV(Y + \beta) = SD(Y)/|E(Y) + \beta|$  [see Box 6.2 (ii), (6.28), and 6.1 (ii)], i. e.,  $CV(Y)$  is not invariant under translations.  $\triangleleft$

The following parameter quantifies, in a sense, the deviation of a distribution from symmetry to its expectation.

**Definition 6.37 (Skewness)**

Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  be a numerical random variable. If  $E(Y^3)$  is finite and  $\text{Var}(Y) > 0$ , then

$$E\left(\frac{[Y - E(Y)]^3}{SD(Y)^3}\right) \quad (6.36)$$

is called the skewness of  $Y$ .

### 6.3 Proofs

#### **Proof of Theorem 6.11**

$$E(Y) = \int Y dP \quad [(6.1)]$$

$$= \int y P_Y(dy) \quad [(3.62)]$$

$$= \int y f_Y(y) \lambda(dy) \quad [(3.72), (5.61)]$$

$$= \int_{-\infty}^{\infty} y f_Y(y) dy. \quad [\text{Riemann integrability of } f_Y, (3.68)]$$

#### **Proof of Lemma 6.18**

For all  $A' \in \mathcal{A}'_X$  and  $g = 1_{A'}$ ,

$$\int 1_{A'}(x) P_X(dx) = P_X(A') \quad [(3.9)]$$

$$= P_{X,Y}(A' \times \Omega'_Y) \quad [(5.21)]$$

$$= \int 1_{A'}(x) \cdot 1_{\Omega'_Y}(y) P_{X,Y}[d(x,y)] \quad [(3.9), (1.38)]$$

$$= \int 1_{A'}(x) P_{X,Y}[d(x,y)]. \quad [1_{\Omega'_Y}(y) = 1]$$

Now the proposition follows applying the standard methods of proofs described in Remark 3.30.

#### **Proof of Corollary 6.22**

$$E[g(Y)] = E_Y(g) = \int g(y) P_Y(dy) \quad [(6.13)]$$

$$= \int g(y) f_Y(y) \lambda(dy) \quad [(3.72), (5.61)]$$

$$= \int_{-\infty}^{\infty} y f_Y(y) dy. \quad [\text{Riemann integrability of } y \cdot f_Y(y), (3.68)]$$

## 6.4 Exercises

▷ **Exercise 6-1** Prove the rules of computation of Box 6.1.

▷ **Exercise 6-2** Show: If, for  $n \in \mathbb{N}$ , the expectation  $E(Y^n)$  exists and is finite, then  $E(Y^m)$  exists and is finite as well for all  $1 \leq m \leq n$ .

▷ **Exercise 6-3** Prove the rules of Box 6.2.

▷ **Exercise 6-4** Prove Equations (6.25) and (6.31).

▷ **Exercise 6-5** Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  be a numerical random variable with finite second moment  $E(Y^2)$  and  $\text{Var}(Y) > 0$ . Show that the expectation of  $Z := [Y - E(Y)]/SD(Y)$  is 0 and its variance is 1.

▷ **Exercise 6-6** Let  $Y_i: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$ ,  $i = 1, 2, \dots, n$ , be a sample with expectations  $E(Y_i) = \mu$  and finite and positive variances  $\text{Var}(Y_i) = \sigma^2$ ,  $i = 1, \dots, n$ . Furthermore, let  $\bar{Y} := S/n$  be the arithmetic mean, where  $S := \sum_{i=1}^n Y_i$ , and

$$\bar{Z} := \frac{(\bar{Y} - \mu) \cdot \sqrt{n}}{\sigma}.$$

Show that  $E(\bar{Z}) = 0$  and  $\text{Var}(\bar{Z}) = 1$ .

▷ **Exercise 6-7** Assume  $E(Y^2) < \infty$  and show that  $\alpha \in \mathbb{R}$  minimizes  $MSE(a) := E[(Y - a)^2]$  if and only if  $\alpha = E(Y)$ .

## Solutions

▷ **Solution 6-1** Because the expectation of a numerical random variable is defined as an integral, we simply can refer to the corresponding propositions of chapter 3.

(i) This is Equation (3.8) with  $A = \Omega$  and  $\mu(A) = P(\Omega) = 1$ .

(ii) This is a special case of Equation (3.34).

(iii) This is a special case of Equation (3.32).

(iv) Note that  $1_A \cdot 1_B = 1_{A \cap B}$ . Hence, this rule follows from Equation (3.8), with  $\alpha = 1$ .

(v) This is a special case of Lemma 3.45.

(vi) If  $Y_1, Y_2$  are nonnegative, then this equation is a special case of Equation (3.65). If  $Y_1$  or  $Y_2$  has a finite expectation, then this equation is a special case of Equation (3.33).

(vii) This rule follows from Equation (3.34) and complete induction.

(viii) This is a special case of Lemma 3.47.

(ix) This is a special case of Theorem 3.48.

(x) This is a special case of Theorem 6.24.

▷ **Solution 6-2** Assume that, for  $n \in \mathbb{N}$ , the expectation  $E(Y^n)$  exists and is finite. Furthermore, let  $A := \{\omega \in \Omega: |Y(\omega)| > 1\}$ . Now, for all  $m$  with  $1 \leq m \leq n$ ,

$$|Y(\omega)^m| \leq |Y(\omega)^n|, \quad \forall \omega \in A \quad \text{and} \quad |Y(\omega)^m| \leq 1, \quad \forall \omega \in A^c.$$

Therefore, applying Corollary 3.38, for  $1 \leq m \leq n$ :

$$E(|Y^m|) = \int |Y^m| dP \quad [(6.1)]$$

$$= \int_A |Y^m| dP + \int_{A^c} |Y^m| dP \quad [(3.36)]$$

$$\leq \int_A |Y^n| dP + \int_{A^c} 1 dP \quad [\text{Lemma 3.26}]$$

$$\leq \int_A |Y^n| dP + 1 \quad [(3.30), 1_{A^c} \leq 1, \text{Lemma 3.26}]$$

$$< \infty.$$

▷ **Solution 6-3** Rules (i), (ii), and (iii) are special cases of rules (i), (ii), and (iii) of Box 7.1, which are proved in Exercise 7-3.

(iv) ⇒: This proposition follows from (vii) of Box 7.1.

⇐:

$$\text{Var}(Y) = 0 \Rightarrow [Y - E(Y)]^2 \stackrel{p}{=} 0 \quad [(6.27), \text{Th. 3.43}]$$

$$\Rightarrow Y - E(Y) \stackrel{p}{=} 0 \quad [[Y - E(Y)]^2 = 0 \Leftrightarrow Y - E(Y) = 0]$$

$$\Rightarrow Y \stackrel{p}{=} E(Y) \in \mathbf{R}. \quad [E(Y)^2 < \infty \Rightarrow E(Y) \in \mathbf{R}, \text{Rem. 6.25 (iii)}]$$

(v)  $X \stackrel{p}{=} Y \Rightarrow X^2 \stackrel{p}{=} Y^2$  [see (2.34)]. Hence,  $X \stackrel{p}{=} Y$  implies  $E(X) = E(Y)$  and  $E(X^2) = E(Y^2)$  (see Cor. 6.17). Therefore, rule (i) of Box 6.2 yields  $\text{Var}(X) = \text{Var}(Y)$ .

(vi) This proposition follows from rules (ix) and (vi) of Box 7.1.

▷ **Solution 6-4**

Equation (6.25) can be derived as follows:

$$E(\bar{Y}) = E\left(\frac{1}{n} \cdot \sum_{i=1}^n Y_i\right) \quad [(6.23)]$$

$$= \frac{1}{n} \cdot \sum_{i=1}^n E(Y_i) \quad [\text{Box 6.1 (vii)}]$$

$$= \frac{1}{n} \cdot n \cdot \mu_Y = \mu_Y. \quad [Y_1, \dots, Y_n \text{ are identically distributed, (6.24)}]$$

Equation (6.31) can be derived as follows:

$$\text{Var}(\bar{Y}) = \text{Var}\left(\frac{1}{n} \cdot \sum_{i=1}^n Y_i\right) \quad [(6.23)]$$

$$= \frac{1}{n^2} \cdot \sum_{i=1}^n \text{Var}(Y_i) \quad [\stackrel{p}{=} Y_1, \dots, Y_n, \text{Box 6.2 (vi)}]$$

$$= \frac{1}{n^2} \cdot n \cdot \sigma_Y^2 = \frac{\sigma_Y^2}{n}. \quad [Y_1, \dots, Y_n \text{ are identically distributed, (6.30)}]$$

▷ **Solution 6-5** Let  $\mu := E(Y)$  and  $\sigma := SD(Y)$ . Then

$$E(Z) = E\left(\frac{Y - \mu}{\sigma}\right) = E\left(\frac{1}{\sigma} \cdot (Y - \mu)\right) = \frac{1}{\sigma} \cdot E(Y - \mu) \quad [\text{Box 6.1 (iii)}]$$

$$= \frac{1}{\sigma} \cdot [E(Y) - E(\mu)] = \frac{1}{\sigma} \cdot (\mu - \mu) = 0 \quad [\text{Box 6.1 (vi), (i)}].$$

$$\begin{aligned} \text{Var}(Z) &= \text{Var}\left(\frac{Y-\mu}{\sigma}\right) = \text{Var}\left(\frac{1}{\sigma} \cdot (Y-\mu)\right) \\ &= \frac{1}{\sigma^2} \cdot \text{Var}(Y) = \frac{1}{\sigma^2} \cdot \sigma^2 = 1 \quad [\text{Box 6.2 (iii), (ii)}]. \end{aligned}$$

▷ **Solution 6-6** Using Equations (6.25), (6.32),  $E(Z_Y) = 0$ , and  $\text{Var}(Z_Y) = 1$  (see Exercise 6-5) yields

$$E(\bar{Z}) = E\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right) = 0 \quad \text{and} \quad \text{Var}(\bar{Z}) = \text{Var}\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right) = 1.$$

▷ **Solution 6-7** For all  $\alpha \in \mathbb{R}$ , using Box 6.1 (iii), (ii),

$$\begin{aligned} E[(Y-\alpha)^2] &= E[(Y-E(Y)+E(Y)-\alpha)^2] \\ &= E[(Y-E(Y))^2] + [E(Y)-\alpha]^2 + 2 \cdot E[(Y-E(Y)) \cdot (E(Y)-\alpha)] \\ &= E[(Y-E(Y))^2] + [E(Y)-\alpha]^2 \\ &\geq E[(Y-E(Y))^2], \end{aligned}$$

and '=' holds if and only if  $E(Y) = \alpha$ .



## Chapter 7

# Linear Quasi-Regression, Covariance, and Correlation

Expectation and variance are parameters that describe important properties of a univariate numerical random variable and its distribution. Now we consider *two* numerical random variables, say  $X$  and  $Y$ , and their joint distribution. In other words, we consider the distribution of the bivariate real-valued random variable  $(X, Y): (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}^2, \mathcal{B}_2)$ . We also introduce a new random variable that can be used to describe a specific kind of dependence of  $Y$  on  $X$ . It is the kind of dependence of  $Y$  on  $X$  that is represented by the best fitting linear function  $\alpha_0 + \alpha_1 X$ , ‘best fitting’ in terms of the minimal mean squared error. This function is the composition of  $X$  and the *linear quasi-regression* or the *linear least-squares regression*. *Covariance* and *correlation* are important parameters quantifying the strength of the kind of dependence that can be described by a linear quasi-regression.

### 7.1 Linear Quasi-Regression

**Remark 7.1 (Implications of Finite Second Moments)** Reading the following definition note that  $E(X^2), E(Y^2) < \infty$  implies that  $E(X), E(Y)$ , and  $E(X \cdot Y)$  are finite (see Klenke, 2013, Remark 5.2, p. 102). Hence, according to Remark 3.42, there is no substantial loss of generality if we additionally assume that  $X$  and  $Y$  are real-valued.  $\triangleleft$

#### Definition 7.2 (Linear Quasi-Regression)

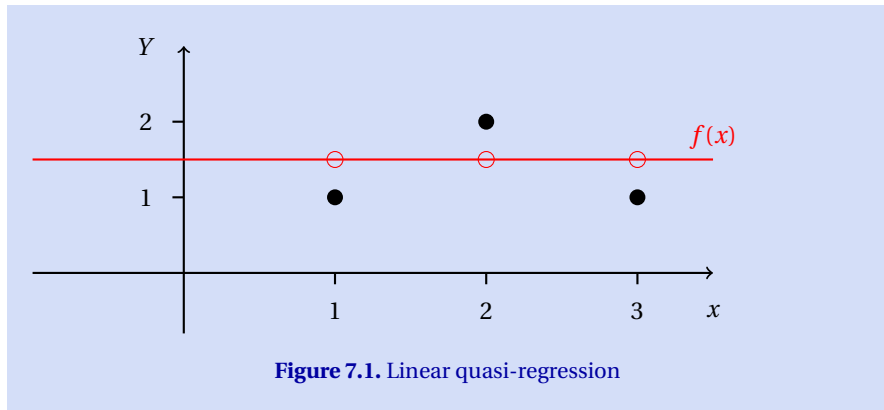
Let  $X, Y: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  be two real-valued random variables, assume  $E(X^2), E(Y^2) < \infty$ , and  $\text{Var}(X) > 0$ . Then the function  $f: \mathbb{R} \rightarrow \mathbb{R}$  defined by

$$f(x) = \alpha_0 + \alpha_1 x, \quad \forall x \in \mathbb{R}, \quad (7.1)$$

where the pair  $(\alpha_0, \alpha_1)$  minimizes the function  $\text{MSE}: \mathbb{R}^2 \rightarrow \mathbb{R}$  with

$$\text{MSE}(a_0, a_1) = E([Y - (a_0 + a_1 X)]^2), \quad \forall a_0, a_1 \in \mathbb{R}, \quad (7.2)$$

is called the *linear quasi-regression of  $Y$  on  $X$* . The composition of  $X$  and  $f$  is denoted by  $Q_{\text{lin}}(Y|X)$ , i. e.,



$$Q_{lin}(Y|X) = f(X) = \alpha_0 + \alpha_1 X. \quad (7.3)$$

**Remark 7.3 (Distinguishing Between  $f$  and  $f(X)$ )** To emphasize, the function  $f: \mathbb{R} \rightarrow \mathbb{R}$  is called a linear quasi-regression. In this context the random variable  $X$  is called the *regressor* and  $Y$  the *regressand*. Note that  $f$  is a function assigning a real number to all real numbers. This applies even if  $X$  only takes on two different real values. In contrast, the number of different values of the composition  $Q_{lin}(Y|X) = f(X): \Omega \rightarrow \mathbb{R}$  is smaller or equal to the number of values of  $X$ , provided that  $X$  takes on a finite number of values only.  $\triangleleft$

**Remark 7.4 (Coefficient of Determination)** Under the assumptions of Definition 7.2 and  $\text{Var}(Y) > 0$ , we define

$$Q_{Y|X}^2 := \frac{\text{Var}[Q_{lin}(Y|X)]}{\text{Var}(Y)} \quad (7.4)$$

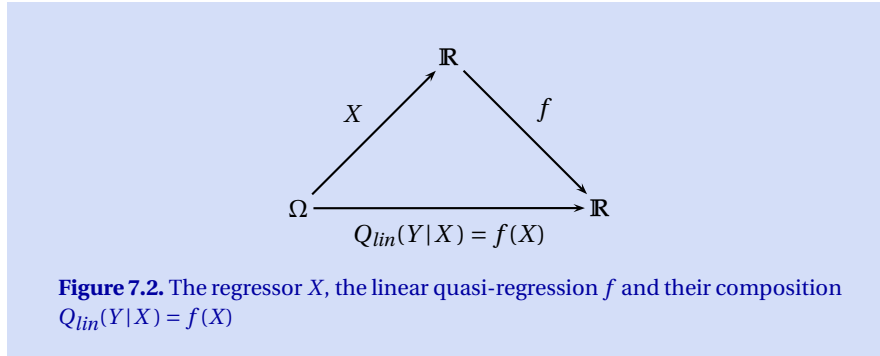
and it call the *coefficient of determination* of the linear quasi-regression  $Q_{lin}(Y|X)$ . In Remark 7.29, this definition is extended to the case in which  $X$  is an  $n$ -dimensional real-valued random variable.  $\triangleleft$

**Example 7.5 (Discrete Regressor With Three Different Values)** Let  $X$  and  $Y$  be real-valued random variables on  $(\Omega, \mathcal{A}, P)$  with values 1, 2, 3 and 1, 2, respectively. Furthermore, assume that their distribution is specified by

$$P(X=1, Y=1) = .25, \quad P(X=2, Y=2) = .5, \quad P(X=3, Y=1) = .25.$$

Then the linear quasi-regression  $f: \mathbb{R} \rightarrow \mathbb{R}$  is specified by

$$f(x) = \alpha_0 + \alpha_1 \cdot x = 1.5 + 0 \cdot x = 1.5, \quad \forall x \in \mathbb{R},$$



and the composition of  $X$  and  $f$  is

$$Q_{lin}(Y|X) = \alpha_0 + \alpha_1 \cdot X = 1.5 + 0 \cdot X = 1.5$$

(see Exercise 7-1). The black points in Figure 7.1 represent the three pairs of values of  $X$  and  $Y$ . All values of the the linear quasi-regression are on the horizontal line, which, in this example, is parallel to  $x$ -axis because its slope is 0. The circles on this line represent the values  $f(x)$  for  $x = 1, 2, 3$ , i. e., for those values of  $X$  with a nonzero probability  $P_X(\{x\}) > 0$ . In contrast, in this example,  $P_X(\{x\}) = 0$ , for all  $x \in \mathbb{R} \setminus \{1, 2, 3\}$ . Nevertheless, as mentioned before, a linear quasi-regression  $f$  is a function assigning a real number *to all real numbers*.  $\triangleleft$

**Remark 7.6 (Linear Quasi-Regression vs. Regression)** As the term ‘linear quasi-regression’ suggests, there is also a ‘genuine’ regression of  $Y$  on  $X$  (see Def. 10.25) and the two terms are not necessarily identical. As will be explained in more detail in Remark 10.27, the ‘genuine’ regression is a function  $g: \mathbb{R} \rightarrow \mathbb{R}$  such that the composition  $g(X)$  is  $X$ -measurable and minimizes the mean squared error  $E([Y - g(X)]^2)$ . In contrast to the conditional expectation  $E(Y|X)$  [see Eq. (10.1)], the composition  $Q_{lin}(Y|X) = f(X)$  minimizes the function  $MSE$  specified in Equation (7.2). Hence,  $f$  has to be a *linear* function even in those cases in which there are no  $a_0, a_1 \in \mathbb{R}$ , such that  $E(Y|X) \stackrel{p}{=} a_0 + a_1 X$  (see Example 7.5).  $\triangleleft$

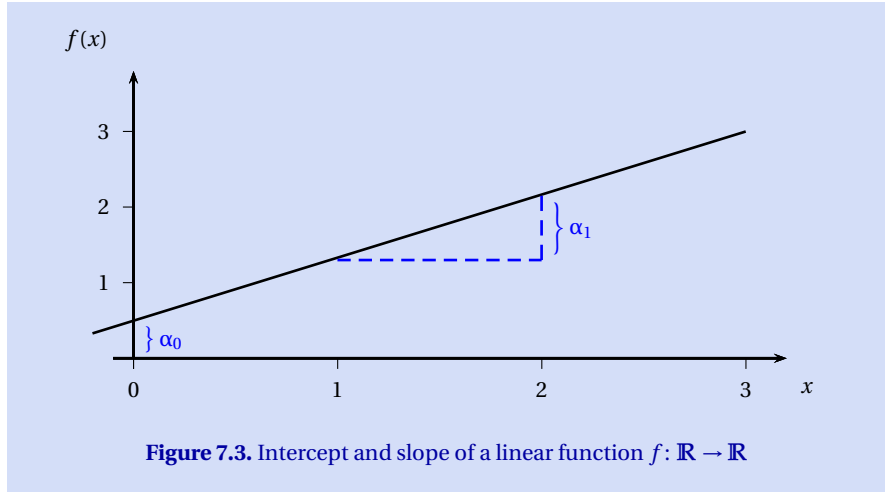
**Remark 7.7 (Intercept and Slope)** Note that the composition  $Q_{lin}(Y|X) = f(X) = \alpha_0 + \alpha_1 X$  is a random variable on  $(\Omega, \mathcal{A}, P)$  (see Fig. 7.2). The coefficient  $\alpha_0$  is called the *intercept* and  $\alpha_1$  the *slope* of (the linear quasi-regression)  $f$  (see Fig. 7.3). Obviously,

$$f(0) = \alpha_0 + \alpha_1 \cdot 0 = \alpha_0. \quad (7.5)$$

Furthermore, if  $x_1, x_2 \in \mathbb{R}$  and  $x_2 > x_1$ , then

$$\alpha_1 = \frac{1}{x_2 - x_1} \cdot [f(x_2) - f(x_1)] \quad (7.6)$$

(see Exercise 7-2). Equation (7.6) yields



$$\alpha_1 = f(x_2) - f(x_1), \quad \text{if } x_2 - x_1 = 1. \quad (7.7)$$

These equations justify calling  $\alpha_0$  the *intercept* and  $\alpha_1$  the *slope* of the linear quasi-regression  $f$  (see Fig. 7.3). Note that these equations also apply if  $P(X=0) = P(X=x_1) = P(X=x_2) = 0$ . They even apply if  $0, x_1, x_2 \notin X(\Omega)$ , because, by definition,  $f$  is a function on  $\mathbb{R}$ .

Figure 7.3 illustrates the intercept and the slope of a linear function such as the linear quasi-regression  $f$ . In this figure,  $\alpha_0 = .5$  and  $\alpha_1 = .85$ . If  $X$  is discrete, then  $Q_{lin}(Y|X) = f(X)$  is discrete as well. More precisely, the number of different values of  $Q_{lin}(Y|X)$  is always smaller than or equal to the number of different values of  $X$ . In contrast, the linear quasi-regression  $f: \mathbb{R} \rightarrow \mathbb{R}$  takes on uncountably many values unless its slope is 0. In this case its sole value is  $\alpha_0$ .  $\triangleleft$

## 7.2 Covariance

While the variance quantifies the variability of a numerical random variable, the covariance quantifies the degree of co-variation of two numerical random variables, i. e., the degree to which the two variables vary together in the following sense: If one variable takes on a large value (i. e., large positive deviation from its expectation), then the other one tends to take on a large value as well. Furthermore, if one variable takes on a small value (i. e., large negative deviation from its expectation), then the other one tends to take on a small value, too. In this case the covariance will be positive. However, the covariance may also be a negative real number. In this case, the two random variables co-vary in the following sense: If one variable takes on a large value, then the other one tends to take on

a small value. Furthermore, if one variable takes on a small value, then the other one tends to take on a large value.

**Definition 7.8 (Covariance)**

Let  $X, Y: (\Omega, \mathcal{A}, P) \rightarrow (\overline{\mathbb{R}}, \mathcal{B})$  be two numerical random variables with  $E(X^2), E(Y^2) < \infty$ . Then the covariance of  $X$  and  $Y$  is defined by

$$\text{Cov}(X, Y) := E([X - E(X)] \cdot [Y - E(Y)]). \quad (7.8)$$

Comparing Equations (7.8) and (6.27) to each other shows that the variance is the covariance of a numerical random variable with itself.

**Remark 7.9 (Correlated Numerical Random Variables)** According to this definition, the *covariance* of  $X$  and  $Y$  is the expectation of the product of the centered variables  $X - E(X)$  and  $Y - E(Y)$ . Hence, a covariance can be negative, zero, or positive. If the covariance is different from zero, then we say that  $X$  and  $Y$  are *correlated*; otherwise, we say that they are *uncorrelated*.  $\triangleleft$

**Remark 7.10 (Rules of Computation)** The most important rules of computation for covariances are summarized in Box 7.1. Proofs are provided in Exercise 7-3. Rule (i) immediately implies

$$E(X \cdot Y) = E(X) \cdot E(Y) + \text{Cov}(X, Y). \quad (7.9)$$

Hence,  $X$  and  $Y$  are uncorrelated if and only if  $E(X \cdot Y) = E(X) \cdot E(Y)$  [see Rule (vi)]. Furthermore, this equation and Theorem 6.24 imply that  $X$  and  $Y$  are uncorrelated if  $X$  and  $Y$  are independent.

Symmetry of the covariance [see Box 7.1 (v)] yields an alternative way to write Rule (viii) of Box 7.1:

$$\text{Var}\left(\sum_{i=1}^n \alpha_i Y_i\right) = \sum_{i=1}^n \alpha_i^2 \text{Var}(Y_i) + 2 \cdot \sum_{i=1}^{n-1} \sum_{j=i+1}^n \alpha_i \alpha_j \text{Cov}(Y_i, Y_j). \quad (7.10)$$

This equation simplifies to

$$\text{Var}\left(\sum_{i=1}^n \alpha_i Y_i\right) = \sum_{i=1}^n \alpha_i^2 \text{Var}(Y_i), \quad (7.11)$$

if  $Y_1, \dots, Y_n$  is a sequence of pairwise uncorrelated random variables. Note that independence of  $Y_1, \dots, Y_n$  implies  $\text{Cov}(Y_i, Y_j) = 0$ ,  $i \neq j$  [see Rule (vi)].

For  $n = 2$ , Rule (viii) simplifies to

$$\begin{aligned} & \text{Var}(\alpha_1 Y_1 + \alpha_2 Y_2) \\ &= \alpha_1^2 \text{Var}(Y_1) + \alpha_2^2 \text{Var}(Y_2) + 2\alpha_1 \alpha_2 \cdot \text{Cov}(Y_1, Y_2). \end{aligned} \quad (7.12)$$

Similarly, for  $n = m = 2$ , Rule (ix) simplifies to

**Box 7.1 Rules of Computation for Covariances**

Let  $X, Y$  be numerical random variables on the probability space  $(\Omega, \mathcal{A}, P)$  with  $E(X^2), E(Y^2) < \infty$ . Furthermore, let  $\alpha, \beta \in \mathbb{R}$ . Then:

$$\text{Cov}(X, Y) = E(X \cdot Y) - E(X) \cdot E(Y). \quad (\text{i})$$

$$\text{Cov}(\alpha + X, \beta + Y) = \text{Cov}(X, Y). \quad (\text{ii})$$

$$\text{Cov}(\alpha X, \beta Y) = \alpha \beta \text{Cov}(X, Y). \quad (\text{iii})$$

$$\text{Cov}(X, X) = \text{Var}(X). \quad (\text{iv})$$

$$\text{Cov}(X, Y) = \text{Cov}(Y, X). \quad (\text{v})$$

$$X \perp\!\!\!\perp Y \Rightarrow \text{Cov}(X, Y) = 0. \quad (\text{vi})$$

$$\exists \alpha \in \mathbb{R}: X \stackrel{p}{=} \alpha \Rightarrow \text{Cov}(X, Y) = 0. \quad (\text{vii})$$

If  $Y_i$  are real-valued random variables on the probability space  $(\Omega, \mathcal{A}, P)$  with  $E(Y_i^2) < \infty$  and  $\alpha_i \in \mathbb{R}$ ,  $i = 1, \dots, n$ , then

$$\text{Var}\left(\sum_{i=1}^n \alpha_i Y_i\right) = \sum_{i=1}^n \alpha_i^2 \text{Var}(Y_i) + \sum_{i=1}^n \sum_{j=1, i \neq j}^n \alpha_i \alpha_j \text{Cov}(Y_i, Y_j). \quad (\text{viii})$$

If  $X_i, Y_j$  are real-valued random variables on the probability space  $(\Omega, \mathcal{A}, P)$  with  $E(X_i^2), E(Y_j^2) < \infty$ , and  $\alpha_i, \beta_j \in \mathbb{R}$ , for all  $i = 1, \dots, n$  and  $j = 1, \dots, m$ , then

$$\text{Cov}\left(\sum_{i=1}^n \alpha_i X_i, \sum_{j=1}^m \beta_j Y_j\right) = \sum_{i=1}^n \sum_{j=1}^m \alpha_i \beta_j \text{Cov}(X_i, Y_j). \quad (\text{ix})$$

If  $X_1 \stackrel{p}{=} X_2$  and  $E(Y^2), E(X_1^2), E(X_2^2) < \infty$ , then

$$\text{Cov}(Y, X_1) = \text{Cov}(Y, X_2). \quad (\text{x})$$

$$\begin{aligned} & \text{Cov}(\alpha_1 X_1 + \alpha_2 X_2, \beta_1 Y_1 + \beta_2 Y_2) \\ &= \alpha_1 \beta_1 \text{Cov}(X_1, Y_1) + \alpha_1 \beta_2 \text{Cov}(X_1, Y_2) \\ & \quad + \alpha_2 \beta_1 \text{Cov}(X_2, Y_1) + \alpha_2 \beta_2 \text{Cov}(X_2, Y_2). \end{aligned} \quad (7.13)$$

&lt;

**Remark 7.11 (Covariance of Indicators)** For  $A, B \in \mathcal{A}$ , Rule (i) of Box 7.1 and Equations (1.33) and (6.4) yield

$$\text{Cov}(1_A, 1_B) = E(1_A \cdot 1_B) - E(1_A) \cdot E(1_B) \quad (7.14)$$

$$= P(A \cap B) - P(A) \cdot P(B). \quad (7.15)$$

&lt;

The following theorem helps to understand the relationship between the covariance and the variances of two numerical random variables  $X$  and  $Y$ .

**Theorem 7.12 (Cauchy-Schwarz Inequality)**

If  $X, Y: (\Omega, \mathcal{A}, P) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  are random variables with  $E(X^2), E(Y^2) < \infty$ , then

$$\text{Cov}(X, Y)^2 \leq \text{Var}(X) \cdot \text{Var}(Y). \quad (7.16)$$

Furthermore, if  $\text{Cov}(X, Y) \neq 0$ , then

$$\text{Cov}(X, Y)^2 = \text{Var}(X) \cdot \text{Var}(Y) \Leftrightarrow \exists a, b \in \mathbb{R}: Y \stackrel{P}{=} a + b \cdot X. \quad (7.17)$$

(Proof p. 233)

**Remark 7.13 (Squared Weighted Sum of Random Variables)** If  $X, Y: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  are random variables with  $E(X^2), E(Y^2) < \infty$  and  $\alpha, \beta \in \mathbb{R}$ , then  $E[(\alpha X + \beta Y)^2] < \infty$  (see Exercise 7-4).  $\triangleleft$

In the following theorem we revisit the linear quasi-regression, studying three equivalent propositions. The first of these propositions deals with the residual variable  $\epsilon := Y - f(X)$ , where  $f$  is the linear quasi-regression of  $Y$  on  $X$ . Note that this residual is not necessarily identical to the residual with respect to a conditional expectation that will be treated in chapters 9 to 11.

**Theorem 7.14 (Three Characterizations of the Linear Quasi-Regression)**

Let  $X, Y: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  be two real-valued random variables with  $E(X^2), E(Y^2) < \infty$ , and  $\text{Var}(X) > 0$ . Furthermore, let  $\alpha_0, \alpha_1 \in \mathbb{R}$ ,  $f(X) = \alpha_0 + \alpha_1 X$  be the composition of  $X$  and  $f: \mathbb{R} \rightarrow \mathbb{R}$ , and define  $\epsilon := Y - f(X)$ . Then the following three propositions are equivalent to each other:

(i)  $E(\epsilon) = \text{Cov}(X, \epsilon) = 0$ .

(ii)  $\alpha_0 = E(Y) - \alpha_1 E(X)$  and  $\alpha_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$ .

(iii)  $f(X) = Q_{lin}(Y|X)$ , i. e.,  $\alpha_0, \alpha_1$  minimize the function  $MSE(a_0, a_1)$  defined by Equation (7.2).

(Proof p. 234)

**Remark 7.15 (Uniqueness)** Suppose that the assumptions of Theorem 7.14 hold and  $f(X) = Q_{lin}(Y|X)$ . Then proposition (ii) of this theorem implies that the coefficients  $\alpha_0$  and  $\alpha_1$  are uniquely defined. Because  $Q_{lin}(Y|X) = \alpha_0 + \alpha_1 X$ , the linear quasi-regression  $f: \mathbb{R} \rightarrow \mathbb{R}$  is uniquely defined as well.  $\triangleleft$

**Remark 7.16 (Relationship Between Slope and Covariance)** According to proposition (ii), a *zero covariance* between  $X$  and  $Y$  implies that the slopes of the linear quasi-regressions of  $Y$  on  $X$  and of  $X$  on  $Y$  are zero, provided that  $\text{Var}(Y) >$

0. A *negative covariance* implies that the slopes of the linear quasi-regressions of  $Y$  on  $X$  and of  $X$  on  $Y$  are negative, and a *positive covariance* implies that the slopes of the linear quasi-regressions of  $Y$  on  $X$  and of  $X$  on  $Y$  are positive.  $\triangleleft$

**Example 7.17 (Discrete Regressor With Three Different Values – continued)** In Example 7.5 we specified the distribution of  $(X, Y)$ . Now we use the equations in Theorem 7.14 (ii) in order to compute the coefficients  $\alpha_0$  and  $\alpha_1$ . For this purpose we have to compute the expectations of  $X$  and  $Y$ , the variance of  $X$ , and the covariance of  $X$  and  $Y$ . Hence, with  $n = 3$ ,  $m = 2$ ,  $x_1 = 1, x_2 = 2, x_3 = 3$ , and  $y_1 = 1, y_2 = 2$ ,

$$E(X) = \sum_{i=1}^n x_i \cdot P(X=x_i) = 1 \cdot \frac{1}{4} + 2 \cdot \frac{1}{2} + 3 \cdot \frac{1}{4} = 2,$$

$$E(Y) = \sum_{i=1}^m y_i \cdot P(Y=y_i) = 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{2} = \frac{3}{2},$$

$$\begin{aligned} \text{Var}(X) &= E(X^2) - E(X)^2 = \sum_{i=1}^n x_i^2 \cdot P(X=x_i) - E(X)^2 \quad [\text{Box 6.2 (i), (6.19)}] \\ &= 1^2 \cdot \frac{1}{4} + 2^2 \cdot \frac{1}{2} + 3^2 \cdot \frac{1}{4} - 2^2 = \frac{1}{2}, \end{aligned}$$

$$\text{Cov}(X, Y) = E(X \cdot Y) - E(X) \cdot E(Y) \quad [\text{Box 7.1 (i)}]$$

$$= \sum_{i=1}^n \sum_{j=1}^m x_i \cdot y_j \cdot P((X, Y)=(x_i, y_j)) - E(X) \cdot E(Y) \quad [(6.3)]$$

$$= 1 \cdot 1 \cdot \frac{1}{4} + 2 \cdot 2 \cdot \frac{1}{2} + 3 \cdot 1 \cdot \frac{1}{4} - 2 \cdot \frac{3}{2}$$

$$= \frac{1}{4} + \frac{8}{4} + \frac{3}{4} - \frac{12}{4} = 0.$$

Using the equations in Theorem 7.14 (ii) yields  $\alpha_1 = \text{Cov}(X, Y) / \text{Var}(X) = \frac{0}{1/2} = 0$  and  $\alpha_0 = E(Y) - \alpha_1 E(X) = \frac{3}{2} - 0 \cdot 2 = 1.5$ , the same result as obtained in Exercise 7-1, in which we minimize the function  $MSE(a_0, a_1)$ .  $\triangleleft$

### 7.3 Correlation

As mentioned before, the covariance between two numerical random variables quantifies the strength of the dependence that can be described by a linear quasi-regression. However, the covariance is not invariant under multiplication with constants [scale transformations; see Box 7.1 (iii)] of the random variables involved. In contrast, the correlation, which quantifies the strength of the same kind of dependence *is invariant* under scale transformations (see Rem. 7.22).

#### Definition 7.18 (Correlation)

Let  $X, Y: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  be two numerical random variables with  $E(X^2)$ ,

$E(Y^2) < \infty$ . Then the correlation of  $X$  and  $Y$  is defined by

$$\text{Corr}(X, Y) := \begin{cases} \frac{\text{Cov}(X, Y)}{SD(X) \cdot SD(Y)}, & \text{if } SD(X), SD(Y) > 0, \\ 0, & \text{otherwise.} \end{cases} \quad (7.18)$$

**Remark 7.19 (Correlation of a Random Variable With Itself)** Assume  $\text{Var}(X) > 0$ . Because  $\text{Cov}(X, X) = \text{Var}(X) = SD(X) \cdot SD(X)$ , Equation (7.18) implies that  $\text{Corr}(X, X) = 1$ . Similarly, because  $\text{Cov}(X, -X) = -\text{Var}(X) = -SD(X) \cdot SD(X)$ , Equation (7.18) implies that  $\text{Corr}(X, -X) = -1$ .  $\triangleleft$

**Remark 7.20 (Range of the Correlation)** An implication of (7.16) is

$$-1 \leq \text{Corr}(X, Y) \leq 1, \quad (7.19)$$

provided that  $\text{Corr}(X, Y)$  exists, i. e., provided that the assumptions hold under which the correlation is defined (see Exercise 7-5).  $\triangleleft$

**Remark 7.21 (Correlation and  $Z$ -Transformed Variables)** If the standard deviations of  $X$  and  $Y$  are positive, then the correlation is also the expectation of the product of the  $Z$ -transformed variables [see Eq. (6.33)], i. e.,

$$\text{Corr}(X, Y) = E\left(\frac{X - E(X)}{SD(X)} \cdot \frac{Y - E(Y)}{SD(Y)}\right) \quad (7.20)$$

(see Exercise 7-6).  $\triangleleft$

**Remark 7.22 (An Invariance Property of the Correlation)** The correlation of linear transformations of  $X$  and  $Y$  is

$$\text{Corr}(a_0 + a_1 X, b_0 + b_1 Y) = \begin{cases} \text{Corr}(X, Y), & \text{if } a_1 \cdot b_1 > 0 \\ -\text{Corr}(X, Y), & \text{if } a_1 \cdot b_1 < 0, \\ 0, & \text{if } a_1 \cdot b_1 = 0 \end{cases} \quad (7.21)$$

where  $a_0, a_1, b_0, b_1 \in \mathbb{R}$  (see Exercise 7-7). This equation implies that the correlation is invariant (up to change of signs) under linear transformations, which include *translations* ( $a_1 = 1$  and  $b_1 = 1$ ) and *scale transformations* ( $a_0 = b_0 = 0$  and  $a_1, b_1 \neq 0$ ).  $\triangleleft$

Theorem 7.12 implies the following corollary about the cases  $\text{Corr}(X, Y) = 1$  or  $\text{Corr}(X, Y) = -1$ , i. e., the ‘perfect’ correlation.

**Corollary 7.23 (Perfect Correlation of Two Random Variables)**

Let the assumptions of Definition 7.8 be satisfied and suppose that  $\text{Var}(X), \text{Var}(Y) > 0$ . Then  $|\text{Corr}(X, Y)| = 1$  if and only if there are  $a_0, a_1 \in \mathbb{R}$ ,  $a_1 \neq 0$ , such that  $Y \stackrel{p}{=} a_0 + a_1 X$ . In this case

$$\text{Corr}(X, Y) = \begin{cases} 1, & \text{if } a_1 > 0 \\ -1, & \text{if } a_1 < 0. \end{cases} \quad (7.22)$$

**Remark 7.24 (Covariance and Standard Deviations)** Let  $Y \stackrel{p}{=} a_0 + a_1 X$ ,  $a_0, a_1 \in \mathbb{R}$ . If  $a_1 \geq 0$ , then  $\text{Cov}(X, Y) = \text{SD}(X) \cdot \text{SD}(Y)$ . If  $a_1 < 0$  then  $\text{Cov}(X, Y) = -\text{SD}(X) \cdot \text{SD}(Y)$  (see Exercise 7-8).  $\triangleleft$

**Remark 7.25 (Slope of a Linear Quasi-Regression and Correlation)** If  $\alpha_1$  is the slope of  $Q_{lin}(Y|X)$  (see Def. 7.2), then

$$\alpha_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \text{Corr}(X, Y) \cdot \frac{\text{SD}(Y)}{\text{SD}(X)} \quad (7.23)$$

[see proposition (ii) of Th. 7.14]. This equation shows that the slope  $\alpha_1$  of the linear quasi-regression has the same sign as the covariance and the correlation. The size of the absolute value of  $\alpha_1$  depends on the ratio  $\text{SD}(Y)/\text{SD}(X)$  of the standard deviations. The smaller the standard deviation of  $X$  compared to the standard deviation of  $Y$ , the larger the absolute value of  $\alpha_1$ . Furthermore, given a fixed variance  $\text{Var}(X)$ , this equation also shows that  $\alpha_1$  is proportional to  $\text{Cov}(X, Y)$  and  $\text{Corr}(X, Y)$ . In this sense, all three parameters  $\alpha_1$ ,  $\text{Cov}(X, Y)$ , and  $\text{Corr}(X, Y)$  quantify the strength of the dependence of  $Y$  on  $X$  described by a linear quasi-regression. Note, however, that  $\alpha_1$  and  $\text{Cov}(X, Y)$  are not invariant under scale transformations of  $X$  and  $Y$ . This can be seen in the following equation for the slope  $\alpha_1^*$  of the linear quasi-regression of  $bY$  on  $aX$ ,  $a, b \in \mathbb{R}$ ,  $a, b \neq 0$ :

$$\alpha_1^* = \frac{\text{Cov}(aX, bY)}{\text{Var}(aX)} = \text{Corr}(aX, bY) \cdot \frac{\text{SD}(bY)}{\text{SD}(aX)} = \frac{b}{a} \cdot \alpha_1. \quad (7.24)$$

Hence, the slope of the linear quasi-regression of  $bY$  on  $aX$  is identical to the slope of the linear quasi-regression of  $Y$  on  $X$  multiplied by  $\frac{b}{a}$ . In contrast, the slope of the linear quasi-regression is invariant under translations  $c + X$ ,  $d + Y$ ,  $c, d \in \mathbb{R}$  (see Exercise 7-9).  $\triangleleft$

**Remark 7.26 (Correlation and Coefficient of Determination)** Under the assumptions of Definition 7.18 and  $\text{Var}(X), \text{Var}(Y) > 0$ ,

$$\begin{aligned} Q_{Y|X}^2 &= \frac{\text{Var}[Q_{lin}(Y|X)]}{\text{Var}(Y)} = \frac{\text{Var}[\alpha_0 + \alpha_1 \text{Var}(X)]}{\text{Var}(Y)} = \frac{\alpha_1^2 \text{Var}(X)}{\text{Var}(Y)} \\ &= \text{Corr}(X, Y)^2. \end{aligned} \quad (7.25)$$

 $\triangleleft$

**Example 7.27 (Joe and Ann With Random Assignment – continued)** Consider the example presented in Table 2.2 (p. 53). In this example, the covariance of  $X$  and  $Y$  is most easily computed using

$$\begin{aligned} \text{Cov}(X, Y) &= E(X \cdot Y) - E(X) \cdot E(Y) && \text{[Box 7.1 (i)]} \\ &= \sum_{(x,y)} (x \cdot y) \cdot P(X=x, Y=y) - P(X=1) \cdot P(Y=1) && \text{[(6.15)]} \\ &= P(X=1, Y=1) - P(X=1) \cdot P(Y=1) \\ &= (.16 + .08) - .40 \cdot .51 = .036, \end{aligned}$$

where  $P(X=1) = E(X) = .40$  and  $P(Y=1) = E(Y) = .51$  have been computed in Example 6.7. Note that  $\sum_{(x,y)}$  is the sum over all pairs  $(x, y)$  of values of  $X$  and  $Y$ . In this example, there are four such pairs, only one of which, namely  $(1, 1)$ , yields a product  $x \cdot y \neq 0$ . Using the results of Example 6.31 on the variances of  $X$  and  $Y$  yields the correlation

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\text{SD}(X) \cdot \text{SD}(Y)} = \frac{.036}{\sqrt{.24} \cdot \sqrt{.2499}} \approx .147$$

Hence, treatment and outcome variables have a positive correlation. This is in accordance with comparing the conditional probability of success given treatment,  $P(C|B) = .60$ , to the conditional probability of success given no treatment,  $P(C|B^c) = .45$  (see Example 4.16).

In this example,

$$\begin{aligned} Q_{lin}(Y|X) &= \alpha_0 + \alpha_1 \cdot X \\ &= \left( E(Y) - \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \cdot E(X) \right) + \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \cdot X && \text{[Th. 7.14 (ii)]} \\ &= \left( .51 - \frac{.036}{.24} \cdot .40 \right) + \frac{.036}{.24} \cdot X \\ &= .45 + .15 \cdot X, \end{aligned}$$

and the linear quasi-regression  $f: \mathbb{R} \rightarrow \mathbb{R}$  of  $Y$  on  $X$  is specified by

$$f(x) = .45 + .15 \cdot x, \quad \forall x \in \mathbb{R}.$$

◁

## 7.4 Expectation Vector and Covariance Matrix

### 7.4.1 Random Vector and Random Matrix

Let  $X = (X_1, \dots, X_n)$  be an  $n$ -variate numerical random variable on a probability space  $(\Omega, \mathcal{A}, P)$ . In order to utilize matrix algebra, we consider the *column random vector*

$$\mathbf{x} := \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix},$$

i. e., the column vector of the random variables  $X_1, \dots, X_n$ . Correspondingly, we consider the *row random vector*  $\mathbf{x}' := [X_1, \dots, X_n]$ , the *transpose* of  $\mathbf{x}$ .

In this section, we also consider a *random matrix*, i. e., a matrix

$$\mathbf{X} := \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1m} \\ X_{21} & X_{22} & \dots & X_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{nm} \end{bmatrix} \quad (7.26)$$

of type  $n \times m$  of numerical random variables  $X_{ij}$  on a probability space  $(\Omega, \mathcal{A}, P)$ ,  $i = 1, \dots, n$ ,  $j = 1, \dots, m$ . Such a random matrix is an  $n \times m$ -array of an  $(n \cdot m)$ -variate random variable (see section 5.3).

#### 7.4.2 Expectation of a Random Vector and a Random Matrix

The *expectation of a (row or column) random vector* is defined as the (row or column) vector of the expectations of its components, i. e.,

$$E(\mathbf{x}') := [E(X_1), \dots, E(X_n)] \quad (7.27)$$

and  $E(\mathbf{x}) := [E(X_1), \dots, E(X_n)]'$ , provided that the expectations exist. Hence,

$$E(\mathbf{x}') = (E(\mathbf{x}))'. \quad (7.28)$$

Analogously to Equation (7.27), the *expectation of an  $n \times m$ -random matrix* is defined as the  $n \times m$ -matrix of the expectations of its components, i. e.,

$$E \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1m} \\ X_{21} & X_{22} & \dots & X_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{nm} \end{bmatrix} := \begin{bmatrix} E(X_{11}) & E(X_{12}) & \dots & E(X_{1m}) \\ E(X_{21}) & E(X_{22}) & \dots & E(X_{2m}) \\ \vdots & \vdots & \ddots & \vdots \\ E(X_{n1}) & E(X_{n2}) & \dots & E(X_{nm}) \end{bmatrix}, \quad (7.29)$$

provided that the expectations exist. Obviously, if  $\mathbf{X}'$  denotes the transpose of the matrix  $\mathbf{X}$ , then

$$E(\mathbf{X}') = (E(\mathbf{X}))'. \quad (7.30)$$

In Box 7.2 we present some rules of computation for the expectations of random vectors and random matrices (for proofs see Exercise 7-10). In this box we use

$$\mathbf{a}' \mathbf{x} := \sum_{i=1}^n a_i \cdot X_i, \quad (7.31)$$

**Box 7.2 Rules of Computation for Expectations of Random Matrices**

Let  $X = (X_1, \dots, X_n)$  be an  $n$ -variate and  $Y = (Y_1, \dots, Y_m)$  an  $m$ -variate real-valued random variable on a probability space  $(\Omega, \mathcal{A}, P)$  such that the expectations of  $X_i$  and  $Y_j$  are finite, for all  $i = 1, \dots, n, j = 1, \dots, m$ . Furthermore, let  $\mathbf{x} = [X_1, \dots, X_n]'$  and  $\mathbf{y} = [Y_1, \dots, Y_m]'$  denote column vectors,  $\mathbf{a} = [a_1, \dots, a_n]'$  a column vector of real numbers, and let  $\mathbf{A}$  and  $\mathbf{B}$  be matrices of types  $k \times n$  and  $k \times m$ , respectively, each of their components being a real number. Furthermore, let  $\mathbf{C}$  and  $\mathbf{D}$  be matrices of real numbers of types  $l \times n$  and  $r \times m$ , respectively. Then

$$\mathbf{x} \stackrel{p}{=} \mathbf{a} \Rightarrow E(\mathbf{x}) = \mathbf{a}. \quad (\text{i})$$

$$E(\mathbf{a} + \mathbf{x}) = \mathbf{a} + E(\mathbf{x}). \quad (\text{ii})$$

$$E(\mathbf{a}'\mathbf{x}) = \mathbf{a}'E(\mathbf{x}) = E(\mathbf{x})'\mathbf{a} = E(\mathbf{x}'\mathbf{a}). \quad (\text{iii})$$

$$E(\mathbf{A}\mathbf{x}) = \mathbf{A}E(\mathbf{x}). \quad (\text{iv})$$

$$E(\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y}) = \mathbf{A}E(\mathbf{x}) + \mathbf{B}E(\mathbf{y}). \quad (\text{v})$$

Let  $\mathbf{X}$  be an  $(n \times k)$ -matrix and  $\mathbf{Y}$  an  $(m \times k)$ -matrix of real-valued random variables on  $(\Omega, \mathcal{A}, P)$ , all with finite second moments. Then

$$\mathbf{X} \stackrel{p}{=} \mathbf{A}' \Rightarrow E(\mathbf{X}) = \mathbf{A}'. \quad (\text{vi})$$

$$E(\mathbf{A}' + \mathbf{X}) = \mathbf{A}' + E(\mathbf{X}). \quad (\text{vii})$$

$$E(\mathbf{C}\mathbf{X}) = \mathbf{C}E(\mathbf{X}). \quad (\text{viii})$$

$$E(\mathbf{C}\mathbf{X}\mathbf{Y}'\mathbf{D}') = \mathbf{C}E(\mathbf{X}\mathbf{Y}')\mathbf{D}'. \quad (\text{ix})$$

the *inner product* of an  $n$ -vector  $\mathbf{a} = [a_1, \dots, a_n]'$  of real numbers and the random vector  $\mathbf{x}$ . Correspondingly,

$$\mathbf{A}\mathbf{x} := \begin{bmatrix} \mathbf{a}'_1 \mathbf{x} \\ \vdots \\ \mathbf{a}'_k \mathbf{x} \end{bmatrix}, \quad (7.32)$$

where  $\mathbf{a}'_l$ ,  $l = 1, \dots, k$ , denotes the  $l$ th row of the  $(k \times n)$ -matrix  $\mathbf{A}$  of real numbers.

**7.4.3 Covariance Matrix of two Multivariate Random Variables**

Now we consider two multivariate numerical random variables  $X = (X_1, \dots, X_n)$  and  $Y = (Y_1, \dots, Y_m)$  on a probability space  $(\Omega, \mathcal{A}, P)$ . In particular, we assume that the second moments of all these random variables are finite and focus on their

covariance matrix, again utilizing the representation of  $X$  and  $Y$  as row or column vectors that has been introduced at the beginning of section 7.4.1.

Note that  $[\mathbf{x} - E(\mathbf{x})][\mathbf{y} - E(\mathbf{y})]'$  is an  $n \times m$ -matrix of the random variables

$$[X_i - E(X_i)] \cdot [Y_j - E(Y_j)], \quad i = 1, \dots, n, \quad j = 1, \dots, m.$$

Therefore, using (7.29), the *covariance matrix*  $\Sigma_{xy}$  is defined by

$$\Sigma_{xy} := E([\mathbf{x} - E(\mathbf{x})][\mathbf{y} - E(\mathbf{y})]'). \quad (7.33)$$

In other words, the covariance matrix of  $\mathbf{x}$  and  $\mathbf{y}$  is the matrix of covariances, i. e.,

$$\Sigma_{xy} = \begin{bmatrix} \sigma_{X_1 Y_1} & \sigma_{X_1 Y_2} & \cdots & \sigma_{X_1 Y_m} \\ \sigma_{X_2 Y_1} & \sigma_{X_2 Y_2} & \cdots & \sigma_{X_2 Y_m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{X_n Y_1} & \sigma_{X_n Y_2} & \cdots & \sigma_{X_n Y_m} \end{bmatrix}, \quad (7.34)$$

where  $\sigma_{X_i Y_j} := Cov(X_i, Y_j) = E([X_i - E(X_i)] \cdot [Y_j - E(Y_j)])$ ,  $i = 1, \dots, n$ ,  $j = 1, \dots, m$ . If we assume that the second moments of the random variables  $X_i$  and  $Y_j$  are finite, then all covariances  $Cov(X_i, Y_j)$  are finite as well, and we say that  $\Sigma_{xy}$  exists.

If we consider a univariate random variable  $Y$ , then  $\mathbf{y} = [Y]$  is also a vector consisting of a single component, the random variable  $Y$ . In this special case,  $\Sigma_{xy}$  is a matrix of type  $n \times 1$ , the column vector

$$\Sigma_{xy} = \begin{bmatrix} \sigma_{X_1 Y} \\ \vdots \\ \sigma_{X_n Y} \end{bmatrix}.$$

Another special case is  $\mathbf{x} = \mathbf{y}$ . The covariance matrix  $\Sigma_{xx}$  of  $\mathbf{x}$  and  $\mathbf{x}$  is called the *variance-covariance matrix of  $\mathbf{x}$*  (or of  $X$ ). Hence,

$$\Sigma_{xx} := E([\mathbf{x} - E(\mathbf{x})][\mathbf{x} - E(\mathbf{x})]') \quad (7.35)$$

and

$$\Sigma_{xx} = \begin{bmatrix} \sigma_{X_1}^2 & \sigma_{X_1 X_2} & \cdots & \sigma_{X_1 X_n} \\ \sigma_{X_2 X_1} & \sigma_{X_2}^2 & \cdots & \sigma_{X_2 X_n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{X_n X_1} & \sigma_{X_n X_2} & \cdots & \sigma_{X_n}^2 \end{bmatrix}. \quad (7.36)$$

The diagonal components of the matrix  $\Sigma_{xx}$  are the variances of the variables  $X_1, \dots, X_n$ , because  $\sigma_{X_i X_i} := Cov(X_i, X_i) = Var(X_i) = \sigma_{X_i}^2$ ,  $i = 1, \dots, n$ .

In Box 7.3 we present some rules of computation for covariance matrices. They are proved in Exercise 7-11.

**Box 7.3 Rules of Computation for Covariance Matrices**

Let  $X = (X_1, \dots, X_n)$  be an  $n$ -variate and  $Y = (Y_1, \dots, Y_m)$  an  $m$ -variate real-valued random variable on a probability space  $(\Omega, \mathcal{A}, P)$  such that the second moments of  $X_i$  and  $Y_j$  are finite, for all  $i = 1, \dots, n, j = 1, \dots, m$ . Furthermore, let  $\mathbf{x} = [X_1, \dots, X_n]'$  and  $\mathbf{y} = [Y_1, \dots, Y_m]'$  denote column vectors,  $\mathbf{a} = [a_1, \dots, a_n]'$  and  $\mathbf{b} = [b_1, \dots, b_m]'$  column vectors of real numbers, and let  $\mathbf{A}$  and  $\mathbf{B}$  be matrices of types  $k \times n$  and  $k \times m$ , respectively, each of their components being a real number. Finally, let  $\mathbf{O}$  denote the  $(n \times m)$ -null matrix. Then:

$$\Sigma_{\mathbf{x}\mathbf{y}} = E(\mathbf{x}\mathbf{y}') - E(\mathbf{x})E(\mathbf{y}'). \quad (\text{i})$$

$$\Sigma_{\mathbf{a}+\mathbf{x}, \mathbf{b}+\mathbf{y}} = \Sigma_{\mathbf{x}\mathbf{y}}. \quad (\text{ii})$$

$$\Sigma_{\mathbf{A}\mathbf{x}, \mathbf{B}\mathbf{y}} = \mathbf{A} \Sigma_{\mathbf{x}\mathbf{y}} \mathbf{B}'. \quad (\text{iii})$$

$$\Sigma_{\mathbf{x}\mathbf{y}} = \Sigma'_{\mathbf{y}\mathbf{x}}. \quad (\text{iv})$$

$$X \perp\!\!\!\perp Y \Rightarrow \Sigma_{\mathbf{x}\mathbf{y}} = \mathbf{O}. \quad (\text{v})$$

$$\mathbf{x} \stackrel{p}{=} \mathbf{a} \Rightarrow \Sigma_{\mathbf{x}\mathbf{y}} = \mathbf{O}. \quad (\text{vi})$$

Additionally, let  $\mathbf{w} = [W_1, \dots, W_r]'$  and  $\mathbf{z} = [Z_1, \dots, Z_s]'$  be real-valued random column vectors on  $(\Omega, \mathcal{A}, P)$  such that all their components have finite second moments. Furthermore, let  $\mathbf{C}$  and  $\mathbf{D}$  be matrices of real numbers of type  $l \times r$  and  $l \times s$ . Then:

$$\Sigma_{\mathbf{A}\mathbf{x}+\mathbf{B}\mathbf{y}, \mathbf{C}\mathbf{w}+\mathbf{D}\mathbf{z}} = \mathbf{A} \Sigma_{\mathbf{x}\mathbf{w}} \mathbf{C}' + \mathbf{A} \Sigma_{\mathbf{x}\mathbf{z}} \mathbf{D}' + \mathbf{B} \Sigma_{\mathbf{y}\mathbf{w}} \mathbf{C}' + \mathbf{B} \Sigma_{\mathbf{y}\mathbf{z}} \mathbf{D}'. \quad (\text{vii})$$

$$n = s \text{ and } \mathbf{x} \stackrel{p}{=} \mathbf{z} \Rightarrow \Sigma_{\mathbf{x}\mathbf{y}} = \Sigma_{\mathbf{z}\mathbf{y}}. \quad (\text{viii})$$

**7.5 Multiple Linear Quasi-Regression**

In the following definition we generalize the concept of a linear quasi-regression considering a multivariate regressor  $X = (X_1, \dots, X_n)$ . We use the notation  $\mathbf{x} = [X_1, \dots, X_n]'$  to denote the column vector of  $X$ ,  $\boldsymbol{\beta}' = [\beta_1, \dots, \beta_n]$  for the row vector of the real numbers  $\beta_1, \dots, \beta_n$ , and  $\mathbf{b}' = [b_1, \dots, b_n]$  for the row vector of the real numbers  $b_1, \dots, b_n$ .

**Definition 7.28 (Multiple Linear Quasi-Regression)**

Let  $X_i: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$ ,  $i = 1, \dots, n$ , and  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  be real-valued random variables, define  $X := (X_1, \dots, X_n)$ , assume  $E(X_i^2), E(Y^2) < \infty$ ,

$i = 1, \dots, n$ , and that the inverse  $\Sigma_{xx}^{-1}$  exists. Define the function  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  by

$$f(x) = \beta_0 + \sum_{i=1}^n \beta_i x_i, \quad \forall x = (x_1, \dots, x_n) \in \mathbb{R}^n, \quad (7.37)$$

where  $\beta_0, \boldsymbol{\beta} = [\beta_1, \dots, \beta_n]'$  minimize the function  $MSE: \mathbb{R}^{n+1} \rightarrow \mathbb{R}$  with

$$MSE(b_0, b_1, \dots, b_n) = E([Y - (b_0 + \mathbf{b}'\mathbf{x})]^2), \quad \forall (b_0, b_1, \dots, b_n) \in \mathbb{R}^{n+1}. \quad (7.38)$$

Then  $f$  is called the *linear quasi-regression of  $Y$  on  $X_1, \dots, X_n$* . The composition of  $X$  and  $f$  is denoted by  $Q_{lin}(Y|X)$  or  $Q_{lin}(Y|X_1, \dots, X_n)$ , i. e.,

$$Q_{lin}(Y|X) := f(X) = \beta_0 + \boldsymbol{\beta}'\mathbf{x} = \beta_0 + \sum_{i=1}^n \beta_i X_i. \quad (7.39)$$

**Remark 7.29 (Coefficient of Determination)** Let the assumptions of Definition 7.28 hold and assume  $Var(Y) > 0$ . Then

$$Q_{Y|X}^2 := \frac{Var[Q_{lin}(Y|X)]}{Var(Y)} \quad (7.40)$$

is called the *coefficient of determination* of the linear quasi-regression  $Q_{lin}(Y|X)$ .  $\triangleleft$

In the following theorem, we generalize Theorem 7.14 considering a multivariate real-valued regressor  $X = (X_1, \dots, X_n)$ . In this theorem  $\Sigma_{x\epsilon}$  denotes the covariance vector of  $\mathbf{x}$  and  $\epsilon$ , which is defined by

$$\epsilon := Y - Q_{lin}(Y|X_1, \dots, X_n) \quad (7.41)$$

and called the *residual of  $Y$  with respect to its linear quasi-regression on  $X$* .

**Theorem 7.30 (Characterizations of the Multiple Linear Quasi-Regression)**

Let  $X_1, \dots, X_n, Y: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  be real-valued random variables, assume  $E(X_i^2), E(Y^2) < \infty$  for all  $i = 1, \dots, n$ , and that the inverse  $\Sigma_{xx}^{-1}$  exists. Furthermore, let  $f: \mathbb{R}^n \rightarrow \mathbb{R}$ , let

$$f(X_1, \dots, X_n) := \beta_0 + \boldsymbol{\beta}'\mathbf{x}, \quad \beta_0 \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^n, \quad (7.42)$$

be the composition of  $(X_1, \dots, X_n)$  and  $f$ , and define  $\epsilon := Y - f(X_1, \dots, X_n)$ . Then the following three propositions are equivalent to each other:

- (i)  $E(\epsilon) = 0$  and  $\Sigma_{x\epsilon} = \mathbf{0}$ .

$$(ii) \beta_0 = E(Y) - \boldsymbol{\beta}'E(\mathbf{x}) \quad \text{and} \quad \boldsymbol{\beta} = \boldsymbol{\Sigma}_{xx}^{-1} \boldsymbol{\Sigma}_{xy}.$$

$$(iii) f(X_1, \dots, X_n) = Q_{lin}(Y|X_1, \dots, X_n).$$

(Proof p. 235)

**Remark 7.31 (Uniqueness)** Suppose that the assumptions of Theorem 7.30 hold and  $f(X_1, \dots, X_n) = Q_{lin}(Y|X)$ . Then proposition (ii) of this theorem implies that the coefficients  $\beta_0, \beta_1, \dots, \beta_n$  are uniquely defined. Because  $f(X_1, \dots, X_n) = \beta_0 + \boldsymbol{\beta}'\mathbf{x} = Q_{lin}(Y|X)$ , the linear quasi-regression  $f$  and  $Q_{lin}(Y|X)$  are uniquely defined as well.  $\triangleleft$

**Corollary 7.32 (No Correlation Between Linear Quasi-Regression and its Residual)**

Under the assumptions of Definition 7.28,

$$\text{Cov}[Q_{lin}(Y|X), \epsilon] = 0. \quad (7.43)$$

and

$$\text{Var}[Q_{lin}(Y|X)] = \text{Cov}[Y, Q_{lin}(Y|X)]. \quad (7.44)$$

(Proof p. 237)

**Remark 7.33 (Correlation Between Linear Quasi-Regression and its Regressand)**

Under the assumptions of Definition 7.28 and  $\text{Var}(Y) > 0$ ,

$$Q_{Y|X}^2 = \text{Corr}[Y, Q_{lin}(Y|X)]^2 \quad (7.45)$$

(see Exercise 7-12).  $\triangleleft$

## 7.6 Proofs

### Proof of Theorem 7.12

Suppose  $\text{Var}(Y) = 0$ . Then rules (iv) of Box 6.2 and (vii) of Box 7.1 imply that  $\text{Cov}(X, Y) = 0$ . This shows that the Inequality (7.16) holds if  $\text{Var}(Y) = 0$ . Now suppose  $\text{Var}(Y) > 0$ . Then

$$\begin{aligned} 0 &\leq \text{Var}\left(X - \frac{\text{Cov}(X, Y)}{\text{Var}(Y)} \cdot Y\right) \cdot \text{Var}(Y) \\ &= \left(\text{Var}(X) - 2 \cdot \frac{\text{Cov}(X, Y)}{\text{Var}(Y)} \text{Cov}(X, Y) + \frac{\text{Cov}(X, Y)^2}{\text{Var}(Y)^2} \text{Var}(Y)\right) \cdot \text{Var}(Y) \quad [(7.12)] \\ &= \text{Var}(X) \cdot \text{Var}(Y) - \text{Cov}(X, Y)^2, \end{aligned}$$

which is equivalent to (7.16). According to the first part of the proof,

$$\text{Cov}(X, Y)^2 = \text{Var}(X) \cdot \text{Var}(Y) \Leftrightarrow \text{Var}(Y) = 0 \quad \text{or} \quad \text{Var}\left(X - \frac{\text{Cov}(X, Y)}{\text{Var}(Y)} \cdot Y\right) = 0.$$

Rule (iv) of Box 6.2 implies that this is equivalent to

$$\exists a \in \mathbb{R} : Y \stackrel{P}{=} a \quad \text{or} \quad \exists c \in \mathbb{R} : X - \frac{\text{Cov}(X, Y)}{\text{Var}(Y)} \cdot Y \stackrel{P}{=} c.$$

If  $\text{Cov}(X, Y) \neq 0$ , this is equivalent to

$$\exists a \in \mathbb{R} : Y \stackrel{P}{=} a + 0 \cdot X \quad \text{or} \quad \exists c \in \mathbb{R} : Y \stackrel{P}{=} \left( -\frac{c \cdot \text{Var}(Y)}{\text{Cov}(X, Y)} \right) + \left( \frac{\text{Var}(Y)}{\text{Cov}(X, Y)} \right) \cdot X.$$

Obviously, in both cases, there is a linear function of  $X$  that is  $P$ -equivalent to  $Y$ . Furthermore, if there are  $a, b \in \mathbb{R}$  with  $Y \stackrel{P}{=} a + bX$  — implying  $\text{Var}(Y) = b^2 \text{Var}(X)$  [see Box 6.2 (ii), (iii)] — then

$$\begin{aligned} \text{Cov}(X, Y)^2 &= \text{Cov}(X, a + bX)^2 && \text{[Box 7.1 (x)]} \\ &= b^2 \text{Cov}(X, X)^2 && \text{[Box 7.1 (ii), (iii)]} \\ &= b^2 \text{Var}(X)^2 && \text{[Box 7.1 (iv)]} \\ &= \text{Var}(X) \cdot \text{Var}(Y). && \text{[Box 6.2 (ii), (iii)]} \end{aligned}$$

### ***Proof of Theorem 7.14***

The proof is organized as follows: (iii)  $\Rightarrow$  (ii)  $\Rightarrow$  (i)  $\Rightarrow$  (iii), which will prove that (i), (ii), and (iii) are equivalent.

(iii)  $\Rightarrow$  (ii). The first partial derivatives of

$$\begin{aligned} &MSE(a_0, a_1) \\ &= E\{[Y - (a_0 + a_1 X)]^2\} \\ &= E(Y^2) + E[(a_0 + a_1 X)^2] - 2E[Y \cdot (a_0 + a_1 X)] && \text{[Box 6.1 (vi)]} \\ &= E(Y^2) + a_0^2 + a_1^2 E(X^2) + 2a_0 a_1 E(X) - 2a_0 E(Y) - 2a_1 E(X \cdot Y) && \text{[Box 6.1 (iii), (vi)]} \end{aligned}$$

with respect to  $a_0$  and  $a_1$  are

$$\frac{\partial MSE(a_0, a_1)}{\partial a_0} = 2a_0 + 2a_1 E(X) - 2E(Y)$$

and

$$\frac{\partial MSE(a_0, a_1)}{\partial a_1} = 2a_1 E(X^2) + 2a_0 E(X) - 2E(X \cdot Y).$$

If  $f(X) = \alpha_0 + \alpha_1 X = Q_{lin}(Y|X)$ , then

$$2\alpha_0 + 2\alpha_1 E(X) - 2E(Y) = 0$$

and

$$2\alpha_1 E(X^2) + 2\alpha_0 E(X) - 2E(X \cdot Y) = 0.$$

Solving the first equation for  $\alpha_0$  yields

$$\alpha_0 = E(Y) - \alpha_1 E(X).$$

Inserting this result into the second equation yields

$$\alpha_1 E(X^2) + E(Y) \cdot E(X) - \alpha_1 E(X)^2 - E(X \cdot Y) = 0.$$

Using  $\text{Cov}(X, Y) = E(X \cdot Y) - E(X) \cdot E(Y)$  [see Box 7.1 (i)] and  $\text{Var}(X) = E(X^2) - E(X)^2$  [see Box 6.2 (i)] we receive

$$\alpha_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}.$$

(ii)  $\Rightarrow$  (i).

$$\begin{aligned} E(\epsilon) &= E[Y - f(X)] \\ &= E[Y - (\alpha_0 + \alpha_1 X)] && \text{[def. of } f(X)\text{]} \\ &= E\{Y - [E(Y) - \alpha_1 E(X) + \alpha_1 X]\} && \text{[(ii)]} \\ &= E(Y) - E(Y) + \alpha_1 E(X) - \alpha_1 \cdot E(X) && \text{[Box 6.1 (vii)]} \\ &= 0. \\ \text{Cov}(X, \epsilon) &= \text{Cov}(X, [Y - (\alpha_0 + \alpha_1 X)]) && \text{[def. of } \epsilon\text{]} \\ &= \text{Cov}(X, Y) - \alpha_1 \text{Var}(X) && \text{[Box 7.1 (ii), (iii)]} \\ &= \text{Cov}(X, Y) - \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \cdot \text{Var}(X) && \text{[(ii)]} \\ &= 0. \end{aligned}$$

(i)  $\Rightarrow$  (iii). Let  $f(X) := \alpha_0 + \alpha_1 X$ ,  $\alpha_0, \alpha_1 \in \mathbb{R}$ , be a linear function of  $X$  with  $E(\epsilon) = 0$  and  $\text{Cov}(X, \epsilon) = 0$ , where  $\epsilon := Y - f(X)$ . Then, for any linear function  $h(X) := a_0 + a_1 X$ ,  $a_0, a_1 \in \mathbb{R}$ ,

$$\begin{aligned} E((Y - f(X))[f(X) - h(X)]) &= E(\epsilon \cdot [f(X) - h(X)]) && \text{[def. of } \epsilon\text{]} \\ &= E(\epsilon \cdot [(\alpha_0 + \alpha_1 X) - (a_0 + a_1 X)]) && \text{[defs. of } f(X), h(X)\text{]} \\ &= E(\epsilon \cdot [(\alpha_0 - a_0) + (\alpha_1 - a_1)X]) \\ &= (\alpha_0 - a_0) \cdot E(\epsilon) + (\alpha_1 - a_1)E(\epsilon \cdot X) && \text{[Box 6.1 (vii)]} \\ &= 0. && \text{[(i), Box 7.1 (i)]} \end{aligned}$$

Using this result and considering

$$\begin{aligned} E([Y - h(X)]^2) &= E\{([Y - f(X)] + [f(X) - h(X)])^2\} \\ &= E([Y - f(X)]^2) + E([f(X) - h(X)]^2) + 2 \cdot E([Y - f(X)][f(X) - h(X)]) \\ &= E([Y - f(X)]^2) + E([f(X) - h(X)]^2) \\ &\geq E([Y - f(X)]^2). \end{aligned}$$

In this inequality '=' holds if and only if  $f(X) \stackrel{p}{=} h(X)$ . Because  $f(X) = \alpha_0 + \alpha_1 X$  and  $h(X) = a_0 + a_1 X$ , the property  $f(X) \stackrel{p}{=} h(X)$  is equivalent to  $f(X) = h(X)$ .

### **Proof of Theorem 7.30**

The proof is organized as follows: (iii)  $\Rightarrow$  (ii)  $\Rightarrow$  (i)  $\Rightarrow$  (iii), which will prove that (i), (ii), and (iii) are equivalent.

(iii)  $\Rightarrow$  (ii). The first partial derivative of

$$\begin{aligned}
& MSE(b_0, b_1, \dots, b_n) \\
&= E\{[Y - (b_0 + \mathbf{b}'\mathbf{x})]^2\} \quad [(7.38)] \\
&= E(Y^2) + E((b_0 + \mathbf{b}'\mathbf{x})^2) - 2E(Y \cdot (b_0 + \mathbf{b}'\mathbf{x})) \quad [\text{Box 6.1 (vi), (iii)}] \\
&= E(Y^2) + b_0^2 + E((\mathbf{b}'\mathbf{x})^2) + 2b_0 \cdot E(\mathbf{b}'\mathbf{x}) - 2b_0 E(Y) - 2E(\mathbf{b}'\mathbf{x} \cdot Y) \quad [\text{Box 6.1 (vi), (i), (iii)}] \\
&= E(Y^2) + b_0^2 + E((\mathbf{b}'\mathbf{x})^2) + 2b_0 \cdot \mathbf{b}'E(\mathbf{x}) - 2b_0 E(Y) - 2\mathbf{b}'E(\mathbf{x} \cdot Y) \quad [\text{Box 7.2 (iii)}]
\end{aligned}$$

with respect to  $b_0$  is

$$\frac{\partial MSE(b_0, b_1, \dots, b_n)}{\partial b_0} = 2b_0 + 2\mathbf{b}'E(\mathbf{x}) - 2E(Y).$$

If  $f(X_1, \dots, X_n) = \beta_0 + \boldsymbol{\beta}'\mathbf{x} = Q_{lin}(Y|X_1, \dots, X_n)$ , then

$$2\beta_0 + 2\boldsymbol{\beta}'E(\mathbf{x}) - 2E(Y) = 0.$$

Dividing both sides by 2, and solving for  $\beta_0$  yields

$$\beta_0 = E(Y) - \boldsymbol{\beta}'E(\mathbf{x}).$$

Gathering the first partial derivatives of  $MSE(b_0, b_1, \dots, b_n)$  with respect to  $b_1, \dots, b_n$  in a vector yields

$$\left[ \frac{\partial MSE(b_0, b_1, \dots, b_n)}{\partial b_1}, \dots, \frac{\partial MSE(b_0, b_1, \dots, b_n)}{\partial b_n} \right]' = 2E(\mathbf{x}\mathbf{x}')\boldsymbol{\beta} + 2b_0E(\mathbf{x}) - 2E(\mathbf{x} \cdot Y).$$

If  $f(X_1, \dots, X_n) = \beta_0 + \boldsymbol{\beta}'\mathbf{x} = Q_{lin}(Y|X_1, \dots, X_n)$ , then  $2E(\mathbf{x}\mathbf{x}')\boldsymbol{\beta} + 2\beta_0E(\mathbf{x}) - 2E(\mathbf{x} \cdot Y) = \mathbf{0}$ , and dividing both sides by 2 yields

$$E(\mathbf{x}\mathbf{x}')\boldsymbol{\beta} + \beta_0E(\mathbf{x}) - E(\mathbf{x} \cdot Y) = \mathbf{0}.$$

Inserting our result  $\beta_0 = E(Y) - \boldsymbol{\beta}'E(\mathbf{x})$ , using  $\boldsymbol{\Sigma}_{xy} = E(\mathbf{x} \cdot Y) - E(\mathbf{x}) \cdot E(Y)$  and  $\boldsymbol{\Sigma}_{xx} = E(\mathbf{x}\mathbf{x}') - E(\mathbf{x})E(\mathbf{x}')$  [see Box 7.3 (i)] yields

$$\begin{aligned}
& E(\mathbf{x}\mathbf{x}')\boldsymbol{\beta} + (E(Y) - \boldsymbol{\beta}'E(\mathbf{x}))E(\mathbf{x}) - E(\mathbf{x} \cdot Y) \\
&= E(\mathbf{x}\mathbf{x}')\boldsymbol{\beta} + E(Y) \cdot E(\mathbf{x}) - E(\mathbf{x})E(\mathbf{x}')\boldsymbol{\beta} - E(\mathbf{x} \cdot Y) \\
&= (E(\mathbf{x}\mathbf{x}') - E(\mathbf{x})E(\mathbf{x}'))\boldsymbol{\beta} - (E(\mathbf{x} \cdot Y) - E(\mathbf{x}) \cdot E(Y)) \\
&= \boldsymbol{\Sigma}_{xx}\boldsymbol{\beta} - \boldsymbol{\Sigma}_{xy} = \mathbf{0},
\end{aligned}$$

which is a necessary condition for a minimum. (Note, in this proof we do not have to check a sufficient condition.) Adding  $\boldsymbol{\Sigma}_{xy}$  on both sides yields  $\boldsymbol{\Sigma}_{xx}\boldsymbol{\beta} = \boldsymbol{\Sigma}_{xy}$ , and multiplying both sides from the left by  $\boldsymbol{\Sigma}_{xx}^{-1}$  we receive

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}_{xx}^{-1} \boldsymbol{\Sigma}_{xy}.$$

(ii)  $\Rightarrow$  (i).

$$\begin{aligned}
E(\epsilon) &= E(Y - f(X)) \quad [\text{def. of } \epsilon] \\
&= E(Y - (\beta_0 + \boldsymbol{\beta}'\mathbf{x})) \quad [\text{def. of } f(X)] \\
&= E(Y - (E(Y) - \boldsymbol{\beta}'E(\mathbf{x}) + \boldsymbol{\beta}'\mathbf{x})) \quad [(ii)] \\
&= E(Y) - E(Y) + \boldsymbol{\beta}'E(\mathbf{x}) - \boldsymbol{\beta}'E(\mathbf{x}) \quad [\text{Box 6.1 (vi), (i), Box 7.2 (iii)}] \\
&= 0.
\end{aligned}$$

$$\begin{aligned}
\Sigma_{x\epsilon} &= \Sigma_{x, Y - (\beta_0 + \beta'x)} && \text{[def. of } \epsilon \text{]} \\
&= \Sigma_{xy} - \Sigma_{xx}\beta && \text{[Box 7.3 (ii), (vii)]} \\
&= \Sigma_{xy} - \Sigma_{xx}\Sigma_{xx}^{-1}\Sigma_{xy} && \text{[(ii)]} \\
&= \mathbf{0}.
\end{aligned}$$

(i)  $\Rightarrow$  (iii). Let  $f(X) := \beta_0 + \beta'x$ ,  $\beta_0 \in \mathbb{R}$ ,  $\beta \in \mathbb{R}^n$  such that  $E(\epsilon) = 0$  and  $\Sigma_{x\epsilon} = \mathbf{0}$ , where  $\epsilon := Y - f(X)$ . Then, for any linear function  $h(X) := b_0 + \mathbf{b}'x$ ,  $b_0 \in \mathbb{R}$ ,  $\mathbf{b} \in \mathbb{R}^n$ ,

$$\begin{aligned}
&E((Y - f(X))[f(X) - h(X)]) \\
&= E(\epsilon \cdot [f(X) - h(X)]) && \text{[def. of } \epsilon \text{]} \\
&= E(\epsilon \cdot [(\beta_0 + \beta'x) - (b_0 + \mathbf{b}'x)]) && \text{[defs. of } f(X), h(X) \text{]} \\
&= E(\epsilon \cdot [(\beta_0 - b_0) + (\beta - \mathbf{b})'x]) \\
&= (\beta_0 - b_0) \cdot E(\epsilon) + (\beta - \mathbf{b})'E(\epsilon \cdot x) && \text{[Box 6.1 (vi), Box 7.2 (iii)]} \\
&= (\beta - \mathbf{b})' \Sigma_{x\epsilon} && \text{[(i), Box 7.3 (i)]} \\
&= \mathbf{0}. && \text{[(i)]}
\end{aligned}$$

Using this result and considering

$$\begin{aligned}
E([Y - h(X)]^2) &= E([(Y - f(X)) + (f(X) - h(X))]^2) \\
&= E([Y - f(X)]^2) + E([f(X) - h(X)]^2) + 2 \cdot E([Y - f(X)][f(X) - h(X)]) \\
&= E([Y - f(X)]^2) + E([f(X) - h(X)]^2) \\
&\geq E([Y - f(X)]^2).
\end{aligned}$$

Hence,  $f(X) = Q_{lin}(Y|X)$ .

### **Proof of Corollary 7.32**

$$\begin{aligned}
Cov [Q_{lin}(Y|X), \epsilon] &= Cov(\beta_0 + \beta'x, \epsilon) && \text{[(7.39)]} \\
&= \beta' Cov(x, \epsilon) && \text{[Box 7.3 (ii), (iii)]} \\
&= \mathbf{0}. && \text{[Cov}(x, \epsilon) = \Sigma_{x\epsilon} = \mathbf{0}, \text{ Th. 7.30]}
\end{aligned}$$

Furthermore,

$$\begin{aligned}
Var [Q_{lin}(Y|X)] &= Cov [Q_{lin}(Y|X), Q_{lin}(Y|X)] && \text{[Box 7.1 (iv)]} \\
&= Cov [Q_{lin}(Y|X), \epsilon] + Cov [Q_{lin}(Y|X), Q_{lin}(Y|X)] && \text{[(7.43)]} \\
&= Cov [Q_{lin}(Y|X) + \epsilon, Q_{lin}(Y|X)] && \text{[Box 7.1 (ix)]} \\
&= Cov [Y, Q_{lin}(Y|X)]. && \text{[(7.41)]}
\end{aligned}$$

## **7.7 Exercises**

$\triangleright$  **Exercise 7-1** Use Definition 7.2 in order to determine the coefficients  $\alpha_0$  and  $\alpha_1$  of the linear quasi-regression in Example 7.5.

▷ **Exercise 7-2** Consider the linear quasi-regression  $f$  with  $f(x) = \alpha_0 + \alpha_1 x$ ,  $x \in \mathbb{R}$ . Prove: If  $x_1, x_2 \in \mathbb{R}$  with  $x_1 \neq x_2$ , then  $\alpha_0 = f(0)$  and  $\alpha_1 = \frac{1}{x_2 - x_1} [f(x_2) - f(x_1)]$ .

▷ **Exercise 7-3** Prove the propositions of Box 7.1.

▷ **Exercise 7-4** Prove the proposition of Remark 7.13.

▷ **Exercise 7-5** Prove the proposition of Remark 7.20.

▷ **Exercise 7-6** Prove the proposition of Remark 7.21.

▷ **Exercise 7-7** Show that  $\text{Corr}(a_0 + a_1 X, b_0 + b_1 Y) = \text{Corr}(X, Y)$ , where  $a_0, a_1, b_0, b_1 \in \mathbb{R}$ .

▷ **Exercise 7-8** Show

$$\text{Cov}(X, Y) = \begin{cases} SD(X) \cdot SD(Y), & \text{if } a_1 > 0 \\ -SD(X) \cdot SD(Y), & \text{if } a_1 < 0, \end{cases} \quad (7.46)$$

provided that there are  $a_0, a_1 \in \mathbb{R}$  with  $Y \stackrel{p}{=} a_0 + a_1 X$ ,  $a_1 \neq 0$ .

▷ **Exercise 7-9** Prove Equation (7.24) and that the slope is invariant under translations  $c + X$ ,  $d + Y$ ,  $c, d \in \mathbb{R}$ .

▷ **Exercise 7-10** Prove the rules of computation of Box 7.2.

▷ **Exercise 7-11** Prove the rules of computation of Box 7.3.

▷ **Exercise 7-12**

Under the assumptions of Definition 7.28 and  $\text{Var}(Y) > 0$ , prove Equation 7.45

## Solutions

▷ **Solution 7-1**

$$\text{MSE}(a_0, a_1) = E([Y - (a_0 + a_1 X)]^2) \quad [(7.2)]$$

$$= \frac{1}{4} \cdot (1 - a_0 - a_1)^2 + \frac{1}{2} \cdot (2 - a_0 - 2a_1)^2 + \frac{1}{4} \cdot (1 - a_0 - 3a_1)^2 \quad [(6.3)]$$

$$= \frac{1}{4} \cdot (1 + a_0^2 + a_1^2 - 2a_0 - 2a_1 + 2a_0 a_1)$$

$$+ \frac{1}{2} \cdot (4 + a_0^2 + 4a_1^2 - 4a_0 - 8a_1 + 4a_0 a_1)$$

$$+ \frac{1}{4} \cdot (1 + a_0^2 + 9a_1^2 - 2a_0 - 6a_1 + 6a_0 a_1)$$

$$= \frac{5}{2} + a_0^2 + \frac{9}{2} \cdot a_1^2 - 3a_0 - 6a_1 + 4a_0 a_1.$$

The partial derivatives are

$$\frac{\partial \text{MSE}(a_0, a_1)}{\partial a_0} = 2a_0 - 3 + 4a_1 \quad \text{and} \quad \frac{\partial \text{MSE}(a_0, a_1)}{\partial a_1} = 9a_1 - 6 + 4a_0.$$

Fixing the partial derivatives to 0 and denoting the solutions by  $\alpha_0$  and  $\alpha_1$  yields

$$2\alpha_0 - 3 + 4\alpha_1 = 0 \quad \text{and} \quad 9\alpha_1 - 6 + 4\alpha_0 = 0.$$

The first equation implies  $\alpha_0 = -2\alpha_1 + \frac{3}{2}$ . Inserting this result into the second equation yields  $9\alpha_1 - 6 - 8\alpha_1 + 6 = 0$ , which implies  $\alpha_1 = 0$  and  $\alpha_0 = \frac{3}{2}$ . The values  $\alpha_1 = 0$  and  $\alpha_0 = \frac{3}{2}$  satisfy a necessary condition for a minimum; now we check if they also satisfy a sufficient condition. The second partial derivatives are

$$\frac{\partial^2 MSE(a_0, a_1)}{\partial a_0^2} = 2 > 0, \quad \frac{\partial^2 MSE(a_0, a_1)}{\partial a_1^2} = 9 > 0, \quad \text{and} \quad \frac{\partial^2 MSE(a_0, a_1)}{\partial a_0 \partial a_1} = 4,$$

i.e., in this case they are constant for all  $a_0, a_1 \in \mathbb{R}$ . Because  $2 \cdot 9 - 4^2 > 0$ , we can conclude that  $MSE(a_0, a_1)$  has its minimum for  $a_0 = \alpha_0 = \frac{3}{2}$  and  $a_1 = \alpha_1 = 0$  (see Ellis & Gulick, 2006, Th. 13.21, p. 877).

▷ **Solution 7-2** The equation  $f(x) = \alpha_0 + \alpha_1 x$ ,  $x \in \mathbb{R}$ , yields  $f(0) = \alpha_0$ ,

$$f(x_1) = \alpha_0 + \alpha_1 x_1, \quad \text{and} \quad f(x_2) = \alpha_0 + \alpha_1 x_2.$$

Hence,

$$f(x_2) - f(x_1) = \alpha_0 + \alpha_1 x_2 - (\alpha_0 + \alpha_1 x_1) = \alpha_1 (x_2 - x_1).$$

Multiplying both sides by  $x_2 - x_1$  yields  $\alpha_1 = \frac{1}{x_2 - x_1} [f(x_2) - f(x_1)]$ , provided that  $x_1 \neq x_2$ .

▷ **Solution 7-3** (i)

$$\begin{aligned} \text{Cov}(X, Y) &= E[(X - E(X)) \cdot (Y - E(Y))] && [(7.8)] \\ &= E[X \cdot Y - X \cdot E(Y) - E(X) \cdot Y + E(X) \cdot E(Y)] \\ &= E(X \cdot Y) - E[X \cdot E(Y)] - E[E(X) \cdot Y] + E(X) \cdot E(Y) && [\text{Box 6.1 (vi), (i)}] \\ &= E(X \cdot Y) - 2E(X) \cdot E(Y) + E(X) \cdot E(Y) && [\text{Box 6.1 (iii)}] \\ &= E(X \cdot Y) - E(X) \cdot E(Y). \end{aligned}$$

(ii)

$$\begin{aligned} \text{Cov}(\alpha X, \beta Y) &= E[(\alpha X - E(\alpha X)) \cdot (\beta Y - E(\beta Y))] && [(7.8)] \\ &= E[(\alpha X - \alpha E(X)) \cdot (\beta Y - \beta E(Y))] && [\text{Box 6.1 (ii)}] \\ &= E[(X - E(X)) \cdot (Y - E(Y))] \\ &= \text{Cov}(X, Y). && [(7.8)] \end{aligned}$$

(iii)

$$\begin{aligned} \text{Cov}(\alpha X, \beta Y) &= E(\alpha X \cdot \beta Y) - E(\alpha X) \cdot E(\beta Y) && [\text{Box 7.1 (i)}] \\ &= \alpha \beta E(X \cdot Y) - \alpha \beta E(X) \cdot E(Y) && [\text{Box 6.1 (iii)}] \\ &= \alpha \beta [E(X \cdot Y) - E(X) \cdot E(Y)] \\ &= \alpha \beta \text{Cov}(X, Y). && [\text{Box 7.1 (i)}] \end{aligned}$$

(iv) This rule immediately follows from Equations (6.27) and (7.8).

(v) This rule immediately follows from Equation (7.8).

(vi) Independence of  $X$  and  $Y$  implies  $E(X \cdot Y) = E(X) \cdot E(Y)$  (see Th. 6.24) and  $\text{Cov}(X, Y) = 0$  [see Rule (i)].

(vii) According to Lemma 5.51,  $X$  and  $Y$  are independent if  $X \stackrel{p}{=} \alpha$ . This implies that  $\text{Cov}(X, Y) = 0$  [see Rule (vi) of Box 7.1].

(viii)

$$\begin{aligned}
& \text{Var}\left(\sum_{i=1}^n \alpha_i Y_i\right) \\
&= E\left[\left(\sum_{i=1}^n \alpha_i Y_i - E\left(\sum_{i=1}^n \alpha_i Y_i\right)\right)^2\right] \quad [(6.27)] \\
&= E\left[\left(\sum_{i=1}^n \alpha_i Y_i - \sum_{i=1}^n \alpha_i E(Y_i)\right)^2\right] \quad [\text{Box 6.1 (vii)}] \\
&= E\left[\left(\sum_{i=1}^n \alpha_i [Y_i - E(Y_i)]\right)^2\right] \\
&= E\left[\sum_{i=1}^n \alpha_i^2 (Y_i - E(Y_i))^2 + \sum_{i=1}^n \sum_{j=1, j \neq i}^n \alpha_i \alpha_j (Y_i - E(Y_i))(Y_j - E(Y_j))\right] \\
&= \sum_{i=1}^n \alpha_i^2 E(Y_i - E(Y_i))^2 + \sum_{i=1}^n \sum_{j=1, j \neq i}^n \alpha_i \alpha_j E[(Y_i - E(Y_i))(Y_j - E(Y_j))] \quad [\text{Box 6.1 (vii)}] \\
&= \sum_{i=1}^n \alpha_i^2 \text{Var}(Y_i) + \sum_{i=1}^n \sum_{j=1, j \neq i}^n \alpha_i \alpha_j \text{Cov}(Y_i, Y_j). \quad [(6.27), (7.8)]
\end{aligned}$$

(ix)

$$\begin{aligned}
& \text{Cov}\left(\sum_{i=1}^n \alpha_i X_i, \sum_{j=1}^m \beta_j Y_j\right) \\
&= E\left[\left(\sum_{i=1}^n \alpha_i X_i - E\left(\sum_{i=1}^n \alpha_i X_i\right)\right)\left(\sum_{j=1}^m \beta_j Y_j - E\left(\sum_{j=1}^m \beta_j Y_j\right)\right)\right] \quad [(7.8)] \\
&= E\left[\left(\sum_{i=1}^n \alpha_i [X_i - E(X_i)]\right)\left(\sum_{j=1}^m \beta_j [Y_j - E(Y_j)]\right)\right] \quad [\text{Box 6.1 (vii)}] \\
&= E\left[\sum_{i=1}^n \sum_{j=1}^m \alpha_i \beta_j [X_i - E(X_i)] [Y_j - E(Y_j)]\right] \\
&= \sum_{i=1}^n \sum_{j=1}^m \alpha_i \beta_j \text{Cov}(X_i, Y_j). \quad [\text{Box 6.1 (vii), (7.8)}]
\end{aligned}$$

(x) If  $X_1 \stackrel{p}{=} X_2$ , then  $[Y - E(Y)] \cdot [X_1 - E(X_1)] \stackrel{p}{=} [Y - E(Y)] \cdot [X_2 - E(X_2)]$ . According to Corollary 5.24, these two product variables have the same distribution and according to Corollary 6.17 the same expectation. However, the expectations of these product variables are the covariances. Hence,  $\text{Cov}(Y, X_1) = \text{Cov}(Y, X_2)$ .

▷ **Solution 7-4** If  $X, Y: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  are random variables, then

$$(\alpha X + \beta Y)^2 = \alpha^2 X^2 + \beta^2 Y^2 + 2\alpha\beta X \cdot Y$$

is also a random variable on  $(\Omega, \mathcal{A}, P)$  (see Def. 5.1, Example 2.61, and Th. 2.57), and this implies

$$\begin{aligned}
E[(\alpha X + \beta Y)^2] &= E(\alpha^2 X^2 + \beta^2 Y^2 + 2\alpha\beta X \cdot Y) \\
&= \alpha^2 E(X^2) + \beta^2 E(Y^2) + 2\alpha\beta E(X \cdot Y). \quad [\text{Box 6.1 (vii)}]
\end{aligned}$$

The terms  $E(X^2)$  and  $E(Y^2)$  are finite by assumption and, according to Remark 7.1,  $E(X \cdot Y)$  is finite as well. This implies  $E[(\alpha X + \beta Y)^2] < \infty$ .

▷ **Solution 7-5** Inequality (7.16) implies

$$-\sqrt{\text{Var}(X) \cdot \text{Var}(Y)} \leq \text{Cov}(X, Y) \leq \sqrt{\text{Var}(X) \cdot \text{Var}(Y)}.$$

If  $\sqrt{\text{Var}(X) \cdot \text{Var}(Y)} = SD(X) \cdot SD(Y) > 0$ , then these inequalities yield  $-1 \leq \text{Corr}(X, Y) \leq 1$ .

If  $\sqrt{\text{Var}(X) \cdot \text{Var}(Y)} = SD(X) \cdot SD(Y) = 0$ , then, by Definition 7.18,  $\text{Corr}(X, Y) = 0$ .

▷ **Solution 7-6** If  $SD(X), SD(Y) > 0$ , then

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{SD(X) \cdot SD(Y)} \quad [(7.18)]$$

$$= \frac{E[(X - E(X)) \cdot (Y - E(Y))]}{SD(X) \cdot SD(Y)} \quad [(7.8)]$$

$$= E\left(\frac{[X - E(X)] \cdot [Y - E(Y)]}{SD(X) \cdot SD(Y)}\right) \quad [\text{Box 6.1 (iii)}]$$

$$= E\left(\frac{X - E(X)}{SD(X)} \cdot \frac{Y - E(Y)}{SD(Y)}\right).$$

▷ **Solution 7-7** According to Rules (ii) and (iii) of Box 7.1,  $\text{Cov}(a_0 + a_1 X, b_0 + b_1 Y) = a_1 b_1 \text{Cov}(X, Y)$ . Similarly, according to Rules (ii) and (iii) of Box 6.2,  $\text{Var}(a_0 + a_1 X) = a_1^2 \text{Var}(X)$ , which implies

$$SD(a_0 + a_1 X) = \begin{cases} a_1 SD(X), & \text{if } a_1 > 0 \\ -a_1 SD(X), & \text{if } a_1 < 0 \\ 0, & \text{if } a_1 = 0. \end{cases}$$

Hence, for  $a_1, b_1 > 0$ ,

$$\begin{aligned} \text{Corr}(a_0 + a_1 X, b_0 + b_1 Y) &= \frac{\text{Cov}(a_0 + a_1 X, b_0 + b_1 Y)}{SD(a_0 + a_1 X) \cdot SD(b_0 + b_1 Y)} \\ &= \frac{a_1 b_1 \text{Cov}(X, Y)}{a_1 SD(X) \cdot b_1 SD(Y)} \\ &= \frac{\text{Cov}(X, Y)}{SD(X) \cdot SD(Y)} \\ &= \text{Corr}(X, Y). \end{aligned}$$

The proofs for the cases of  $a_1, b_1 \neq 0$  are analogous. Note that we defined  $\text{Corr}(X, Y) = 0$  if  $SD(X) = 0$  or  $SD(Y) = 0$ .

▷ **Solution 7-8** According to Box 7.1 (x),  $\text{Cov}(X, Y) = \text{Cov}(X, \alpha_0 + \alpha_1 X)$  if  $Y \stackrel{p}{=} \alpha_0 + \alpha_1 X$ . Box 6.2 (v), (ii), (iii) imply  $\text{Var}(Y) = \alpha_1^2 \text{Var}(X)$ . Therefore,

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, \alpha_0 + \alpha_1 X)}{SD(X) \cdot \sqrt{\alpha_1^2 \text{Var}(X)}} = \frac{\alpha_1 \text{Var}(X)}{SD(X) \cdot |\alpha_1| \cdot SD(X)} = \frac{\alpha_1}{|\alpha_1|}.$$

Hence,  $\text{Corr}(X, Y) = 1$  if  $\alpha_1 > 0$  and  $\text{Corr}(X, Y) = -1$  if  $\alpha_1 < 0$ . Therefore,

$$\text{Corr}(X, Y) := \frac{\text{Cov}(X, Y)}{SD(X) \cdot SD(Y)}$$

yields (7.46).

▷ **Solution 7-9**

$$\begin{aligned}
\alpha_1^* &= \frac{\text{Cov}(aX, bY)}{\text{Var}(aX)} && \text{[Th. 7.14 (ii)]} \\
&= \frac{ab \cdot \text{Cov}(X, Y)}{a^2 \cdot \text{Var}(X)} && \text{[Box 7.1 (iii), Box 6.2 (iii)]} \\
&= \frac{b}{a} \cdot \alpha_1. && \text{[Th. 7.14 (ii)]}
\end{aligned}$$

According to Theorem 7.14 (ii), the slope of the linear quasi-regression of  $c+Y$  on  $d+X$ , is

$$\begin{aligned}
\frac{\text{Cov}(c+X, d+Y)}{\text{Var}(c+X)} &= \frac{\text{Cov}(X, Y)}{\text{Var}(X)} && \text{[Box 7.1 (ii), Box 6.2 (ii)]} \\
&= \alpha_1. && \text{[Th. 7.14 (ii)]}
\end{aligned}$$

▷ **Solution 7-10** (i) Equation (7.27) and Rule (i) of Box 6.1 imply

$$E(\mathbf{x}) = [E(X_1), \dots, E(X_n)]' = [a_1, \dots, a_n]' = \mathbf{a}.$$

(ii) Equation (7.27) and Rule (ii) of Box 6.1 imply

$$E(\mathbf{a} + \mathbf{x}) = \begin{bmatrix} E(a_1 + X_1) \\ \vdots \\ E(a_n + X_n) \end{bmatrix} = \begin{bmatrix} a_1 + E(X_1) \\ \vdots \\ a_n + E(X_n) \end{bmatrix} = \mathbf{a} + E(\mathbf{x}).$$

(iii) Equation (7.31) and Rule (vii) of Box 6.1 imply

$$E(\mathbf{a}'\mathbf{x}) = E\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i \cdot E(X_i) = \mathbf{a}'E(\mathbf{x}).$$

The other equations summarized in (iii) follow from the fact that  $\mathbf{a}'\mathbf{x}$  is a one-dimensional random variable (see Example 2.61) and  $\mathbf{a}'\mathbf{x} = \mathbf{x}'\mathbf{a}$ .

(v) Let  $\mathbf{a}'_l$  and  $\mathbf{b}'_l$ ,  $l = 1, \dots, k$ , denote the row vectors of  $\mathbf{A}$  and  $\mathbf{B}$ , respectively. Applying Equation (7.27), Rule (vi) of Box 6.1, and Rule (iii) to the terms  $\mathbf{a}'_l\mathbf{x}$  and  $\mathbf{b}'_l\mathbf{y}$ ,  $l = 1, 2, \dots, k$ , respectively, yields

$$\begin{aligned}
E(\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y}) &= E \begin{bmatrix} \mathbf{a}'_1\mathbf{x} + \mathbf{b}'_1\mathbf{y} \\ \vdots \\ \mathbf{a}'_k\mathbf{x} + \mathbf{b}'_k\mathbf{y} \end{bmatrix} = \begin{bmatrix} E(\mathbf{a}'_1\mathbf{x} + \mathbf{b}'_1\mathbf{y}) \\ \vdots \\ E(\mathbf{a}'_k\mathbf{x} + \mathbf{b}'_k\mathbf{y}) \end{bmatrix} = \begin{bmatrix} E(\mathbf{a}'_1\mathbf{x}) + E(\mathbf{b}'_1\mathbf{y}) \\ \vdots \\ E(\mathbf{a}'_k\mathbf{x}) + E(\mathbf{b}'_k\mathbf{y}) \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{a}'_1 E(\mathbf{x}) + \mathbf{b}'_1 E(\mathbf{y}) \\ \vdots \\ \mathbf{a}'_k E(\mathbf{x}) + \mathbf{b}'_k E(\mathbf{y}) \end{bmatrix} = \mathbf{A}E(\mathbf{x}) + \mathbf{B}E(\mathbf{y}).
\end{aligned}$$

(iv) This rule is a special case of Rule (v) with  $\mathbf{B} = \mathbf{0}$ .

(vi) Let  $\mathbf{a}_l$  and  $\mathbf{x}_l$ ,  $l = 1, \dots, k$ , denote the column vectors of  $\mathbf{A}'$  and  $\mathbf{X}$ , respectively. Then Equations (7.27), (7.29), and Rule (i) imply

$$E(\mathbf{X}) = [E(\mathbf{x}_1), \dots, E(\mathbf{x}_k)] = [E(\mathbf{a}_1), \dots, E(\mathbf{a}_k)] = [\mathbf{a}_1, \dots, \mathbf{a}_k] = \mathbf{A}'.$$

(vii) Let  $\mathbf{a}_l$  and  $\mathbf{x}_l$ ,  $l = 1, \dots, k$ , denote the column vectors of  $\mathbf{A}'$  and  $\mathbf{X}$ , respectively. Then Equations (7.27), (7.29) and Rule (ii) imply

$$\begin{aligned} E(\mathbf{A}' + \mathbf{X}) &= E(\mathbf{a}_1 + \mathbf{x}_1, \dots, \mathbf{a}_k + \mathbf{x}_k) = [E(\mathbf{a}_1 + \mathbf{x}_1), \dots, E(\mathbf{a}_k + \mathbf{x}_k)] \\ &= [\mathbf{a}_1 + E(\mathbf{x}_1), \dots, \mathbf{a}_k + E(\mathbf{x}_k)] = \mathbf{A}' + E(\mathbf{X}). \end{aligned}$$

(viii) Let  $\mathbf{c}'_i$ ,  $i = 1, \dots, l$ , denote the row vectors of  $\mathbf{C}$ . Then Equations (7.27) and (7.29) imply

$$E(\mathbf{C}\mathbf{X}) = E \begin{bmatrix} \mathbf{c}'_1 \mathbf{X} \\ \vdots \\ \mathbf{c}'_l \mathbf{X} \end{bmatrix} = \begin{bmatrix} E(\mathbf{c}'_1 \mathbf{X}) \\ \vdots \\ E(\mathbf{c}'_l \mathbf{X}) \end{bmatrix} = \begin{bmatrix} \mathbf{c}'_1 E(\mathbf{X}) \\ \vdots \\ \mathbf{c}'_l E(\mathbf{X}) \end{bmatrix} = \mathbf{C}E(\mathbf{X}).$$

(ix). Rule (viii), Equation (7.30), and the rules for the transpose of a matrix yield

$$\begin{aligned} E(\mathbf{C}\mathbf{X}\mathbf{Y}'\mathbf{D}') &= \mathbf{C}E(\mathbf{X}\mathbf{Y}'\mathbf{D}') && \text{[(viii)]} \\ &= \mathbf{C}(E(\mathbf{D}\mathbf{Y}\mathbf{X}'))' && \text{[(7.30)]} \\ &= \mathbf{C}(\mathbf{D}E(\mathbf{Y}\mathbf{X}'))' && \text{[(viii)]} \\ &= \mathbf{C}E(\mathbf{Y}\mathbf{X}')'\mathbf{D}' && \text{[(7.30)]} \\ &= \mathbf{C}E(\mathbf{X}\mathbf{Y}')\mathbf{D}'. && \text{[(7.30)]} \end{aligned}$$

▷ **Solution 7-11** (i)

$$\begin{aligned} \Sigma_{\mathbf{x}\mathbf{y}} &= E([\mathbf{x} - E(\mathbf{x})][\mathbf{y} - E(\mathbf{y})]') && \text{[(7.33)]} \\ &= E(\mathbf{x}\mathbf{y}' - \mathbf{x}E(\mathbf{y})' - E(\mathbf{x})\mathbf{y}' + E(\mathbf{x})E(\mathbf{y})') \\ &= E(\mathbf{x}\mathbf{y}') - E(\mathbf{x})E(\mathbf{y})' - E(\mathbf{x})E(\mathbf{y}') + E(\mathbf{x})E(\mathbf{y})' && \text{[Box 7.2 (iii)]} \\ &= E(\mathbf{x}\mathbf{y}') - E(\mathbf{x})E(\mathbf{y}') \end{aligned}$$

(ii)

$$\begin{aligned} \Sigma_{\mathbf{a}+\mathbf{x}, \mathbf{b}+\mathbf{y}} &= E([\mathbf{a} + \mathbf{x} - E(\mathbf{a} + \mathbf{x})][\mathbf{b} + \mathbf{y} - E(\mathbf{b} + \mathbf{y})]') && \text{[(7.33)]} \\ &= E([\mathbf{a} + \mathbf{x} - \mathbf{a} - E(\mathbf{x})][\mathbf{b} + \mathbf{y} - \mathbf{b} - E(\mathbf{y})]') && \text{[Box 7.2 (ii)]} \\ &= E([\mathbf{x} - E(\mathbf{x})][\mathbf{y} - E(\mathbf{y})]') \\ &= \Sigma_{\mathbf{x}\mathbf{y}}. && \text{[(7.33)]} \end{aligned}$$

(iii)

$$\begin{aligned} \Sigma_{\mathbf{A}\mathbf{x}, \mathbf{B}\mathbf{y}} &= E([\mathbf{A}\mathbf{x} - E(\mathbf{A}\mathbf{x})][\mathbf{B}\mathbf{y} - E(\mathbf{B}\mathbf{y})]') && \text{[(7.33)]} \\ &= E([\mathbf{A}\mathbf{x} - \mathbf{A}E(\mathbf{x})][\mathbf{B}\mathbf{y} - \mathbf{B}E(\mathbf{y})]') && \text{[Box 7.2 (iv)]} \\ &= E[\mathbf{A}[\mathbf{x} - E(\mathbf{x})][\mathbf{B}[\mathbf{y} - E(\mathbf{y})]']'] \\ &= E(\mathbf{A}[\mathbf{x} - E(\mathbf{x})][\mathbf{y} - E(\mathbf{y})]'\mathbf{B}') \\ &= \mathbf{A}E([\mathbf{x} - E(\mathbf{x})][\mathbf{y} - E(\mathbf{y})]')\mathbf{B}' && \text{[Box 7.2 (ix)]} \\ &= \mathbf{A}\Sigma_{\mathbf{x}\mathbf{y}}\mathbf{B}'. && \text{[(7.33)]} \end{aligned}$$

(iv)

$$\begin{aligned} \Sigma_{\mathbf{x}\mathbf{y}} &= E([\mathbf{x} - E(\mathbf{x})][\mathbf{y} - E(\mathbf{y})]') && \text{[(7.33)]} \\ &= \left( E([\mathbf{x} - E(\mathbf{x})][\mathbf{y} - E(\mathbf{y})]')' \right)' && \text{[(\mathbf{A}')' = \mathbf{A}]} \\ &= E([\mathbf{x} - E(\mathbf{x})][\mathbf{y} - E(\mathbf{y})]')' && \text{[(7.30)]} \\ &= E([\mathbf{y} - E(\mathbf{y})][\mathbf{x} - E(\mathbf{x})]')' && \text{[(\mathbf{a}\mathbf{b}')' = \mathbf{b}\mathbf{a}']} \end{aligned}$$

$$= \boldsymbol{\Sigma}'_{y\mathbf{x}}. \quad [(7.33)]$$

(v) Independence of the multivariate random variables  $X$  and  $Y$  implies  $X_i \perp\!\!\!\perp Y_j$ , for all  $i = 1, \dots, n$  and  $j = 1, \dots, m$ . Therefore, Rule (vi) of Box 7.1 implies  $\text{Cov}(X_i, Y_j) = 0$ , for all  $i = 1, \dots, n$  and  $j = 1, \dots, m$ . Equation (7.34) then implies  $\boldsymbol{\Sigma}_{\mathbf{x}\mathbf{y}} = \mathbf{0}$ .

(vi) If  $\mathbf{x} \stackrel{p}{=} \mathbf{a}$ , then Rule (vii) of Box (7.1) yields  $\text{Cov}(X_i, Y_j) = 0$ , for all  $i = 1, \dots, n$ ,  $j = 1, \dots, m$ . Equation (7.34) then implies  $\boldsymbol{\Sigma}_{\mathbf{x}\mathbf{y}} = \mathbf{0}$ .

(vii)

$$\begin{aligned} & \boldsymbol{\Sigma}_{\mathbf{A}\mathbf{x}+\mathbf{B}\mathbf{y}, \mathbf{C}\mathbf{w}+\mathbf{D}\mathbf{z}} \\ &= E\{[\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} - E(\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y})] [\mathbf{C}\mathbf{w} + \mathbf{D}\mathbf{z} - E(\mathbf{C}\mathbf{w} + \mathbf{D}\mathbf{z})]'\} \quad [(7.33)] \\ &= E\{[\mathbf{A}\mathbf{x} - E(\mathbf{A}\mathbf{x}) + \mathbf{B}\mathbf{y} - E(\mathbf{B}\mathbf{y})] [\mathbf{C}\mathbf{w} - E(\mathbf{C}\mathbf{w}) + \mathbf{D}\mathbf{z} - E(\mathbf{D}\mathbf{z})]'\} \\ &= E\{[\mathbf{A}\mathbf{x} - \mathbf{A}E(\mathbf{x})] [\mathbf{C}\mathbf{w} - \mathbf{C}E(\mathbf{w})]'\} + E\{[\mathbf{A}\mathbf{x} - \mathbf{A}E(\mathbf{x})] [\mathbf{D}\mathbf{z} - \mathbf{D}E(\mathbf{z})]'\} \\ &\quad + E\{[\mathbf{B}\mathbf{y} - \mathbf{B}E(\mathbf{y})] [\mathbf{C}\mathbf{w} - \mathbf{C}E(\mathbf{w})]'\} + E\{[\mathbf{B}\mathbf{y} - \mathbf{B}E(\mathbf{y})] [\mathbf{D}\mathbf{z} - \mathbf{D}E(\mathbf{z})]'\} \quad [\text{Box 7.2 (iv)}] \\ &= E\{\mathbf{A}[\mathbf{x} - E(\mathbf{x})] [\mathbf{w} - E(\mathbf{w})]'\mathbf{C}'\} + E\{\mathbf{A}[\mathbf{x} - E(\mathbf{x})] [\mathbf{z} - E(\mathbf{z})]'\mathbf{D}'\} \\ &\quad + E\{\mathbf{B}[\mathbf{y} - E(\mathbf{y})] [\mathbf{w} - E(\mathbf{w})]'\mathbf{C}'\} + E\{\mathbf{B}[\mathbf{y} - E(\mathbf{y})] [\mathbf{z} - E(\mathbf{z})]'\mathbf{D}'\} \\ &= \mathbf{A}E\{[\mathbf{x} - E(\mathbf{x})] [\mathbf{w} - E(\mathbf{w})]'\}\mathbf{C}' + \mathbf{A}E\{[\mathbf{x} - E(\mathbf{x})] [\mathbf{z} - E(\mathbf{z})]'\}\mathbf{D}' \\ &\quad + \mathbf{B}E\{[\mathbf{y} - E(\mathbf{y})] [\mathbf{w} - E(\mathbf{w})]'\}\mathbf{C}' + \mathbf{B}E\{[\mathbf{y} - E(\mathbf{y})] [\mathbf{z} - E(\mathbf{z})]'\}\mathbf{D}' \quad [\text{Box 7.2 (ix)}] \\ &= \mathbf{A}\boldsymbol{\Sigma}_{x\mathbf{w}}\mathbf{C}' + \mathbf{A}\boldsymbol{\Sigma}_{x\mathbf{z}}\mathbf{D}' + \mathbf{B}\boldsymbol{\Sigma}_{y\mathbf{w}}\mathbf{C}' + \mathbf{B}\boldsymbol{\Sigma}_{y\mathbf{z}}\mathbf{D}'. \quad [(7.33)] \end{aligned}$$

(viii) If  $\mathbf{x} \stackrel{p}{=} \mathbf{z}$ , then Rule (x) of Box 7.1 implies that  $\text{Cov}(X_i, Y_j) = \text{Cov}(Z_i, Y_j)$ , for all  $i = 1, \dots, n$ ,  $j = 1, \dots, m$ . Equation (7.34) then implies  $\boldsymbol{\Sigma}_{\mathbf{x}\mathbf{y}} = \boldsymbol{\Sigma}_{\mathbf{z}\mathbf{y}}$ .

▷ **Solution 7-12**

$$Q_{Y|X}^2 = \frac{\text{Var}[Q_{\text{lin}}(Y|X)]}{\text{Var}(Y)} \quad [(7.40)]$$

$$= \frac{\text{Cov}[Y, Q_{\text{lin}}(Y|X)]}{\text{Var}(Y)} \quad [(7.44)]$$

$$= \frac{\text{Cov}[Y, Q_{\text{lin}}(Y|X)] \cdot \text{Var}[Q_{\text{lin}}(Y|X)]}{\text{Var}[Q_{\text{lin}}(Y|X)] \cdot \text{Var}(Y)}$$

$$= \frac{\text{Cov}[Y, Q_{\text{lin}}(Y|X)] \cdot \text{Cov}[Y, Q_{\text{lin}}(Y|X)]}{\text{Var}[Q_{\text{lin}}(Y|X)] \cdot \text{Var}(Y)} \quad [(7.44)]$$

$$= \frac{\text{Cov}[Y, Q_{\text{lin}}(Y|X)]^2}{\text{SD}[Q_{\text{lin}}(Y|X)]^2 \cdot \text{SD}(Y)^2}$$

$$= \text{Corr}[Y, Q_{\text{lin}}(Y|X)]^2.$$

Note that  $\text{Corr}[Y, Q_{\text{lin}}(Y|X)] = 0$  if  $\text{Var}[Q_{\text{lin}}(Y|X)] = 0$  [see Eq. (7.18)].



**Part III**  
**Conditional Expectation and Regression**



## Chapter 9

# Conditional Expectation Value and Discrete Conditional Expectation

In chapter 6 we introduced the concepts covariance and correlation, which quantify the strength of the kind of dependence that can be described by a linear quasi-regression. In the next five chapters we introduce the concept of a *conditional expectation* and a ‘genuine’ regression. These concepts can be used to describe how the  $(X=x)$ -conditional expectation values of a numerical random variable  $Y$  depend on the values of a (numerical, non-numerical, multivariate) random variable  $X$ . In this chapter we start with the concepts  $(X=x)$ -*conditional expectation value* and *discrete conditional expectation*, presuming that  $X$  is a discrete random variable. In this case, the conditional expectation  $E(Y|X)$  is easily defined as that random variable whose values are the conditional expectation values  $E(Y|X=x)$ . In chapter 10 we introduce the general concept of a conditional expectation, dropping the assumption that  $X$  is discrete. Chapters 12 and 13 deal with parametrizations of a conditional expectation, and Chapter 11 is devoted to the concepts *residual with respect to a conditional expectation*, *conditional variance*, *conditional covariance*, and partial correlation.

### 9.1 Conditional Expectation Value

Remember, the expectation of a numerical random variable  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  is defined by  $E(Y) = \int Y dP$ , using the probability measure  $P$ . Now we choose an event  $B \in \mathcal{A}$  with  $P(B) > 0$  and, instead of  $P$ , we consider the  $B$ -*conditional probability measure*  $P^B: \mathcal{A} \rightarrow [0, 1]$  defined by

$$P^B(A) := P(A|B), \quad \forall A \in \mathcal{A}, \quad (9.1)$$

(see Def. 4.29). Referring to this measure,

$$E^B(Y) := \int Y dP^B, \quad (9.2)$$

defines the  $P^B$ -*expectation of*  $Y$ , i. e., the expectation of  $Y$  with respect to the measure  $P^B$ . Reading the following definition, also remember that

$$\{X=x\} = X^{-1}(\{x\}) = \{\omega \in \Omega: X(\omega) = x\}$$

denotes the event that a random variable  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  takes on the value  $x$  and that we defined  $P(X=x) := P(\{X=x\})$  (see Rem. 5.4). Assuming  $P(X=x) > 0$  and  $A \in \mathcal{A}$ , we denote

$$P(A|X=x) := P(A|\{X=x\}), \quad (9.3)$$

Furthermore, we denote  $P^{X=x} := P^{\{X=x\}}$ . Hence, according to Equation (9.1),

$$P^{X=x}(A) := P(A|X=x), \quad \forall A \in \mathcal{A}. \quad (9.4)$$

**Remark 9.1 ( $P^{X=x}$ -Equivalence of  $f(X)$  and  $f(x)$ )** Let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  be a random variable,  $x \in \Omega'_X$ ,  $\{x\} \in \mathcal{A}'_X$ , and  $P(X=x) > 0$ . Furthermore, let  $f: (\Omega'_X, \mathcal{A}'_X) \rightarrow (\Omega', \mathcal{A}')$  be a measurable mapping, then  $f(X) \stackrel{P^{X=x}}{=} f(x)$  (see Exercise 9-1).  $\triangleleft$

The probability measures  $P^B$  and  $P^{X=x}$  are now used to define the conditional expectation value.

**Definition 9.2 (Conditional Expectation Value)**

Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  be a random variable.

- (i) If  $B \in \mathcal{A}$  with  $P(B) > 0$  and  $Y$  is quasi-integrable with respect to  $P^B$ , then we define

$$E(Y|B) := E^B(Y) = \int Y dP^B, \quad (9.5)$$

call it the conditional expectation value of  $Y$  given the event  $B$  (or the  $B$ -conditional expectation value of  $Y$ ), and say that it exists.

- (ii) If  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  is a random variable,  $x \in \Omega'_X$  with  $\{x\} \in \mathcal{A}'_X$  and  $P(X=x) > 0$ , and  $Y$  is quasi-integrable with respect to  $P^{X=x}$ , then we define

$$E(Y|X=x) := E(Y|\{X=x\}) \quad (9.6)$$

and call it the conditional expectation value of  $Y$  given  $X=x$  (or the  $(X=x)$ -conditional expectation value of  $Y$ ), and say that it exists.

Note that  $E(Y|B)$  can be infinite. The only restriction is  $B \in \mathcal{A}$  with  $P(B) > 0$  and that  $Y$  is quasi-integrable with respect to  $P^B$ . Otherwise the integral  $\int Y dP^B$  is not defined.

**Remark 9.3 (Multivariate Numerical or Qualitative  $X$ )** Also note that the random variable  $X$  in Definition 9.2 (ii) can be numerical, i. e.,  $\Omega'_X \subset \overline{\mathbb{R}}$ , multivariate (see section 5.3), or nonnumerical. If  $X = (X_1, \dots, X_n)$  is a multivariate random variable, then we also use the notation  $E(Y|X_1=x_1, \dots, X_n=x_n)$  instead of  $E(Y|X=x)$  for the conditional expectation value of  $Y$  given  $(X_1, \dots, X_n) = (x_1, \dots, x_n)$  and call it the  $(X_1=x_1, \dots, X_n=x_n)$ -conditional expectation value of  $Y$  or the conditional expectation value of  $Y$  given  $X_1=x_1, \dots, X_n=x_n$ .  $\triangleleft$

The following theorem addresses the relationship between the  $B$ -conditional expectation value of  $Y$  and the expectation of  $1_B \cdot Y$  with respect to  $P$ .

**Theorem 9.4 (B-Conditional Expectation Value and the P-Expectation)**

Let the assumptions of Definition 9.2 (i) hold. Then:

- (i)  $E(Y)$  exists  $\Rightarrow E(Y|B)$  exists.
- (ii)  $E(Y)$  is finite  $\Rightarrow E(Y|B)$  is finite.
- (iii) Furthermore, if  $E(Y^2)$  is finite, then

$$E(Y|B) = \frac{1}{P(B)} \cdot \int 1_B \cdot Y \, dP = \frac{1}{P(B)} \cdot E(1_B \cdot Y) \quad (9.7)$$

$$= \frac{1}{P(B)} \cdot \text{Cov}(Y, 1_B) + E(Y). \quad (9.8)$$

(Proof p. 291)

Recall the following notation:

$$1_{X=x} := 1_{\{X=x\}}, \quad P(X=x) := P(\{X=x\}), \quad (9.9)$$

and

$$E^{X=x}(Y) := E^{\{X=x\}}(Y). \quad (9.10)$$

Using this notation, Equations (9.5), (9.6), and (9.7) yield the following corollary.

**Corollary 9.5 ((X=x)-Conditional Expectation Value and P-Expectation)**

If the assumptions of Definition 9.2 (ii) hold, then

$$\begin{aligned} E(Y|X=x) &= E^{X=x}(Y) = \int Y \, dP^{X=x} \\ &= \frac{1}{P(X=x)} \cdot \int 1_{X=x} \cdot Y \, dP = \frac{1}{P(X=x)} \cdot E(1_{X=x} \cdot Y). \end{aligned} \quad (9.11)$$

**Remark 9.6 (B-Conditional Probability)** If  $A \in \mathcal{A}$  and  $P(B) > 0$ , then

$$E(1_A|B) = \frac{1}{P(B)} \cdot E(1_B \cdot 1_A) \quad [(9.7)]$$

$$= \frac{1}{P(B)} \cdot P(A \cap B) \quad [\text{Box 6.1 (iv)}] \quad (9.12)$$

$$= P(A|B) = P^B(A). \quad [(4.2), (9.1)]$$

Because  $P^B$  is a probability measure, these equations imply  $0 \leq E(1_A|B) \leq 1$ .  $\triangleleft$

**Remark 9.7 ((X=x)-Conditional Probability)** For  $B = \{X=x\}$ , using the notation (9.3) and (9.4), Equation (9.12) implies

$$\begin{aligned} E(1_A|X=x) &= \frac{1}{P(X=x)} \cdot E(1_{X=x} \cdot 1_A) \\ &= \frac{1}{P(X=x)} \cdot P(A \cap \{X=x\}) \\ &= P(A|X=x) = P^{X=x}(A), \end{aligned} \quad (9.13)$$

provided that  $A \in \mathcal{A}$  and  $P(X=x) > 0$ . Equations (9.13) show that  $E(1_A | X=x)$  is identical to the conditional probability  $P(A | \{X=x\})$  of  $A$  given the event  $\{X=x\}$  (see Def. 4.12). The term  $P(A | X=x)$  is also called the *conditional probability of  $A$  given  $X=x$*  or the  $(X=x)$ -*conditional probability of  $A$* .  $\triangleleft$

**Remark 9.8 (( $X=x$ )-Conditional Probability of  $\{Y=y\}$ )** If  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Y, \mathcal{A}'_Y)$  is a random variable and the assumptions of Definition 9.2 (ii) hold, then we also use the notation

$$P(Y=y | X=x) := P(\{Y=y\} | X=x) = P(1_{Y=y}=1 | X=x) = E(1_{Y=y} | X=x) \quad (9.14)$$

and call it the  $(X=x)$ -*conditional probability of  $\{Y=y\}$*  [see Eqs. (9.9)]. Hence, Equations (9.3), (9.14), and (4.2) yield

$$P(Y=y | X=x) = \frac{P(Y=y, X=x)}{P(X=x)}. \quad (9.15)$$

$\triangleleft$

## 9.2 Transformation Theorem

If  $P_Y^{X=x}: \mathcal{A}'_Y \rightarrow [0, 1]$  denotes the distribution of  $Y$  with respect to the  $(X=x)$ -conditional-probability measure  $P^{X=x}$  and  $E_Y^{X=x}(g)$  the expectation of  $g$  with respect to the distribution  $P_Y^{X=x}$ , then the transformation theorem (cf. Th. 6.13) for the conditional expectation value  $E(Y | X=x)$  can be formulated as follows:

### Theorem 9.9 (Transformation Theorem for $E(Y | X=x)$ )

Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Y, \mathcal{A}'_Y)$ ,  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  be random variables, and  $g: (\Omega'_Y, \mathcal{A}'_Y) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  be a measurable function. Furthermore, let  $x \in \Omega'_X$  with  $\{x\} \in \mathcal{A}'_X$  and  $P(X=x) > 0$ .

(i) If  $g$  is nonnegative or with finite expectation  $E_Y^{X=x}(g)$ , then

$$\begin{aligned} E_Y^{X=x}(g) &= \int g \, dP_Y^{X=x} = \int g(y) P_Y^{X=x}(dy) \\ &= \int g(Y) \, dP^{X=x} = E^{X=x}[g(Y)] = E[g(Y) | X=x]. \end{aligned} \quad (9.16)$$

(ii)  $E_Y^{X=x}(g)$  is finite if and only if  $E^{X=x}[g(Y)]$  is finite.

There are two important points in Equations (9.16). *First*, these equations show the relationship between integrals of the composition  $g(Y)$  with respect to the conditional-probability measure  $P^{X=x}$  on  $(\Omega, \mathcal{A})$  on one side, and the distribution  $P_Y^{X=x}$  of  $Y$  with respect to  $P^{X=x}$  on the other side. *Second*,  $E[g(Y) | X=x]$  is identical to the expectation of  $g$  with respect to the distribution  $P_Y^{X=x}$ , i. e., the

distribution of  $Y$  with respect to the probability measure  $P^{X=x}$ . Thus, using the distribution  $P_{g(Y)}^{X=x}$  of  $g(Y)$  with respect to  $P^{X=x}$  is not necessary.

**Remark 9.10 (( $X=x$ )-Conditional Expectation Value of  $g(Y)$  and  $P$ -Expectation)** Equations (9.16) and Equations (9.11) imply

$$E[g(Y)|X=x] = \frac{1}{P(X=x)} \cdot \int 1_{X=x} \cdot g(Y) dP = \frac{1}{P(X=x)} \cdot E[1_{X=x} \cdot g(Y)]. \quad (9.17)$$

&lt;

**Remark 9.11 (A Special Case of the Transformation Theorem)** Let  $(\Omega'_Y, \mathcal{A}'_Y) = (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  and  $g$  be the identity function  $id: \bar{\mathbb{R}} \rightarrow \bar{\mathbb{R}}$ , defined by  $id(y) = y$  for all  $y \in \bar{\mathbb{R}}$ , which implies  $id(Y) = Y$ . If we assume that  $x \in \Omega'_X$  with  $\{x\} \in \mathcal{A}'_X$  and  $P(X=x) > 0$ , and that  $Y$  is nonnegative or with finite expectation  $E^{X=x}(Y)$ , then Equations (9.16) yield

$$E(Y|X=x) = E^{X=x}(Y) = \int Y dP^{X=x} = \int y P_Y^{X=x}(dy) = \int id dP_Y^{X=x}. \quad (9.18)$$

&lt;

Using the notation introduced in Equation (9.14), Equations (9.16), Theorem 9.9 (i), Equations (6.3) and (6.6) imply the following corollary.

**Corollary 9.12 ( $Y$  Discrete,  $g(Y)$  Real-Valued)**

Let the assumptions of Theorem 9.9 (i) hold and let  $g: (\Omega'_Y, \mathcal{A}'_Y) \rightarrow (\mathbb{R}, \mathcal{B})$ .

- (i) If  $Y$  is discrete and we assume that there is a finite set  $\{y_1, \dots, y_n\} \subset \Omega'_Y$ ,  $\{y_1\}, \dots, \{y_n\} \in \mathcal{A}'_Y$  with  $P_Y(\{y_1, \dots, y_n\}) = 1$ , then

$$E[g(Y)|X=x] = \sum_{i=1}^n g(y_i) \cdot P(Y=y_i|X=x). \quad (9.19)$$

- (ii) If  $Y$  is discrete and we assume that there is a countably infinite set  $\{y_1, y_2, \dots\} \subset \Omega'_Y$ ,  $\{y_i\} \in \mathcal{A}'_Y$ ,  $i \in \mathbb{N}$ , with  $P_Y(\{y_1, y_2, \dots\}) = 1$ , then

$$E[g(Y)|X=x] = \sum_{i=1}^{\infty} g(y_i) \cdot P(Y=y_i|X=x). \quad (9.20)$$

Note that, in this corollary,  $Y$  does not have to be real-valued or numerical. We only assume that  $g(Y)$  is real-valued. In contrast, in the following theorem we have to assume that  $Y$  itself is real-valued. Remember that  $\{y\} \in \bar{\mathcal{B}}$  if  $y \in \mathbb{R}$  [see (1.22)].

**Corollary 9.13 (Discrete and Real-Valued  $Y$ )**

Let the assumptions of Definition 9.2 (ii) hold.

- (i) If  $Y$  is discrete and there is a finite set  $\{y_1, \dots, y_n\} \subset \mathbb{R}$  of real numbers with  $P_Y(\{y_1, \dots, y_n\}) = 1$ . Then

$$E(Y|X=x) = \sum_{i=1}^n y_i \cdot P(Y=y_i | X=x). \quad (9.21)$$

- (ii) If  $Y$  is discrete and there is a countably infinite set  $\{y_1, y_2, \dots\} \subset \mathbb{R}$  of real numbers with  $P_Y(\{y_1, y_2, \dots\}) = 1$ , then

$$E(Y|X=x) = \sum_{i=1}^{\infty} y_i \cdot P(Y=y_i | X=x). \quad (9.22)$$

**9.3 Other Properties**

Because  $E(Y|B)$  is defined as the expectation  $E^B(Y)$  of  $Y$  with respect to the probability measure  $P^B$ , all properties of the expectation with respect to  $P$  can be translated to  $E(Y|B)$ , simply replacing  $P$  by  $P^B$  and  $E(Y)$  by  $E^B(Y) = E(Y|B)$ . Box 9.1 is such a translation of Box 6.1. Note that, according to Theorem 9.4 (i), the expectation  $E(Y|B)$  exists if  $E(Y)$  exists, provided that  $P(B) > 0$ . Of course, the rules for  $E(Y|B)$  also apply to the  $(X=x)$ -conditional expectation value  $E(Y|X=x)$  [see Def. 9.2 (ii)].

However, there are additional properties dealing with the relationship between the expectation and the conditional expectation value. Some of these have already been formulated in Theorem 9.4. Other additional properties are summarized in Box 9.2 and proved in Exercise 9-2.

Rule (ii) shows how the  $(X=x)$ -conditional expectation values  $E(Y|X=x)$  can be computed from the conditional expectation values  $E(Y|X=x, Z=z_i)$  and the conditional probabilities  $P(Z=z_i | X=x)$ . Hence, considering Equation (9.21) and Rule (ii) in Box 9.2 shows that we have two different equations for computing the conditional expectation value  $E(Y|X=x)$ . Finally, note that a special case of Rule (ii) is

$$E(Y) = \sum_{i=1}^m E(Y|Z=z_i) \cdot P(Z=z_i) \quad (9.23)$$

(see Exercise 9-3). According to this equation, we can also compute the expectation of  $Y$  from the conditional expectations  $E(Y|Z=z_i)$  and the probabilities  $P(Z=z_i)$ .

**Box 9.1 Rules of Computation for  $B$ -Conditional Expectation Values**

Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  be a random variable, let  $A, B, C \in \mathcal{A}$  with  $P(B) > 0$ , let the conditional expectation value  $E(Y|B)$  exist, and let  $\alpha \in \mathbb{R}$ . Then:

$$Y \stackrel{p^B}{=} \alpha \Rightarrow E(Y|B) = \alpha. \quad (\text{i})$$

$$E(\alpha + Y|B) = \alpha + E(Y|B). \quad (\text{ii})$$

$$E(\alpha \cdot Y|B) = \alpha \cdot E(Y|B). \quad (\text{iii})$$

$$E(1_A \cdot 1_C|B) = P(A \cap C|B). \quad (\text{iv})$$

For  $i = 1, \dots, n$ , let  $Y_i: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  be random variables with finite  $B$ -conditional expectation values  $E(Y_i|B)$  and  $\alpha_i \in \mathbb{R}$ . Then

$$E\left(\sum_{i=1}^n \alpha_i \cdot Y_i \mid B\right) = \sum_{i=1}^n \alpha_i \cdot E(Y_i|B). \quad (\text{v})$$

Let  $X, Y: (\Omega, \mathcal{A}, P) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  be random variables that are nonnegative or with finite  $B$ -conditional expectation values. Then:

$$X \stackrel{p^B}{=} Y \Rightarrow E(X|B) = E(Y|B). \quad (\text{vi})$$

$$X \stackrel{p^B}{=} Y \Leftrightarrow \forall A \in \mathcal{A}: E(1_A X|B) = E(1_A Y|B). \quad (\text{vii})$$

$$X \stackrel{p^B}{\perp\!\!\!\perp} Y \Rightarrow E(X \cdot Y|B) = E(X|B) \cdot E(Y|B). \quad (\text{viii})$$

**9.4 Discrete Conditional Expectation**

The *discrete conditional expectation*  $E(Y|X)$  of a numerical random variable  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  given a random variable  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  is now defined as that random variable on  $(\Omega, \mathcal{A}, P)$  whose values are identical to the conditional expectation values  $E(Y|X=x)$ . In this definition we have to assume that  $X$  is *discrete*, i. e., we assume that there is a finite or countable set  $\Omega'_0 \subset \Omega'_X$  such that  $P_X(\Omega'_0) = 1$  and  $P(X=x) > 0$  for all  $x \in \Omega'_0$  (see Def. 5.56). In chapter 10 this limitation is dropped.

**Definition 9.14 (Discrete Conditional Expectation)**

Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  be a numerical random variable that is nonnegative or has a finite expectation and let the random variable  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  be discrete.

- (i) If  $\{x_1, \dots, x_m\} = \Omega'_0 \subset \Omega'_X$  such that  $P_X(\Omega'_0) = 1$  and, for all  $i = 1, \dots, m$ ,  $\{x_i\} \in \mathcal{A}'_X$  and  $P(X=x_i) > 0$ , then the *discrete conditional expectation*

**Box 9.2 Rules of Computation for  $(X=x)$ -Conditional Expectation Values**

Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  be a numerical random variable. Furthermore, let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  be a random variable, and let  $x \in \Omega'_X$  with  $\{x\} \in \mathcal{A}'_X$  and  $P(X=x) > 0$ . If  $E(Y|X=x)$  exists and  $f: (\Omega'_X, \mathcal{A}'_X) \rightarrow (\mathbb{R}, \mathcal{B})$  is a measurable function and  $E(Y^2), E[f(X)^2] < \infty$ , then

$$E[f(X) \cdot Y | X=x] = f(x) \cdot E(Y | X=x) = E[f(x) \cdot Y | X=x]. \quad (\text{i})$$

If  $Z: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Z, \mathcal{A}'_Z)$  is a random variable and  $z_1, \dots, z_m \in \Omega'_Z$  such that  $P_Z(\{z_1, \dots, z_m\}) = 1$  and, for all  $i = 1, \dots, m$ ,  $\{z_i\} \in \mathcal{A}'_Z$  and  $P(X=x, Z=z_i) > 0$ , then

$$E(Y | X=x) = \sum_{i=1}^m E(Y | X=x, Z=z_i) \cdot P(Z=z_i | X=x). \quad (\text{ii})$$

Correspondingly, if  $z_1, z_2, \dots \in \Omega'_Z$  such that  $P_Z(\{z_1, z_2, \dots\}) = 1$  and, for all  $i = 1, 2, \dots$ ,  $\{z_i\} \in \mathcal{A}'_Z$  and  $P(X=x, Z=z_i) > 0$ , then

$$E(Y | X=x) = \sum_{i=1}^{\infty} E(Y | X=x, Z=z_i) \cdot P(Z=z_i | X=x). \quad (\text{iii})$$

*tation of  $Y$  given  $X$  is defined by*

$$E(Y | X) := \sum_{i=1}^m E(Y | X=x_i) \cdot \mathbf{1}_{X=x_i}. \quad (9.24)$$

(ii) *If  $\{x_1, x_2, \dots\} = \Omega'_0 \subset \Omega'_X$  such that  $P_X(\Omega'_0) = 1$  and, for all  $i = 1, 2, \dots$ ,  $\{x_i\} \in \mathcal{A}'_X$  and  $P(X=x_i) > 0$ , then the discrete conditional expectation of  $Y$  given  $X$  is defined by*

$$E(Y | X) := \sum_{i=1}^{\infty} E(Y | X=x_i) \cdot \mathbf{1}_{X=x_i}. \quad (9.25)$$

Hence, in contrast to a conditional expectation value  $E(Y|X=x)$ , a discrete conditional expectation  $E(Y|X)$  is a discrete *random variable* (see Def. 5.56) on  $(\Omega, \mathcal{A}, P)$ . Note that  $E(Y|X)$  is a random variable taking a numerical value for each  $\omega \in \Omega$ . This means that we might look at its expectation, variance, covariance and correlation with other random variables (see, e.g., Box 10.2 and section 11.2).

**Remark 9.15 ( $X$ -Conditional Probability)** If  $A \in \mathcal{A}$ , then we use the notation

$$P(A|X) := E(\mathbf{1}_A | X) \quad (9.26)$$

and call it the *X*-conditional probability of *A*. If *Y* is dichotomous with values 0 and 1, we also use the notation  $P(Y=1|X)$  for the *X*-conditional probability of the event  $\{Y=1\}$ . If *Y* is dichotomous with values 0 and 1, Equations (9.21) and (9.26) then yield

$$P(Y=1|X) = E(Y|X). \quad (9.27)$$

◁

**Remark 9.16 (Uniqueness and Values of the Conditional Expectation)** An alternative way to write Equations (9.24) and (9.25) is

$$E(Y|X)(\omega) = \begin{cases} E(Y|X=x), & \text{if } \omega \in X^{-1}(\{x\}), \quad \forall x \in \Omega'_0, \\ 0, & \text{otherwise.} \end{cases} \quad (9.28)$$

Hence, the values of the conditional expectation  $E(Y|X)$  are uniquely defined by Equations (9.24) and (9.25) for all  $\omega \in \Omega$  (see Example 9.22). Assigning the value  $E(Y|X)(\omega) = 0$  if  $\omega \in \Omega \setminus X^{-1}(\Omega'_0)$  is arbitrary, but note that  $P(\Omega \setminus X^{-1}(\Omega'_0)) = 0$ . According to Equation (9.28) and Definition 5.56 this arbitrary assignment does not occur if  $P(X=x) > 0$  for all  $x \in X(\Omega)$ , i. e., if  $\Omega'_0 = X(\Omega)$ . ◁

## 9.5 Discrete Regression

**Remark 9.17 (Measurability and Factorization)** Definition 9.14 implies that the discrete conditional expectation  $E(Y|X)$  is a random variable on  $(\Omega, \mathcal{A}, P)$  that is measurable with respect to *X*. In more formal terms,  $E(Y|X): (\Omega, \mathcal{A}, P) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  and  $\sigma[E(Y|X)] \subset \sigma(X)$ . The reason is that there is a measurable function  $g: (\Omega'_X, \mathcal{A}'_X) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  that is defined by

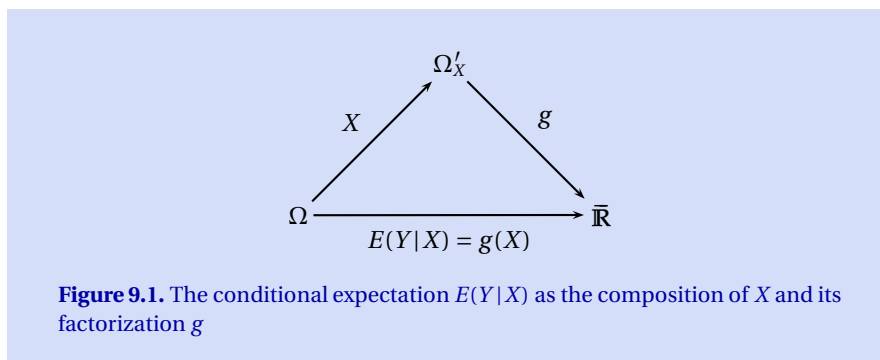
$$g(x) = \begin{cases} E(Y|X=x), & \forall x \in \Omega'_0, \\ 0, & \text{otherwise.} \end{cases} \quad (9.29)$$

Hence,  $E(Y|X) = g(X)$  (see Fig. 9.1) and Lemma 2.52 implies that  $E(Y|X)$  is measurable with respect to *X*. The function *g* is called the *factorization of  $E(Y|X)$*  or the *discrete regression of  $Y$  on  $X$* . ◁

### Definition 9.18 (Discrete Regression)

Under the assumptions specified in Definition 9.14, the function  $g: \Omega'_X \rightarrow \mathbb{R}$  defined by Equation (9.29) is called the *discrete regression of  $Y$  on  $X$* .

**Remark 9.19 (Regressand and Regressor)** Considering the conditional expectation  $E(Y|X)$ , we call *Y* the *regressand* and *X* the *regressor*. Whereas the regressand *Y* has to be numerical, the regressor *X* can be *any* random variable on  $(\Omega, \mathcal{A}, P)$  as long as it is discrete and satisfies  $P(X=x) > 0$  for all  $x \in \Omega'_0$ . ◁



**Figure 9.1.** The conditional expectation  $E(Y|X)$  as the composition of  $X$  and its factorization  $g$

**Table 9.1.** Joe and Ann With Random Assignment: Conditional Expectations

Elements of $\Omega$			Random variables			Conditional expectations		
Unit	Treatment	Success	Person variable $U$	Treatment variable $X$	Outcome variable $Y$	$E(Y X, U)$	$E(Y X)$	$P(X=1 U)$
(Joe, no, -)			Joe	0	0	.70	.45	.40
(Joe, no, +)			Joe	0	1	.70	.45	.40
(Joe, yes, -)			Joe	1	0	.80	.60	.40
(Joe, yes, +)			Joe	1	1	.80	.60	.40
(Ann, no, -)			Ann	0	0	.20	.45	.40
(Ann, no, +)			Ann	0	1	.20	.45	.40
(Ann, yes, -)			Ann	1	0	.40	.60	.40
(Ann, yes, +)			Ann	1	1	.40	.60	.40

**Remark 9.20 (Multivariate Numerical or Qualitative Regressors)** Note that the codomain  $\Omega'_X$  of  $X$  can be *any* set as long as there is a subset  $\Omega'_0 \subset \Omega'_X$  such that  $\Omega'_0$  is finite, or countable with  $P(X \in \Omega'_0) = 1$  and  $P(X=x) > 0$  for all  $x \in \Omega'_0$ . Hence, the *regressor*  $X$  can be uni- or multivariate (see Examples 9.21 and 9.22). If  $X = (X_1, \dots, X_n)$  is a discrete multivariate random variable, then we also use the notation  $E(Y|X_1, \dots, X_n)$  instead of  $E(Y|X)$  for the conditional expectation of  $Y$  given  $X$ . ◁

## 9.6 Examples

We treat two examples in some detail. Example 9.21 is straightforward, whereas Example 9.22 exemplifies that the values of a conditional expectation are uniquely defined by Equation (9.24) for all  $\omega \in \Omega$ .

**Example 9.21 (Joe and Ann With Random Assignment – continued)** Table 9.1 contains three conditional expectations we may consider in the example already used in Example 1.9. All of them are random variables taking a numerical value for each  $\omega \in \Omega$ . According to Remark 9.16, the values of  $E(Y|X)$  are the conditional expectation values  $E(Y|X=x)$  for all  $x \in \Omega'_0$ , and they are 0 for all  $x \in \Omega'_X \setminus \Omega'_0$ .

We start illustrating the conditional expectation of  $Y$  given  $X$ . Both random variables,  $X$  and  $Y$ , are specified in Table 9.1. We consider the random variable  $X: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$ . In this case,  $\Omega'_X = \mathbb{R}$  and  $\Omega'_0 = \{0, 1\}$ . Because  $X$  takes on a value in  $\Omega'_0 = \{0, 1\}$  for *all*  $\omega \in \Omega$ , the conditional expectation  $E(Y|X)$  either takes on the value  $E(Y|X=0)$  or the value  $E(Y|X=1)$ . It does not take on the value 0, because  $\{X=x\} = \emptyset$  for all  $x \in \mathbb{R} \setminus \{0, 1\}$  [see Eq. (9.28)].

Because  $Y$  is an indicator,  $E(Y|X=x) = P(Y=1|X=x)$  [see Eqs. (9.14) and (9.27)]. Hence, if we want to compute the values of  $E(Y|X) = P(Y=1|X)$ , we have to compute the conditional probabilities  $P(Y=1|X=x)$ . For  $x=0$  we receive

$$P(Y=1|X=0) = \frac{P(Y=1, X=0)}{P(X=0)} = \frac{.21 + .06}{.09 + .21 + .24 + .06} = \frac{.27}{.60} = .45,$$

and for  $x=1$ ,

$$P(Y=1|X=1) = \frac{P(Y=1, X=1)}{P(X=1)} = \frac{.16 + .08}{.04 + .16 + .12 + .08} = \frac{.24}{.40} = .60.$$

Now we consider the conditional expectation  $E(Y|X, U)$ . The regressor is the random variable  $(X, U): (\Omega, \mathcal{A}, P) \rightarrow [\mathbb{R} \times \Omega_U, \mathcal{B} \otimes \mathcal{P}(\Omega_U)]$ , where  $\Omega_U = \{Joe, Ann\}$ , and  $\Omega'_0 = \{0, 1\} \times \Omega_U$ . Because,

$$\forall (x, u) \in (\mathbb{R} \times \Omega_U) \setminus (\{0, 1\} \times \Omega_U): \{(X=x, U=u)\} = \emptyset,$$

$E(Y|X, U)$  takes on a value in  $\mathbb{R}$  for *all*  $\omega \in \Omega$ . Furthermore, because  $Y$  is an indicator,  $E(Y|X, U) = P(Y=1|X, U)$ , and this conditional expectation has only four different values: the conditional probabilities  $P(Y=1|X=x, U=u)$ . For  $x=0$ ,  $u=Joe$  we receive

$$P(Y=1|X=0, U=Joe) = \frac{P(Y=1, X=0, U=Joe)}{P(X=0, U=Joe)} = \frac{.21}{.09 + .21} = .70,$$

for  $x=1$ ,  $u=Joe$ ,

$$P(Y=1|X=1, U=Joe) = \frac{P(Y=1, X=1, U=Joe)}{P(X=1, U=Joe)} = \frac{.16}{.04 + .16} = .80,$$

for  $x=0$ ,  $u=Ann$ ,

$$P(Y=1|X=0, U=Ann) = \frac{P(Y=1, X=0, U=Ann)}{P(X=0, U=Ann)} = \frac{.06}{.24 + .06} = .20,$$

and for  $x=1, u=Ann$ ,

$$P(Y=1|X=1, U=Ann) = \frac{P(Y=1, X=1, U=Ann)}{P(X=1, U=Ann)} = \frac{.08}{.12 + .08} = .40.$$

◁

**Example 9.22 (No Treatment for Joe)** Let us use a second example in order to illustrate the concepts introduced above. Again the random experiment consists of sampling a person, observing the value  $x$  of the treatment variable  $X$ , i. e., whether or not the sampled person receives a treatment ( $x=1$  vs.  $x=0$ ), and observing whether or not a success criterion is reached some time after treatment. In this new example, we fixed new probabilities of the elementary events. For instance, now the probability that Joe receives treatment is zero. This is useful to illustrate some general properties of conditional expectations. Also note that the probabilities of the other elementary events have been changed as well. The only restriction on the probabilities of the elementary events in such a hypothetical example is that they sum up to one.

Using the probabilities displayed in Table 9.2, Equation (9.21) yields:

$$\begin{aligned} E(Y|X=1) &= P(Y=1|X=1) = \frac{P(Y=1, X=1)}{P(X=1)} \\ &= \frac{0 + .152}{0 + 0 + .228 + .152} = .40 \end{aligned}$$

for the treatment condition  $x=1$ . Applying the corresponding formula to the control condition  $x=0$  yields  $E(Y|X=0) = (.348 + .024) / (.152 + .348 + .096 + .024) = .60$ . Note that the conditional probabilities  $P(Y=1|X=1)$  and  $P(Y=1|X=0)$  do *not* necessarily add up to 1. In contrast, the sum of  $P(Y=1|X=1)$  and  $P(Y=0|X=1)$  and the sum of  $P(Y=1|X=0)$  and  $P(Y=0|X=0)$  is always equal to 1, provided that  $Y$  is dichotomous with values 0 and 1.

Table 9.2 also displays the conditional probability  $P(X=1|U)$ , whose values are the treatment probabilities of Joe and Ann. For Joe, the treatment probability is  $P(X=1|U=Joe) = 0$ , and for Ann it is

$$P(X=1|U=Ann) = (.228 + .152) / (.096 + .024 + .228 + .152) = .76.$$

Finally, we compute the conditional expectations  $E(Y|X=x, U=u) = P(Y=1|X=x, U=u)$  [see Eq. (9.14)]:

$$\begin{aligned} P(Y=1|X=0, U=Joe) &= \frac{P(Y=1, X=0, U=Joe)}{P(X=0, U=Joe)} = \frac{.348}{.152 + .348} = .696, \\ P(Y=1|X=0, U=Ann) &= \frac{P(Y=1, X=0, U=Ann)}{P(X=0, U=Ann)} = \frac{.024}{.096 + .024} = .20, \end{aligned}$$

**Table 9.2.** No Treatment for Joe With Conditional Expectations

Elements of $\Omega$			Observables			Conditional expectations			
Unit	Treatment	Success	$P(\{\omega\})$	Person variable $U$	Treatment variable $X$	Outcome variable $Y$	$E(Y X,U)$	$E(Y X)$	$P(X=1 U)$
(Joe, no, -)			.152	Joe	0	0	.696	.60	0
(Joe, no, +)			.348	Joe	0	1	.696	.60	0
(Joe, yes, -)			0	Joe	1	0	0	.40	0
(Joe, yes, +)			0	Joe	1	1	0	.40	0
(Ann, no, -)			.096	Ann	0	0	.20	.60	.76
(Ann, no, +)			.024	Ann	0	1	.20	.60	.76
(Ann, yes, -)			.228	Ann	1	0	.40	.40	.76
(Ann, yes, +)			.152	Ann	1	1	.40	.40	.76

$$P(Y=1|X=1, U=Ann) = \frac{P(Y=1, X=1, U=Ann)}{P(X=1, U=Ann)} = \frac{.152}{.228 + .152} = .40.$$

Note that  $E(Y|X=1, U=Joe)$  is not defined, because  $P(X=1, U=Joe) = 0$  (see Def. 9.2). However, according to Definition 9.14 (i),  $E(Y|X,U)(\omega) = 0$  for  $\omega \in \Omega \setminus \Omega'_0 = \{(Joe, yes, -), (Joe, yes, +)\}$ . Thus  $E(Y|X,U)$  is uniquely defined, i.e., the values  $E(Y|X,U)(\omega)$  are uniquely defined for all  $\omega \in \Omega$ .  $\triangleleft$

**Example 9.23 (No Treatment for Joe – continued)** Using the results obtained in Example 9.22, Equation (9.24) yields:

$$\begin{aligned} E(Y|X) &= \sum_{x=0}^1 E(Y|X=x) \cdot 1_{X=x} \\ &= E(Y|X=0) \cdot 1_{X=0} + E(Y|X=1) \cdot 1_{X=1} \\ &= .60 \cdot 1_{X=0} + .40 \cdot 1_{X=1}. \end{aligned}$$

Hence, the values  $E(Y|X)(\omega)$  of the  $X$ -conditional expectation of  $Y$  are

$$E(Y|X)(\omega) = .60 \cdot 1_{X=0}(\omega) + .40 \cdot 1_{X=1}(\omega) = .60, \quad \text{if } X(\omega) = 0,$$

and

$$E(Y|X)(\omega) = .60 \cdot 1_{X=0}(\omega) + .40 \cdot 1_{X=1}(\omega) = .40, \quad \text{if } X(\omega) = 1.$$

These are the only two values that  $E(Y|X)$  takes on depending on the outcome  $\omega \in \Omega$  of the random experiment considered (see the first column of Table 9.2). This example illustrates that  $E(Y|X)$  is a random variable on  $(\Omega, \mathcal{A}, P)$  just like  $X$ ,

$Y$ , and  $U$ . Note again that the two values of  $E(Y|X) = P(Y=1|X)$  add up to 1 only by coincidence.

Table 9.2 shows two additional conditional expectations,  $E(X|U) = P(X=1|U)$  as well as  $E(Y|X, U) = P(Y=1|X, U)$ . Again using Equation (9.24) and the results obtained in Example 9.22, the conditional expectation  $P(X=1|U)$  is

$$\begin{aligned} P(X=1|U) &= \sum_u P(X=1|U=u) \cdot 1_{U=u} \\ &= P(X=1|U=Joe) \cdot 1_{U=Joe} + P(X=1|U=Ann) \cdot 1_{U=Ann} \\ &= .00 \cdot 1_{U=Joe} + .76 \cdot 1_{U=Ann} = .76 \cdot 1_{U=Ann}. \end{aligned}$$

Hence, the values  $P(X=1|U)(\omega)$  of the  $U$ -conditional expectation of  $X$  are

$$P(X=1|U)(\omega) = .00 \cdot 1_{U=Joe}(\omega) + .76 \cdot 1_{U=Ann}(\omega) = .00, \quad \text{if } U(\omega) = Joe$$

and

$$P(X=1|U)(\omega) = .00 \cdot 1_{U=Joe}(\omega) + .76 \cdot 1_{U=Ann}(\omega) = .76, \quad \text{if } U(\omega) = Ann.$$

These are the only two values that  $P(X=1|U)$  takes on. Again, Table 9.2 shows how the values of  $P(X=1|U)$  are assigned to the outcomes  $\omega \in \Omega$ .

Finally, let us turn to the conditional expectation  $E(Y|X, U)$ . Using the results obtained in Example 9.22, its values are

$$E(Y|X, U)(\omega) = E(Y|X=0, U=Joe) = .696, \quad \text{if } X(\omega) = 0 \text{ and } U(\omega) = Joe$$

$$E(Y|X, U)(\omega) = E(Y|X=0, U=Ann) = .20, \quad \text{if } X(\omega) = 0 \text{ and } U(\omega) = Ann$$

and

$$E(Y|X, U)(\omega) = E(Y|X=1, U=Ann) = .40, \quad \text{if } X(\omega) = 1 \text{ and } U(\omega) = Ann,$$

whereas

$$E(Y|X, U)(\omega) = 0, \quad \text{if } X(\omega) = 1 \text{ and } U(\omega) = Joe.$$

Note that the value  $E(Y|X, U)(\omega)$  is defined for  $\omega \in \{X=1, U=Joe\}$  although the conditional expectation value  $E(Y|X=1, U=Joe)$  is *not* defined. Also note that in this case the value  $E(Y|X, U)(\omega) = 0$  is arbitrarily fixed and has no substantive meaning. However, because  $P(X=1, U=Joe) = 0$ , this has no disadvantageous consequences. [In chapter 10, we show that the values of a conditional expectation can arbitrarily be fixed for all elements  $\omega$  of a subset  $A$  of  $\Omega$  for which  $P(A) = 0$ .]

According to Equation (9.24), the conditional expectation  $E(Y|X, U)$  is

$$\begin{aligned} E(Y|X, U) &= \sum_{(x,u)} E(Y|X=x, U=u) \cdot 1_{X=x, U=u} \\ &= .696 \cdot 1_{X=0, U=Joe} + .20 \cdot 1_{X=0, U=Ann} + .40 \cdot 1_{X=1, U=Ann}. \end{aligned} \tag{9.30}$$

The pair  $(1, Joe)$  is not an element of the set  $\Omega'_0$  (see Def. 9.14), and therefore the corresponding indicator  $1_{X=1, U=Joe}$  does not occur in this sum. Hence, if

$$\omega \in \{X=1, U=Joe\} = \{(Joe, yes, -), (Joe, yes, +)\},$$

then all three indicators occurring in Equation (9.30) take on the value 0, implying that  $E(Y|X, U)(\omega) = 0$  for these two elements of  $\Omega$ .  $\triangleleft$

## 9.7 Proofs

### Proof of Theorem 9.4

$$\begin{aligned} E(Y|B) &= E^B(Y) = \int Y dP^B && [(9.5)] \\ &= \int Y d\left(\frac{1}{P(B)} \cdot 1_B\right) \circ P && [\text{Lem. 4.32 (ii)}] \\ &= \frac{1}{P(B)} \cdot \int 1_B \cdot Y dP && [(3.72), (3.32)] \\ &= \frac{1}{P(B)} \cdot E(1_B \cdot Y), && [(6.1)] \end{aligned}$$

which yields Equation (9.7). Using Lemma 3.33 (i) yields Theorem 9.4 (i) and Lemma 3.33 (ii) implies Theorem 9.4 (ii).

If  $E(Y^2)$  is finite, then  $E(Y)$  and  $E(1_B \cdot Y)$  are finite as well [see Rem. 6.25 (iii) and Lemma 3.33 (ii)]. Therefore, according to Box 7.1 (i),  $Cov(1_B, Y) = E(1_B \cdot Y) - E(1_B) \cdot E(Y)$  is finite. Hence,

$$\begin{aligned} E(Y|B) &= \frac{1}{P(B)} \cdot E(1_B \cdot Y) && [(9.7)] \\ &= \frac{1}{P(B)} \cdot [Cov(1_B, Y) + E(1_B) \cdot E(Y)] && [\text{Box 7.1 (i)}] \\ &= \frac{1}{P(B)} \cdot Cov(1_B, Y) + E(Y). && [(6.4)] \end{aligned}$$

## 9.8 Exercises

- ▷ **Exercise 9-1** Prove the proposition of Remark 9.1
- ▷ **Exercise 9-2** Prove the rules of computation of conditional expectation values displayed in Box 9.2.
- ▷ **Exercise 9-3** Show that Equation (9.23) is a special case of Rule (ii) of Box 9.2.

## Solutions

▷ **Solution 9-1** It is sufficient to prove  $P^{X=x}(\{f(X) = f(x)\}^c) = 0$ , which is equivalent to  $P^{X=x}(\{f(X) = f(x)\}) = 1$ .

$$P^{X=x}[f(X) = f(x)] = \frac{P(f(X) = f(x), X=x)}{P(X=x)} \quad [(9.4), (9.15)]$$

$$\begin{aligned}
&= \frac{P(X=x)}{P(X=x)} && \{X=x\} \subset \{f(X)=f(x)\}, (9.9) \\
&= 1.
\end{aligned}$$

▷ **Solution 9-2**

$$\begin{aligned}
\text{(i)} \quad E[f(X) \cdot Y | X=x] &= \frac{1}{P(X=x)} \cdot E[1_{X=x} \cdot f(X) \cdot Y] && [(9.11)] \\
&= \frac{1}{P(X=x)} \cdot E[1_{X=x} \cdot f(x) \cdot Y] && [1_{X=x} \cdot f(X) = 1_{X=x} \cdot f(x)] \\
&= f(x) \cdot \frac{1}{P(X=x)} \cdot E(1_{X=x} \cdot Y) && [\text{Box 6.1 (iii)}] \\
&= f(x) \cdot E(Y | X=x) && [(9.7)] \\
&= E[f(x) \cdot Y | X=x]. && [\text{Box 9.1 (iii)}]
\end{aligned}$$

$$\begin{aligned}
\text{(ii)} \quad E(Y | X=x) &= \frac{1}{P(X=x)} \cdot E(1_{X=x} \cdot Y) && [(9.11)] \\
&= \frac{1}{P(X=x)} \cdot E\left(1_{X=x} \cdot \sum_{i=1}^m 1_{Z=z_i} \cdot Y\right) && \left[1 \stackrel{p}{=} \sum_{i=1}^m 1_{Z=z_i}, (5.33), \text{Rem. 6.27}\right] \\
&= \sum_{i=1}^m \frac{1}{P(X=x)} \cdot E(1_{X=x} \cdot 1_{Z=z_i} \cdot Y) && [\text{Box 6.1 (vii)}] \\
&= \sum_{i=1}^m \frac{P(X=x, Z=z_i)}{P(X=x)} \cdot \frac{1}{P(X=x, Z=z_i)} \cdot E(1_{X=x, Z=z_i} \cdot Y) \\
&= \sum_{i=1}^m E(Y | X=x, Z=z_i) \cdot P(Z=z_i | X=x). && [(9.11), (9.15)]
\end{aligned}$$

$$\begin{aligned}
\text{(iii)} \quad &\sum_{i=1}^{\infty} E(Y | X=x, Z=z_i) \cdot P(Z=z_i | X=x) \\
&= \sum_{i=1}^{\infty} P(Z=z_i | X=x) \cdot \frac{1}{P(X=x, Z=z_i)} \cdot E(1_{X=x, Z=z_i} \cdot Y) && [(9.11)] \\
&= \sum_{i=1}^{\infty} \frac{1}{P(X=x)} \cdot E(1_{X=x, Z=z_i} \cdot Y) && [(9.15)] \\
&= \frac{1}{P(X=x)} \sum_{i=1}^{\infty} \int 1_{X=x, Z=z_i} \cdot Y \, dP && [\text{Def. 6.1}] \\
&= \frac{1}{P(X=x)} \sum_{i=1}^{\infty} \left[ \int 1_{Z=z_i} \cdot 1_{X=x} \cdot Y^+ \, dP - \int 1_{Z=z_i} \cdot 1_{X=x} \cdot Y^- \, dP \right] && [\text{Def. 3.28}] \\
&= \frac{1}{P(X=x)} \left[ \sum_{i=1}^{\infty} \int 1_{Z=z_i} \cdot 1_{X=x} \cdot Y^+ \, dP - \sum_{i=1}^{\infty} \int 1_{Z=z_i} \cdot 1_{X=x} \cdot Y^- \, dP \right] \\
& && [1_{X=x} \cdot Y \text{ is quasi-integrable}] \\
&= \frac{1}{P(X=x)} \left[ \int 1_{X=x} \cdot Y^+ \, dP - \int 1_{X=x} \cdot Y^- \, dP \right] && \left[ (3.65), 1 \stackrel{p}{=} \sum_{i=1}^{\infty} 1_{Z=z_i}, (5.33), (3.44) \right]
\end{aligned}$$

$$\begin{aligned} &= \frac{1}{P(X=x)} \int 1_{X=x} \cdot Y \, dP && \text{[Def. 3.28]} \\ &= E(Y|X=x). && \text{[(9.11)]} \end{aligned}$$

▷ **Solution 9-3** This is easily seen considering the special case  $X=\alpha$ ,  $\alpha \in \Omega'_X$ , i. e., the case in which  $X$  is a constant. Then  $\{X=\alpha\} = \Omega$ , and Equation (9.11) yields  $E(Y|X=\alpha) = E(Y)$ , and  $E(Y|X=\alpha, Z=z_i) = E(Y|Z=z_i)$ . Hence, Rule (ii) of Box 9.2 yields

$$E(Y) = \sum_{i=1}^m E(Y|Z=z_i) \cdot P(Z=z_i).$$



## Chapter 10

# Conditional Expectation

In chapter 9 we treated the conditional expectation value given an event and the conditional expectation  $E(Y|X)$  assuming that  $X$  is discrete taking on each of its values  $x$  with a positive probability  $P(X=x) > 0$ . In this chapter we introduce the general concept of a conditional expectation given a  $\sigma$ -algebra  $\mathcal{C}$ . The price of this generalization is that a  $\mathcal{C}$ -conditional expectation is uniquely defined only up to  $P$ -equivalence, i. e., if there are two versions of such a  $\mathcal{C}$ -conditional expectation of a numerical random variable  $Y$ , then they are not necessarily identical, but they are  $P$ -equivalent. Furthermore, if  $\mathcal{C}$  is generated by a random variable  $X$ , then a  $\mathcal{C}$ -conditional expectation is also called a *conditional expectation of  $Y$  given  $X$* . This definition also applies if  $X$  is continuous. Hence, it also holds if  $P(X=x) = 0$  for all values  $x$  of  $X$ . In this chapter, we also define the general concept of a regression as a factorization  $g$  of a conditional expectation  $E(Y|X) = g(X)$ , and an  $(X=x)$ -conditional expectation value  $E(Y|X=x)$  as a value  $g(x)$  of  $g$ . This means that  $E(Y|X=x)$  is defined even if  $P(X=x) = 0$ . However,  $E(Y|X=x)$  is not uniquely defined. Nevertheless, we can formulate propositions about the conditional expectation values  $E(Y|X=x)$  for  $P_X$ -almost all values  $x$  of  $X$ . Finally, we introduce the concepts of *mean independence*, and study their relationship to stochastic independence and uncorrelatedness.

### 10.1 Assumptions and Definitions

Throughout this chapter, we will make the following assumptions and use the following notation.

#### Notation and Assumptions 10.1

$Y: (\Omega, \mathcal{A}, P) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  is a numerical random variable that is nonnegative or has a finite expectation  $E(Y)$ . Furthermore,  $\mathcal{C} \subset \mathcal{A}$  is a  $\sigma$ -algebra, and  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  is a random variable.

The definition of a conditional expectation given a  $\sigma$ -algebra is already found in Kolmogorov (1933/1977) (see also Kolmogorov, 1956). Reading the following definition, remember that, for a random variable  $V: (\Omega, \mathcal{A}, P) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$ , we use  $\sigma(V) = V^{-1}(\bar{\mathcal{B}})$  to denote the  $\sigma$ -algebra generated by  $V$ , and  $\sigma(V) \subset \mathcal{C}$  means that  $V$  is  $(\mathcal{C}, \bar{\mathcal{B}})$ -measurable (see Def. 2.26 and Cor. 2.28).

**Definition 10.2 (Conditional Expectation Given a  $\sigma$ -Algebra)**

Let the assumptions 10.1 hold. A random variable  $V: (\Omega, \mathcal{A}, P) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  is called a (version of the)  $\mathcal{C}$ -conditional expectation of  $Y$  with respect to  $P$ , if the following two conditions hold:

- (a)  $\sigma(V) \subset \mathcal{C}$ .
- (b)  $E(1_C \cdot V) = E(1_C \cdot Y), \quad \forall C \in \mathcal{C}$ .

If  $V$  satisfies (a) and (b), then we also use the notation  $E(Y|\mathcal{C}) := V$ .

**Remark 10.3 ( $X$ -Conditional Expectation)** If the assumptions 10.1 hold, then we define

$$E(Y|X) := E[Y|\sigma(X)] \quad (10.1)$$

and call it a version of the  $X$ -conditional expectation of  $Y$  with respect to  $P$ . If there is no ambiguity, we will omit the reference to the measure  $P$ . Otherwise, we also use the terms  $X$ -conditional  $P$ -expectation of  $Y$ .  $\triangleleft$

**Remark 10.4 (Conditional Probability Given a  $\sigma$ -Algebra)** Let the assumptions 10.1 hold and let  $A \in \mathcal{A}$ . Then we define

$$P(A|\mathcal{C}) := E(1_A|\mathcal{C}) \quad (10.2)$$

and call it a version of the  $\mathcal{C}$ -conditional probability of  $A$  with respect to  $P$ . Similarly, we define

$$P(A|X) := E[1_A|\sigma(X)] \quad (10.3)$$

and call it a version of the  $X$ -conditional probability of  $A$  with respect to  $P$ .

Furthermore, considering the event  $\{Y=y\}$ , we also use the notation

$$P(Y=y|X) := P(\{Y=y\}|X) = E(1_{Y=y}|X). \quad (10.4)$$

$\triangleleft$

**Remark 10.5 (Conditioning on the Smallest  $\sigma$ -Algebra)** If  $\mathcal{C} = \{\Omega, \emptyset\}$ , then Definition 10.2 (a) implies that  $E(Y|\mathcal{C})$  is a constant (see Example 2.14), and 10.2 (b) implies

$$E(Y|\mathcal{C}) = E(Y),$$

because  $E[1_\Omega \cdot E(Y)] = E(Y) = E(1_\Omega \cdot Y)$  and  $E[1_\emptyset \cdot E(Y)] = 0 = E(1_\emptyset \cdot Y)$ . In fact, if  $\mathcal{C} = \{\Omega, \emptyset\}$ , then  $E(Y)$  is the only version of the  $\mathcal{C}$ -conditional expectation of  $Y$ . Correspondingly, if  $X$  is a constant, i. e., if  $X = \alpha, \alpha \in \Omega'_X$ , then

$$E(Y|X) = E(Y).$$

$\triangleleft$

**Remark 10.6 (C-Conditional Expectation of a  $\mathcal{C}$ -Conditional Expectation)**

Consider an event  $C \in \mathcal{C}$  with  $P(C) > 0$  and a version  $V$  of the  $\mathcal{C}$ -conditional expectation of  $Y$  defined in Definition 10.2. Then the  $C$ -conditional expectation value of  $V$  (see Def. 9.2) is

$$\begin{aligned} E(V|C) &= \frac{1}{P(C)} \cdot E(1_C \cdot V) && [(9.7)] \\ &= \frac{1}{P(C)} \cdot E(1_C \cdot Y) && [\text{Def. 10.2 (b)}] \\ &= E(Y|C). && [(9.7)] \end{aligned} \tag{10.5}$$

Inserting  $E(Y|\mathcal{C}) = V$  into this equation shows that with condition (b) of Definition 10.2 we implicitly require

$$E[E(Y|\mathcal{C})|C] = E(Y|C), \quad \forall C \in \mathcal{C} \text{ with } P(C) > 0. \tag{10.6}$$

&lt;

**Remark 10.7 (Multivariate  $X$ )** If  $X = (X_1, \dots, X_n)$  is an  $n$ -variate random variable on  $(\Omega, \mathcal{A}, P)$  (see section 5.3), then  $E(Y|X)$  is also denoted by  $E(Y|X_1, \dots, X_n)$ . <

**Remark 10.8 ( $\mathcal{C}$ -Conditional and  $X$ -Conditional Expectation)** If  $X: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  is a *nonnegative* real-valued random variable, then  $\sigma(X^2) = \sigma(X)$  [see Example 2.56 (i)]. Therefore, in this case  $E(Y|X)$  and  $E(Y|X^2)$  are just two different notations of the conditional expectation  $E[Y|\sigma(X)]$ . If  $X$  takes on also negative real numbers, then  $\sigma(X^2) \subset \sigma(X)$ , but  $\sigma(X^2) = \sigma(X)$  does not necessarily hold [see Example 2.56 (ii)]. <

**10.2 Existence and Uniqueness**

By its definition it is not obvious that a conditional expectation exists and that it is well-defined. These issues are addressed in the following theorem.

**Theorem 10.9 (Existence and Uniqueness of a  $\mathcal{C}$ -Conditional Expectation)**

Let the assumptions 10.1 hold. Then:

- (i) There is a  $\mathcal{C}$ -measurable random variable  $V: (\Omega, \mathcal{A}, P) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  that is nonnegative (if  $Y$  is nonnegative) or has a finite expectation  $E(V)$  (if  $E(Y)$  is finite) satisfying

$$E(1_C \cdot V) = E(1_C \cdot Y), \quad \forall C \in \mathcal{C}. \tag{10.7}$$

- (ii) If  $V, V^*: (\Omega, \mathcal{A}, P) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  satisfy (10.7) and  $\sigma(V), \sigma(V^*) \subset \mathcal{C}$ , then  $V \stackrel{\overline{P}}{=} V^*$ .

For a proof see Bauer (1996, Theorem 15.1, p. 111). Using the term ‘version of a conditional expectation’ already hints at the fact that a conditional expectation, even if it exists, is not necessarily uniquely defined. However, according to Theorem 10.9 (ii), different versions of a conditional expectations are  $P$ -equivalent (see Remark 5.17).

**Remark 10.10 (The Sets  $\mathcal{E}(Y|\mathcal{C})$  and  $\mathcal{E}(Y|X)$ )** We define  $\mathcal{E}(Y|\mathcal{C})$  to be the set of all  $\mathcal{C}$ -measurable random variables satisfying Equation (10.7). Hence,  $\mathcal{E}(Y|\mathcal{C})$  is the set of all versions of the  $\mathcal{C}$ -conditional expectation of  $Y$  with respect to the measure  $P$ . Similarly,  $\mathcal{E}(Y|X)$  denotes the set of all versions of the  $X$ -conditional expectation of  $Y$ . The sets  $\mathcal{P}(A|\mathcal{C})$  and  $\mathcal{P}(A|X)$  are defined correspondingly for conditional probabilities.  $\triangleleft$

**Remark 10.11 (Consistency of Definitions)** If  $X$  is discrete such that there is a finite or countable set  $\Omega'_0 \subset \Omega'_X$  with  $P_X(\Omega'_0) = 1$  and  $P(X=x) > 0$  for all  $x \in \Omega'_0$ , then the discrete conditional expectation introduced in Definition 9.14 is a version of the conditional expectation of  $Y$  given  $X$  defined in Remark 10.3 (see Exercises 10-1 and 10-2). This implies: For all versions  $E(Y|X) \in \mathcal{E}(Y|X)$ ,

$$E(Y|X) \stackrel{P}{=} \sum_{x \in \Omega'_0} E(Y|X=x) \cdot 1_{X=x} \quad (10.8)$$

and

$$\forall x \in \Omega'_0: E(Y|X)(\omega) = E(Y|X=x), \quad \text{if } \omega \in \{X=x\}. \quad (10.9)$$

This equation shows that the conditional expectation  $E(Y|X)$  describes how the  $(X=x)$ -conditional expectation values of  $Y$  depend on the values  $x \in \Omega'_0$ .  $\triangleleft$

**Remark 10.12 (Uniqueness of  $E(Y|X)$ )** If we add  $\Omega'_0 = X(\Omega)$  to the assumptions of Remark 10.11, then

$$E(Y|X) = \sum_{x \in X(\Omega)} E(Y|X=x) \cdot 1_{X=x} \quad (10.10)$$

Hence, under these assumptions,  $V = V^*$  for all  $V, V^* \in \mathcal{E}(Y|X)$ , i. e., there is only one single version of the  $X$ -conditional expectation of  $Y$ . The corresponding results for  $E(Y|\mathcal{C})$  are formulated in the following section.  $\triangleleft$

### 10.2.1 Uniqueness With Respect to a Probability Measure

**Remark 10.13 (Uniqueness of  $E(Y|\mathcal{C})$  With Respect to a Probability Measure)**

Let the assumptions 10.1 hold and let  $Q$  be a probability measure on  $(\Omega, \mathcal{A})$ . Then we define

$$E(Y|\mathcal{C}) \text{ is } Q\text{-unique} \quad :\Leftrightarrow \quad \forall V, V^* \in \mathcal{E}(Y|\mathcal{C}): V \stackrel{Q}{=} V^*. \quad (10.11)$$

This term is convenient not only for  $Q = P$ . According to the following remark  $E(Y|\mathcal{C})$  is  $P$ -unique.  $\triangleleft$

**Remark 10.14 ( $\mathcal{E}(Y|\mathcal{C})$  is a  $P$ -Equivalence Class)** Let  $V \in \mathcal{E}(Y|\mathcal{C})$  and suppose that the random variable  $V^*: (\Omega, \mathcal{A}, P) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  is  $\mathcal{C}$ -measurable with  $V \stackrel{P}{=} V^*$ . Then, according to Theorem 3.68 (i), Equation (10.7) also holds for  $V^*$ , and this implies  $V^* \in \mathcal{E}(Y|\mathcal{C})$ . Hence, if  $V \in \mathcal{E}(Y|\mathcal{C})$ , then Theorem 10.9 (ii) implies:

$$V^* \in \mathcal{E}(Y|\mathcal{C}) \Leftrightarrow V^* \stackrel{P}{=} V \text{ and } V^* \text{ is } \mathcal{C}\text{-measurable.} \quad (10.12)$$

Therefore,  $\mathcal{E}(Y|\mathcal{C})$  is the  $P$ -equivalence class of  $V$  in the set of all  $\mathcal{C}$ -measurable random variables (see Def. 2.74).  $\triangleleft$

**Remark 10.15 ( $P$ -Equivalence and  $\mathcal{C}$ -Measurability)** Suppose that  $V$  is a version of the  $\mathcal{C}$ -conditional expectation, i. e.,  $V \in \mathcal{E}(Y|\mathcal{C})$ , and that  $V^*: (\Omega, \mathcal{A}, P) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  is a random variable. Then (10.12) implies

$$V^* \in \mathcal{E}(Y|\mathcal{C}) \Rightarrow V^* \stackrel{P}{=} V.$$

However,  $V^* \stackrel{P}{=} E(Y|\mathcal{C})$  may be true and yet  $V^* \notin \mathcal{E}(Y|\mathcal{C})$ , because  $V^* \stackrel{P}{=} E(Y|\mathcal{C})$  does not imply that  $V^*$  is  $\mathcal{C}$ -measurable.  $\triangleleft$

**Remark 10.16 (Versions of  $E(Y|\mathcal{C})$ )** For simplicity, we also say that  $V$  is a *version* of  $E(Y|\mathcal{C})$ , meaning  $V \in \mathcal{E}(Y|\mathcal{C})$ .  $\triangleleft$

### 10.2.2 A Necessary and Sufficient Condition of Uniqueness

Now we present a necessary and sufficient condition for uniqueness of a conditional expectation for special  $\sigma$ -algebras  $\mathcal{C}$ . We consider a finite or countable partition of  $\Omega$ , i. e., a finite or countable set  $\mathcal{E}$  of pairwise disjoint nonempty subsets of  $\Omega$  with  $\bigcup_{A_i \in \mathcal{E}} A_i = \Omega$ .

**Theorem 10.17 (Uniqueness of  $E(Y|\mathcal{C})$ )**

Let the assumptions 10.1 hold and let  $\mathcal{C} = \sigma(\mathcal{E})$ , where  $\mathcal{E} = \{A_1, A_2, \dots\}$  is a finite or countable partition of  $\Omega$ . Then  $V = V^*$  for all  $V, V^* \in \mathcal{E}(Y|\mathcal{C})$  if and only if

$$P(A_i) > 0, \quad \forall A_i \in \mathcal{E}. \quad (10.13)$$

(Proof p. 318)

**Remark 10.18 (Values of a  $\mathcal{C}$ -Conditional Expectation)** Under the assumptions of Theorem 10.17,  $E(Y|\mathcal{C})$  is uniquely defined if and only if (10.13) holds, and in this case we can write

$$E(Y|\mathcal{C}) = \sum_{A_i \in \mathcal{E}} E(Y|A_i) \cdot 1_{A_i} \quad (10.14)$$

and

$$\forall A_i \in \mathcal{E}: \quad E(Y|\mathcal{C})(\omega) = E(Y|A_i), \quad \text{if } \omega \in A_i. \quad (10.15)$$

This equation shows that the conditional expectation  $E(Y|\mathcal{C})$  describes how the conditional expectation values  $E(Y|A_i)$  depend on the events  $A_i \in \mathcal{E}$  (see Exercise 10-3).  $\triangleleft$

### 10.2.3 Examples

**Example 10.19 (No Treatment for Joe – continued)** In Table 9.2 (p. 289), the conditional expectation  $E(Y|X)$  of the outcome variable  $Y$  on the treatment variable  $X$  has only two different values, the conditional expectations

$$E(Y|X=0) = .60 \quad \text{and} \quad E(Y|X=1) = .40.$$

The last but one column of Table 9.2 shows how these values are assigned to the eight possible outcomes  $\omega \in \Omega$ . The values and  $E(Y|X)$  itself are uniquely defined.

In contrast, the conditional expectation  $E(Y|X, U)$  of  $Y$  on the treatment variable  $X$  and the person variable  $U$  has four different values, .696, .20, .40, and 0 (see Example 9.23). Note that these four values define only one element, say  $V$ , of  $\mathcal{E}(Y|X, U)$ . If, instead of  $E(Y|X, U)(\omega) = 0$  for  $\omega \in \{(\text{Joe}, \text{yes}, -), (\text{Joe}, \text{yes}, +)\}$ , we define

$$E(Y|X, U)(\omega) = \alpha, \quad \alpha \neq 0, \quad \alpha \in \mathbb{R}, \quad \text{for } \omega \in \{(\text{Joe}, \text{yes}, -), (\text{Joe}, \text{yes}, +)\},$$

then we have a new element, say  $V^*$ , of  $\mathcal{E}(Y|X, U)$ . Because  $\alpha$  can be any zero real number, in this example, the set  $\mathcal{E}(Y|X, U)$  is uncountably infinite. However, because  $P(X=1, U=\text{Joe}) = 0$ , two elements  $V$  and  $V^*$  of  $\mathcal{E}(Y|X, U)$  are always identical with probability 1, i. e.,  $V$  and  $V^*$  are  $P$ -equivalent. Also note that  $E(Y|X)$  and  $E(Y|X, U)$  are random variables on the same probability space as the other random variables such as  $Y$ ,  $X$ , and  $U$ .  $\triangleleft$

**Example 10.20 (No Treatment for Joe – continued)** In Example 9.23, we specified the conditional expectations  $E(Y|X)$ ,  $E(Y|X, U)$ , and  $P(X=1|U)$ . Now we check whether  $E(Y|X)$  satisfies conditions (a) and (b) of Definition 10.2. First of all,

$$\mathcal{C} = \{\Omega, \emptyset, \{X=0\}, \{X=1\}\} = \sigma(X)$$

is the  $\sigma$ -algebra generated by  $X$ , where, in this example,  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  with  $\Omega'_X = \{0, 1\}$  and  $\mathcal{A}'_X = \{\Omega'_X, \emptyset, \{0\}, \{1\}\}$ . If  $V \in \mathcal{E}(Y|X)$  and  $\overline{\mathcal{B}}$  is the Borel  $\sigma$ -algebra on  $\overline{\mathbb{R}}$ , then

$$V^{-1}(B) = \begin{cases} \Omega, & \text{if } .40 \in B \text{ and } .60 \in B, \\ \emptyset, & \text{if } .40 \notin B \text{ and } .60 \notin B, \\ \{X=0\}, & \text{if } .40 \notin B \text{ and } .60 \in B, \\ \{X=1\}, & \text{if } .40 \in B \text{ and } .60 \notin B, \end{cases} \quad \forall B \in \overline{\mathcal{B}}.$$

Hence, in this example,  $V^{-1}(\overline{\mathcal{B}}) = \mathcal{C} = \sigma(X)$ . Therefore, condition (a) of Definition 10.2 is satisfied for  $V = E(Y|X)$  specified in Example 9.23.

Now we check condition (b) of Definition 10.2. For  $C = \Omega$  this condition requires  $E(1_\Omega \cdot V) = E(1_\Omega \cdot Y)$ . The expectation of  $1_\Omega \cdot Y$  is

$$E(1_\Omega \cdot Y) = E(Y) = P(Y=1) = .348 + 0 + .024 + .152 = .524,$$

and the expectation of  $1_{\Omega} \cdot V$  is:

$$\begin{aligned} E(1_{\Omega} \cdot V) &= E(V) = .60 \cdot P(X=0) + .40 \cdot P(X=1) \\ &= .60 \cdot (.152 + .348 + .096 + .024) + .40 \cdot (0 + 0 + .228 + .152) = .524 \end{aligned}$$

(see Table 9.2, p. 289). For  $C = \emptyset$ , condition (b) of Definition 10.2 requires that  $E(1_{\emptyset} \cdot V) = E(1_{\emptyset} \cdot Y)$ , which is always true [see Box 6.1 (v)]. For  $C = \{X=0\}$ , condition (b) of Definition 10.2 requires  $E(1_{X=0} \cdot V) = E(1_{X=0} \cdot Y)$ , and the expectations of  $1_{X=0} \cdot Y$  and  $1_{X=0} \cdot V$  are

$$E(1_{X=0} \cdot Y) = E(1_{X=0} \cdot 1_{Y=1}) = P(X=0, Y=1) = .348 + .024 = .372$$

and

$$\begin{aligned} E(1_{X=0} \cdot V) &= .60 \cdot P(1_{X=0} = 1, X=0) + 0 \cdot P(1_{X=0} = 1, X=1) \\ &= .60 \cdot P(X=0) = .60 \cdot (.152 + .348 + .096 + .024) = .372. \end{aligned}$$

Note that the random variable  $1_{X=0} \cdot E(Y|X)$  has two values, .60 and 0 (see Table 9.2).

Finally, for  $C = \{X=1\}$ , condition (b) of Definition 10.2 requires that  $E(1_{X=1} \cdot V) = E(1_{X=1} \cdot Y)$  and the expectation of  $1_{X=1} \cdot Y$  and  $1_{X=1} \cdot V$  are

$$E(1_{X=1} \cdot Y) = E(1_{X=1} \cdot 1_{Y=1}) = P(X=1, Y=1) = 0 + .152 = .152$$

and

$$\begin{aligned} E(1_{X=1} \cdot V) &= 0 \cdot P(1_{X=1} = 1, X=0) + .40 \cdot P(1_{X=1} = 1, X=1) \\ &= .40 \cdot P(X=1) = .40 \cdot (0 + 0 + .228 + .152) = .152 \end{aligned}$$

(see Table 9.2). Hence,  $V = E(Y|X)$  satisfies conditions (a) and (b) of Definition 10.2. Therefore,  $V = E(Y|X)$  is in fact an element of  $\mathcal{E}(Y|X) = \mathcal{E}[Y|\sigma(X)]$ . As mentioned before, in this example,  $E(Y|X)$  is uniquely defined. This means that it is the only element of the set  $\mathcal{E}(Y|X)$ .  $\triangleleft$

## 10.3 Rules of Computation and Other Properties

### 10.3.1 Rules of Computation

Some rules of computation for  $\mathcal{C}$ -conditional expectations are presented in Box 10.1, some of which are analog to the rules for expectations (see Box 6.1) and to the rules for  $B$ -conditional expectation values (see Box 9.1). In Rule (iv), the term  $E[E(Y|\mathcal{C})]$  denotes the expectation (with respect to the measure  $P$ ) of a  $\mathcal{C}$ -conditional expectation of  $Y$ . Similarly, in Rule (v), the term  $E[E(Y|\mathcal{C})|\mathcal{C}_0]$  denotes the  $\mathcal{C}_0$ -conditional expectation of the  $\mathcal{C}$ -conditional expectation of  $Y$ , where we presume  $\mathcal{C}_0 \subset \mathcal{C}$ . (For a proof of these rules see Exercise 10-4).

For convenience, in Box 10.2 we translate these rules to  $X$ -conditional expectations, i. e., to the case in which  $\mathcal{C} = \sigma(X)$ . Hence, these properties are special cases of those listed in Box 10.1, and they do not need proofs of their own.

In some of these rules we refer to the composition  $f(X) = f \circ X$ , where the function  $f: (\Omega'_X, \mathcal{A}'_X) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  is assumed to be  $(\mathcal{A}'_X, \bar{\mathcal{B}})$ -measurable. Remember, according Lemma 2.52, the composition  $f(X)$  is measurable with respect to the  $\sigma$ -algebra  $\mathcal{C} = X^{-1}(\mathcal{A}'_X) = \sigma(X)$  if  $f: \Omega'_X \rightarrow \bar{\mathbb{R}}$  is  $(\mathcal{A}'_X, \bar{\mathcal{B}})$ -measurable. Furthermore, according Corollary 2.53, the composition  $f(X)$  is measurable with respect to the  $\sigma$ -algebra  $\sigma(X)$  if  $f: \Omega'_X \rightarrow \Omega'$ , where  $\Omega'$  is finite or countable but not necessarily a subset of  $\bar{\mathbb{R}}$  and  $f$  is  $(\mathcal{A}'_X, \mathcal{P}(\Omega'))$ -measurable. According to Equation (2.29),

$$\begin{aligned} \sigma[f(X)] &= [f(X)]^{-1}(\bar{\mathcal{B}}) = X^{-1}[f^{-1}(\bar{\mathcal{B}})] \\ &\subset X^{-1}(\mathcal{A}'_X) = \sigma(X). \end{aligned}$$

In other words, we assume that the composition  $f(X)$  is measurable with respect to the  $\sigma$ -algebra  $\mathcal{C} = X^{-1}(\mathcal{A}'_X) = \sigma(X)$ . Therefore,  $\sigma[f(X)]$  can take the role of  $\mathcal{C}_0$  in Rule (v) of Box 10.1. Furthermore,  $f(X)$  takes the role of  $Y$  in Rule (vii) of Box 10.1, and the role of  $Y_1$  in Rules (xiii) and (xiv) of Box 10.1.

### 10.3.2 Monotonicity

Box 10.3 displays some monotonicity properties that are proved in Exercise 10-5. Of course, if  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  is a random variable, then these properties also hold for  $\mathcal{C} = \sigma(X)$  and  $E(Y|X) = E[Y|\sigma(X)]$ . For example, Rule (ix) can then be written

$$Y \underset{P}{\leq} Z \quad \Rightarrow \quad E(Y|X) \underset{P}{\leq} E(Z|X) \quad (10.16)$$

and Rule (v) yields

$$Y \underset{P}{\geq} \alpha \quad \Rightarrow \quad \exists V \in \mathcal{E}(Y|X): V \geq \alpha. \quad (10.17)$$

### 10.3.3 Convergence Theorems

Now we turn to convergence of  $\mathcal{C}$ -conditional expectations. Theorems 10.21 and 10.22 provide sufficient conditions that allow to exchange taking the limit and taking the conditional expectation, e. g.,

$$\lim_{i \rightarrow \infty} E(Y_i | \mathcal{C}) \underset{P}{=} E(\lim_{i \rightarrow \infty} Y_i | \mathcal{C}).$$

This is not only of technical interest for many proofs; instead, it also describes a continuity property of the conditional expectation: If  $Y_i$  is a good approximation to  $Y$ , then  $E(Y_i | \mathcal{C})$  is a good approximation to  $E(Y | \mathcal{C})$ .

The first theorem deals with *monotone convergence*, the second with *dominated convergence*. Reading these theorems, note that  $\lim_{i \rightarrow \infty} Y_i \underset{P}{=} Y$  means

$$P\left\{\{\omega \in \Omega: \lim_{i \rightarrow \infty} Y_i(\omega) = Y(\omega)\}\right\} = 1, \quad (10.18)$$

i. e., the sequence  $(Y_i, i \in \mathbb{N})$ , converges  $P$ -almost surely pointwise to  $Y$ .

**Box 10.1 Rules of Computation for  $\mathcal{C}$ -Conditional Expectations**

Let  $Y, Y_1, Y_2 : (\Omega, \mathcal{A}, P) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  be numerical random variables that are nonnegative or have a finite expectation,  $\mathcal{C} \subset \mathcal{A}$  a  $\sigma$ -algebra, and  $\alpha \in \mathbb{R}$ . Then :

$$E(\alpha | \mathcal{C}) \stackrel{p}{=} \alpha. \quad (\text{i})$$

$$E(\alpha + Y | \mathcal{C}) \stackrel{p}{=} \alpha + E(Y | \mathcal{C}). \quad (\text{ii})$$

$$E(\alpha \cdot Y | \mathcal{C}) \stackrel{p}{=} \alpha \cdot E(Y | \mathcal{C}). \quad (\text{iii})$$

$$E[E(Y | \mathcal{C})] = E(Y). \quad (\text{iv})$$

$$E[E(Y | \mathcal{C}) | \mathcal{C}_0] \stackrel{p}{=} E(Y | \mathcal{C}_0), \quad \text{if } \mathcal{C}_0 \subset \mathcal{C} \text{ is a } \sigma\text{-algebra.} \quad (\text{v})$$

$$E(Y | \mathcal{C}) \stackrel{p}{=} E(Y), \quad \text{if } Y \perp\!\!\!\perp \mathcal{C}. \quad (\text{vi})$$

$$E(Y | \mathcal{C}) \stackrel{p}{=} Y, \quad \text{if } Y \text{ is } \mathcal{C}\text{-measurable.} \quad (\text{vii})$$

$$E(Y | \mathcal{C}) \stackrel{p}{=} E[Y | E(Y | \mathcal{C})]. \quad (\text{viii})$$

$$E(Y_1 | \mathcal{C}) \stackrel{p}{=} E(Y_2 | \mathcal{C}), \quad \text{if } Y_1 \stackrel{p}{=} Y_2. \quad (\text{ix})$$

$$E(Y) \text{ is finite} \Rightarrow \exists V \in \mathcal{E}(Y | \mathcal{C}) : V \text{ is real-valued.} \quad (\text{x})$$

$$E(Y^2) < \infty \Rightarrow E[E(Y | \mathcal{C})^2] < \infty. \quad (\text{xi})$$

$$\text{Cov}[Y, E(Y | \mathcal{C})] = \text{Var}[E(Y | \mathcal{C})], \quad \text{if } E(Y^2) < \infty. \quad (\text{xii})$$

$$\text{Cov}[Y_1, E(Y_2 | \mathcal{C})] = \text{Cov}(Y_1, Y_2), \quad \text{if } \sigma(Y_1) \subset \mathcal{C}, E(Y_1^2), E(Y_2^2) < \infty. \quad (\text{xiii})$$

Let  $E(Y_1^2), E(Y_2^2) < \infty$  or  $Y_1, Y_2$  nonnegative. Then  $\sigma(Y_1) \subset \mathcal{C}$  implies:

$$E(Y_1 \cdot Y_2 | \mathcal{C}) \stackrel{p}{=} Y_1 \cdot E(Y_2 | \mathcal{C}). \quad (\text{xiv})$$

If  $Y_1, Y_2$  are nonnegative or real-valued with finite expectations, then there is a nonnegative or real-valued version  $E(Y_1 | \mathcal{C}) \in \mathcal{E}(Y_1 | \mathcal{C})$  and a nonnegative or real-valued version  $E(Y_2 | \mathcal{C}) \in \mathcal{E}(Y_2 | \mathcal{C})$  such that

$$E(Y_1 + Y_2 | \mathcal{C}) \stackrel{p}{=} E(Y_1 | \mathcal{C}) + E(Y_2 | \mathcal{C}). \quad (\text{xv})$$

If  $Y_1, \dots, Y_n$  are real-valued with finite expectations and  $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ , then

$$E\left(\sum_{i=1}^n \alpha_i \cdot Y_i \mid \mathcal{C}\right) \stackrel{p}{=} \sum_{i=1}^n \alpha_i \cdot E(Y_i | \mathcal{C}). \quad (\text{xvi})$$

**Box 10.2 Rules of Computation for  $X$ -Conditional Expectations**

Let  $Y, Y_1, Y_2: (\Omega, \mathcal{A}, P) \rightarrow (\bar{\mathbb{R}}, \mathcal{B})$  be numerical random variables that are non-negative or have a finite expectation, let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  be a random variable, let  $f: (\Omega'_X, \mathcal{A}'_X) \rightarrow (\bar{\mathbb{R}}, \mathcal{B})$  be measurable, and let  $\alpha \in \mathbb{R}$ . Then:

$$E(\alpha | X) \stackrel{\text{p}}{=} \alpha. \quad (\text{i})$$

$$E(\alpha + Y | X) \stackrel{\text{p}}{=} \alpha + E(Y | X). \quad (\text{ii})$$

$$E(\alpha \cdot Y | X) \stackrel{\text{p}}{=} \alpha \cdot E(Y | X). \quad (\text{iii})$$

$$E[E(Y | X)] = E(Y). \quad (\text{iv})$$

$$E[E(Y | X) | f(X)] \stackrel{\text{p}}{=} E[Y | f(X)]. \quad (\text{v})$$

$$E(Y | X) \stackrel{\text{p}}{=} E(Y), \quad \text{if } Y \perp\!\!\!\perp X. \quad (\text{vi})$$

$$E[f(X) | X] \stackrel{\text{p}}{=} f(X), \quad \text{if } f(X) \geq 0 \text{ or } E[f(X)] < \infty. \quad (\text{vii})$$

$$E(Y | X) \stackrel{\text{p}}{=} E[Y | E(Y | X)]. \quad (\text{viii})$$

$$E(Y_1 | X) \stackrel{\text{p}}{=} E(Y_2 | X), \quad \text{if } Y_1 \stackrel{\text{p}}{=} Y_2. \quad (\text{ix})$$

$$E(Y) \text{ is finite} \Rightarrow \exists V \in \mathcal{E}(Y | X): V \text{ is real-valued.} \quad (\text{x})$$

$$E(Y^2) < \infty \Rightarrow E[E(Y | X)^2] < \infty. \quad (\text{xi})$$

$$\text{Cov}[Y, E(Y | X)] = \text{Var}[E(Y | X)], \quad \text{if } E(Y^2) < \infty. \quad (\text{xii})$$

$$\text{Cov}[f(X), E(Y | X)] = \text{Cov}[f(X), Y], \quad \text{if } E[f(X)^2], E(Y^2) < \infty. \quad (\text{xiii})$$

Let  $E(Y^2), E[f(X)^2] < \infty$  or  $Y, f(X)$  nonnegative. Then:

$$E[f(X) \cdot Y | X] \stackrel{\text{p}}{=} f(X) \cdot E(Y | X). \quad (\text{xiv})$$

If  $Y_1, Y_2$  are nonnegative or real-valued with finite expectations, then there is a nonnegative or real-valued version  $E(Y_1 | X) \in \mathcal{E}(Y_1 | X)$  and a nonnegative or real-valued version  $E(Y_2 | X) \in \mathcal{E}(Y_2 | X)$  such that

$$E(Y_1 + Y_2 | X) \stackrel{\text{p}}{=} E(Y_1 | X) + E(Y_2 | X). \quad (\text{xv})$$

If  $Y_1, \dots, Y_n$  are real-valued with finite expectations and  $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ , then

$$E\left(\sum_{i=1}^n \alpha_i \cdot Y_i \mid X\right) \stackrel{\text{p}}{=} \sum_{i=1}^n \alpha_i \cdot E(Y_i | X). \quad (\text{xvi})$$

**Box 10.3 Monotonicity of Conditional Expectations**

Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  be a numerical random variable that is nonnegative or with finite expectation,  $\mathcal{C} \subset \mathcal{A}$  be a  $\sigma$ -algebra, and  $\alpha \in \mathbb{R}$ . Then:

$$Y \geq_{\bar{P}} 0 \text{ and } E(Y) = 0 \Rightarrow E(Y|\mathcal{C}) \stackrel{\bar{P}}{=} 0. \quad (\text{i})$$

$$Y \geq_{\bar{P}} \alpha \Rightarrow E(Y|\mathcal{C}) \geq_{\bar{P}} \alpha. \quad (\text{ii})$$

$$Y \leq_{\bar{P}} \alpha \Rightarrow E(Y|\mathcal{C}) \leq_{\bar{P}} \alpha. \quad (\text{iii})$$

$$Y \stackrel{\bar{P}}{=} \alpha \Rightarrow E(Y|\mathcal{C}) \stackrel{\bar{P}}{=} \alpha. \quad (\text{iv})$$

$$Y \geq_{\bar{P}} \alpha \Rightarrow \exists V \in \mathcal{E}(Y|\mathcal{C}): V \geq \alpha. \quad (\text{v})$$

$$Y \stackrel{\bar{P}}{=} \alpha \Rightarrow \exists V \in \mathcal{E}(Y|\mathcal{C}): V = \alpha. \quad (\text{vi})$$

$$Y \leq_{\bar{P}} \alpha \Rightarrow \exists V \in \mathcal{E}(Y|\mathcal{C}): V \leq \alpha. \quad (\text{vii})$$

$$Y \stackrel{\bar{P}}{=} \alpha \Rightarrow E(1_{Y=\alpha}|\mathcal{C}) \stackrel{\bar{P}}{=} P(Y=\alpha|\mathcal{C}) \stackrel{\bar{P}}{=} 1. \quad (\text{viii})$$

Let  $Y, Z: (\Omega, \mathcal{A}, P) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  be numerical random variables that are both nonnegative or such that both are real-valued with finite expectations and let  $\mathcal{C} \subset \mathcal{A}$  be a  $\sigma$ -algebra. Then:

$$Y \leq_{\bar{P}} Z \Rightarrow E(Y|\mathcal{C}) \leq_{\bar{P}} E(Z|\mathcal{C}). \quad (\text{ix})$$

If  $A \in \mathcal{A}$  with  $P(A) = 0$ , then

$$E(1_A|\mathcal{C}) \stackrel{\bar{P}}{=} P(A|\mathcal{C}) \stackrel{\bar{P}}{=} 0. \quad (\text{x})$$

$$E(1_{A^c}|\mathcal{C}) \stackrel{\bar{P}}{=} 1 - E(1_A|\mathcal{C}) \stackrel{\bar{P}}{=} P(A^c|\mathcal{C}) \stackrel{\bar{P}}{=} 1. \quad (\text{xi})$$

**Theorem 10.21 (Monotone Convergence)**

Let  $Y, Y_i: (\Omega, \mathcal{A}, P) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$ ,  $i \in \mathbb{N}$ , be random variables.

- (i) If the sequence  $(Y_i, i \in \mathbb{N})$  is increasing with  $\lim_{i \rightarrow \infty} Y_i \stackrel{\bar{P}}{=} Y$ , and  $Y_i$  is nonnegative  $\forall i \in \mathbb{N}$ , then

$$\lim_{i \rightarrow \infty} E(Y_i|\mathcal{C}) \stackrel{\bar{P}}{=} E(Y|\mathcal{C}). \quad (10.19)$$

- (ii) If  $Y_i \geq 0$ ,  $\forall i \in \mathbb{N}$ , then

$$E\left(\sum_{i=1}^{\infty} Y_i|\mathcal{C}\right) \stackrel{\bar{P}}{=} \sum_{i=1}^{\infty} E(Y_i|\mathcal{C}). \quad (10.20)$$

(Proof p. 319)

If the sequence  $(Y_i, i \in \mathbb{N})$  is not increasing, then we need an additional assumption in order to guarantee convergence of the conditional expectations. A sufficient condition is that all  $|Y_i|$  are dominated by the same  $P$ -integrable function  $Z$ .

**Theorem 10.22 (Dominated Convergence)**

Let  $Y_i: (\Omega, \mathcal{A}, P) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ ,  $i \in \mathbb{N}$ , and  $Z: (\Omega, \mathcal{A}, P) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  be random variables. If

- (a)  $Z$  has a finite expectation and  $|Y_i| \leq Z$ ,  $\forall i \in \mathbb{N}$ ,
- (b)  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  is a random variable such that  $\lim_{i \rightarrow \infty} Y_i \stackrel{p}{=} Y$ ,

then

$$\lim_{i \rightarrow \infty} E(Y_i | \mathcal{C}) \stackrel{p}{=} E(Y | \mathcal{C}). \quad (10.21)$$

For a proof see Klenke (2013, Theorem 8.14 (viii), p. 174).

## 10.4 Factorization, Regression, and Conditional Expectation Value

As already mentioned in the Preface, much empirical research uses some kind of regression in order to investigate how the expectation of one random variable depends on the values of one or more other random variables. This applies to analysis of variance, regression analysis, the general linear model, the generalized linear model, factor analysis, structural equation models, hierarchical linear models, and analysis of qualitative data. Using these methods we aim at learning about specific regressions. A regression is a special case of a factorization of an  $X$ -conditional expectation, in which  $X$  is numerical. This concept is also used for a general definition of a conditional expectation value.

### 10.4.1 Existence of a Factorization

Lemma 2.52 can be applied to  $E(Y|X)$ , which, by definition, is measurable with respect to  $X$ . This immediately implies the following corollary:

**Corollary 10.23 (Existence of a Factorization of a Conditional Expectation)**

Let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  and  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  be random variables, where  $Y$  is nonnegative or with finite expectation. If  $E(Y|X) \in \mathcal{E}(Y|X)$ , then there is a measurable function  $g: (\Omega'_X, \mathcal{A}'_X) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  such that

$$E(Y|X) = g \circ X. \quad (10.22)$$

**Remark 10.24 (Notation)** Instead of  $g \circ X$  we also use the notation  $g(X)$ . Figure 9.1 displays the random variable  $X$ , a factorization  $g$ , and a version  $E(Y|X) = g(X) \in \mathcal{E}(Y|X)$ .  $\triangleleft$

**Definition 10.25 (Factorization and Regression)**

Under the assumptions of Corollary 10.23, the function  $g: \Omega'_X \rightarrow \bar{\mathbb{R}}$  is called a factorization of  $E(Y|X)$ . If  $(\Omega'_X, \mathcal{A}'_X) = (\bar{\mathbb{R}}^n, \mathcal{B}_n)$ , for  $n \in \mathbb{N}$ , then  $g$  is also called a regression of  $Y$  on  $X$ . In this case,  $Y$  is called the regressand and  $X$  the regressor.

Note that the definition of a regression does not refer to any parametric function. In contrast, these requirements were made in the definition of a linear quasi-regression (see 7.1). The relationship between the regression and the linear quasi-regression is considered in the following section.

### 10.4.2 Conditional Expectation and Mean-Squared Error

In Definition 7.2 we introduced the linear quasi-regression as that linear function  $f: \mathbb{R} \rightarrow \mathbb{R}$  defined by  $f(x) = \alpha_0 + \alpha_1 x$ ,  $x \in \mathbb{R}$ , that minimizes  $E([Y - (a_0 + a_1 X)]^2)$ , where  $X$  is a real-valued random variable. In a sense, the linear quasi-regression is a function  $f$  such that the composition  $Q_{lin}(Y|X) = f(X)$  is the best approximation of  $Y$  by a linear function of  $X$ . Now consider the approximation (with respect to the mean squared error) of  $Y$  by a more general function that is  $\mathcal{C}$ -measurable or  $X$ -measurable, respectively. Intuitively speaking, we ask for the best approximation of  $Y$  based on the information contained in  $\mathcal{C}$  or in  $X$ .

Reading the following theorem, note that the right-hand sides of (10.23) and (10.24) do not depend on the particular choice of a version  $E(Y|X) \in \mathcal{E}(Y|X)$  [see Th. 10.9 (ii) and Rule (viii) of Box 6.1].

**Theorem 10.26 (Conditional Expectation and Mean-Squared Error)**

Let  $Y, Z: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  be real-valued random variables with  $E(Y^2), E(Z^2) < \infty$ , let  $\mathcal{C} \subset \mathcal{A}$  be a  $\sigma$ -algebra, and suppose that  $Z$  is  $\mathcal{C}$ -measurable. Then

$$E[(Y - Z)^2] \geq E[(Y - E(Y|\mathcal{C}))^2]. \quad (10.23)$$

and

$$Z \stackrel{\bar{P}}{=} E(Y|\mathcal{C}) \Leftrightarrow E[(Y - Z)^2] = E[(Y - E(Y|\mathcal{C}))^2]. \quad (10.24)$$

For a proof see Klenke (2013, Cor. 8.17, p. 176).

**Remark 10.27 (Regression vs. Linear Quasi-Regression)** If  $E(Y^2) < \infty$ , then Theorem 10.26 implies that  $V \in \mathcal{E}(Y|X)$  is an  $X$ -measurable random variable with

$E(V^2) < \infty$  minimizing the mean squared error  $E[(Y - Z)^2]$  for all  $X$ -measurable random variables  $Z$  with  $E(Z^2) < \infty$ . Vice versa, if  $V$  is an  $X$ -measurable random variable with finite second moment minimizing  $E[(Y - Z)^2]$  for all  $X$ -measurable random variables  $Z$  with  $E(Z^2) < \infty$ , then  $V \in \mathcal{E}(Y|X)$ , provided that  $E(Y^2) < \infty$ .

In contrast,  $Q_{lin}(Y|X)$  is a *linear function* of  $X$  minimizing  $E[(Y - (a_0 + a_1 X))^2]$ ,  $a_0, a_1 \in \mathbb{R}$ , provided that  $E(Y^2) < \infty$ . Hence,  $Q_{lin}(Y|X)$  is the best (with respect to the mean squared error) approximation of  $Y$  in the set of all linear functions of  $X$ , whereas  $E(Y|X)$  is the best approximation of  $Y$  in the set of all  $X$ -measurable functions, provided that the second moment of  $Y$  is finite.

According to Definition 7.2,  $Q_{lin}(Y|X) = f(X)$ , where  $f: \mathbb{R} \rightarrow \mathbb{R}$  defined by  $f(x) = \alpha_0 + \alpha_1 x$ ,  $x \in \mathbb{R}$ , is the linear quasi-regression of  $Y$  on  $X$ . In contrast, the regression of  $Y$  on  $X$  is a function  $g: \Omega'_X \rightarrow \bar{\mathbb{R}}$  such that there is a version  $E(Y|X) \in \mathcal{E}(Y|X)$  with  $E(Y|X) = g(X)$ . Even if we consider a real-valued random variable  $X$  such that  $\Omega'_X = \mathbb{R}$ , a regression  $g$  does not require that it is a linear function. If  $g$  is a linear function with domain  $\Omega'_X = \mathbb{R}$  and there is a version  $E(Y|X) \in \mathcal{E}(Y|X)$  with  $E(Y|X) = g(X)$ , then  $f = g$  and  $Q_{lin}(Y|X) = E(Y|X)$ .  $\triangleleft$

### 10.4.3 Uniqueness of a Factorization

A factorization of  $E(Y|X)$  or synonymously, a regression of  $Y$  on  $X$ , is not necessarily uniquely defined. This even applies if we consider a fixed version  $E(Y|X)$  of the conditional expectation.

**Remark 10.28 (Uniqueness of a Factorization)** For two elements  $V$  and  $V^*$  of  $\mathcal{E}(Y|X)$  there can be different factorizations  $g$  and  $g^*$  with  $V = g(X)$  and  $V^* = g^*(X)$ . This is true even if  $V = V^*$ . Hence, there can be different factorizations of a single element  $V \in \mathcal{E}(Y|X)$  (see Example 10.32). In other words,  $V = g(X) = g^*(X)$ , with  $g \neq g^*$  is not necessarily contradictory. In this case  $g(x) = g^*(x)$  for all  $x \in X(\Omega)$ , whereas  $g(x) = g^*(x)$  does *not* hold for all  $x \in \Omega'_X$ . However, Theorem 10.9 (ii) and Corollary 5.25 (i) imply the following corollary:  $\triangleleft$

#### Corollary 10.29 ( $P_X$ -Equivalence of Factorizations)

Let the assumptions 10.1 hold and let  $g, g^*: (\Omega'_X, \mathcal{A}'_X) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  be  $(\mathcal{A}'_X, \bar{\mathcal{B}})$ -measurable functions. If  $g(X), g^*(X) \in \mathcal{E}(Y|X)$ , then

$$g \stackrel{P_X}{=} g^*. \quad (10.25)$$

**Remark 10.30 ( $P_X$ -Equivalence)** Note that, according to (5.12), Equation (10.25) is equivalent to

$$g(x) = g^*(x), \quad \text{for } P_X\text{-a.a. } x \in \Omega'_X. \quad (10.26)$$

$\triangleleft$

According to Remark 10.12,  $P(X=x) > 0$  for all  $x \in X(\Omega)$  implies that  $E(Y|X)$  is uniquely defined. According to the following corollary, this also applies to the factorization of  $E(Y|X)$  if we additionally assume  $\Omega'_X = X(\Omega)$ .

**Corollary 10.31 (Uniqueness of the Factorization)**

Let the assumptions 10.1 hold and assume  $P(X=x) > 0$  for all  $x \in \Omega'_X$ . Then the function  $g: \Omega'_X \rightarrow \bar{\mathbb{R}}$  satisfying  $E(Y|X) = g(X)$  is uniquely defined.

**Example 10.32 (No Treatment for Joe – continued)** In Example 9.23 we specified  $E(Y|X)$  with its two values

$$E(Y|X)(\omega) = E(Y|X=0) = .60, \quad \forall \omega \in \Omega \text{ with } X(\omega) = 0,$$

and

$$E(Y|X)(\omega) = E(Y|X=1) = .40, \quad \forall \omega \in \Omega \text{ with } X(\omega) = 1.$$

If we consider the treatment variable  $X: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$ , then  $g: \mathbb{R} \rightarrow \mathbb{R}$  defined by

$$g(x) = \begin{cases} .60, & \text{if } x = 0, \\ .40, & \text{if } x = 1, \\ \alpha, & \text{otherwise,} \end{cases}$$

is a factorization of the conditional expectation of  $Y$  given  $X$  for any choice of  $\alpha \in \mathbb{R}$ . This implies that there are different factorizations  $g, g^*$  with  $g(X), g^*(X) \in \mathcal{E}(Y|X)$ . However,  $g$  and  $g^*$  are  $P_X$ -equivalent. In contrast, if we consider  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$ , with  $\Omega'_X = X(\Omega) = \{0, 1\}$  and  $\mathcal{A}'_X = \mathcal{P}(\Omega'_X)$ , then  $P(X=x) > 0$  for all  $x \in \Omega'_X$  and there is only one single factorization  $g: \Omega'_X \rightarrow \mathbb{R}$  with the two values  $g(0) = .60$  and  $g(1) = .40$  (see Cor. 10.31).  $\triangleleft$

**10.4.4 Conditional Expectation Value**

The concepts of a conditional expectation value  $E(Y|X=x)$  and a conditional probability  $P(A|X=x)$  have been introduced in Definition 9.2 and Remark 9.7 only for  $P(X=x) > 0$ . Now we drop this assumption and define these concepts more generally, again using the factorization of a conditional expectation.

**Definition 10.33 (( $X=x$ )-Conditional Expectation Value)**

Let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  be a random variable, let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  be nonnegative or with finite expectation  $E(Y)$ , and let  $g: (\Omega'_X, \mathcal{A}'_X) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  be a function satisfying (10.22). Then the value  $g(x)$  of  $g$  is called an ( $X=x$ )-conditional expectation value of  $Y$  and is denoted by  $E(Y|X=x)$ , i. e.,

$$E(Y|X=x) := g(x). \quad (10.27)$$

**Remark 10.34 (( $X=x$ )-Conditional Probability)** If  $1_A$  is the indicator of  $A \in \mathcal{A}$ , then  $E(1_A|X=x)$  is also called an ( $X=x$ )-conditional probability of  $A$  and it is denoted by  $P(A|X=x)$ , i. e.,

$$P(A|X=x) := E(1_A|X=x). \quad (10.28)$$

Furthermore, considering the event  $\{Y=y\}$ , we also use the notation

$$P(Y=y|X=x) := P(\{Y=y\}|X=x) = E(1_{Y=y}|X=x). \quad (10.29)$$

◁

**Remark 10.35 (Uniqueness and Consistency of Definitions)** If  $P(X=x) > 0$ , then the  $(X=x)$ -conditional expectation value of  $Y$  is uniquely defined and it is identical to the term introduced in Definition 9.2, i. e.,

$$E(Y|X=x) = E^{X=x}(Y), \quad \text{if } P(X=x) > 0 \quad (10.30)$$

(see Exercise 10-6). In the general case,  $E(Y|X=x)$  is not uniquely defined. However,  $g$  is uniquely defined up to  $P_X$ -equivalence (see Cor. 10.29). ◁

**Remark 10.36 (Versions of a Conditional Expectation With a Discrete  $X$ )** Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  be a random variable that is nonnegative or with finite expectation and suppose that  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  is discrete implying that  $\Omega'_0 \subset \Omega'_X$  is finite or countable with  $P_X(\Omega'_0) = 1$  and  $\{x\} \in \mathcal{A}'_X$  for all  $x \in \Omega'_0$ . Then, for all  $\alpha \in \mathbb{R}$ , the function  $E(Y|X): \Omega \rightarrow \mathbb{R}$  defined by

$$E(Y|X)(\omega) := \begin{cases} E(Y|X=x), & \text{if } X(\omega) = x \text{ and } P(X=x) > 0 \\ \alpha, & \text{otherwise.} \end{cases} \quad (10.31)$$

is a version  $E(Y|X) \in \mathcal{E}(Y|X)$ . This proposition follows from Remark 10.35 and proposition (10.12). ◁

**Remark 10.37 (Values of the Conditional Expectation)** Assume that  $E(Y|X) = g(X) \in \mathcal{E}(Y|X)$ . Then

$$E(Y|X)(\omega) = g(x) = E(Y|X=x), \quad \forall \omega \in \Omega: X(\omega) = x. \quad (10.32)$$

(see Exercise 10-7). This also implies that the value of  $E(Y|X)$  is constant on all sets  $\{X=x\} = \{\omega \in \Omega: X(\omega) = x\}$ . Note that this also holds if  $\Omega$  is finite or countable and some  $\omega \in \{X=x\}$  have probability  $P(\{\omega\}) = 0$ . As an example, see  $E(Y|X)(\omega)$  for  $\omega \in \{(Joe, yes, -), (Joe, yes, +)\}$  in Table 9.2. These two values are equal to  $E(Y|X=1) = .40$ , although  $P(\{(Joe, yes, -)\}) = P(\{(Joe, yes, +)\}) = 0$ . ◁

**Remark 10.38 (Equivalent Propositions)** Let the assumptions 10.1 hold, let  $g(X) \in \mathcal{E}(Y|X)$  and  $g^*: (\Omega'_X, \mathcal{A}'_X) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  be an  $(\mathcal{A}'_X, \bar{\mathcal{B}})$ -measurable function. Then proposition (10.12) and Theorem 2.49 yield

$$g(X) \stackrel{p}{=} g^*(X) \Leftrightarrow g^*(X) \in \mathcal{E}(Y|X). \quad (10.33)$$

◁

Definition 10.33, Corollary 5.25, Remark 5.26, Corollary 10.29, and Equation (10.22) imply the following corollary, according to which we may either formulate propositions in terms of  $(X=x)$ -conditional expectation values or, equivalently, in terms of the corresponding conditional expectations.

**Corollary 10.39 (Equivalent Propositions)**

Let  $Y_1, Y_2: (\Omega, \mathcal{A}, P) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  be two numerical random variables that are nonnegative or with finite expectations and let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  be a random variable. Then:

(i)  $E(Y_1|X) \stackrel{\bar{P}}{=} E(Y_2|X)$  is equivalent to

$$E(Y_1|X=x) = E(Y_2|X=x), \quad \text{for } P_X\text{-a.a. } x \in \Omega'_X. \quad (10.34)$$

(ii)  $E(Y_1|X) \stackrel{\bar{P}}{>} E(Y_2|X)$  is equivalent to

$$E(Y_1|X=x) > E(Y_2|X=x), \quad \text{for } P_X\text{-a.a. } x \in \Omega'_X. \quad (10.35)$$

(iii)  $E(Y_1|X) \stackrel{\bar{P}}{\geq} E(Y_2|X)$  is equivalent to

$$E(Y_1|X=x) \geq E(Y_2|X=x), \quad \text{for } P_X\text{-a.a. } x \in \Omega'_X. \quad (10.36)$$

**Remark 10.40 ( $(X=x)$ -Conditional Expectation Value of  $E(Y|X)$ )** Suppose that  $f: (\Omega'_X, \mathcal{A}'_X) \rightarrow (\Omega'', \mathcal{A}'')$  is a measurable mapping,  $w \in \Omega''$ ,  $\{w\} \in \mathcal{A}''$ , and  $C = \{f(X)=w\}$ , then Equation (10.6) implies

$$\begin{aligned} E[E(Y|X) | f(X)=w] &= E[Y | f(X)=w], \\ &\text{if } w \in \Omega'' \text{ and } P[f(X)=w] > 0. \end{aligned} \quad (10.37)$$

As a special case, this equation yields

$$E[E(Y|X, Z) | X=x] = E(Y|X=x), \quad \text{if } x \in \Omega'_X \text{ with } P(X=x) > 0, \quad (10.38)$$

provided that  $Z: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Z, \mathcal{A}'_Z)$  is a random variable, too. If  $Z$  takes on only a finite number of values  $z_1, \dots, z_m$  and  $P(X=x, Z=z_i) > 0$ , for all  $i = 1, \dots, m$ , then Rule (ii) of Box 9.2 follows from Equations (10.38) and (9.21). Similarly, applying (9.20), Equation (10.38) yields Rule (iii) of Box 9.2 (see Exercise 10-8).  $\triangleleft$

Proposition (i) of Corollary 10.39 and Rule (v) of Box 10.2 yield:

$$E[E(Y|X) | f(X)=w] = E[Y | f(X)=w], \quad \text{for } P_{f(X)}\text{-a.a. } w \in \Omega'', \quad (10.39)$$

which generalizes Equation (10.37). A special case of Equation (10.39) is

$$E[E(Y|X, Z) | X=x] = E(Y|X=x), \quad \text{for } P_X\text{-a.a. } x \in \Omega'_X. \quad (10.40)$$

**Remark 10.41 (Expectation of a Conditional Expectation)** Applying Definition 10.33 and Equation (6.13), Rule (iv) of Box 10.2 is equivalent to

$$E(Y) = \int E(Y|X=x) P_X(dX). \quad (10.41)$$

Furthermore, if  $X$  only takes on a finite number of values  $x_1, \dots, x_n$ , then Equation (10.41) can be written

$$E(Y) = \sum_{i=1}^n E(Y|X=x_i) \cdot P(X=x_i) \quad (10.42)$$

[see Eq. (6.3)]. ◁

## 10.5 Characterizing a Conditional Expectation by the Joint Distribution

Using the factorization and Equation (3.59) yields two conditions that are equivalent to those occurring in Definition 10.2. In these conditions we refer to the joint distribution  $P_{X,Y}$  of  $X$  and  $Y$ .

### Theorem 10.42 (Conditional Expectation and the Joint Distribution)

Let the assumptions 10.1 hold and assume that  $g: (\Omega'_X, \mathcal{A}'_X) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  is measurable. Then  $g(X) \in \mathcal{E}(Y|X)$  if and only if

$$\int 1_{C'}(x) \cdot g(x) P_X(dx) = \int 1_{C'}(x) \cdot y P_{X,Y}[d(x,y)], \quad \forall C' \in \mathcal{A}'_X. \quad (10.43)$$

(Proof p. 319)

**Remark 10.43 (Two Alternative Formulations)** According to (3.28) and (3.59), Equation (10.43) is equivalent to

$$\int_{C'} g dP_X = \int_{\{X \in C'\}} Y dP, \quad \forall C' \in \mathcal{A}'_X. \quad (10.44)$$

If  $E_X(\cdot)$  denotes the expectation with respect to the distribution  $P_X$  and  $1_{X \in C'}$  denotes the indicator of  $\{X \in C'\}$ , then Equation (10.44) may also be written:

$$E_X(1_{C'} g) = E(1_{X \in C'} Y), \quad \forall C' \in \mathcal{A}'_X. \quad (10.45)$$

◁

### Remark 10.44 (Conditional Expectation With Respect to a Joint Distribution)

Note that in Equation (10.43) we do not explicitly refer to the measure  $P$ . Instead we refer to  $P_{X,Y}$ , the joint distribution of  $X$  and  $Y$ . Therefore  $g(X) \in \mathcal{E}(Y|X)$  may also be called a version of the *conditional expectation* of  $Y$  on  $X$  with respect to  $P_{X,Y}$ . This can be used, e. g., to consider a conditional expectation with respect to the  $(Z=z)$ -conditional distribution  $P_{X,Y|Z=z}$  (see Def. 17.7). ◁

## 10.6 Conditional Mean Independence

In section 5.4 we defined *independence* of two random variables  $X$  and  $Y$ . Furthermore, in chapter 7, introducing the covariance  $Cov(X, Y)$  and the correlation  $Corr(X, Y)$  of two numerical random variables, we also defined *uncorrelatedness* by  $Cov(X, Y) = 0$ . Now we add two other related concepts: *mean independence* and *conditional mean independence*.

### Definition 10.45 ( $\mathcal{C}$ -Conditional Mean Independence)

Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  be a numerical random variable that is nonnegative or has a finite expectation  $E(Y)$  and let  $\mathcal{D} \subset \mathcal{A}$  be a  $\sigma$ -algebra.

(i) Then  $Y$  is called *mean independent from  $\mathcal{D}$* , if

$$E(Y|\mathcal{D}) \stackrel{p}{=} E(Y). \quad (10.46)$$

(ii) Let also  $\mathcal{C} \subset \mathcal{A}$  be a  $\sigma$ -algebra and let  $E(Y|\mathcal{C}, \mathcal{D})$  denote the conditional expectation of  $Y$  given  $\sigma(\mathcal{C} \cup \mathcal{D})$ . Then  $Y$  is called  *$\mathcal{C}$ -conditionally mean independent from  $\mathcal{D}$* , if

$$E(Y|\mathcal{C}, \mathcal{D}) \stackrel{p}{=} E(Y|\mathcal{C}). \quad (10.47)$$

Analogously to  $E(Y|\mathcal{C}, \mathcal{D}) := E[Y|\sigma(\mathcal{C} \cup \mathcal{D})]$ , we use the notation  $P(A|\mathcal{C}, \mathcal{D}) := P[A|\sigma(\mathcal{C} \cup \mathcal{D})]$  for the  $\sigma(\mathcal{C} \cup \mathcal{D})$ -conditional probability of  $A \in \mathcal{A}$ .

**Remark 10.46 ( $X$ -Conditional Mean Independence)** Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  be a numerical random variable that is nonnegative or has a finite expectation  $E(Y)$  and let  $Z: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Z, \mathcal{A}'_Z)$  be a random variable.

(i) Then  $Y$  is called *mean independent from  $Z$*  if

$$E(Y|Z) \stackrel{p}{=} E(Y). \quad (10.48)$$

(ii) Let also  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  be a random variable. Then  $Y$  is called  *$X$ -conditionally mean independent from  $Z$*  if

$$E(Y|X, Z) \stackrel{p}{=} E(Y|X). \quad (10.49)$$

◁

**Remark 10.47 (A Special Case)** Of course, if  $\mathcal{D} \subset \mathcal{C}$ , then  $Y$  is  $\mathcal{C}$ -conditionally mean independent from  $\mathcal{D}$ . In this case,  $\sigma(\mathcal{C} \cup \mathcal{D}) = \sigma(\mathcal{C}) = \mathcal{C}$  and  $E(Y|\mathcal{C}, \mathcal{D})$  is just a different notation of  $E(Y|\mathcal{C})$ . Correspondingly, assume that  $Z$  is measurable with respect to  $X$ . Then  $\sigma(X, Z) = \sigma(X)$  and therefore,

$$E(Y|X, Z) \stackrel{p}{=} E(Y|X), \quad \text{if } \sigma(Z) \subset \sigma(X). \quad (10.50)$$

**Table 10.1.** Joe and Ann With no Treatment Effects

	$P(\{\omega\})$	Person variable $U$	Treatment variable $X$	Outcome variable $Y$	$E(Y X, U)$	$E(Y U)$	$E(Y X)$	$P(X=1 U)$
$(Joe, no, -)$	0.16	Joe	0	0	0.2	0.2	0.56	0.6
$(Joe, no, +)$	0.04	Joe	0	1	0.2	0.2	0.56	0.6
$(Joe, yes, -)$	0.24	Joe	1	0	0.2	0.2	0.44	0.6
$(Joe, yes, +)$	0.06	Joe	1	1	0.2	0.2	0.44	0.6
$(Ann, no, -)$	0.06	Ann	0	0	0.8	0.8	0.56	0.4
$(Ann, no, +)$	0.24	Ann	0	1	0.8	0.8	0.56	0.4
$(Ann, yes, -)$	0.04	Ann	1	0	0.8	0.8	0.44	0.4
$(Ann, yes, +)$	0.16	Ann	1	1	0.8	0.8	0.44	0.4

Hence,  $Y$  is  $X$ -conditionally mean independent from all random variables  $Z$  that are measurable with respect to  $X$ . In these cases  $Z$  does not carry any information additional to  $X$ . In more formal terms,  $Z$  does not represent any event that is not already represented by  $X$ , i. e.,  $\{Z \in A'\} \in \sigma(X)$ , for all  $A' \in \mathcal{A}'_Z$ .  $\triangleleft$

**Example 10.48 (Joe and Ann With no Treatment Effect)** Table 10.1 displays an example for  $U$ -conditional mean independence of  $Y$  from  $X$ , i. e.,

$$E(Y|X, U) \stackrel{\bar{P}}{=} E(Y|U).$$

The values of the conditional expectations  $E(Y|X, U)$  and  $E(Y|U)$  can be computed in the same way as in Example 9.21. This new example shows that  $E(Y|X, U) \stackrel{\bar{P}}{=} E(Y|U)$  does not imply  $E(Y|X) \stackrel{\bar{P}}{=} E(Y)$ . Hence, although  $E(Y|X) \stackrel{\bar{P}}{=} E(Y)$  *does not hold* and the conditional expectation values  $E(Y|X=x)$  *do depend* on the values  $x$  of  $X$ , in a sense, the treatment variable  $X$  is irrelevant once we condition on  $U$ . In other words, for *Joe* success does not depend on whether or not he receives treatment, and the same is true for *Ann* [see the column headed  $E(Y|X, U)$ ]. This example shows that the conditional expectation  $E(Y|X)$  can be completely misleading if used for the evaluation of the effect of the treatment variable  $X$  on the outcome variable  $Y$ .  $\triangleleft$

**Remark 10.49 (Implication Structure Among Different Kinds of Independence)** According to Rule (vi) of Box 10.1, independence of  $Y$  and  $\mathcal{C}$  implies that  $Y$  is mean independent from  $\mathcal{C}$ . Analogously, according to Rule (vi) of Box 10.2 independence of  $Y$  and  $X$  implies that  $Y$  is mean independent from  $X$ . Furthermore,

mean independence of  $Y$  from  $X$  implies that  $X$  and  $Y$  are uncorrelated, provided that  $X$  and  $Y$  are numerical and  $E(X^2), E(Y^2) < \infty$  (see Exercise 10-9). Hence, if  $E(X^2), E(Y^2) < \infty$ , then

$$Y \perp\!\!\!\perp_p X \Rightarrow E(Y|X) \stackrel{\bar{p}}{=} E(Y) \tag{10.51}$$

and

$$E(Y|X) \stackrel{\bar{p}}{=} E(Y) \Rightarrow \text{Corr}(X, Y) = \text{Cov}(X, Y) = 0. \tag{10.52}$$

<

**Remark 10.50 (Dichotomous  $Y$ )** If  $Y$  is dichotomous with values 0 and 1 (see Example 5.10), then

$$Y \perp\!\!\!\perp_p X \Leftrightarrow E(Y|X) \stackrel{\bar{p}}{=} E(Y), \tag{10.53}$$

(see Exercise 10-10), which can equivalently be written

$$Y \perp\!\!\!\perp_p X \Leftrightarrow P(Y=1|X) \stackrel{\bar{p}}{=} P(Y=1) \tag{10.54}$$

[see (10.4)].

If  $Y$  is dichotomous with values  $y_1$  and  $y_2$ , then

$$Y \perp\!\!\!\perp_p X \Leftrightarrow P(Y=y_1|X) \stackrel{\bar{p}}{=} P(Y=y_1), \tag{10.55}$$

because  $P(1_{Y=y_1}=1|X) = P(Y=y_1|X)$  and  $P(1_{Y=y_1}=1) = P(Y=y_1)$  [see again (10.4) and Def. 1.49]. <

Now we turn to conditions that are equivalent to conditional mean independence. We start with a theorem that only applies to a *nonnegative* numerical random variable  $Y$  that *also has a finite expectation*. A second theorem also applies to a numerical random variable  $Y$  with *finite second moment*.

**Theorem 10.51 ( $\mathcal{C}$ -Conditional Mean Independence if  $E(Y)$  is Finite)**

Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  be a nonnegative random variable that has a finite expectation  $E(Y)$  and let  $\mathcal{C}, \mathcal{D} \subset \mathcal{A}$  be  $\sigma$ -algebras. Then the following two propositions are equivalent to each other:

- (a)  $E(Y|\mathcal{C}, \mathcal{D}) \stackrel{\bar{p}}{=} E(Y|\mathcal{C})$ .
- (b) For all random variables  $W: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  that are nonnegative and  $\mathcal{D}$ -measurable,

$$E(W \cdot Y|\mathcal{C}) \stackrel{\bar{p}}{=} E(W|\mathcal{C}) \cdot E(Y|\mathcal{C}). \tag{10.56}$$

(Proof p. 320)

**Remark 10.52 (Mean Independence From a  $\sigma$ -Algebra)** For  $\mathcal{C} = \{\Omega, \emptyset\}$ , Theorem 10.51 and Remark 10.5 immediately yield the following proposition: If  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  is a nonnegative random variable that has a finite expectation  $E(Y)$  and  $\mathcal{D} \subset \mathcal{A}$  is a  $\sigma$ -algebra, then the following two propositions are equivalent to each other:

- (a)  $E(Y|\mathcal{D}) \stackrel{p}{=} E(Y)$ .  
 (b) For all random variables  $W: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  that are nonnegative and  $\mathcal{D}$ -measurable,

$$E(W \cdot Y) = E(W) \cdot E(Y). \quad (10.57)$$

&lt;

In Theorem 10.51 we required that  $Y$  is nonnegative and has a finite expectation. The implication of conditional mean independence formulated in proposition (i) of the following theorem is *not* restricted to nonnegative random variables  $Y$ . Instead we assume that  $Y$  has a finite second moment.

**Theorem 10.53 ( $\mathcal{C}$ -Conditional Mean Independence if  $E(Y^2)$  is Finite)**

Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  be a random variable that has a finite second moment  $E(Y^2)$ , let  $\mathcal{C}, \mathcal{D} \subset \mathcal{A}$  be  $\sigma$ -algebras, and consider:

- (a)  $E(Y|\mathcal{C}, \mathcal{D}) \stackrel{p}{=} E(Y|\mathcal{C})$ .  
 (b) For all random variables  $W: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  that are  $\mathcal{D}$ -measurable and have a finite second moment  $E(W^2)$ ,

$$E(W \cdot Y|\mathcal{C}) \stackrel{p}{=} E(W|\mathcal{C}) \cdot E(Y|\mathcal{C}). \quad (10.58)$$

Then:

- (i) (a) implies (b).  
 (ii) If, additionally,  $Y$  is nonnegative, then (a) and (b) are equivalent to each other.

(Proof p. 320)

**Remark 10.54 (Mean Independence if  $E(Y^2)$  is Finite)** For  $\mathcal{C} = \{\Omega, \emptyset\}$ , Theorem 10.53 and Remark 10.5 immediately yield the following proposition. Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  be a random variable that has a finite second moment  $E(Y^2)$ , let  $\mathcal{D} \subset \mathcal{A}$  be a  $\sigma$ -algebra, and consider:

- (a)  $E(Y|\mathcal{D}) \stackrel{p}{=} E(Y)$ .  
 (b) For all random variables  $W: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  that are  $\mathcal{D}$ -measurable and have a finite second moment  $E(W^2)$ ,

$$E(W \cdot Y) = E(W) \cdot E(Y). \quad (10.59)$$

Then:

- (i) (a) implies (b).

(ii) If  $Y$  is also nonnegative, then (a) and (b) are equivalent to each other.

◁

For  $\mathcal{C} = \sigma(X)$  and  $\mathcal{D} = \sigma(Z)$ , Theorem 10.51 immediately implies the following corollary.

**Corollary 10.55 (X-Conditional Mean Independence if  $E(Y)$  is Finite)**

Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  be a nonnegative random variable that has a finite expectation  $E(Y)$  and let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$ ,  $Z: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Z, \mathcal{A}'_Z)$  be random variables. Then the following two propositions are equivalent to each other:

(a)  $E(Y|X, Z) \stackrel{P}{=} E(Y|X)$ ;

(b) For all random variables  $W: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  that are nonnegative and  $Z$ -measurable,

$$E(W \cdot Y | X) \stackrel{P}{=} E(W | X) \cdot E(Y | X). \quad (10.60)$$

Similarly, for  $\mathcal{C} = \sigma(X)$  and  $\mathcal{D} = \sigma(Z)$ , Theorem 10.53 immediately implies the following corollary.

**Corollary 10.56 (X-Conditional Mean Independence if  $E(Y^2)$  is Finite)**

Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  be a random variable that has a finite second moment  $E(Y^2)$ , let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$ ,  $Z: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Z, \mathcal{A}'_Z)$  be random variables, and consider:

(a)  $E(Y|X, Z) \stackrel{P}{=} E(Y|X)$ .

(b) For all random variables  $W: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  that are  $Z$ -measurable and have a finite second moment  $E(W^2)$ ,

$$E(W \cdot Y | X) \stackrel{P}{=} E(W | X) \cdot E(Y | X). \quad (10.61)$$

Then:

(i) (a) implies (b).

(ii) If  $Y$  is also nonnegative, then (a) and (b) are equivalent to each other.

**Remark 10.57 (Mean Independence From a Random Variable)** For  $X = \alpha$ ,  $\alpha \in \Omega'_X$ , this corollary and Remark 10.5 immediately yield the following proposition. Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  and  $Z: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Z, \mathcal{A}'_Z)$  be random variables with  $E(Y^2) < \infty$ , and consider:

(a)  $E(Y|Z) \stackrel{P}{=} E(Y)$ .

(b) For all random variables  $W: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  that are  $Z$ -measurable and have a finite second moment  $E(W^2)$ ,

$$E(W \cdot Y) = E(W) \cdot E(Y). \quad (10.62)$$

Then:

- (i) (a) implies (b).
- (ii) If  $Y$  is also nonnegative, then (a) and (b) are equivalent to each other.

◁

**Remark 10.58 (Mean Independence and Uncorrelatedness)** For  $Z = W$ , Proposition (i) of Remark 10.57 yields

$$E(Y|Z) \stackrel{\bar{p}}{=} E(Y) \Rightarrow E(Z \cdot Y) = E(Z) \cdot E(Y), \quad (10.63)$$

provided that  $E(Z^2), E(Y^2) < \infty$ . Proposition (10.63) is equivalent to (10.52), because  $E(Z \cdot Y) = E(Z) \cdot E(Y)$  is equivalent to  $Cov(Z, Y) = 0$ , provided that  $E(Z^2), E(Y^2) < \infty$  [see Box 7.1 (i)]. However,  $Cov(Z, Y) = 0$  does *not* imply  $E(Y|Z) \stackrel{\bar{p}}{=} E(Y)$ . In other words, uncorrelatedness does not imply mean independence. ◁

**Remark 10.59 (Dichotomous Variables)** If  $Y$  and  $Z$  are dichotomous with values 0 and 1, then  $Y \stackrel{\bar{p}}{=} 1_{Y=1}$  (see Example 5.10), and hence Equation (10.63) can equivalently be written

$$P(Y=1|Z) \stackrel{\bar{p}}{=} P(Y=1) \Rightarrow P(Y=1, Z=1) = P(Y=1) \cdot P(Z=1) \quad (10.64)$$

[see Eqs. (10.2), (6.5), and (1.33)]. Applying Remark 5.46 yields

$$P(Y=1|Z) \stackrel{\bar{p}}{=} P(Y=1) \Rightarrow Y \perp\!\!\!\perp Z. \quad (10.65)$$

◁

Further properties of conditional mean independence are treated in section 16.3, in particular the relationship between conditional independence and conditional mean independence.

## 10.7 Proofs

### *Proof of Theorem 10.17*

Under the assumptions about  $\mathcal{E} = \{A_1, A_2, \dots\}$  and  $\mathcal{C}$ , a function  $V: (\Omega, \mathcal{A}) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  is  $\mathcal{C}$ -measurable if and only if there are  $\alpha_i \in \bar{\mathbb{R}}, i = 1, 2, \dots$ , such that  $V = \sum_{i=1}^{\infty} \alpha_i 1_{A_i}$  (see Lemma 2.19). Hence, if  $V, V^* \in \mathcal{E}(Y|\mathcal{C})$ , then

$$V = \sum_{i=1}^{\infty} \alpha_i 1_{A_i} \quad \text{and} \quad V^* = \sum_{i=1}^{\infty} \alpha_i^* 1_{A_i}, \quad \alpha_i, \alpha_i^* \in \bar{\mathbb{R}}. \quad (10.66)$$

This implies

$$P(\{V \neq V^*\}) = \sum_{i: \alpha_i \neq \alpha_i^*} P(A_i).$$

Because  $P(\{V \neq V^*\}) = 0$  (see Th. 10.9 and Def. 2.68), we can conclude

$$\sum_{i: \alpha_i \neq \alpha_i^*} P(A_i) = 0. \quad (10.67)$$

Hence if (10.13) holds, then Equation (10.67) implies that there is no  $i$  with  $\alpha_i \neq \alpha_i^*$ , which implies  $\{V \neq V^*\} = \emptyset$ .

Now assume that there is an  $A_j \in \mathcal{E}$  with  $P(A_j) = 0$ . Then define  $V \in \mathcal{E}(Y|\mathcal{C})$  as in Equation (10.66) and  $V^* = \sum_{i=1}^{\infty} \alpha_i^* 1_{A_i}$ , with  $\alpha_i^* = \alpha_i$  for  $i \neq j$  and  $\alpha_j^* := \alpha_j + 1$ . This implies  $V \stackrel{P}{=} V^*$  and  $V^* \in \mathcal{E}(Y|\mathcal{C})$ , but  $\emptyset \neq A_j \subset \{V \neq V^*\}$ . By contraposition, this proves that  $\{V \neq V^*\} = \emptyset$  implies Equation (10.13). For  $\mathcal{E} = \{A_1, \dots, A_n\}$ , the proof is analogous.

### **Proof of Theorem 10.21**

(i) This proof is found in Bauer (1996, (15.13), p. 115). Because the sequence  $Y_i, i \in \mathbb{N}$ , is increasing and the conditional expectation is monotone [see Rule (ix) of Box 10.3], we can conclude:  $\lim_{i \rightarrow \infty} Y_i = \sup_{i \in \mathbb{N}} Y_i$  and  $\lim_{i \rightarrow \infty} E(Y_i|\mathcal{C}) = \sup_{i \in \mathbb{N}} E(Y_i|\mathcal{C})$ .

(ii) If  $Y_i \geq 0$ , for all  $i \in \mathbb{N}$ , then  $\tilde{Y}_n := \sum_{i=1}^n Y_i, n \in \mathbb{N}$ , is increasing and  $\lim_{n \rightarrow \infty} \tilde{Y}_n = \sum_{i=1}^{\infty} Y_i$ . Hence,

$$\begin{aligned} E\left(\sum_{i=1}^{\infty} Y_i \mid \mathcal{C}\right) &\stackrel{P}{=} E\left(\lim_{n \rightarrow \infty} \tilde{Y}_n \mid \mathcal{C}\right) \\ &\stackrel{P}{=} \lim_{n \rightarrow \infty} E\left(\tilde{Y}_n \mid \mathcal{C}\right) \quad [(10.19)] \\ &\stackrel{P}{=} \lim_{n \rightarrow \infty} E\left(\sum_{i=1}^n Y_i \mid \mathcal{C}\right) \\ &\stackrel{P}{=} \lim_{n \rightarrow \infty} \sum_{i=1}^n E(Y_i|\mathcal{C}) \quad [\text{Box 10.1 (xvi)}] \\ &\stackrel{P}{=} \sum_{i=1}^{\infty} E(Y_i|\mathcal{C}). \end{aligned}$$

### **Proof of Theorem 10.42**

According to Lemma 2.52,  $g(X)$  is measurable with respect to  $\sigma(X)$ . Therefore, according to Definition 10.2 (b), we only have to show that

$$E[1_C \cdot g(X)] = E(1_C \cdot Y), \quad \forall C \in \sigma(X), \quad (10.68)$$

and Equation (10.43) are equivalent to each other.

(10.43)  $\Leftrightarrow$  (10.68). By definition,  $\sigma(X) = \{X^{-1}(C') : C' \in \mathcal{A}'_X\}$ . Therefore,

$$\begin{aligned} \int 1_{C'}(x) \cdot g(x) P_X(dx) &= \int 1_{C'}(x) \cdot y P_{X,Y}(d(x,y)), \quad \forall C' \in \mathcal{A}'_X, \\ \Leftrightarrow \int 1_C \cdot g(X) dP &= \int 1_C \cdot Y dP, \quad \forall C = X^{-1}(C') \in \sigma(X), \quad [(3.59), (3.28)] \\ \Leftrightarrow E[1_C \cdot g(X)] &= E(1_C \cdot Y), \quad \forall C \in \sigma(X). \quad [(6.1)] \end{aligned}$$

**Proof of Theorem 10.51**

(a)  $\Rightarrow$  (b) If  $Y$  and  $W$  are nonnegative, then  $Y \cdot W$  is nonnegative as well and, for  $\mathcal{D}$ -measurable  $W$ ,

$$\begin{aligned} E(W \cdot Y | \mathcal{C}) &\stackrel{P}{=} E[E(W \cdot Y | \mathcal{C}, \mathcal{D}) | \mathcal{C}] && \text{[Box 10.1 (v)]} \\ &\stackrel{P}{=} E[W \cdot E(Y | \mathcal{C}, \mathcal{D}) | \mathcal{C}] && \text{[Box 10.1 (xiv)]} \\ &\stackrel{P}{=} E[W \cdot E(Y | \mathcal{C}) | \mathcal{C}] && \text{[(a), Box 10.1 (ix)]} \\ &\stackrel{P}{=} E(Y | \mathcal{C}) \cdot E(W | \mathcal{C}). && \text{[Box 10.1 (xiv)]} \end{aligned} \quad (10.69)$$

(b)  $\Rightarrow$  (a)

$$\begin{aligned} E[W \cdot E(Y | \mathcal{C}, \mathcal{D}) | \mathcal{C}] &\stackrel{P}{=} E[E(W \cdot Y | \mathcal{C}, \mathcal{D}) | \mathcal{C}] && \text{[Box 10.1 (xiv)]} \\ &\stackrel{P}{=} E(W \cdot Y | \mathcal{C}) && \text{[Box 10.1 (v)]} \\ &\stackrel{P}{=} E(Y | \mathcal{C}) \cdot E(W | \mathcal{C}) && \text{[(b)]} \\ &\stackrel{P}{=} E[W \cdot E(Y | \mathcal{C}) | \mathcal{C}]. && \text{[Box 10.1 (xiv)]} \end{aligned} \quad (10.70)$$

Choosing in this equation  $W = 1_D$ ,  $D \in \mathcal{D}$ , Definition 10.2 (b) yields,

$$E[1_C 1_D E(Y | \mathcal{C}, \mathcal{D})] = E[1_C 1_D E(Y | \mathcal{C})], \quad \forall C \in \mathcal{C},$$

which is equivalent to

$$\int 1_{C \cap D} E(Y | \mathcal{C}, \mathcal{D}) dP = \int 1_{C \cap D} E(Y | \mathcal{C}) dP, \quad \forall C \in \mathcal{C}. \quad (10.71)$$

The set  $\{C \cap D: C \in \mathcal{C}, D \in \mathcal{D}\}$  is  $\cap$ -stable and generates  $\sigma(\mathcal{C} \cup \mathcal{D})$ . Furthermore, finiteness of  $E(Y) = E[E(Y | \mathcal{C}, \mathcal{D})] = E[E(Y | \mathcal{C})]$  [see Box 10.1, (iv)] implies that  $E(Y | \mathcal{C}, \mathcal{D})$  and  $E(Y | \mathcal{C})$  are integrable with respect to  $P$ . Hence, according to Theorem 3.68 (iv), we can conclude  $E(Y | \mathcal{C}, \mathcal{D}) \stackrel{P}{=} E(Y | \mathcal{C})$ .

**Proof of Theorem 10.53**

(a)  $\Rightarrow$  (b) If  $E(Y^2), E(W^2) < \infty$ , then  $E(Y)$  and  $E(Y \cdot W)$  are finite as well. Then, for  $\mathcal{D}$ -measurable  $W$ , (10.69) also applies to this case.

(b)  $\Rightarrow$  (a) Choosing  $W = 1_D$ ,  $D \in \mathcal{D}$  in Equation (10.70), Definition 10.2 (b) yields,

$$E[1_C 1_D E(Y | \mathcal{C}, \mathcal{D})] = E[1_C 1_D E(Y | \mathcal{C})], \quad \forall C \in \mathcal{C},$$

which is equivalent to

$$\int 1_{C \cap D} E(Y | \mathcal{C}, \mathcal{D}) dP = \int 1_{C \cap D} E(Y | \mathcal{C}) dP, \quad \forall C \in \mathcal{C}. \quad (10.72)$$

The set  $\{C \cap D: C \in \mathcal{C}, D \in \mathcal{D}\}$  is  $\cap$ -stable and generates  $\sigma(\mathcal{C} \cup \mathcal{D})$ . Furthermore, finiteness of  $E(Y) = E[E(Y | \mathcal{C}, \mathcal{D})] = E[E(Y | \mathcal{C})]$  [see Box 10.1, (iv)] implies that  $E(Y | \mathcal{C}, \mathcal{D})$  and  $E(Y | \mathcal{C})$  are integrable with respect to  $P$ . Hence, if additionally  $Y \geq 0$ , then there are versions  $E(Y | \mathcal{C}), E(Y | \mathcal{C}, \mathcal{D}) \geq 0$  [see Box 10.3 v] and, according to Theorem 3.68 (iv), we can conclude  $E(Y | \mathcal{C}, \mathcal{D}) \stackrel{P}{=} E(Y | \mathcal{C})$ .

## 10.8 Exercises

- ▷ **Exercise 10-1** Show that, according to Definition 10.2, Equations (9.24) and (9.25) define an element of  $\mathcal{E}(Y|X)$  provided that the assumptions of Definition 9.14 hold.
- ▷ **Exercise 10-2** Table 9.2 (p. 289) presents an element, say  $V$ , of  $\mathcal{E}(Y|X, U)$ . Define an alternative element  $V^* \in \mathcal{E}(Y|X, U)$  and show that the two elements are  $P$ -equivalent.
- ▷ **Exercise 10-3** Under the assumptions of Theorem 10.17 and Equation (10.13), prove Equation (10.14).
- ▷ **Exercise 10-4** Prove the propositions of Box 10.1.
- ▷ **Exercise 10-5** Prove the propositions of Box 10.3.
- ▷ **Exercise 10-6** Show that  $P(X=x) > 0$  implies that the  $(X=x)$ -conditional expectation value of  $Y$  defined by Equation (10.27) is uniquely defined and identical to the term introduced in Definition 9.2.
- ▷ **Exercise 10-7** Prove Equation (10.32).
- ▷ **Exercise 10-8** Show that Equation (10.38) implies Rule (iii) of Box 9.2.
- ▷ **Exercise 10-9** Show that mean independence of  $Y$  from  $X$  implies that  $X$  and  $Y$  are uncorrelated, provided that the second moments of  $X$  and  $Y$  are finite [see (10.52)].
- ▷ **Exercise 10-10** Prove: If  $Y$  is dichotomous with values 0 and 1, then

$$Y \perp\!\!\!\perp_P X \Leftrightarrow E(Y|X) \stackrel{P}{=} E(Y),$$

(see Rem. 10.50).

## Solutions

- ▷ **Solution 10-1** Let  $\Omega'_0 \subset \Omega'_X$  denote the finite or countable set introduced in Definitions 9.14 (i) or (ii). Then  $\{X=x\} \in \sigma(X)$  for all  $x \in \Omega'_0$ . This implies:
- (a) For all  $x \in \Omega'_0$ , the indicator  $1_{X=x}$  is  $X$ -measurable (see Example 2.12), which implies that  $\sum_{x \in \Omega'_0} E(Y|X=x) \cdot 1_{X=x}$  is  $X$ -measurable as well (see Lemma 2.19).
- (b) For  $C \in \sigma(X)$ , define  $C_0 := C \cap X^{-1}(\Omega'_0)$ . Because, by definition,  $P[X^{-1}(\Omega'_0)] = P_X(\Omega'_0) = 1$ , this implies  $P(C \setminus C_0) = 0$  (see Box 4.1 (v), (vii)) and

$$\begin{aligned} & E\left(1_C \cdot \sum_{x \in \Omega'_0} E(Y|X=x) \cdot 1_{X=x}\right) \\ &= E\left(1_{C_0} \cdot \sum_{x \in \Omega'_0} E(Y|X=x) \cdot 1_{X=x}\right) + E\left(1_{C \setminus C_0} \cdot \sum_{x \in \Omega'_0} E(Y|X=x) \cdot 1_{X=x}\right) \quad [(6.21)] \\ &= E\left(\sum_{x \in \Omega'_0} E(Y|X=x) \cdot 1_{C_0} \cdot 1_{X=x}\right). \quad [\text{Box 6.1 (iii), (v)}] \end{aligned}$$

Furthermore, denote  $C'_0 := X(C_0) = \{x \in \Omega'_0 : X^{-1}(\{x\}) \subset C\}$ . Then

$$\begin{aligned}
E\left(1_C \cdot \sum_{x \in \Omega'_0} E(Y|X=x) \cdot 1_{X=x}\right) &= E\left(\sum_{x \in \Omega'_0} E(Y|X=x) \cdot 1_{C_0} \cdot 1_{X=x}\right) \\
&= \sum_{x \in \Omega'_0} E[E(Y|X=x) \cdot 1_{C_0} \cdot 1_{X=x}] && \text{[Box 6.1 (vii), (3.65)]} \\
&= \sum_{x \in C'_0} E[E(Y|X=x) \cdot 1_{X=x}] && \text{[def. of } C_0\text{]} \\
&= \sum_{x \in C'_0} E(Y|X=x) \cdot E(1_{X=x}) && \text{[Box 6.1 (iii)]} \\
&= \sum_{x \in C'_0} E(Y|X=x) \cdot P(X=x) && \text{[(6.4)]} \\
&= \sum_{x \in C'_0} \left(\frac{1}{P(X=x)} \cdot E(1_{X=x} \cdot Y)\right) \cdot P(X=x) && \text{[(9.11)]} \\
&= \sum_{x \in C'_0} E(1_{X=x} \cdot Y) \\
&= E\left(\sum_{x \in C'_0} 1_{X=x} \cdot Y\right) && \text{[Box 6.1 (vii), (3.65)]} \\
&= E(1_{C_0} Y) && [1_{C_0} = \sum_{x \in C'_0} 1_{X=x}] \\
&= E(1_{C_0} Y) + E(1_{C \setminus C_0} Y) && \text{[(6.20)]} \\
&= E(1_C Y). && \text{[(6.21)]}
\end{aligned}$$

▷ **Solution 10-2** Another element  $V^* \in \mathcal{E}(Y|X, U)$  is obtained defining

$$V^*(\omega) = \begin{cases} 9, & \text{if } \omega = \omega_3 \text{ or } \omega = \omega_4 \\ V(\omega), & \text{if } \omega \in \Omega, \omega \neq \omega_3, \omega \neq \omega_4, \end{cases}$$

where  $\omega_3 = (Joe, yes, -)$  and  $\omega_4 = (Joe, yes, +)$ . For  $V$  and  $V^*$ ,  $P(A_1) = 1$ , where  $A_1 = \{\omega \in \Omega : V(\omega) = V^*(\omega)\}$ . The probability  $P(A_1) = 1$  is obtained from adding the probabilities of all six outcomes  $\omega$  for which  $P(\{\omega\}) > 0$  (see the second column of Table 9.2, p. 289).

▷ **Solution 10-3** (a) If  $\mathcal{E}$  is a finite or countable partition of  $\Omega$  and  $\sigma(\mathcal{E}) = \mathcal{C}$ , then, for all  $A_i \in \mathcal{E}$ , the indicator  $1_{A_i}$  is  $\mathcal{C}$ -measurable. This implies that  $\sum_{A_i \in \mathcal{E}} E(Y|A_i) \cdot 1_{A_i}$  is  $\mathcal{C}$ -measurable as well (see Lemma 2.19). Hence, condition (a) of Definition 10.2 is satisfied.

(b) According to Lemma 1.20 and Equations (1.36), (1.37), for all  $C \in \mathcal{C}$ ,

$$1_C = \sum_{A_i \in \mathcal{E}, A_i \subset C} 1_{A_i}. \tag{10.73}$$

Hence,

$$\begin{aligned}
E\left(1_C \cdot \sum_{A_i \in \mathcal{E}} E(Y|A_i) \cdot 1_{A_i}\right) &= E\left(\sum_{A_i \in \mathcal{E}} E(Y|A_i) \cdot 1_{A_i} 1_C\right) \\
&= \sum_{A_i \in \mathcal{E}} E[E(Y|A_i) \cdot 1_{A_i} 1_C] && \text{[Box 6.1 (vii), (3.65)]} \\
&= \sum_{A_i \in \mathcal{E}, A_i \subset C} E[E(Y|A_i) \cdot 1_{A_i}] && \text{[(10.73)]} \\
&= \sum_{A_i \in \mathcal{E}, A_i \subset C} E(Y|A_i) \cdot E(1_{A_i}) && \text{[Box 6.1 (iii)]} \\
&= \sum_{A_i \in \mathcal{E}, A_i \subset C} \frac{1}{P(A_i)} \cdot E(1_{A_i} \cdot Y) \cdot P(A_i) && \text{[(9.7), (6.4)]} \\
&= \sum_{A_i \in \mathcal{E}, A_i \subset C} E(1_{A_i} \cdot Y) \\
&= E\left(\sum_{A_i \in \mathcal{E}, A_i \subset C} 1_{A_i} \cdot Y\right) && \text{[Box 6.1 (vii)]} \\
&= E(1_C Y). && \text{[(10.73)]}
\end{aligned}$$

This shows that condition (b) of Definition 10.2 is satisfied and that  $\sum_{A_i \in \mathcal{E}} E(Y|A_i) \cdot 1_{A_i}$  is a version of  $E(Y|\mathcal{C})$ . Equation (10.14) then follows from Proposition (10.12) and the assumption that  $\mathcal{E}$  is a countable partition of  $\Omega$  and  $P(A_i) > 0$  for all  $A_i \in \mathcal{E}$  (see Theorem 10.17).

▷ **Solution 10-4**

(ii) Both sides are  $\mathcal{C}$ -measurable. Furthermore, for all  $C \in \mathcal{C}$ ,

$$\begin{aligned}
E[1_C \cdot (\alpha + Y)] &= E(1_C \alpha + 1_C Y) = E(1_C \alpha) + E(1_C Y) && \text{[Box 6.1 (vi)]} \\
&= E(1_C \alpha) + E[1_C E(Y|\mathcal{C})] && \text{[Def. 10.2 (b)]} \\
&= E(1_C \alpha + 1_C E(Y|\mathcal{C})) && \text{[Box 6.1 (vi)]} \\
&= E(1_C \cdot [\alpha + E(Y|\mathcal{C})]).
\end{aligned}$$

Hence, according to conditions (a) and (b) of Definition 10.2,  $\alpha + E(Y|\mathcal{C}) \in \mathcal{E}(\alpha + Y|\mathcal{C})$  and Proposition (10.12) yields (ii).

(iii) Both sides are  $\mathcal{C}$ -measurable. For all  $C \in \mathcal{C}$ ,

$$\begin{aligned}
E(1_C \cdot \alpha Y) &= \alpha E(1_C \cdot Y) && \text{[Box 6.1 (iii)]} \\
&= \alpha E[1_C \cdot E(Y|\mathcal{C})]. && \text{[Def. 10.2 (b)]}
\end{aligned}$$

Hence, according to Definition 10.2,  $\alpha E(Y|\mathcal{C}) \in \mathcal{E}(\alpha Y|\mathcal{C})$  and Proposition (10.12) yields (iii).

(iv) This rule immediately follows from condition (b) of Definition 10.2 for  $C = \Omega$ , because

$$\begin{aligned}
E[E(Y|\mathcal{C})] &= E[1_\Omega E(Y|\mathcal{C})] && \text{[(3.31)]} \\
&= E(1_\Omega Y) && \text{[Def. 10.2 (b)]} \\
&= E(Y). && \text{[(3.31)]}
\end{aligned}$$

(v) The terms on both sides of this equation are  $\mathcal{C}_0$ -measurable, because of Definition 10.2 (a). Furthermore, for all  $C \in \mathcal{C}_0 \subset \mathcal{C}$ ,

$$\begin{aligned}
E(1_C E[E(Y|\mathcal{C})|\mathcal{C}_0]) &= E(1_C E(Y|\mathcal{C})) && \text{[Def. 10.2 (b)]} \\
&= E(1_C Y). && \text{[Def. 10.2 (b)]}
\end{aligned}$$

In the first equation, we apply Definition 10.2 (b) to  $E[E(Y|\mathcal{C})|\mathcal{C}_0]$  and  $\mathcal{C}_0$ , whereas in the second, we apply it to  $E(Y|\mathcal{C})$  and  $\mathcal{C}$ . The last equation shows that  $E[E(Y|\mathcal{C})|\mathcal{C}_0] \in \mathcal{E}(Y|\mathcal{C}_0)$  and Proposition (10.12) yields (v).

(vi) The constant  $E(Y)$  is measurable with respect to any  $\sigma$ -algebra  $\mathcal{C}$  on  $\Omega$  (see Example 2.10). Furthermore, if  $Y$  and  $\mathcal{C}$  are independent, then  $Y$  and  $1_C$  are independent for all  $C \in \mathcal{C}$  (see Rem. 5.46). Hence, for  $C \in \mathcal{C}$ ,  $E(1_C Y) = E(1_C)E(Y)$  for all  $C \in \mathcal{C}$  (see Th. 6.24). Therefore, Rules (ii) and (iii) of Box 6.1 yield

$$E(1_C Y) = E(1_C)E(Y) = E[1_C E(Y)], \quad \forall C \in \mathcal{C}.$$

(vii) We assume that  $Y$  is  $\mathcal{C}$ -measurable. Furthermore,

$$E(1_C Y) = E(1_C Y), \quad \forall C \in \mathcal{C}.$$

obviously holds. Hence, according to Definition 10.2 (b) and Proposition (10.12), this implies that  $Y \in \mathcal{E}(Y|\mathcal{C})$  and  $E(Y|\mathcal{C}) \stackrel{P}{=} Y$ .

(i) The  $\sigma$ -algebra generated by  $\alpha$  is  $\{\Omega, \emptyset\}$ , which is a subset of every  $\sigma$ -algebra on  $\Omega$ . Hence,  $\sigma(\alpha) \subset \mathcal{C}$ , which shows that Rule (i) is a special case of Rule (vii).

(viii)

$$\begin{aligned} E[Y|E(Y|\mathcal{C})] &\stackrel{P}{=} E[E(Y|\mathcal{C})|E(Y|\mathcal{C})] && \text{[(v) with } \mathcal{C}_0 = \sigma[E(Y|\mathcal{C})]] \\ &\stackrel{P}{=} E(Y|\mathcal{C}). && [\sigma[E(Y|\mathcal{C})] \subset \sigma[E(Y|\mathcal{C})], \text{ (vii)}] \end{aligned}$$

(ix)

$$\begin{aligned} Y_1 &\stackrel{P}{=} Y_2 \\ \Rightarrow \forall C \in \mathcal{C}: E(1_C \cdot Y_1) &= E(1_C \cdot Y_2) && \text{[Th. 3.48, (6.1), } \mathcal{C} \subset \mathcal{A}] \\ \Rightarrow \forall C \in \mathcal{C}: E[1_C \cdot E(Y_1|\mathcal{C})] &= E(1_C \cdot Y_1) = E(1_C \cdot Y_2) = E[1_C \cdot E(Y_2|\mathcal{C})] && \text{[Def. 10.2 (b)]} \\ \Rightarrow E(Y_2|\mathcal{C}) &\in \mathcal{E}(Y_1|\mathcal{C}). && \text{[Def. 10.2]} \end{aligned}$$

According to Proposition (10.12) this implies (ix).

(x)

$$\begin{aligned} E(Y) \text{ finite} &\Rightarrow \forall V \in \mathcal{E}(Y|\mathcal{C}): E(V) \text{ finite} && [E(V) = E(Y), \text{ (iv)}] \\ &\Rightarrow \forall V \in \mathcal{E}(Y|\mathcal{C}): V \text{ is real-valued } P\text{-a.s.} && \text{[Lemma 3.41]} \end{aligned}$$

Now let  $V^* \in \mathcal{E}(Y|\mathcal{C})$  and  $A := \{\omega \in \Omega: V(\omega) \notin \mathbb{R}\}$ . Then  $A \in \mathcal{C}$  and  $P(A) = 0$ . Define  $V := 1_{\Omega \setminus A} \cdot V^*$ . Then  $V$  is real-valued,  $\mathcal{C}$ -measurable, and  $V \stackrel{P}{=} V^*$ , which implies  $V \in \mathcal{E}(Y|\mathcal{C})$ .

(xv) If  $Y_1$  (or  $Y_2$ ) is real-valued and with finite expectation, then there is a real-valued version  $E(Y_1|\mathcal{C}) \in \mathcal{E}(Y_1|\mathcal{C})$  [or  $E(Y_2|\mathcal{C}) \in \mathcal{E}(Y_2|\mathcal{C})$ ] [see (x)]. If  $Y_1$  (or  $Y_2$ ) is nonnegative, then there is a nonnegative version  $E(Y_1|\mathcal{C}) \in \mathcal{E}(Y_1|\mathcal{C})$  [or  $E(Y_2|\mathcal{C}) \in \mathcal{E}(Y_2|\mathcal{C})$ ] [see Box 10.3 (v) for  $\alpha = 0$ ]. [Note that the proof of Box 10.3 (v) uses Box 10.1 (iv) and (ii).]

For versions  $E(Y_1|\mathcal{C}) \in \mathcal{E}(Y_1|\mathcal{C})$ ,  $E(Y_2|\mathcal{C}) \in \mathcal{E}(Y_2|\mathcal{C})$  that are nonnegative or with finite expectations, the sum  $E(Y_1|\mathcal{C}) + E(Y_2|\mathcal{C})$  is  $\mathcal{C}$ -measurable (see Th. 2.57), and for all  $C \in \mathcal{C}$ ,

$$\begin{aligned} &E(1_C [E(Y_1|\mathcal{C}) + E(Y_2|\mathcal{C})]) \\ &= E(1_C E(Y_1|\mathcal{C})) + E(1_C E(Y_2|\mathcal{C})) && \text{[Box 6.1 (vi)]} \\ &= E(1_C Y_1) + E(1_C Y_2) && \text{[Def. 10.2 (b)]} \\ &= E(1_C (Y_1 + Y_2)). && \text{[Box 6.1 (vi)]} \end{aligned}$$

(xvi)

$$\begin{aligned}
& E(\alpha_1 Y_1 + \alpha_2 Y_2 | \mathcal{C}) \\
& \stackrel{P}{=} E(\alpha_1 Y_1 | \mathcal{C}) + E(\alpha_2 Y_2 | \mathcal{C}) \quad \text{[(xv)]} \\
& \stackrel{P}{=} \alpha_1 E(Y_1 | \mathcal{C}) + \alpha_2 E(Y_2 | \mathcal{C}) \quad \text{[(iii)]}
\end{aligned}$$

The equation for  $n$  summands follows by induction.

(xi) Cor. 8.17 of Klenke, 2013, p. 176).

(xiv) If  $Y_1$  is  $\mathcal{C}$ -measurable, then  $Y_1 \cdot E(Y_2 | \mathcal{C})$  is  $\mathcal{C}$ -measurable as well [see Def. 10.2 (a), Th. 2.57]. First, consider the case  $Y_1 = 1_{C^*}$ , for  $C^* \in \mathcal{C}$ . Then, for all  $C \in \mathcal{C}$ ,

$$\begin{aligned}
\int 1_C \cdot E(1_{C^*} \cdot Y_2 | \mathcal{C}) dP &= \int 1_C \cdot 1_{C^*} \cdot Y_2 dP && \text{[Def. 10.2 (b)]} \\
&= \int 1_{C \cap C^*} \cdot Y_2 dP && \text{[(1.33)]} \\
&= \int 1_{C \cap C^*} \cdot E(Y_2 | \mathcal{C}) dP && [C \cap C^* \in \mathcal{C}, \text{ Def. 10.2 (b)}] \\
&= \int 1_C \cdot 1_{C^*} \cdot E(Y_2 | \mathcal{C}) dP. && \text{[(1.33)]}
\end{aligned}$$

Hence, according to Definition 10.2 (b) and Proposition (10.12),

$$E(1_{C^*} \cdot Y_2 | \mathcal{C}) \stackrel{P}{=} 1_{C^*} \cdot E(Y_2 | \mathcal{C}). \quad (10.74)$$

If  $E(Y_1^2), E(Y_2^2) < \infty$  or  $Y_1, Y_2$  nonnegative, then Remark 3.30, Box 10.1 (xi), (xvi), and Theorem 10.21 imply, for all  $\mathcal{C}$ -measurable  $Y_1$ ,

$$\int 1_C \cdot E(Y_1 \cdot Y_2 | \mathcal{C}) dP = \int 1_C \cdot Y_1 \cdot E(Y_2 | \mathcal{C}) dP, \quad \forall C \in \mathcal{C}.$$

Now, according to Definition 10.2 (b),  $Y_1 \cdot E(Y_2 | \mathcal{C}) \in \mathcal{E}(Y_1 \cdot Y_2 | \mathcal{C})$ , and Proposition (10.12) yields

$$E(Y_1 \cdot Y_2 | \mathcal{C}) \stackrel{P}{=} Y_1 \cdot E(Y_2 | \mathcal{C}).$$

(xii) Note that  $E(Y^2) < \infty$  implies  $E[E(Y | \mathcal{C})^2] < \infty$  [see Box 10.1 (xi)].

$$\begin{aligned}
& \text{Cov}[Y, E(Y | \mathcal{C})] \\
& = E[Y \cdot E(Y | \mathcal{C})] - E(Y) \cdot E[E(Y | \mathcal{C})] && \text{[Box 7.1 (i)]} \\
& = E[E(Y \cdot E(Y | \mathcal{C}) | \mathcal{C})] - E(Y) \cdot E[E(Y | \mathcal{C})] && \text{[(iv)]} \\
& = E[E(Y | \mathcal{C}) \cdot E(Y | \mathcal{C})] - E[E(Y | \mathcal{C})] \cdot E[E(Y | \mathcal{C})] && \text{[(xiv), (iv)]} \\
& = \text{Var}[E(Y | \mathcal{C})]. && \text{[Box 6.2 (i)]}
\end{aligned}$$

(xiii) Note that  $E(Y_2^2) < \infty$  implies  $E[E(Y_2 | \mathcal{C})^2] < \infty$  [see Box 10.1 (xi)]. Hence,

$$\begin{aligned}
\text{Cov}(Y_1, Y_2) &= E(Y_1 \cdot Y_2) - E(Y_1) \cdot E(Y_2) && \text{[Box 7.1 (i)]} \\
&= E[E(Y_1 \cdot Y_2 | \mathcal{C})] - E(Y_1) \cdot E[E(Y_2 | \mathcal{C})] && \text{[(iv)]} \\
&= E[Y_1 \cdot E(Y_2 | \mathcal{C})] - E(Y_1) \cdot E[E(Y_2 | \mathcal{C})] && [\sigma(Y_1) \subset \mathcal{C}, \text{ (xiv)}] \\
&= \text{Cov}[Y_1, E(Y_2 | \mathcal{C})]. && \text{[Box 7.1 (i)]}
\end{aligned}$$

▷ **Solution 10-5** (v) First, we prove

$$Y \geq_{\overline{P}} 0 \Rightarrow \exists V \in \mathcal{E}(Y|\mathcal{C}): V \geq 0. \quad (10.75)$$

Let  $V^* \in \mathcal{E}(Y|\mathcal{C})$ . Define  $A := \{\omega \in \Omega: V^*(\omega) < 0\} \in \mathcal{C}$  [see Rem. 2.67 (a)]. Applying Definition 10.2 (b) and Equation (3.50) to  $1_A \cdot Y$  yields

$$\int 1_A \cdot V^* dP = \int 1_A \cdot Y dP \geq 0, \quad \text{if } Y \geq_{\overline{P}} 0. \quad (10.76)$$

If we assume  $P(A) > 0$ , then Lemma 3.44 with  $f := -V^*$  and Equation (3.32) yield

$$-\int 1_A \cdot V^* dP = \int 1_A \cdot (-V^*) dP > 0$$

and hence  $\int 1_A \cdot V^* dP < 0$ , which contradicts (10.76). Therefore,  $P(A) = 0$  and thus  $V^* \geq_{\overline{P}} 0$ .

If we define  $V := 1_{\Omega \setminus A} \cdot V^* \geq 0$ , then  $V$  is  $\mathcal{C}$ -measurable (see Th. 2.57),  $V \overline{=} V^*$ , and  $V \in \mathcal{E}(Y|\mathcal{C})$  [see (10.12)]. For  $\alpha \in \mathbb{R}$ , applying (10.75),

$$Y \geq_{\overline{P}} \alpha \Rightarrow Y - \alpha \geq_{\overline{P}} 0 \Rightarrow \exists V_\alpha \in \mathcal{E}(Y - \alpha|\mathcal{C}): V_\alpha \geq 0.$$

Now Rule (ii) of Box 10.1 implies that there is a  $V := V_\alpha + \alpha$  such that  $V \in \mathcal{E}(Y|\mathcal{C})$  and  $V \geq \alpha$ .

(vii) If  $Y$  is nonnegative and there is an  $\alpha \in \mathbb{R}$  such that  $Y \leq_{\overline{P}} \alpha$ , then (3.25) and (3.44) imply  $0 \leq E(Y) \leq \alpha$ . If  $E(Y)$  is finite, then  $E(-Y) = -E(Y)$  is finite as well. Furthermore,

$$\begin{aligned} Y \leq_{\overline{P}} \alpha &\Rightarrow -Y \geq_{\overline{P}} -\alpha \\ &\Rightarrow \exists V^* \in \mathcal{E}(-Y|\mathcal{C}): V^* \geq -\alpha && \text{[(v)]} \\ &\Rightarrow \exists V \in \mathcal{E}(Y|\mathcal{C}): V \leq \alpha. && \text{[Box 10.1 (iii), } V := -V^*] \end{aligned}$$

(vi)

$$\begin{aligned} Y \overline{=} \alpha & \\ \Rightarrow Y \geq_{\overline{P}} \alpha \wedge Y \leq_{\overline{P}} \alpha & \\ \Rightarrow \exists V_1 \in \mathcal{E}(Y|\mathcal{C}): V_1 \geq \alpha \wedge \exists V_2 \in \mathcal{E}(Y|\mathcal{C}): V_2 \leq \alpha && \text{[(v), (vii)]} \\ \Rightarrow \exists V_1, V_2 \in \mathcal{E}(Y|\mathcal{C}): \exists A \in \mathcal{C}: P(A) = 0 \wedge \forall \omega \in \Omega \setminus A: V_1(\omega) = V_2(\omega) = \alpha && \text{[(10.12)]} \\ \Rightarrow \exists V \in \mathcal{E}(Y|\mathcal{C}): V = \alpha. && \text{[} V := \alpha \cdot 1_A + V_1 \cdot 1_{\Omega \setminus A} \text{]} \end{aligned}$$

(ii), (iv), and (iii) are direct implications of (v), (vi), and (vii).

(x), (xi) These rules follow from (iv) and (5.11).

(i) is a straightforward implication of Theorem 3.43 and (iv).

(ix) The proof is given in Bauer (1996, p. 114, 115).

▷ **Solution 10-6** Let  $g(X) \in \mathcal{E}(Y|X)$ . Then, according to Equation (10.26), for all  $g^*(X) \in \mathcal{E}(Y|X)$ ,

$$g(x) = g^*(x), \quad \text{for } P_X\text{-a.a. } x \in \Omega'_X.$$

Hence, if  $P(X=x) > 0$  for an  $x \in \Omega'_X$ , then according to Remark 2.71,  $g(x) = g^*(x)$ , i. e.,  $g(x)$  is uniquely defined. Furthermore, Equation (9.6) yields

$$E(Y|X=x) = \int Y dP^{X=x} \quad [(9.5)]$$

$$= \frac{1}{P(X=x)} \cdot \int 1_{X=x} \cdot Y dP \quad [(9.7)]$$

$$= \frac{1}{P(X=x)} \cdot \int 1_{X=x} \cdot g(X) dP \quad [\text{Def. 10.2 (b), (10.22)}]$$

$$= \frac{1}{P(X=x)} \cdot \int 1_{X=x} \cdot g(x) dP \quad [1_{X=x} \cdot g(X) = 1_{X=x} \cdot g(x)]$$

$$= \frac{1}{P(X=x)} \cdot g(x) \cdot \int 1_{X=x} dP \quad [(3.32)]$$

$$= \frac{1}{P(X=x)} \cdot g(x) \cdot P(X=x) = g(x). \quad [(3.8)]$$

▷ **Solution 10-7** If  $g$  is a factorization of  $E(Y|X) \in \mathcal{E}(Y|X)$ , then, for all  $\omega \in \{X=x\}$ ,

$$E(Y|X)(\omega) = (g \circ X)(\omega) \quad [(10.22)]$$

$$= g[X(\omega)] = g(x) \quad [\omega \in \{X=x\}]$$

$$= E(Y|X=x). \quad [(10.27)]$$

▷ **Solution 10-8** Assume that  $Z$  is a discrete random variable with values  $z_1, z_2, \dots \in \Omega'_Z$  such that  $P_Z(\{z_1, z_2, \dots\}) = 1$  and, for all  $i = 1, 2, \dots$ ,  $\{z_i\} \in \mathcal{A}'_Z$ , and let  $g$  be a factorization of  $E(Y|X, Z) = g(X, Z)$ . Then, for all  $x \in \Omega'_X$  with  $P(X=x, Z=z_i) > 0$  for all  $i \in \mathbb{N}$ ,

$$E(Y|X=x) = E[E(Y|X, Z)|X=x] \quad [(10.38)]$$

$$= E[g(X, Z)|X=x] \quad [(10.22)]$$

$$= \sum_{i=1}^{\infty} g(x, z_i) \cdot P(X=x, Z=z_i | X=x) \quad [\text{Rem. 10.35, (9.16), (9.22)}]$$

$$= \sum_{i=1}^{\infty} E(Y|X=x, Z=z_i) \cdot \frac{P(X=x, Z=z_i, X=x)}{P(X=x)} \quad [(10.27), (4.2)]$$

$$= \sum_{i=1}^{\infty} E(Y|X=x, Z=z_i) \cdot \frac{P(X=x, Z=z_i)}{P(X=x)}$$

$$= \sum_{i=1}^{\infty} E(Y|X=x, Z=z_i) \cdot P(Z=z_i | X=x). \quad [(4.2)]$$

▷ **Solution 10-9** If  $E(Y|X) \stackrel{p}{=} E(Y)$ , then

$$\text{Cov}(X, Y) = \text{Cov}[X, E(Y|X)] \quad [\text{Box 10.2 (xiii)}]$$

$$= \text{Cov}[X, E(Y)] \quad [E(Y|X) \stackrel{p}{=} E(Y), \text{ Box 7.1 (x)}]$$

$$= 0. \quad [\text{Box 7.1 (vii)}]$$

▷ **Solution 10-10** If  $Y$  is dichotomous with values 0 and 1, then  $E(Y) = P(Y=1)$  (see Example 5.10 and Eq. (6.5)). Hence,  $E(Y)$  is finite, and Box 10.2 (vi) yields

$$Y \perp\!\!\!\perp_p X \Rightarrow E(Y|X) \stackrel{p}{=} E(Y).$$

For all  $C' \in \mathcal{A}'_X$ , the event  $\{X \in C'\} = X^{-1}(C') \in \sigma(X)$  [see Eq. (2.14)]. Hence, if  $E(Y|X) \stackrel{p}{=} E(Y)$ , and therefore  $P(Y=1|X) \stackrel{p}{=} P(Y=1)$  [see Eq. (10.4)], then for all  $C' \in \mathcal{A}'_X$ ,

$$P(Y=1, X \in C') = \int 1_{X \in C'} \cdot 1_{Y=1} dP \quad [(3.9), (1.33)]$$

$$\begin{aligned} &= \int 1_{X \in C'} \cdot P(Y=1|X) dP && \text{[Def. 10.2 (b)]} \\ &= \int 1_{X \in C'} \cdot P(Y=1) dP && [P(Y=1|X) \stackrel{\bar{P}}{=} P(Y=1)] \\ &= P(Y=1) \cdot P(X \in C'), && \text{[(3.9)]} \end{aligned}$$

which, according to Remark 5.46, implies  $Y \perp\!\!\!\perp X$ .

## Chapter 11

# Residual, Conditional Variance, and Conditional Covariance

In chapters 9 and 10 we introduced the concepts conditional expectation and regression. In this chapter we turn to the *residual* of a conditional expectation. Its properties supplement the properties of conditional expectations. Oftentimes a residual is what econometricians call a ‘disturbance’, applied statisticians call an ‘error term’, and psychometricians call a ‘measurement error’. Furthermore, we define the *coefficient of determination*, which represents the proportion of variance of a regressand explained by the regressor. It appears under different names in special areas of applied statistics, ranging from ‘intra-class correlation’ to ‘reliability’ in psychometrics. The square root of the coefficient of determination is known as the ‘multiple correlation’. Next, we will define the concepts of a *conditional variance* and a *conditional covariance* given a  $\sigma$ -algebra and given a random variable, as well as the *partial correlation*. Just like the expectation has been used to define variance, covariance, and correlation, the conditional expectation can be used to define conditional variance, conditional covariance, and the partial correlation.

### 11.1 Residual With Respect to a Conditional Expectation

In section 10.4.2 we have shown that a conditional expectation  $E(Y|\mathcal{C})$  is the best approximation of  $Y$  in the sense of minimizing the mean-squared error function. Now we will study the properties of the deviation of  $Y$  from  $E(Y|\mathcal{C})$ . Defining this deviation, we refer to the following assumptions.

#### Assumptions 11.1

Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  be a real-valued random variable with finite expectation,  $\mathcal{C} \subset \mathcal{A}$  a  $\sigma$ -algebra, and  $E(Y|\mathcal{C})$  a real-valued version of the  $\mathcal{C}$ -conditional expectation of  $Y$ .

Note that, according to Rule (x) of Box 10.1, finiteness of  $E(Y)$  implies that there is a real-valued version  $E(Y|\mathcal{C})$ . Hence, assuming that  $E(Y|\mathcal{C})$  is real-valued is no substantial loss of generality. Referring to a real-valued version  $E(Y|\mathcal{C})$  avoids the subtraction of  $\infty$  and  $\infty$  for values of  $Y$  and  $E(Y|\mathcal{C})$ , respectively.

**Definition 11.2 (Residual With Respect to a Conditional Expectation)**

Under the assumptions 11.1,

$$\varepsilon := Y - E(Y|\mathcal{C}) \quad (11.1)$$

is called a version of the residual of  $Y$  with respect to  $E(Y|\mathcal{C})$ .

**Remark 11.3 (Versions of the Residual)** If  $E(Y|\mathcal{C}), E(Y|\mathcal{C})^* \in \mathcal{E}(Y|\mathcal{C})$  are real-valued and  $\varepsilon, \varepsilon^*$  are the respective residuals, then  $\varepsilon \stackrel{\mathbb{P}}{=} \varepsilon^*$ .  $\triangleleft$

Box 11.1 summarizes some properties of the residual, which are proved in Exercise 11-1. All these properties follow from the definition of a residual and the assumption that  $E(Y|\mathcal{C})$  is a real-valued version of the  $\mathcal{C}$ -conditional expectation of  $Y$ .

**Remark 11.4 (Some Special Cases)** Because  $E(Y|\mathcal{C})$  is  $\mathcal{C}$ -measurable, the following equations are special cases of Rules (vii) and (viii) of Box 11.1, respectively.

$$E[\varepsilon | E(Y|\mathcal{C})] \stackrel{\mathbb{P}}{=} 0, \quad (11.2)$$

$$\text{Cov}[\varepsilon, E(Y|\mathcal{C})] = 0, \quad \text{if } E(Y^2) < \infty. \quad (11.3)$$

According to Equation (11.2), the conditional expectation of the residual  $\varepsilon$  given  $E(Y|\mathcal{C})$  is 0 with probability 1. According to the second equation, the residual  $\varepsilon = Y - E(Y|\mathcal{C})$  is uncorrelated with  $E(Y|\mathcal{C})$  if  $E(Y^2) < \infty$ . [Note that finiteness of  $E(E(Y|\mathcal{C})^2)$  follows from  $E(Y^2) < \infty$ ; see Box 10.1 (xi)].

Now consider a random variable  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  with  $\sigma(X) = \mathcal{C}$  and the residual  $\varepsilon := Y - E(Y|X)$ . Then a special case of Rule (vii) is

$$E(\varepsilon | X) \stackrel{\mathbb{P}}{=} 0. \quad (11.4)$$

This property is illustrated in Figure 11.1. In this figure the black points represent the values of the conditional expectation  $E(\varepsilon | X)$ , whereas the circles are possible values of  $\varepsilon$ . If  $f(X)$  denotes the composition of  $X$  and a function  $f: (\Omega'_X, \mathcal{A}'_X) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  that is  $(\mathcal{A}'_X, \bar{\mathcal{B}})$ -measurable, then

$$E[\varepsilon | f(X)] \stackrel{\mathbb{P}}{=} 0 \quad (11.5)$$

with the special case

$$E[\varepsilon | E(Y|X)] \stackrel{\mathbb{P}}{=} 0. \quad (11.6)$$

Furthermore, if  $E(Y^2), E[f(X)^2] < \infty$ , then

$$\text{Cov}[\varepsilon, f(X)] = 0 \quad (11.7)$$

is a special case of Rule (viii). Hence, if  $(\Omega'_X, \mathcal{A}'_X) = (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  and  $E(X^2), E(Y^2) < \infty$ , then

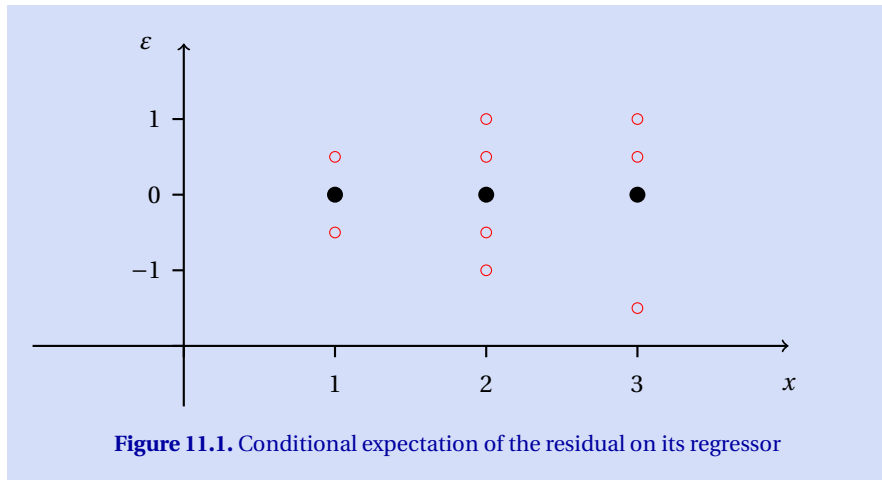


Figure 11.1. Conditional expectation of the residual on its regressor

$$\text{Cov}(\varepsilon, X) = 0, \tag{11.8}$$

is another special case of Rule (viii).

Now consider the residual  $\varepsilon := Y - E(Y|X, Z)$ , where  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  and  $Z: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Z, \mathcal{A}'_Z)$  are not necessarily real-valued. In this case,

$$E(\varepsilon|X, Z) \stackrel{p}{=} E(\varepsilon|X) \stackrel{p}{=} E(\varepsilon|Z) \stackrel{p}{=} 0 \tag{11.9}$$

are special cases of Rule (vii), where  $\mathcal{C} = \sigma(X, Z)$ . If we additionally assume  $X$  and  $Z$  to be numerical and  $E(Y^2), E(X^2), E(Z^2) < \infty$ , then

$$\text{Cov}(X, \varepsilon) = E(X \cdot \varepsilon) = \text{Cov}(Z, \varepsilon) = E(Z \cdot \varepsilon) = 0 \tag{11.10}$$

are special cases of Rule (viii) (see Exercise 11-2). ◁

**Example 11.5 (No Treatment for Joe – continued)** Table 11.1 displays the conditional expectations  $E(Y|X)$ ,  $E(Y|X, U)$ , and  $P(X=1|U)$ , which have been computed in Examples 9.22 and 9.23. Additionally, it contains the residuals of  $Y$  with respect to these conditional expectations. First, we illustrate the property  $E(\varepsilon) = 0$  for  $\varepsilon = Y - E(Y|X)$ . Looking at the table reveals that  $\varepsilon = Y - E(Y|X)$  has four different values:  $-.60, .40, .60$ , and  $-.40$ . Hence, according to Equation (6.3),

$$\begin{aligned} E(\varepsilon) &= -.60 \cdot (.152 + .096) + .40 \cdot (.348 + .024) + .60 \cdot (0 + .152) - .40 \cdot (0 + .228) \\ &= 0. \end{aligned}$$

Second, we illustrate the property

$$E(\varepsilon|X) = 0,$$

**Box 11.1 Rules of Computation for a Residual**

Let the assumptions 11.1 hold. Then the following properties hold for all real-valued versions of  $E(Y|\mathcal{C})$  and all versions of the residual  $\varepsilon$  defined in (11.1):

$$\varepsilon \stackrel{\bar{P}}{=} Y - E(Y|\mathcal{C}). \quad (\text{i})$$

$$Y \stackrel{\bar{P}}{=} E(Y|\mathcal{C}) + \varepsilon. \quad (\text{ii})$$

$$E(\varepsilon) = 0. \quad (\text{iii})$$

$$\text{Var}(Y) = \text{Var}[E(Y|\mathcal{C})] + \text{Var}(\varepsilon), \quad \text{if } E(Y^2) < \infty. \quad (\text{iv})$$

$$\varepsilon \stackrel{\bar{P}}{=} 0, \quad \text{if } Y \stackrel{\bar{P}}{=} E(Y|\mathcal{C}). \quad (\text{v})$$

Additionally, let  $\mathcal{C}_0$  be a  $\sigma$ -algebra and  $W: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_W, \mathcal{A}'_W)$  be a random variable. Then

$$E(\varepsilon|\mathcal{C}_0) \stackrel{\bar{P}}{=} 0, \quad \text{if } \mathcal{C}_0 \subset \mathcal{C}. \quad (\text{vi})$$

$$E(\varepsilon|W) \stackrel{\bar{P}}{=} 0, \quad \text{if } \sigma(W) \subset \mathcal{C}. \quad (\text{vii})$$

If  $W$  is real-valued,  $\sigma(W) \subset \mathcal{C}$ , and  $E(W^2), E(Y^2) < \infty$ , then

$$\text{Cov}(\varepsilon, W) = E(\varepsilon \cdot W) = 0. \quad (\text{viii})$$

$$\text{Cov}[W, E(Y|\mathcal{C})] = \text{Cov}[W, E(Y|\mathcal{C}) + \varepsilon] = \text{Cov}(W, Y). \quad (\text{ix})$$

[see Eq. (11.4)]. Because  $X$  is an indicator variable with values 0 and 1, according to Equation (9.24) and Remark 10.35, it suffices to show that  $E(\varepsilon|X=0) = 0$  and  $E(\varepsilon|X=1) = 0$ . The four values of  $\varepsilon = Y - E(Y|X)$  occur with  $(X=0)$ -conditional probabilities

$$P(\varepsilon = -.60|X=0) = \frac{.152 + .096}{.152 + .348 + .096 + .024} = .40,$$

$$P(\varepsilon = .40|X=0) = \frac{.348 + .024}{.152 + .348 + .096 + .024} = .60,$$

$$P(\varepsilon = .60|X=0) = 0, \quad P(\varepsilon = -.40|X=0) = 0$$

and with  $(X=1)$ -conditional probabilities

$$P(\varepsilon = -.60|X=1) = 0, \quad P(\varepsilon = .40|X=1) = 0$$

$$P(\varepsilon = .60|X=1) = \frac{0 + .152}{0 + 0 + .152 + .228} = .40,$$

**Table 11.1.** No Treatment for Joe With Conditional Expectations and Residuals

Elements of $\Omega$	$P(\{\omega\})$	Observables			Conditional expectations			Residuals		
		Person variable $U$	Treatment variable $X$	Outcome variable $Y$	$E(Y X, U)$	$E(Y X)$	$P(X=1 U)$	$Y - E(Y X, U)$	$Y - E(Y X)$	$X - P(X=1 U)$
(Joe, no, -)	.152	Joe	0	0	.696	.60	0	-.696	-.60	0
(Joe, no, +)	.348	Joe	0	1	.696	.60	0	.304	.40	0
(Joe, yes, -)	0	Joe	1	0	0	.40	0	0	-.40	1
(Joe, yes, +)	0	Joe	1	1	0	.40	0	1	.60	1
(Ann, no, -)	.096	Ann	0	0	.20	.60	.76	-.20	-.60	-.76
(Ann, no, +)	.024	Ann	0	1	.20	.60	.76	.80	.40	-.76
(Ann, yes, -)	.228	Ann	1	0	.40	.40	.76	-.40	-.40	.24
(Ann, yes, +)	.152	Ann	1	1	.40	.40	.76	.60	.60	.24

$$P(\varepsilon = -.40 | X=1) = \frac{0 + .228}{0 + 0 + .152 + .228} = .60,$$

respectively. Hence, according to Equation (9.21),

$$E(\varepsilon | X=0) = -.60 \cdot .40 + .40 \cdot .60 + .60 \cdot 0 - .40 \cdot 0 = 0$$

and

$$E(\varepsilon | X=1) = -.60 \cdot 0 + .40 \cdot 0 + .60 \cdot .40 - .40 \cdot .60 = 0.$$

Because  $X$  is dichotomous with values 0 and 1, and  $P(X=0), P(X=1) > 0$ , we can conclude:

$$E(\varepsilon | X) = E(\varepsilon | X=0) \cdot 1_{X=0} + E(\varepsilon | X=1) \cdot 1_{X=1} = 0 \cdot 1_{X=0} + 0 \cdot 1_{X=1} = 0$$

[see Eq. (9.24) and Rem. 10.35].

◁

### 11.2 Coefficient of Determination and Multiple Correlation

The *coefficient of determination* quantifies the strength of the dependence of a numerical random variable  $Y$  on a  $\sigma$ -algebra  $\mathcal{C}$ , where we refer to the dependence described by the conditional expectation  $E(Y|\mathcal{C})$ . As we will see, the *multiple correlation* is a closely related concept. Reading the following definition, remember that  $E(Y^2) < \infty$  implies  $Var(Y) < \infty$  and  $E(Y) < \infty$  (see Rem. 6.25). It also implies that the conditional expectation  $E(Y|\mathcal{C})$  is defined.

**Definition 11.6 (Coefficient of Determination)**

Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  be a real-valued random variable with  $E(Y^2) < \infty$  and  $\mathcal{C} \subset \mathcal{A}$  be a  $\sigma$ -algebra. Then

$$R_{Y|\mathcal{C}}^2 := \begin{cases} \frac{\text{Var}[E(Y|\mathcal{C})]}{\text{Var}(Y)}, & \text{if } \text{Var}(Y) > 0, \\ 0, & \text{if } \text{Var}(Y) = 0, \end{cases} \quad (11.11)$$

is called the coefficient of determination of  $E(Y|\mathcal{C})$ .

**Remark 11.7 (The Case  $\text{Var}(Y) = 0$ )** If  $E(Y|\mathcal{C}) \stackrel{P}{=} E(Y)$  and  $\text{Var}(Y) > 0$ , then  $R_{Y|\mathcal{C}}^2 = \text{Var}[E(Y)]/\text{Var}(Y) = 0$  [see Eq. (11.11) and Box 6.2 (iv)]. Defining  $R_{Y|\mathcal{C}}^2 = 0$  if  $\text{Var}(Y) = 0$  is arbitrary. However,  $\text{Var}(Y) = 0$  if and only if there is an  $\alpha \in \mathbb{R}$  such that  $Y \stackrel{P}{=} \alpha$  [see again Box 6.2 (iv)], and  $Y \stackrel{P}{=} \alpha$  implies  $E(Y|\mathcal{C}) \stackrel{P}{=} E(Y)$  [see Box 6.1 (i) and Box 10.1 (i)]. Hence, Definition (11.11) implies  $R_{Y|\mathcal{C}}^2 = 0$  whenever  $E(Y|\mathcal{C}) \stackrel{P}{=} E(Y)$ , i. e., whenever  $Y$  is mean independent from  $\mathcal{C}$ . Because  $\mathcal{C} \perp\!\!\!\perp Y \Rightarrow E(Y|\mathcal{C}) \stackrel{P}{=} E(Y)$  [see (10.51)], independence of  $\mathcal{C}$  and  $Y$  implies mean independence of  $Y$  from  $\mathcal{C}$ , which itself implies  $R_{Y|\mathcal{C}}^2 = 0$  and this implication holds irrespective of whether or not  $\text{Var}(Y) > 0$ .  $\triangleleft$

**Remark 11.8 (Range of the Coefficient of Determination)** Using Rule (iv) of Box 11.1 yields

$$R_{Y|\mathcal{C}}^2 = \frac{\text{Var}[E(Y|\mathcal{C})]}{\text{Var}[E(Y|\mathcal{C})] + \text{Var}(\varepsilon)}, \quad (11.12)$$

provided that  $\text{Var}(Y) > 0$ . Because  $\text{Var}(\varepsilon)$  is nonnegative,  $0 \leq R_{Y|\mathcal{C}}^2 \leq 1$ . The number  $R_{Y|\mathcal{C}}^2$  is close to 1 if the variance of the residual  $\varepsilon = Y - E(Y|\mathcal{C})$  is small compared to the variance of the conditional expectation  $E(Y|\mathcal{C})$ . In contrast,  $R_{Y|\mathcal{C}}^2$  is close to 0 if the variance of the residual is large compared to the variance of  $E(Y|\mathcal{C})$ .  $\triangleleft$

**Remark 11.9 (Conditions Implying  $R_{Y|\mathcal{C}}^2 = 1$ )** If  $\text{Var}(Y) > 0$  and we assume that  $Y \stackrel{P}{=} E(Y|\mathcal{C})$ , then  $R_{Y|\mathcal{C}}^2 = 1$  [see Eq. (11.11) and Box 6.2 (v)]. Note that this does not necessarily mean that  $Y$  is  $\mathcal{C}$ -measurable. However, if  $Y$  is  $\mathcal{C}$ -measurable, then  $Y \stackrel{P}{=} E(Y|\mathcal{C})$  already follows from Box 10.1 (vii).  $\triangleleft$

**Remark 11.10 (Alternative Notation)** Suppose that  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  is a random variable and  $\mathcal{C} = \sigma(X)$ , then we also use the notation  $R_{Y|X}^2$  instead of  $R_{Y|\mathcal{C}}^2$ , i. e.,

$$R_{Y|X}^2 := R_{Y|\sigma(X)}^2. \quad (11.13)$$

Equations (10.1) and (11.11) yield

$$R_{Y|X}^2 := \begin{cases} \frac{\text{Var}[E(Y|X)]}{\text{Var}(Y)}, & \text{if } \text{Var}(Y) > 0, \\ 0, & \text{if } \text{Var}(Y) = 0. \end{cases} \quad (11.14)$$

If we consider the multivariate regressor  $X = (X_1, \dots, X_n)$ , then we also use the notation

$$R_{Y|X_1, \dots, X_n}^2 := R_{Y|X}^2. \tag{11.15}$$

◁

**Remark 11.11 (Correlation and the Coefficient of Determination)** Assume that  $X: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  is a real-valued random variable,  $E(X^2), E(Y^2) < \infty$ , and that there is a version  $E(Y|X) \in \mathcal{E}(Y|X)$  with

$$E(Y|X) = Q_{lin}(Y|X) = \beta_0 + \beta_1 X. \tag{11.16}$$

(see Def. 7.2). Then

$$\beta_1 = \frac{Cov(X, Y)}{Var(X)} \tag{11.17}$$

[see Th. 7.14 (ii)] and

$$R_{Y|X}^2 = Corr(X, Y)^2, \tag{11.18}$$

which implies

$$R_{Y|X}^2 = 0 \Leftrightarrow Corr(X, Y) = 0 \tag{11.19}$$

(see Exercise 11-3). Hence, under these assumptions, the correlation  $Corr(X, Y)$  also quantifies the strength of the dependence of  $Y$  on  $X$  described by  $E(Y|X)$ . Both,  $R_{Y|X}^2$  and  $Corr(X, Y)$  are normed quantities. The first takes on its values in the interval  $[0, 1]$ , the latter in the interval  $[-1, 1]$ . In contrast, the slope  $\beta_1$  as well as  $Cov(X, Y)$  quantify the strength of the dependence described by (11.16) by real numbers without bounds. ◁

**Remark 11.12 (Quantifying the Strength of Dependence)** The term  $R_{Y|X}^2$  quantifies the strength of the dependence of  $Y$  on  $X$  described by  $E(Y|X)$ , irrespective on whether or not Equation (11.16) holds. While  $E(Y|X) = g(X)$  describes how the conditional expectation values  $E(Y|X=x)$  of  $Y$  depend on the values  $x$  of  $X$ , the coefficient of determination  $R_{Y|X}^2$  quantifies the strength of this dependence by a single real number between 0 and 1. Similarly,  $R_{Y|\mathcal{C}}^2$  quantifies the strength of the dependence of  $Y$  on  $\mathcal{C}$  described by  $E(Y|\mathcal{C})$  (see Rem. 10.18). ◁

**Remark 11.13 (Uniqueness of  $R_{Y|\mathcal{C}}^2$ )** If  $V, V^* \in \mathcal{E}(Y|\mathcal{C})$ , then  $V$  and  $V^*$  are  $P$ -equivalent and, according to Rule (v) of Box 6.2, this implies  $Var(V) = Var(V^*)$ . Hence, Equation (11.11) implies that  $R_{Y|\mathcal{C}}^2$  is identical for all versions  $V \in \mathcal{E}(Y|\mathcal{C})$ . ◁

**Remark 11.14 (Correlation of  $Y$  and the Conditional Expectation)** If  $E(Y^2)$  is finite, then the coefficient of determination  $R_{Y|\mathcal{C}}^2$  is identical to the squared correlation of  $Y$  and  $E(Y|\mathcal{C})$ , i. e.,

$$R_{Y|\mathcal{C}}^2 = Corr[Y, E(Y|\mathcal{C})]^2 \tag{11.20}$$

(see Exercise 11-4). Correspondingly,

$$R_{Y|X}^2 = Corr[Y, E(Y|X)]^2. \tag{11.21}$$

Note that this equation does not rely on any parametrization of  $E(Y|X)$ . ◁

**Definition 11.15 (Multiple Correlation)**

Let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega_X, \mathcal{A}_X)$  and  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  be random variables and assume  $E(Y^2) < \infty$ . Then

$$R_{Y|X} := \sqrt{R_{Y|X}^2} \quad (11.22)$$

is called the multiple correlation of  $Y$  and  $X$ .

Equations (11.21) and (11.22) immediately imply

$$R_{Y|X} = \text{Corr}[Y, E(Y|X)]. \quad (11.23)$$

**Remark 11.16 (Multivariate  $X$ )** If  $X = (X_1, \dots, X_n)$  is a multivariate random variable, then we also use the notation

$$R_{Y|X_1, \dots, X_n} := R_{Y|X}. \quad (11.24)$$

<

**Remark 11.17 (The Multiple Correlation is Not Symmetric)** Note that, in contrast to a correlation of two numerical random variables, the multiple correlation of  $Y$  and  $X$  is not symmetric. In general, the roles of  $Y$  and  $X$  cannot be exchanged.

<

**Example 11.18 (No Treatment for Joe – continued)** In Table 9.2 (p. 289) we displayed the conditional expectations  $E(Y|X)$ ,  $E(Y|X, U)$ , and  $P(X=1|U)$ . Now we compute  $R_{Y|X}^2$  for the conditional expectation  $E(Y|X)$ . Looking at the table reveals that  $E(Y|X)$  has two different values: .60, which occurs with probability

$$P[E(Y|X) = .60] = .152 + .348 + .096 + .024 = .62,$$

and .40, occurring with probability

$$P[E(Y|X) = .40] = 0 + 0 + .228 + .152 = .38.$$

Furthermore, the expectation of  $Y$  is

$$E(Y) = P(Y=1) = .348 + 0 + .024 + .152 = .524$$

Hence, according to Equation (i) of Box 6.2,

$$\begin{aligned} \text{Var}[E(Y|X)] &= E[E(Y|X)^2] - E[E(Y|X)]^2 \\ &= E[E(Y|X)^2] - E(Y)^2 \quad [\text{Box 10.2 (iv)}] \\ &= (.60^2 \cdot .62 + .40^2 \cdot .38) - .524^2 \\ &\approx .284 - .2746 = .0094. \end{aligned}$$

According to Equation (6.29), the variance of  $Y$  is  $\text{Var}(Y) = P(Y=1) \cdot [1 - P(Y=1)] = .524 \cdot (1 - .524) \approx 0.2494$ . This yields

$$R_{Y|X}^2 = \frac{\text{Var}[E(Y|X)]}{\text{Var}(Y)} \approx \frac{.0094}{.2494} \approx .0377 \quad \text{and} \quad R_{Y|X} \approx .1941.$$

Similarly, the conditional expectation  $E(Y|U)$  takes on each of the two values .696 and .352 with probability .50, i. e.,

$$P[E(Y|U) = .696] = P[E(Y|U) = .352] = .50.$$

Hence,

$$\begin{aligned} \text{Var}[E(Y|U)] &= E[E(Y|U)^2] - E[E(Y|U)]^2 && \text{[Box 6.2 (i)]} \\ &= E[E(Y|U)^2] - E(Y)^2 && \text{[Box 10.2 (iv)]} \\ &= (.696^2 \cdot .50 + .352^2 \cdot .50) - .524^2 \\ &\approx .3042 - .2746 = .0296. \end{aligned}$$

Hence,

$$R_{Y|U}^2 = \frac{\text{Var}[E(Y|U)]}{\text{Var}(Y)} \approx \frac{.0296}{.2494} \approx .1187 \quad \text{and} \quad R_{Y|U} \approx .3445.$$

Finally,

$$\begin{aligned} \text{Var}[E(Y|X,U)] &= E[E(Y|X,U)^2] - E[E(Y|X,U)]^2 && \text{[Box 6.2 (i)]} \\ &= E[E(Y|X,U)^2] - E(Y)^2 && \text{[Box 10.2 (iv)]} \\ &= (.696^2 \cdot .50 + .20^2 \cdot .12 + .40^2 \cdot .38) - .524^2 \\ &\approx .3078 - .2746 = .0332. \end{aligned}$$

Hence,

$$R_{Y|X,U}^2 = \frac{\text{Var}[E(Y|X,U)]}{\text{Var}(Y)} \approx \frac{.0332}{.2494} \approx .1331 \quad \text{and} \quad R_{Y|X,U} \approx .3649.$$

Note that, in this example,  $R_{Y|X,U}^2$  is smaller than the sum of  $R_{Y|X}^2$  and  $R_{Y|U}^2$ .  $\triangleleft$

In the following theorem we present a condition under which the coefficients of determination are additive [see Eq. (11.31)]. This theorem also contains a condition under which the coefficient  $\alpha_1$  of  $X$  in the equation  $E(Y|X) = \alpha_0 + \alpha_1 X$  is identical to the coefficient  $\beta_1$  of  $X$  in the equation  $E(Y|X,Z) = \beta_0 + \beta_1 X + \beta_2 Z$  (see section 12.8 for a generalization).

**Theorem 11.19 (Additivity of the Coefficients of Determination)**

Let  $X, Y, Z: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  be real-valued random variables with finite second moments and positive variances and assume that there are  $\beta_0, \beta_1, \beta_2, \gamma_0, \gamma_1 \in \mathbb{R}$ ,  $E(Y|X,Z) \in \mathcal{E}(Y|X,Z)$ , and  $E(Z|X) \in \mathcal{E}(Z|X)$  such that

$$E(Y|X,Z) = \beta_0 + \beta_1 X + \beta_2 Z, \quad (11.25)$$

$$E(Z|X) = \gamma_0 + \gamma_1 X. \quad (11.26)$$

(i) Then there are  $\alpha_0, \alpha_1 \in \mathbb{R}$  such that

$$E(Y|X) = \alpha_0 + \alpha_1 X. \quad (11.27)$$

(ii) If  $E(Z|X) \stackrel{p}{=} E(Z)$  or  $\beta_2 = 0$ , then for  $\alpha_0, \alpha_1$  occurring in (i),

$$\alpha_0 = \beta_0 + \beta_2 E(Z) \quad (11.28)$$

and

$$\alpha_1 = \beta_1 \quad (11.29)$$

(iii) If

$$\left( E(Z|X) \stackrel{p}{=} E(Z) \text{ or } \beta_2 = 0 \right) \quad \text{and} \quad \left( E(X|Z) \stackrel{p}{=} E(X) \text{ or } \beta_1 = 0 \right),$$

then

$$\text{Var}[E(Y|X, Z)] = \text{Var}[E(Y|X)] + \text{Var}[E(Y|Z)] \quad (11.30)$$

and

$$R_{Y|X, Z}^2 = R_{Y|X}^2 + R_{Y|Z}^2. \quad (11.31)$$

(Proof p. 347)

**Remark 11.20 (Independence of  $X$  and  $Z$ )** Note that the condition specified in proposition (iii) of Theorem 11.19 is satisfied, e. g., if  $X$  and  $Z$  are independent [see Rule (vi) of Box 10.2].

<

### 11.3 Conditional Variance and Covariance Given a $\sigma$ -Algebra

The covariance  $\text{Cov}(Y_1, Y_2)$  has been defined as the expectation of the product of the mean centered random variables  $Y_1 - E(Y_1)$  and  $Y_2 - E(Y_2)$ , i. e.,

$$\text{Cov}(Y_1, Y_2) = E([Y_1 - E(Y_1)] \cdot [Y_2 - E(Y_2)]) \quad (11.32)$$

(cf. Def. 7.8). Similarly, we define the  $\mathcal{C}$ -conditional covariance  $\text{Cov}(Y_1, Y_2|\mathcal{C})$  as the  $\mathcal{C}$ -conditional expectation of the product  $[Y_1 - E(Y_1|\mathcal{C})] \cdot [Y_2 - E(Y_2|\mathcal{C})] = \varepsilon_1 \cdot \varepsilon_2$  of the residuals of  $Y_1$  and  $Y_2$  with respect to their  $\mathcal{C}$ -conditional expectations.

#### Definition 11.21 (Conditional Covariance Given a $\sigma$ -Algebra)

For  $i = 1, 2$ , let  $Y_i: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  be real-valued random variables with  $E(Y_i^2) < \infty$ , let  $\mathcal{C} \subset \mathcal{A}$  be a  $\sigma$ -algebra, and define  $\varepsilon_i := Y_i - E(Y_i|\mathcal{C})$ . Then

$$\text{Cov}(Y_1, Y_2|\mathcal{C}) := E(\varepsilon_1 \cdot \varepsilon_2|\mathcal{C}) \quad (11.33)$$

*is called a version of the  $\mathcal{C}$ -conditional covariance of  $Y_1$  and  $Y_2$ .*

**Remark 11.22 (X-Conditional Covariance)** Let the assumptions of Definition 11.21 hold and let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  be a random variable. Then

$$\text{Cov}(Y_1, Y_2 | X) := \text{Cov}(Y_1, Y_2 | \sigma(X)) \tag{11.34}$$

is called a version of the *X-conditional covariance* of  $Y_1$  and  $Y_2$ . ◁

The  $\mathcal{C}$ -conditional variance is defined analogously.

**Definition 11.23 (Conditional Variance Given a  $\sigma$ -Algebra)**  
 Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  be a real-valued random variable with  $E(Y^2) < \infty$ , let  $\mathcal{C} \subset \mathcal{A}$  be a  $\sigma$ -algebra, and define  $\varepsilon := Y - E(Y|\mathcal{C})$ .

(i) Then

$$\text{Var}(Y|\mathcal{C}) := E(\varepsilon^2|\mathcal{C}) \tag{11.35}$$

*is called a version of the  $\mathcal{C}$ -conditional variance of  $Y$ .*

(ii) Let  $\text{Var}(Y|\mathcal{C})$  be a nonnegative version of the  $\mathcal{C}$ -conditional variance of  $Y$ . Then

$$SD(Y|\mathcal{C}) := \sqrt{\text{Var}(Y|\mathcal{C})} \tag{11.36}$$

*is called a version of the  $\mathcal{C}$ -conditional standard deviation of  $Y$ .*

**Remark 11.24 (Conditional Variance Given X)** Let the assumptions of Definition 11.23 hold and let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  be a random variable. Then

$$\text{Var}(Y|X) := \text{Var}(Y|\sigma(X)) \tag{11.37}$$

is called a version of the *X-conditional variance* of  $Y$ . Correspondingly, let  $\text{Var}(Y|X)$  be a nonnegative version of the *X-conditional variance* of  $Y$ . Then we call

$$SD(Y|X) := \sqrt{\text{Var}(Y|X)} \tag{11.38}$$

a version of the *X-conditional standard deviation* of  $Y$ . ◁

### 11.4 Conditional Variance and Covariance Given a Value of a Random Variable

While the concepts defined above are random variables, the  $(X=x)$ -conditional covariance is a number. It is defined using the  $(X=x)$ -conditional expectation value  $E(\varepsilon_1 \cdot \varepsilon_2 | X=x)$  that has been introduced as a value  $g(x)$  of a factorization  $g$  of an  $X$ -conditional expectation  $E(\varepsilon_1 \cdot \varepsilon_2 | X) = g(X)$  (see section 10.4.4).

**Definition 11.25 (( $X=x$ )-Conditional Variance and Covariance)**

Let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  be a random variable.

- (i) For  $i = 1, 2$ , let  $Y_i: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  be real-valued random variables with  $E(Y_i^2) < \infty$ , and let  $\varepsilon_i := Y_i - E(Y_i|X)$ . Then we call

$$\text{Cov}(Y_1, Y_2 | X=x) := E(\varepsilon_1 \cdot \varepsilon_2 | X=x) \quad (11.39)$$

an ( $X=x$ )-conditional covariance of  $Y_1$  and  $Y_2$ .

- (ii) Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  be a real-valued random variable with  $E(Y^2) < \infty$  and let  $\varepsilon := Y - E(Y|X)$ . Then we call

$$\text{Var}(Y|X=x) := E(\varepsilon^2 | X=x) \quad (11.40)$$

an ( $X=x$ )-conditional variance of  $Y$ .

- (iii) If, under the assumptions of (ii),  $\text{Var}(Y|X)$  is a nonnegative version of the  $X$ -conditional variance of  $Y$ , then we call

$$\text{SD}(Y|X=x) := \sqrt{\text{Var}(Y|X=x)} \quad (11.41)$$

an ( $X=x$ )-conditional standard deviation of  $Y$ .

**Remark 11.26 (Equivalent Propositions)** Note that  $\text{Cov}(Y_1, Y_2 | X=x)$  is uniquely defined only if  $P(X=x) > 0$ . However, even if  $P(X=x) = 0$  for all  $x \in \Omega'_X$ , then we can still make propositions such as

$$\text{Cov}(Y_1, Y_2 | X=x) = \text{Cov}(Z_1, Z_2 | X=x), \quad \text{for } P_X\text{-almost } x \in \Omega'_X, \quad (11.42)$$

provided that  $Z_1, Z_2$  are real-valued random variables on  $(\Omega, \mathcal{A}, P)$  with finite second moments. According to Corollary 10.39, this proposition is equivalent to

$$\text{Cov}(Y_1, Y_2 | X) \stackrel{p}{=} \text{Cov}(Z_1, Z_2 | X). \quad (11.43)$$

Of course, the same applies to the  $X$ -conditional variance.  $\triangleleft$

**Remark 11.27 (Values of the Conditional Covariance)** As mentioned before, the term defined in Remark 11.22 is a *random variable*. Its values are

$$\text{Cov}(Y_1, Y_2 | X)(\omega) = \text{Cov}(Y_1, Y_2 | X=x), \quad \text{if } X(\omega) = x \quad (11.44)$$

(see Rem. 10.37). This also implies that the value of  $\text{Cov}(Y_1, Y_2 | X)$  is constant on all sets  $\{X=x\}$ . Similarly,

$$\text{Var}(Y|X)(\omega) = \text{Var}(Y|X=x), \quad \text{if } X(\omega) = x. \quad (11.45)$$

$\triangleleft$

**Table 11.2. Joe and Ann With Self-Selection and Residuals**

Unit Treatment Success	Outcomes $\omega$ $P(\{\omega\})$	Observables			Conditional expectations				Residuals	
		Person variable $U$	Treatment variable $X$	Outcome variable $Y$	$E(Y X,U)$	$E(Y X)$	$E(Y U)$	$P(X=1 U)$	$\varepsilon_Y = Y - E(Y U)$	$\varepsilon_X = X - P(X=1 U)$
(Joe, no, -)	.144	Joe	0	0	.70	.60	.704	.04	-.704	-.04
(Joe, no, +)	.336	Joe	0	1	.70	.60	.704	.04	.296	-.04
(Joe, yes, -)	.004	Joe	1	0	.80	.42	.704	.04	-.704	.96
(Joe, yes, +)	.016	Joe	1	1	.80	.42	.704	.04	.296	.96
(Ann, no, -)	.096	Ann	0	0	.20	.60	.352	.76	-.352	-.76
(Ann, no, +)	.024	Ann	0	1	.20	.60	.352	.76	.648	-.76
(Ann, yes, -)	.228	Ann	1	0	.40	.42	.352	.76	-.352	.24
(Ann, yes, +)	.152	Ann	1	1	.40	.42	.352	.76	.648	.24

**Example 11.28 (Joe and Ann With Self-Selection)** Table 11.2 presents a new example with Joe and Ann. In this example, the probabilities of the elementary events differ from the example with randomized assignment (see Table 4.1, p. 131). The values of the conditional expectations are computed analogously as in Example 9.21. In this new example, all individual treatment effects

$$E(Y | X=1, U=Joe) - E(Y | X=0, U=Joe)$$

and

$$E(Y | X=1, U=Ann) - E(Y | X=0, U=Ann)$$

are *positive*, whereas the difference  $E(Y | X=1) - E(Y | X=0)$  is *negative*. Hence, in this example, this difference *can not* be used to evaluate the treatment effect.

Let us consider the (unconditional) covariance of the treatment variable  $X$  and the outcome variable  $Y$ . Note that  $X$  and  $Y$  are indicator variables with values 0 and 1. Therefore,  $E(X) = P(X=1)$ ,  $E(Y) = P(Y=1)$ ,  $E(X \cdot Y) = P(X=1, Y=1)$ , and

$$\begin{aligned} Cov(X, Y) &= E(X \cdot Y) - E(X) \cdot E(Y) = P(X=1, Y=1) - P(X=1) \cdot P(Y=1) \\ &= (.016 + .152) - (.004 + .016 + .228 + .152) \cdot (.336 + .016 + .024 + .152) \\ &= .168 - .40 \cdot .528 = -0.0432. \end{aligned}$$

Hence, the treatment variable and the outcome variable have a *negative* covariance.

Now let us compute the  $(U=u)$ -conditional covariances of  $X$  and  $Y$  for  $u=Joe$  and for  $u=Ann$ . First of all, note that  $P(U=Joe) = .144 + .336 + .004 + .016 = .50$  and  $P(U=Ann) = .096 + .024 + .228 + .152 = .50$ . According to Equation (9.21) we have to sum the values of the product variable  $\varepsilon_X \cdot \varepsilon_Y$  weighted by their  $(U=u)$ -conditional probabilities. Hence,

$$\begin{aligned} Cov(Y, X | U=Joe) &= E(\varepsilon_X \cdot \varepsilon_Y | U=Joe) \\ &= .04 \cdot .704 \cdot \frac{.144}{.50} - .04 \cdot .296 \cdot \frac{.336}{.50} - .96 \cdot .704 \cdot \frac{.004}{.50} + .96 \cdot .296 \cdot \frac{.016}{.50} \\ &= .00384 \end{aligned}$$

and

$$\begin{aligned} Cov(Y, X | U=Ann) &= E(\varepsilon_X \cdot \varepsilon_Y | U=Ann) \\ &= .76 \cdot .352 \cdot \frac{.096}{.50} - .76 \cdot .648 \cdot \frac{.024}{.50} - .24 \cdot .352 \cdot \frac{.228}{.50} + .24 \cdot .648 \cdot \frac{.152}{.50} \\ &= .03648. \end{aligned}$$

Both conditional covariances are *positive*. Hence, in this example, the (unconditional) covariance of  $X$  and  $Y$  (which is negative) is highly misleading if used to evaluate the effects of the treatment on success, because for both persons the  $(U=u)$ -conditional (or person-specific) covariances of  $X$  and  $Y$  are positive.  $\triangleleft$

## 11.5 Properties of Conditional Variances and Covariances

Boxes 11.2 and 11.3 summarize some important properties of conditional covariances and conditional variances. The rules for conditional variances are special cases of the corresponding rules for conditional covariances with  $Y_1 = Y_2 = Y$  and  $A = B$ , respectively [see Box 11.2 (xiv)]. Hence we only have to prove the rules for the conditional covariances (see Exercise 11-5).

For  $n=2$  variables  $Y_i$  and  $m=2$  variables  $Z_j$ , Equation (xiii) of Box 11.2 can be written

$$\begin{aligned} &Cov(\alpha_1 Y_1 + \alpha_2 Y_2, \beta_1 Z_1 + \beta_2 Z_2 | \mathcal{C}) \\ &\stackrel{\text{P}}{=} \alpha_1 \beta_1 Cov(Y_1, Z_1 | \mathcal{C}) + \alpha_1 \beta_2 Cov(Y_1, Z_2 | \mathcal{C}) + \\ &\quad \alpha_2 \beta_1 Cov(Y_2, Z_1 | \mathcal{C}) + \alpha_2 \beta_2 Cov(Y_2, Z_2 | \mathcal{C}). \end{aligned} \tag{11.46}$$

Similarly, for two random variables  $Y_1$  and  $Y_2$ , Rule (xiii) of Box 11.3 can also be written

$$\begin{aligned} &Var(\alpha_1 Y_1 + \alpha_2 Y_2 | \mathcal{C}) \\ &\stackrel{\text{P}}{=} \alpha_1^2 Var(Y_1 | \mathcal{C}) + \alpha_2^2 Var(Y_2 | \mathcal{C}) + 2\alpha_1 \alpha_2 Cov(Y_1, Y_2 | \mathcal{C}). \end{aligned} \tag{11.47}$$

**Example 11.29 (Conditional Variance of an Indicator)** Let  $(\Omega, \mathcal{A}, P)$  be a probability space, let  $1_A$  denote the indicator variable of  $A \in \mathcal{A}$ , and consider the random variable  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$ . Then, according to Rule (xiv) of Box 11.3,

**Box 11.2 Rules of Computation for  $\mathcal{C}$ -Conditional Covariances**

For  $i = 1, 2$ , let  $Y_i: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  be real-valued random variables with  $E(Y_i^2) < \infty$ . Furthermore, let  $\alpha, \beta \in \mathbb{R}$  and let  $\mathcal{C}_0, \mathcal{C} \subset \mathcal{A}$  be  $\sigma$ -algebras. Then the following properties hold for all real-valued versions of  $E(Y_i | \mathcal{C})$  and all versions of the residuals  $\varepsilon_i := Y_i - E(Y_i | \mathcal{C})$ ,  $i = 1, 2$ :

$$\text{Cov}(Y_1, Y_2 | \mathcal{C}) \stackrel{\text{p}}{=} E(Y_1 \cdot Y_2 | \mathcal{C}) - E(Y_1 | \mathcal{C}) \cdot E(Y_2 | \mathcal{C}). \quad (\text{i})$$

$$\text{Cov}(Y_1, Y_2 | \mathcal{C}) \stackrel{\text{p}}{=} \text{Cov}(\varepsilon_1, \varepsilon_2 | \mathcal{C}). \quad (\text{ii})$$

$$\text{Cov}(Y_1, Y_2 | \mathcal{C}) \stackrel{\text{p}}{=} 0, \quad \text{if } Y_1 \stackrel{\text{p}}{=} \alpha. \quad (\text{iii})$$

$$\text{Cov}(\alpha + Y_1, \beta + Y_2 | \mathcal{C}) \stackrel{\text{p}}{=} \text{Cov}(Y_1, Y_2 | \mathcal{C}). \quad (\text{iv})$$

$$\text{Cov}(\alpha Y_1, \beta Y_2 | \mathcal{C}) \stackrel{\text{p}}{=} \alpha \beta \text{Cov}(Y_1, Y_2 | \mathcal{C}). \quad (\text{v})$$

$$E[\text{Cov}(Y_1, Y_2 | \mathcal{C}) | \mathcal{C}_0] \stackrel{\text{p}}{=} E(\varepsilon_1 \cdot \varepsilon_2 | \mathcal{C}_0) \stackrel{\text{p}}{=} \text{Cov}(Y_1, Y_2 | \mathcal{C}_0), \quad \text{if } \mathcal{C}_0 \subset \mathcal{C}. \quad (\text{vi})$$

$$E[\text{Cov}(Y_1, Y_2 | \mathcal{C})] = E(\varepsilon_1 \cdot \varepsilon_2) = \text{Cov}(\varepsilon_1, \varepsilon_2). \quad (\text{vii})$$

$$\text{Cov}(Y_1, Y_2) = \text{Cov}[E(Y_1 | \mathcal{C}), E(Y_2 | \mathcal{C})] + E[\text{Cov}(Y_1, Y_2 | \mathcal{C})] \quad (\text{viii})$$

$$= \text{Cov}[E(Y_1 | \mathcal{C}), E(Y_2 | \mathcal{C})] + \text{Cov}(\varepsilon_1, \varepsilon_2). \quad (\text{ix})$$

If we additionally assume  $\mathcal{C}_0 \subset \mathcal{C}$ , then

$$\text{Cov}(Y_1, Y_2 | \mathcal{C}_0) \stackrel{\text{p}}{=} \text{Cov}[E(Y_1 | \mathcal{C}), E(Y_2 | \mathcal{C}) | \mathcal{C}_0] + E[\text{Cov}(Y_1, Y_2 | \mathcal{C}) | \mathcal{C}_0] \quad (\text{x})$$

$$\stackrel{\text{p}}{=} \text{Cov}[E(Y_1 | \mathcal{C}), E(Y_2 | \mathcal{C}) | \mathcal{C}_0] + \text{Cov}(\varepsilon_1, \varepsilon_2 | \mathcal{C}_0). \quad (\text{xi})$$

Let  $W_1, W_2, Y_1, Y_2: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  be random variables such that  $E(W_1^4), E(W_2^4), E(Y_1^4), E(Y_2^4) < \infty$ . If  $W_1$  and  $W_2$  are  $\mathcal{C}$ -measurable, then

$$\text{Cov}(W_1 \cdot Y_1, W_2 \cdot Y_2 | \mathcal{C}) \stackrel{\text{p}}{=} W_1 \cdot W_2 \cdot \text{Cov}(Y_1, Y_2 | \mathcal{C}). \quad (\text{xii})$$

For  $i = 1, \dots, n$ ,  $j = 1, \dots, m$ , let  $Y_i, Z_j: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  be random variables with  $E(Y_i^2), E(Z_j^2) < \infty$ , and let  $\alpha_i, \beta_j \in \mathbb{R}$ . Then:

$$\text{Cov}\left(\sum_{i=1}^n \alpha_i Y_i, \sum_{j=1}^m \beta_j Z_j \mid \mathcal{C}\right) \stackrel{\text{p}}{=} \sum_{i=1}^n \sum_{j=1}^m \alpha_i \beta_j \text{Cov}(Y_i, Z_j | \mathcal{C}). \quad (\text{xiii})$$

If  $A, B \in \mathcal{A}$ , then

$$\text{Cov}(1_A, 1_B | \mathcal{C}) \stackrel{\text{p}}{=} P(A \cap B | \mathcal{C}) - P(A | \mathcal{C}) \cdot P(B | \mathcal{C}). \quad (\text{xiv})$$

**Box 11.3 Rules of Computation for Conditional Variances**

Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  denote a real-valued random variable with  $E(Y^2) < \infty$ . Furthermore, let  $\alpha \in \mathbb{R}$  and let  $\mathcal{C} \subset \mathcal{A}$  be a  $\sigma$ -algebra. Then the following properties hold for all real-valued versions of  $E(Y|\mathcal{C})$  and all versions of the residual  $\varepsilon := Y - E(Y|\mathcal{C})$ :

$$\text{Var}(Y|\mathcal{C}) \stackrel{\text{p}}{=} E(Y^2|\mathcal{C}) - E(Y|\mathcal{C})^2 \quad (\text{i})$$

$$\text{Var}(Y|\mathcal{C}) \stackrel{\text{p}}{=} \text{Var}(\varepsilon|\mathcal{C}). \quad (\text{ii})$$

$$\text{Var}(Y|\mathcal{C}) \stackrel{\text{p}}{=} 0, \quad \text{if } Y \stackrel{\text{p}}{=} \alpha \quad (\text{iii})$$

$$\text{Var}(\alpha + Y|\mathcal{C}) \stackrel{\text{p}}{=} \text{Var}(Y|\mathcal{C}) \quad (\text{iv})$$

$$\text{Var}(\alpha Y|\mathcal{C}) \stackrel{\text{p}}{=} \alpha^2 \text{Var}(Y|\mathcal{C}) \quad (\text{v})$$

$$E[\text{Var}(Y|\mathcal{C})|\mathcal{C}_0] \stackrel{\text{p}}{=} E(\varepsilon^2|\mathcal{C}_0) \stackrel{\text{p}}{=} \text{Var}(\varepsilon|\mathcal{C}_0), \quad \text{if } \mathcal{C}_0 \subset \mathcal{C} \quad (\text{vi})$$

$$E[\text{Var}(Y|\mathcal{C})] = E(\varepsilon^2) = \text{Var}(\varepsilon) \quad (\text{vii})$$

$$\text{Var}(Y) = \text{Var}[E(Y|\mathcal{C})] + E[\text{Var}(Y|\mathcal{C})] \quad (\text{viii})$$

$$= \text{Var}[E(Y|\mathcal{C})] + \text{Var}(\varepsilon). \quad (\text{ix})$$

If we additionally assume that the random variable  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  is  $\mathcal{C}$ -measurable, then

$$\text{Var}(Y|X) \stackrel{\text{p}}{=} \text{Var}[E(Y|\mathcal{C})|X] + E[\text{Var}(Y|\mathcal{C})|X] \quad (\text{x})$$

$$\stackrel{\text{p}}{=} \text{Var}[E(Y|\mathcal{C})|X] + \text{Var}(\varepsilon|X). \quad (\text{xi})$$

Let  $X, Y: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  be random variables with  $E(X^4), E(Y^4) < \infty$ . If  $X$  is  $\mathcal{C}$ -measurable, then

$$\text{Var}(X \cdot Y|\mathcal{C}) \stackrel{\text{p}}{=} X^2 \cdot \text{Var}(Y|\mathcal{C}). \quad (\text{xii})$$

Let  $Y_i: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  be random variables with  $E(Y_i^2) < \infty$  and  $\alpha_i \in \mathbb{R}$ , where  $i = 1, \dots, n$ . Then:

$$\text{Var}\left(\sum_{i=1}^n \alpha_i Y_i \mid \mathcal{C}\right) \stackrel{\text{p}}{=} \sum_{i=1}^n \alpha_i^2 \text{Var}(Y_i|\mathcal{C}) + \sum_{i=1}^n \sum_{j=1, j \neq i}^n \alpha_i \alpha_j \text{Cov}(Y_i, Y_j|\mathcal{C}). \quad (\text{xiii})$$

If  $A \in \mathcal{A}$ , then

$$\text{Var}(1_A|\mathcal{C}) \stackrel{\text{p}}{=} P(A|\mathcal{C}) \cdot [1 - P(A|\mathcal{C})]. \quad (\text{xiv})$$

$$\text{Var}(1_A|X) \stackrel{p}{=} P(A|X) \cdot [1 - P(A|X)]. \quad (11.48)$$

According to Corollary 10.29 and 10.26, Equation (11.48) is equivalent to

$$\text{Var}(Y|X=x) = P(Y=1|X=x) \cdot [1 - P(Y=1|X=x)]. \quad \text{for } P_X\text{-a.a. } x \in \Omega'_X. \quad (11.49)$$

Hence, the  $X$ -conditional variance of an indicator variable does not contain any information additional to the  $X$ -conditional expectation  $E(1_A|X) \stackrel{p}{=} P(A|X)$  (cf. Exercise 11-7).  $\triangleleft$

**Remark 11.30 (Two More Properties of Conditional Covariances)** If  $W: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  is a random variable and  $E(W^2) < \infty$ , then

$$\text{Cov}(\varepsilon_1, W | \mathcal{C}_0) \stackrel{p}{=} E(\varepsilon_1 \cdot W | \mathcal{C}_0), \quad \text{if } \mathcal{C}_0 \subset \mathcal{C}. \quad (11.50)$$

and

$$\text{Cov}(Y_1, W | \mathcal{C}) \stackrel{p}{=} 0, \quad \text{if } \sigma(W) \subset \mathcal{C}. \quad (11.51)$$

For a proof see Exercise 11-6.  $\triangleleft$

## 11.6 Partial Correlation

Another concept used to describe a certain kind of dependence between two random variables  $Y_1$  and  $Y_2$  is the *partial correlation*, which is the correlation of the residuals of  $Y_1$  and  $Y_2$  with respect to the conditional expectations  $E(Y_1|\mathcal{C})$  and  $E(Y_2|\mathcal{C})$ , respectively.

### Definition 11.31 (Partial Correlation)

For  $i = 1, 2$ , let  $Y_i: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  be random variables with  $E(Y_i^2) < \infty$ , let  $\mathcal{C} \subset \mathcal{A}$  be a  $\sigma$ -algebra, and define  $\varepsilon_i := Y_i - E(Y_i | \mathcal{C})$  for real-valued versions  $E(Y_i | \mathcal{C})$ . Then we call

$$\text{Corr}(Y_1, Y_2; \mathcal{C}) := \text{Corr}(\varepsilon_1, \varepsilon_2) \quad (11.52)$$

the *partial correlation of  $Y_1$  and  $Y_2$  given  $\mathcal{C}$* . If  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  is a random variable, then we call

$$\text{Corr}(Y_1, Y_2; X) := \text{Corr}[Y_1, Y_2; \sigma(X)] \quad (11.53)$$

the *partial correlation of  $Y_1$  and  $Y_2$  given  $X$* .

**Remark 11.32 (Formulas for the Partial Correlation)** If  $\text{Var}(Y_1), \text{Var}(Y_2) > 0$  and  $R_{Y_1|\mathcal{C}}^2, R_{Y_2|\mathcal{C}}^2 < 1$ , then

$$\text{Corr}(Y_1, Y_2; \mathcal{C}) = \frac{\text{Corr}(Y_1, Y_2) - R_{Y_1|\mathcal{C}} \cdot R_{Y_2|\mathcal{C}} \cdot \text{Corr}[E(Y_1|\mathcal{C}), E(Y_2|\mathcal{C})]}{\sqrt{1 - R_{Y_1|\mathcal{C}}^2} \cdot \sqrt{1 - R_{Y_2|\mathcal{C}}^2}}, \quad (11.54)$$

where  $R_{Y_i|\mathcal{C}}^2 = \text{Var}[E(Y_i|\mathcal{C})]/\text{Var}(Y_i)$ ,  $i = 1, 2$ , denotes the coefficient of determination (see Def. 11.6 and Exercise 11-8). Similarly, if, for  $i = 1, 2$ , there are versions  $E(Y_i|X)$  with  $E(Y_i|X) = \beta_{i0} + \beta_{i1}X$  with slopes  $\beta_{i1} \neq 0$  and  $\text{Corr}(Y_i, X)^2 < 1$ , then Equation (11.54) simplifies to

$$\text{Corr}(Y_1, Y_2; X) = \frac{\text{Corr}(Y_1, Y_2) - \text{Corr}(Y_1, X) \cdot \text{Corr}(Y_2, X)}{\sqrt{1 - \text{Corr}(Y_1, X)^2} \cdot \sqrt{1 - \text{Corr}(Y_2, X)^2}} \quad (11.55)$$

(see Exercise 11-9). ◁

**Remark 11.33 (( $X=x$ )-Conditional Correlation)** If  $P(X=x) > 0$  and, for  $i = 1, 2$ ,  $\text{Var}(Y_i|X=x) > 0$ , then we define

$$\text{Corr}(Y_1, Y_2|X=x) = \frac{\text{Cov}(Y_1, Y_2|X=x)}{\sqrt{\text{Var}(Y_1|X=x)} \cdot \sqrt{\text{Var}(Y_2|X=x)}}, \quad (11.56)$$

and call it the ( $X=x$ )-conditional correlation of  $Y_1$  and  $Y_2$ . If  $\text{Var}(Y_1|X=x) = 0$  or  $\text{Var}(Y_2|X=x) = 0$  we define  $\text{Corr}(Y_1, Y_2|X=x) = 0$ . ◁

**Remark 11.34 (Interpretation of the Partial Correlation)** The definition of the partial correlation  $\text{Corr}(Y_1, Y_2; X)$ , Rule (vii) of Box 11.2, and Rule (vii) of Box 11.3 imply

$$\text{Corr}(Y_1, Y_2; X) = \frac{E[\text{Cov}(Y_1, Y_2|X)]}{\sqrt{E[\text{Var}(Y_1|X)]} \cdot \sqrt{E[\text{Var}(Y_2|X)]}}. \quad (11.57)$$

◁

**Example 11.35 (Joe and Ann With Self-Selection – continued)** We compute the partial correlation  $\text{Corr}(Y, X; U)$  in the example presented in Table 11.2. For this purpose we use Equations (11.52), (11.53), and (7.18). The covariance of the two residuals is

$$\begin{aligned} \text{Cov}(\varepsilon_Y, \varepsilon_X) &= E(\varepsilon_Y \cdot \varepsilon_X) \\ &= (-.704 \cdot (-.04)) \cdot .144 + (.296 \cdot (-.04)) \cdot .336 + (-.704 \cdot .96) \cdot .004 \\ &\quad + (.296 \cdot .96) \cdot .016 + (-.352 \cdot (-.76)) \cdot .096 + (.648 \cdot (-.76)) \cdot .024 \\ &\quad + (-.352 \cdot .24) \cdot .228 + (.648 \cdot .24) \cdot .152 = 0.02016, \end{aligned}$$

the variance of  $\varepsilon_Y$  is

$$\begin{aligned} \text{Var}(\varepsilon_Y) = E(\varepsilon_Y^2) &= (-.704)^2 \cdot .144 + .296^2 \cdot .336 + (-.704)^2 \cdot .004 \\ &\quad + .296^2 \cdot .016 + (-.352)^2 \cdot .096 + .648^2 \cdot .024 \\ &\quad + (-.352)^2 \cdot .228 + .648^2 \cdot .152 = 0.21824, \end{aligned}$$

and the variance of  $\varepsilon_X$  is

$$\begin{aligned} \text{Var}(\varepsilon_X) = E(\varepsilon_X^2) &= (-.04)^2 \cdot .144 + (-.04)^2 \cdot .336 + .96^2 \cdot .004 + .96^2 \cdot .016 \\ &\quad + (-.76)^2 \cdot .096 + (-.76)^2 \cdot .024 + .24^2 \cdot .228 + .24^2 \cdot .152 = 0.1104. \end{aligned}$$

Hence,

$$\text{Corr}(Y, X; U) = \frac{\text{Cov}(\varepsilon_Y, \varepsilon_X)}{\text{SD}(\varepsilon_Y) \cdot \text{SD}(\varepsilon_X)} = \frac{0.02016}{\sqrt{0.21824} \cdot \sqrt{0.1104}} \approx 0.1299,$$

which is a *positive* number. Again, this indicates that using the (unconditional) covariance of  $X$  and  $Y$  or the (unconditional) correlation for the evaluation of the effects of the treatment on success would be highly misleading, because in Example 11.28 we showed that  $\text{Cov}(X, Y)$  is negative although the treatment effects are positive for both persons. For Joe it is  $.80 - .70 = .10$  and for Ann  $.40 - .20 = .20$  (see the column headed  $E(Y|X, U)$  in Table 11.2).  $\triangleleft$

## 11.7 Proofs

### *Proof of Theorem 11.19*

(i) Equation (11.27) can be derived as follows:

$$\begin{aligned} E(Y|X) &\stackrel{p}{=} E[E(Y|X, Z)|X] && \text{[Box 10.2 (v)]} \\ &\stackrel{p}{=} E(\beta_0 + \beta_1 X + \beta_2 Z|X) && \text{[(11.25)]} \\ &\stackrel{p}{=} \beta_0 + \beta_1 X + \beta_2 E(Z|X) && \text{[Box 10.2 (xvi), (vii)]} \\ &\stackrel{p}{=} \beta_0 + \beta_1 X + \beta_2 (\gamma_0 + \gamma_1 X) && \text{[(11.26)]} \\ &\stackrel{p}{=} (\beta_0 + \beta_2 \gamma_0) + (\beta_1 + \gamma_1) X \\ &\stackrel{p}{=} \alpha_0 + \alpha_1 X, \end{aligned}$$

with  $\alpha_0 := \beta_0 + \beta_2 \gamma_0$  and  $\alpha_1 := \beta_1 + \gamma_1$ .

(ii) If  $E(Z|X) \stackrel{p}{=} E(Z)$ , then the third line of the equations above yields Equations (11.28) and (11.29). If  $\beta_2 = 0$ , then the proof of (i) shows that  $E(Y|X) \stackrel{p}{=} \beta_0 + \beta_1 X \stackrel{p}{=} \alpha_0 + \alpha_1 X$  which proves the proposition.

(iii) If  $E(X|Z) \stackrel{p}{=} E(X)$  or  $\beta_1 = 0$ , then

$$\begin{aligned} E(Y|Z) &\stackrel{p}{=} E[E(Y|X, Z)|Z] && \text{[Box 10.2 (v)]} \\ &\stackrel{p}{=} E(\beta_0 + \beta_1 X + \beta_2 Z|Z) && \text{[(11.25)]} \\ &\stackrel{p}{=} \beta_0 + \beta_1 E(X|Z) + \beta_2 Z && \text{[Box 10.2 (i), (xvi), (vii)]} \\ &\stackrel{p}{=} \beta_0 + \beta_1 E(X) + \beta_2 Z. && \text{[}E(X|Z) \stackrel{p}{=} E(X) \text{ or } \beta_1 = 0\text{]} \end{aligned}$$

Hence, our assumption implies

$$\text{Var}[E(Y|Z)] = \beta_2^2 \text{Var}(Z) \tag{11.58}$$

[see Box 6.2 (ii), (iii)]. Now consider

$$\begin{aligned} \text{Var}[E(Y|X, Z)] &= \text{Var}(\beta_0 + \beta_1 X + \beta_2 Z) && [(11.25)] \\ &= \beta_1^2 \text{Var}(X) + \beta_2^2 \text{Var}(Z) + 2\beta_1\beta_2 \text{Cov}(X, Z). && [\text{Box 6.2 (ii), Box 7.1 (viii)}] \end{aligned}$$

Assuming  $[E(Z|X) \stackrel{p}{=} E(Z)$  or  $\beta_2=0]$  and  $[E(X|Z) \stackrel{p}{=} E(X)$  or  $\beta_1=0]$  yields  $2\beta_1\beta_2 \text{Cov}(X, Z) = 0$ , because  $E(Z|X) \stackrel{p}{=} E(Z)$  and  $E(X|Z) \stackrel{p}{=} E(X)$  both imply  $\text{Cov}(X, Z) = 0$  [see Eq. (10.52)]. Hence,

$$\begin{aligned} \text{Var}[E(Y|X, Z)] &= \beta_1^2 \text{Var}(X) + \beta_2^2 \text{Var}(Z) \\ &= \alpha_1^2 \text{Var}(X) + \beta_2^2 \text{Var}(Z) && [(11.29)] \\ &= \text{Var}[E(Y|X)] + \text{Var}[E(Y|Z)], \end{aligned}$$

because  $\text{Var}[E(Y|X)] = \alpha_1^2 \text{Var}(X)$  and  $\text{Var}[E(Y|Z)] = \beta_2^2 \text{Var}(Z)$  [see (i), Eq. (11.58), and Box 6.2 (ii), (iii)]. Now Equation (11.31) follows dividing both sides by  $\text{Var}(Y)$  and using the definition of the coefficient of determination.

## 11.8 Exercises

▷ **Exercise 11-1** Prove the rules of computation for the residual  $\varepsilon := Y - E(Y|\mathcal{C})$  summarized in Box 11.1.

▷ **Exercise 11-2** Show that for  $\varepsilon := Y - E(Y|X, Z)$ , the equations  $\text{Cov}(X, \varepsilon) = \text{Cov}(Z, \varepsilon) = 0$  are special cases of Rule (viii) of Box 11.1 if we consider the conditional expectation  $E(Y|X, Z)$ , assume  $X$  and  $Z$  to be numerical, and  $E(Y^2), E(X^2), E(Z^2) < \infty$ .

▷ **Exercise 11-3** Prove Equations (11.17) and (11.18).

▷ **Exercise 11-4** Show that  $R_{Y|\mathcal{C}}^2 = \text{Corr}[Y, E(Y|\mathcal{C})]^2$ .

▷ **Exercise 11-5** Prove the rules of Box 11.2.

▷ **Exercise 11-6** Prove the propositions of Remark 11.30.

▷ **Exercise 11-7** Show: If  $Y$  is a dichotomous random variable on  $(\Omega, \mathcal{A}, P)$  with values 0 and 1, and we assume  $P(X=x) > 0$ , then  $\text{Var}(Y|X=x) = P(Y=1|X=x) \cdot [1 - P(Y=1|X=x)]$ . Furthermore, if  $P(X=x_1), P(X=x_2) > 0$ ,  $P(Y=1|X=x_1) \neq P(Y=1|X=x_2)$  and  $P(Y=1|X=x_1) \neq 1 - P(Y=1|X=x_2)$ , then  $\text{Var}(Y|X=x_1) \neq \text{Var}(Y|X=x_2)$ .

▷ **Exercise 11-8** Show that Equation (11.54) holds for  $\text{Corr}(Y_1, Y_2; \mathcal{C})$ .

▷ **Exercise 11-9** Prove Equation (11.55).

## Solutions

▷ **Solution 11-1** (i) This rule directly follows from Theorem 10.9 (ii) and Proposition (2.36).

(ii) This rule directly follows from (i) and the assumption that  $Y$  and  $E(Y|\mathcal{C})$  are real-valued.

(v)

$$\begin{aligned} \varepsilon &\stackrel{p}{=} Y - E(Y|\mathcal{C}) && \text{[(i)]} \\ &\stackrel{p}{=} Y - Y = 0. && [Y \stackrel{p}{=} E(Y|\mathcal{C})] \end{aligned}$$

(vi) This rule can be derived as follows:

$$\begin{aligned} E(\varepsilon|\mathcal{C}_0) &\stackrel{p}{=} E[Y - E(Y|\mathcal{C})|\mathcal{C}_0] && \text{[def. of } \varepsilon \text{]} \\ &\stackrel{p}{=} E(Y|\mathcal{C}_0) - E[E(Y|\mathcal{C})|\mathcal{C}_0] && \text{[Box 10.1 (xvi)]} \\ &\stackrel{p}{=} E(Y|\mathcal{C}_0) - E(Y|\mathcal{C}_0) && \text{[Box 10.1 (v)]} \\ &\stackrel{p}{=} 0. \end{aligned}$$

(iii) This rule is a special case of Rule (vi) for  $\mathcal{C}_0 = \{\Omega, \emptyset\}$  (see Rem. 10.5).

(vii) This rule is a special case of Rule (vi) for  $\mathcal{C}_0 := \sigma(W) \subset \mathcal{C}$ .

(viii) Note that  $E(Y^2) < \infty$  implies  $E[E(Y|\mathcal{C})^2] < \infty$  [see Box 10.1 (xi)], which in turn implies  $E(\varepsilon^2) = E(Y^2) + E[E(Y|\mathcal{C})^2] - 2E[Y \cdot E(Y|\mathcal{C})] < \infty$  (see Rem. 7.1).

$$\begin{aligned} \text{Cov}(\varepsilon, W) &= E(\varepsilon \cdot W) - E(\varepsilon) \cdot E(W) && \text{[Box 7.1 (i)]} \\ &= E(\varepsilon \cdot W) - 0 && \text{[(iii)]} \\ &= E[E(\varepsilon \cdot W|\mathcal{C})] && \text{[Box 10.1 (iv)]} \\ &= E[W \cdot E(\varepsilon|\mathcal{C})] && \text{[Box 10.1 (xiv)]} \\ &= E(W \cdot 0) && \text{[(vi)]} \\ &= E(0) = 0. && \text{[Box 6.1 (i)]} \end{aligned}$$

(iv)

$$\begin{aligned} \text{Var}(Y) &= \text{Var}[E(Y|\mathcal{C}) + \varepsilon] && \text{[(ii)]} \\ &= \text{Var}[E(Y|\mathcal{C})] + \text{Var}(\varepsilon) + 2 \cdot \text{Cov}[E(Y|\mathcal{C}), \varepsilon] && \text{[Box 7.1 (viii)]} \\ &= \text{Var}[E(Y|\mathcal{C})] + \text{Var}(\varepsilon). && \text{[(viii)]} \end{aligned}$$

In the last equation we used the fact that  $E(Y|\mathcal{C})$  is  $\mathcal{C}$ -measurable, thus taking the role of  $W$  in Rule (viii) of Box 11.1.

(ix) Note that  $E(Y^2) < \infty$  implies  $E[E(Y|\mathcal{C})^2] < \infty$  [see Box 10.1 (xi)]. Hence,

$$\begin{aligned} \text{Cov}(W, Y) &= \text{Cov}[W, E(Y|\mathcal{C}) + \varepsilon] && \text{[(ii)]} \\ &= \text{Cov}[W, E(Y|\mathcal{C})] + \text{Cov}(W, \varepsilon) && \text{[Box 7.1 (ix)]} \\ &= \text{Cov}[W, E(Y|\mathcal{C})] + 0. && \text{[(viii)]} \end{aligned}$$

▷ **Solution 11-2**  $E(Y|X, Z) \stackrel{p}{=} E(Y|\mathcal{C})$  with  $\mathcal{C} := \sigma(X, Z)$ . Hence,  $X$  and  $Z$  are both  $\mathcal{C}$ -measurable [see Eq. (2.17)] and Rule (viii) of Box 11.1 applies if we assume  $E(Y^2)$ ,  $E(X^2)$ ,  $E(Z^2) < \infty$ .

▷ **Solution 11-3** Equation (11.17) follows from Theorem 7.14 (ii). Now assume  $\text{Var}(X)$ ,  $\text{Var}(Y) > 0$ . Then

$$R_{Y|X}^2 = \frac{\text{Var}[E(Y|X)]}{\text{Var}(Y)} \tag{11.14}$$

$$= \frac{\text{Var}(\beta_0 + \beta_1 X)}{\text{Var}(Y)} \tag{11.16}$$

$$\begin{aligned}
&= \frac{\beta_1^2 \text{Var}(X)}{\text{Var}(Y)} && \text{[Box 6.2 (ii), (iii)]} \\
&= \frac{\left(\frac{\text{Cov}(X, Y)}{\text{Var}(X)}\right)^2 \text{Var}(X)}{\text{Var}(Y)} && \text{[(11.17)]} \\
&= \frac{\text{Cov}(X, Y)^2}{\text{Var}(X) \cdot \text{Var}(Y)} \\
&= \text{Corr}(X, Y)^2. && \text{[(7.18)]}
\end{aligned}$$

If  $\text{Var}(Y) = 0$ , then  $R_{Y|X}^2 = \text{Corr}(X, Y) = 0$  by the definitions of the two terms. If  $\text{Var}(X) = 0$ ,  $\text{Var}(Y) > 0$ , then  $\text{Corr}(X, Y) = 0$  by definition and  $\text{Var}[E(Y|X)] = \text{Var}(\beta_0 + \beta_1 X) = \beta_1^2 \text{Var}(X) = 0$ , and Equation (11.14) implies  $R_{Y|X}^2 = 0$  as well. Hence, in both cases Equation (11.18) holds.

▷ **Solution 11-4** Assume that  $\text{Var}(Y), \text{Var}[E(Y|\mathcal{C})] > 0$ . Then

$$\begin{aligned}
\text{Corr}[Y, E(Y|\mathcal{C})] &= \frac{\text{Cov}[Y, E(Y|\mathcal{C})]}{\text{SD}(Y) \cdot \text{SD}[E(Y|\mathcal{C})]} && \text{[(7.18)]} \\
&= \frac{\text{Var}[E(Y|\mathcal{C})]}{\text{SD}(Y) \cdot \text{SD}[E(Y|\mathcal{C})]} && \text{[Box 10.1 (xii)]} \\
&= \frac{\text{SD}[E(Y|\mathcal{C})] \cdot \text{SD}[E(Y|\mathcal{C})]}{\text{SD}(Y) \cdot \text{SD}[E(Y|\mathcal{C})]} \\
&= \frac{\text{SD}[E(Y|\mathcal{C})]}{\text{SD}(Y)}.
\end{aligned}$$

Squaring both sides and inserting the definition of  $R_{Y|\mathcal{C}}^2$  yields  $R_{Y|\mathcal{C}}^2 = \text{Corr}[Y, E(Y|\mathcal{C})]^2$ . If  $\text{Var}(Y) = 0$ , then  $\text{Corr}[Y, E(Y|\mathcal{C})]^2 = 0 = R_{Y|\mathcal{C}}^2$  by the definitions of  $\text{Corr}[Y, E(Y|\mathcal{C})]$  and  $R_{Y|\mathcal{C}}^2$ . If  $\text{Var}[E(Y|\mathcal{C})] = 0$ , then  $\text{Corr}[Y, E(Y|\mathcal{C})]^2 = 0$  by definition of  $\text{Corr}[Y, E(Y|\mathcal{C})]$ , and  $R_{Y|\mathcal{C}}^2 = 0$ , either because  $\text{Var}[E(Y|\mathcal{C})] = 0$ , or, if  $\text{Var}(Y) = 0$ , by definition of  $R_{Y|\mathcal{C}}^2$ .

▷ **Solution 11-5**

(i)

$$\begin{aligned}
&\text{Cov}(Y_1, Y_2 | \mathcal{C}) \\
&\stackrel{p}{=} E(\varepsilon_1 \cdot \varepsilon_2 | \mathcal{C}) && \text{[(11.33)]} \\
&\stackrel{p}{=} E\left([Y_1 - E(Y_1|\mathcal{C})] \cdot [Y_2 - E(Y_2|\mathcal{C})] | \mathcal{C}\right) && \text{[(11.1)]} \\
&\stackrel{p}{=} E\left[Y_1 \cdot Y_2 - Y_1 \cdot E(Y_2|\mathcal{C}) - E(Y_1|\mathcal{C}) \cdot Y_2 + E(Y_1|\mathcal{C}) \cdot E(Y_2|\mathcal{C}) | \mathcal{C}\right] \\
&\stackrel{p}{=} E(Y_1 \cdot Y_2 | \mathcal{C}) - E[Y_1 \cdot E(Y_2|\mathcal{C}) | \mathcal{C}] - E[E(Y_1|\mathcal{C}) \cdot Y_2 | \mathcal{C}] \\
&\quad + E[E(Y_1|\mathcal{C}) \cdot E(Y_2|\mathcal{C}) | \mathcal{C}] && \text{[Box 10.1 (xvi)]} \\
&\stackrel{p}{=} E(Y_1 \cdot Y_2 | \mathcal{C}) - E(Y_1|\mathcal{C}) \cdot E(Y_2|\mathcal{C}). && \text{[Box 10.1 (xiv)]}
\end{aligned}$$

(ii)

$$\begin{aligned}
\text{Cov}(\varepsilon_1, \varepsilon_2 | \mathcal{C}) &\stackrel{p}{=} E(\varepsilon_1 \cdot \varepsilon_2 | \mathcal{C}) - E(\varepsilon_1 | \mathcal{C}) \cdot E(\varepsilon_2 | \mathcal{C}) && \text{[(i)]} \\
&\stackrel{p}{=} E(\varepsilon_1 \cdot \varepsilon_2 | \mathcal{C}) && \text{[Box 11.1 (vi)]} \\
&\stackrel{p}{=} \text{Cov}(Y_1, Y_2 | \mathcal{C}). && \text{[(11.33)]}
\end{aligned}$$

(iii)

$$\begin{aligned}
\text{Cov}(Y_1, Y_2 | \mathcal{C}) &\stackrel{\text{p}}{=} E(Y_1 \cdot Y_2 | \mathcal{C}) - E(Y_1 | \mathcal{C}) \cdot E(Y_2 | \mathcal{C}) && \text{[(i)]} \\
&\stackrel{\text{p}}{=} E(\alpha Y_2 | \mathcal{C}) - \alpha E(Y_2 | \mathcal{C}) && [Y_1 \stackrel{\text{p}}{=} \alpha, \text{Box 10.3 (iv)}] \\
&\stackrel{\text{p}}{=} \alpha E(Y_2 | \mathcal{C}) - \alpha E(Y_2 | \mathcal{C}) \stackrel{\text{p}}{=} 0. && \text{[Box 10.1 (iii)]}
\end{aligned}$$

(iv)

$$\begin{aligned}
&\text{Cov}(\alpha + Y_1, \beta + Y_2 | \mathcal{C}) \\
&\stackrel{\text{p}}{=} E\left([\alpha + Y_1 - E(\alpha + Y_1 | \mathcal{C})] \cdot [\beta + Y_2 - E(\beta + Y_2 | \mathcal{C})] \mid \mathcal{C}\right) && \text{[(11.33)]} \\
&\stackrel{\text{p}}{=} E\left([Y_1 - E(Y_1 | \mathcal{C})] \cdot [Y_2 - E(Y_2 | \mathcal{C})] \mid \mathcal{C}\right) && \text{[Box 10.1 (ii)]} \\
&\stackrel{\text{p}}{=} E(\varepsilon_1 \cdot \varepsilon_2 | \mathcal{C}) && \text{[(11.1)]} \\
&\stackrel{\text{p}}{=} \text{Cov}(Y_1, Y_2 | \mathcal{C}). && \text{[(11.33)]}
\end{aligned}$$

(v)

$$\begin{aligned}
&\text{Cov}(\alpha Y_1, \beta Y_2 | \mathcal{C}) \\
&\stackrel{\text{p}}{=} E\left([\alpha Y_1 - E(\alpha Y_1 | \mathcal{C})] \cdot [\beta Y_2 - E(\beta Y_2 | \mathcal{C})] \mid \mathcal{C}\right) && \text{[(11.33)]} \\
&\stackrel{\text{p}}{=} E\left([\alpha \cdot [Y_1 - E(Y_1 | \mathcal{C})]] \cdot [\beta \cdot [Y_2 - E(Y_2 | \mathcal{C})]] \mid \mathcal{C}\right) && \text{[Box 10.1 (iii)]} \\
&\stackrel{\text{p}}{=} E(\alpha \varepsilon_1 \cdot \beta \varepsilon_2 | \mathcal{C}) && \text{[(11.1)]} \\
&\stackrel{\text{p}}{=} \alpha \beta E(\varepsilon_1 \cdot \varepsilon_2 | \mathcal{C}) && \text{[Box 10.1 (iii)]} \\
&\stackrel{\text{p}}{=} \alpha \beta \text{Cov}(Y_1, Y_2 | \mathcal{C}). && \text{[(11.33)]}
\end{aligned}$$

(vi)

$$\begin{aligned}
\text{Cov}(Y_1, Y_2 | \mathcal{C}_0) &\stackrel{\text{p}}{=} E(\varepsilon_1 \cdot \varepsilon_2 | \mathcal{C}_0) && \text{[(11.33)]} \\
&\stackrel{\text{p}}{=} E[E(\varepsilon_1 \cdot \varepsilon_2 | \mathcal{C}) | \mathcal{C}_0] && \text{[Box 10.1 (v)]} \\
&\stackrel{\text{p}}{=} E[\text{Cov}(Y_1, Y_2 | \mathcal{C}) | \mathcal{C}_0]. && \text{[(11.33)]}
\end{aligned}$$

(vii)

$$\begin{aligned}
E[\text{Cov}(Y_1, Y_2 | \mathcal{C})] &= E[E(\varepsilon_1 \cdot \varepsilon_2 | \mathcal{C})] && \text{[(11.33)]} \\
&= E(\varepsilon_1 \cdot \varepsilon_2) && \text{[Box 10.1 (iv)]} \\
&= \text{Cov}(\varepsilon_1, \varepsilon_2). && \text{[Box 7.1 (i), Box 11.1 (iii)]}
\end{aligned}$$

(ix), (viii)

$$\begin{aligned}
\text{Cov}(Y_1, Y_2) &= \text{Cov}(E(Y_1 | \mathcal{C}) + \varepsilon_1, E(Y_2 | \mathcal{C}) + \varepsilon_2) && \text{[Box 11.1 (ii)]} \\
&= \text{Cov}(E(Y_1 | \mathcal{C}), E(Y_2 | \mathcal{C})) + \text{Cov}(E(Y_1 | \mathcal{C}), \varepsilon_2) \\
&\quad + \text{Cov}(\varepsilon_1, E(Y_2 | \mathcal{C})) + \text{Cov}(\varepsilon_1, \varepsilon_2) && \text{[Box 7.1 (ix)]} \\
&= \text{Cov}(E(Y_1 | \mathcal{C}), E(Y_2 | \mathcal{C})) + \text{Cov}(\varepsilon_1, \varepsilon_2). && \text{[Box 11.1 (viii)]} \\
&= \text{Cov}(E(Y_1 | \mathcal{C}), E(Y_2 | \mathcal{C})) + E[\text{Cov}(Y_1, Y_2 | \mathcal{C})]. && \text{[(vii)]}
\end{aligned}$$

(xii) For  $i = 1, 2$ , consider the residual of  $W_i \cdot Y_i$  with respect to its  $\mathcal{C}$ -conditional expectation,

$$\begin{aligned} W_i \cdot Y_i - E(W_i \cdot Y_i | \mathcal{C}) &\stackrel{\bar{p}}{=} W_i \cdot Y_i - W_i E(Y_i | \mathcal{C}) && \text{[Box 10.1 (xiv)]} \\ &\stackrel{\bar{p}}{=} W_i \cdot [Y_i - E(Y_i | \mathcal{C})] \stackrel{\bar{p}}{=} W_i \cdot \varepsilon_i. \end{aligned}$$

This equation implies

$$\begin{aligned} \text{Cov}(W_1 \cdot Y_1, W_2 \cdot Y_2 | \mathcal{C}) &\stackrel{\bar{p}}{=} E(W_1 \cdot \varepsilon_1 \cdot W_2 \cdot \varepsilon_2 | \mathcal{C}) && \text{[(11.33)]} \\ &\stackrel{\bar{p}}{=} W_1 \cdot W_2 \cdot E(\varepsilon_1 \cdot \varepsilon_2 | \mathcal{C}) && \text{[Box 10.1 (xiv)]} \\ &\stackrel{\bar{p}}{=} W_1 \cdot W_2 \cdot \text{Cov}(Y_1, Y_2 | \mathcal{C}). && \text{[(11.33)]} \end{aligned}$$

(xiii) Define

$$\varepsilon_i := Y_i - E(Y_i | \mathcal{C}), \quad i = 1, \dots, n, \quad \text{and} \quad \delta_j := Z_j - E(Z_j | \mathcal{C}), \quad j = 1, \dots, m.$$

Then

$$\begin{aligned} \sum_{i=1}^n \alpha_i Y_i - E\left(\sum_{i=1}^n \alpha_i Y_i \mid \mathcal{C}\right) &\stackrel{\bar{p}}{=} \sum_{i=1}^n \alpha_i [Y_i - E(Y_i | \mathcal{C})] && \text{[Box 10.1 (xvi)]} \\ &\stackrel{\bar{p}}{=} \sum_{i=1}^n \alpha_i \varepsilon_i, \end{aligned}$$

and, analogously,

$$\sum_{j=1}^m \beta_j Z_j - E\left(\sum_{j=1}^m \beta_j Z_j \mid \mathcal{C}\right) = \sum_{j=1}^m \beta_j \delta_j.$$

Hence,

$$\begin{aligned} \text{Cov}\left(\sum_{i=1}^n \alpha_i Y_i, \sum_{j=1}^m \beta_j Z_j \mid \mathcal{C}\right) &\stackrel{\bar{p}}{=} E\left[\left(\sum_{i=1}^n \alpha_i \varepsilon_i\right) \cdot \left(\sum_{j=1}^m \beta_j \delta_j\right) \mid \mathcal{C}\right] && \text{[(11.33)]} \\ &\stackrel{\bar{p}}{=} E\left(\sum_{i=1}^n \sum_{j=1}^m \alpha_i \beta_j \cdot \varepsilon_i \delta_j \mid \mathcal{C}\right) \\ &\stackrel{\bar{p}}{=} \sum_{i=1}^n \sum_{j=1}^m \alpha_i \beta_j E(\varepsilon_i \cdot \delta_j | \mathcal{C}) && \text{[Box 10.1 (xvi)]} \\ &\stackrel{\bar{p}}{=} \sum_{i=1}^n \sum_{j=1}^m \alpha_i \beta_j \text{Cov}(Y_i, Z_j | \mathcal{C}). && \text{[(11.33)]} \end{aligned}$$

(x), (xi)

$$\begin{aligned} \text{Cov}(Y_1, Y_2 | \mathcal{C}_0) &= \text{Cov}(E(Y_1 | \mathcal{C}) + \varepsilon_1, E(Y_2 | \mathcal{C}) + \varepsilon_2 | \mathcal{C}_0) && \text{[Box 11.1 (ii)]} \\ &= \text{Cov}(E(Y_1 | \mathcal{C}), E(Y_2 | \mathcal{C}) | \mathcal{C}_0) + \text{Cov}(E(Y_1 | \mathcal{C}), \varepsilon_2 | \mathcal{C}_0) \\ &\quad + \text{Cov}(\varepsilon_1, E(Y_2 | \mathcal{C}) | \mathcal{C}_0) + \text{Cov}(\varepsilon_1, \varepsilon_2 | \mathcal{C}_0) && \text{[(xiii)]} \\ &= \text{Cov}(E(Y_1 | \mathcal{C}), E(Y_2 | \mathcal{C}) | \mathcal{C}_0) + \text{Cov}(\varepsilon_1, \varepsilon_2 | \mathcal{C}_0). && \text{[Box 11.1 (viii)]} \\ &= \text{Cov}(E(Y_1 | \mathcal{C}), E(Y_2 | \mathcal{C}) | \mathcal{C}_0) + E[\text{Cov}(Y_1, Y_2 | \mathcal{C}) | \mathcal{C}_0]. && \text{[(vii)]} \end{aligned}$$

Hence,

$$\text{Cov}(Y_1, Y_2, | \mathcal{C}_0) \stackrel{\bar{p}}{=} \text{Cov}[E(Y_1 | \mathcal{C}), E(Y_2 | \mathcal{C}) | \mathcal{C}_0] + E[\text{Cov}(Y_1, Y_2 | \mathcal{C}) | \mathcal{C}_0]$$

$$\stackrel{\bar{P}}{=} \text{Cov}[E(Y_1|\mathcal{C}), E(Y_2|\mathcal{C})|\mathcal{C}_0] + \text{Cov}(\varepsilon_1, \varepsilon_2|\mathcal{C}_0). \quad \text{[(ii), (vi), (ii)]}$$

(xiv)

$$\begin{aligned} \text{Cov}(1_A, 1_B|\mathcal{C}) &\stackrel{\bar{P}}{=} E(1_A 1_B|\mathcal{C}) - E(1_A|\mathcal{C}) \cdot E(1_B|\mathcal{C}) && \text{[(i)]} \\ &\stackrel{\bar{P}}{=} P(A \cap B|\mathcal{C}) - P(A|\mathcal{C}) \cdot P(B|\mathcal{C}). && \text{[(1.33), (10.2)]} \end{aligned}$$

▷ **Solution 11-6**

(11.50)

$$\begin{aligned} \text{Cov}(\varepsilon_1, W|\mathcal{C}_0) &\stackrel{\bar{P}}{=} E[(\varepsilon_1 - E(\varepsilon_1|\mathcal{C}_0)) \cdot (W - E(W|\mathcal{C}_0))|\mathcal{C}_0] && \text{[(11.33)]} \\ &\stackrel{\bar{P}}{=} E(\varepsilon_1 \cdot (W - E(W|\mathcal{C}_0))|\mathcal{C}_0) && [\mathcal{C}_0 \subset \mathcal{C}, \text{Box 11.1 (vi)}] \\ &\stackrel{\bar{P}}{=} E(\varepsilon_1 \cdot W - \varepsilon_1 \cdot E(W|\mathcal{C}_0)|\mathcal{C}_0) \\ &\stackrel{\bar{P}}{=} E(\varepsilon_1 \cdot W|\mathcal{C}_0) - E(\varepsilon_1 \cdot E(W|\mathcal{C}_0)|\mathcal{C}_0) && \text{[Box 10.1 (xvi)]} \\ &\stackrel{\bar{P}}{=} E(\varepsilon_1 \cdot W|\mathcal{C}_0) - E(\varepsilon_1|\mathcal{C}_0) \cdot E(W|\mathcal{C}_0) && \text{[Box 10.1 (xiv)]} \\ &\stackrel{\bar{P}}{=} E(\varepsilon_1 \cdot W|\mathcal{C}_0). && \text{[Box 11.1 (vi)]} \end{aligned}$$

(11.51)

$$\begin{aligned} &\text{Cov}(Y_1, W|\mathcal{C}) \\ &\stackrel{\bar{P}}{=} E\left[\left(Y_1 - E(Y_1|\mathcal{C})\right) \cdot \left(W - E(W|\mathcal{C})\right)\right]|\mathcal{C} && \text{[(11.33)]} \\ &\stackrel{\bar{P}}{=} E\left(Y_1 - E(Y_1|\mathcal{C})\right)|\mathcal{C} \cdot \left(W - E(W|\mathcal{C})\right) && [\sigma(W) \subset \mathcal{C}, \text{Box 10.1 (xiv)}] \\ &\stackrel{\bar{P}}{=} 0. && \text{[Box 11.1 (vi)]} \end{aligned}$$

▷ **Solution 11-7** If  $P(X=x) > 0$ , then (11.48) and Remark 10.35 imply

$$\text{Var}(Y|X=x) = P(Y=1|X=x) \cdot [1 - P(Y=1|X=x)].$$

Now, define  $a := P(Y=1|X=x_1)$  and  $b := P(Y=1|X=x_2)$ . Then

$$\begin{aligned} a \cdot (1 - a) &= b \cdot (1 - b) \\ \Rightarrow a - a^2 &= b - b^2 \\ \Rightarrow a - b &= (a - b) \cdot (a + b). \end{aligned}$$

If  $a - b = 0$ , then  $a = b$ . Furthermore, if  $a - b \neq 0$ , then  $a + b = 1$  and hence  $a = 1 - b$ . Thus we have shown that  $a \cdot (1 - a) = b \cdot (1 - b)$  implies  $a = b$  or  $a = 1 - b$ . The contraposition of this implication and substituting  $a$  and  $b$  by the conditional probabilities proves the proposition.

▷ **Solution 11-8** Because  $E(Y_1|\mathcal{C})$  and  $E(Y_2|\mathcal{C})$  are  $\mathcal{C}$ -measurable, Rule (ix) of Box 11.1 implies

$$\text{Cov}[Y_1, E(Y_2|\mathcal{C})] = \text{Cov}[E(Y_1|\mathcal{C}), E(Y_2|\mathcal{C})] = \text{Cov}[Y_2, E(Y_1|\mathcal{C})].$$

Therefore,

$$\begin{aligned}
& \text{Cov}(\varepsilon_1, \varepsilon_2) \\
&= \text{Cov}[Y_1 - E(Y_1|\mathcal{C}), Y_2 - E(Y_2|\mathcal{C})] \\
&= \text{Cov}(Y_1, Y_2) + \text{Cov}[E(Y_1|\mathcal{C}), E(Y_2|\mathcal{C})] - \\
&\quad \text{Cov}[Y_1, E(Y_2|\mathcal{C})] - \text{Cov}[Y_2, E(Y_1|\mathcal{C})] \quad \text{[Box 7.1 (ix)]} \\
&= \text{Cov}(Y_1, Y_2) - \text{Cov}[E(Y_1|\mathcal{C}), E(Y_2|\mathcal{C})] \\
&= \text{SD}(Y_1) \cdot \text{SD}(Y_2) \cdot \text{Corr}(Y_1, Y_2) - \text{Cov}[E(Y_1|\mathcal{C}), E(Y_2|\mathcal{C})] \quad \text{[(7.18)]} \\
&= \text{SD}(Y_1) \cdot \text{SD}(Y_2) \cdot \left[ \text{Corr}(Y_1, Y_2) - \frac{\text{Cov}[E(Y_1|\mathcal{C}), E(Y_2|\mathcal{C})]}{\text{SD}(Y_1) \cdot \text{SD}(Y_2)} \right] \\
&= \text{SD}(Y_1) \cdot \text{SD}(Y_2) \cdot \left[ \text{Corr}(Y_1, Y_2) - R_{Y_1|\mathcal{C}} R_{Y_2|\mathcal{C}} \cdot \frac{\text{Cov}[E(Y_1|\mathcal{C}), E(Y_2|\mathcal{C})]}{\text{SD}[E(Y_1|\mathcal{C})] \cdot \text{SD}[E(Y_2|\mathcal{C})]} \right] \quad \text{[(11.11), (11.22)]} \\
&= \text{SD}(Y_1) \cdot \text{SD}(Y_2) \cdot [\text{Corr}(Y_1, Y_2) - R_{Y_1|\mathcal{C}} R_{Y_2|\mathcal{C}} \cdot \text{Corr}[E(Y_1|\mathcal{C}), E(Y_2|\mathcal{C})]].
\end{aligned}$$

Furthermore, for  $i = 1, 2$ ,

$$\begin{aligned}
\text{SD}(\varepsilon_i) &= \sqrt{\text{Var}(\varepsilon_i)} = \sqrt{\text{Var}(Y_i) - \text{Var}[E(Y_i|\mathcal{C})]} \quad \text{[Box 11.1 (iv)]} \\
&= \sqrt{\text{Var}(Y_i) - \text{Var}(Y_i) \cdot R_{Y_i|\mathcal{C}}^2} = \sqrt{\text{Var}(Y_i) \cdot (1 - R_{Y_i|\mathcal{C}}^2)} \quad \text{[(11.11)]} \\
&= \text{SD}(Y_i) \cdot \sqrt{1 - R_{Y_i|\mathcal{C}}^2},
\end{aligned}$$

which implies

$$\text{SD}(\varepsilon_1) \cdot \text{SD}(\varepsilon_2) = \text{SD}(Y_1) \cdot \text{SD}(Y_2) \cdot \sqrt{1 - R_{Y_1|\mathcal{C}}^2} \cdot \sqrt{1 - R_{Y_2|\mathcal{C}}^2}.$$

Using these results, Definition (11.52) yields

$$\begin{aligned}
\text{Corr}(Y_1, Y_2; \mathcal{C}) &:= \text{Corr}(\varepsilon_1, \varepsilon_2) = \frac{\text{Cov}(\varepsilon_1, \varepsilon_2)}{\text{SD}(\varepsilon_1) \cdot \text{SD}(\varepsilon_2)} \\
&= \frac{\text{SD}(Y_1) \cdot \text{SD}(Y_2) \cdot [\text{Corr}(Y_1, Y_2) - R_{Y_1|\mathcal{C}} R_{Y_2|\mathcal{C}} \cdot \text{Corr}[E(Y_1|\mathcal{C}), E(Y_2|\mathcal{C})]]}{\text{SD}(Y_1) \cdot \text{SD}(Y_2) \cdot \sqrt{1 - R_{Y_1|\mathcal{C}}^2} \cdot \sqrt{1 - R_{Y_2|\mathcal{C}}^2}} \\
&= \frac{\text{Corr}(Y_1, Y_2) - R_{Y_1|\mathcal{C}} R_{Y_2|\mathcal{C}} \cdot \text{Corr}[E(Y_1|\mathcal{C}), E(Y_2|\mathcal{C})]}{\sqrt{1 - R_{Y_1|\mathcal{C}}^2} \cdot \sqrt{1 - R_{Y_2|\mathcal{C}}^2}},
\end{aligned}$$

which is Equation (11.54).

▷ **Solution 11-9** If there are  $\beta_{i0}, \beta_{i1} \in \mathbb{R}$ ,  $i = 1, 2$ , such that

$$E(Y_1|X) \stackrel{p}{=} \beta_{10} + \beta_{11}X \quad \text{and} \quad E(Y_2|X) \stackrel{p}{=} \beta_{20} + \beta_{21}X, \quad \beta_{11}, \beta_{21} \neq 0,$$

then

$$E(Y_1|X) \stackrel{p}{=} a_0 + a_1 E(Y_2|X), \quad \text{with} \quad a_0 = \beta_{10} - \beta_{20} \cdot \frac{\beta_{11}}{\beta_{21}} \quad \text{and} \quad a_1 = \frac{\beta_{11}}{\beta_{21}}.$$

Furthermore,  $\beta_{i1} \neq 0$ ,  $\text{Corr}(Y_i, X)^2 < 1$ ,  $i = 1, 2$ , and Equations (7.18) and (11.17) imply  $0 < \text{Corr}(Y_i, X)^2 < 1$ . Now we consider two cases.

Case 1: The slopes  $\beta_{11}$  and  $\beta_{21}$  have *identical signs*. Then  $a_1 > 0$  and the correlations  $\text{Corr}(Y_1, X)$ ,  $\text{Corr}(Y_2, X)$  have identical signs as well [see (11.17)], and

$$R_{Y_1|X} \cdot R_{Y_2|X} = \text{Corr}(Y_1, X) \cdot \text{Corr}(Y_2, X)$$

[see Eqs. (11.18) and (11.22)] and  $\text{Corr}[E(Y_1|X), E(Y_2|X)] = 1$  [see Cor. 7.23]. This implies that Equation (11.54) simplifies to

$$\text{Corr}(Y_1, Y_2; X) = \frac{\text{Corr}(Y_1, Y_2) - \text{Corr}(Y_1, X) \cdot \text{Corr}(Y_2, X)}{\sqrt{1 - \text{Corr}(Y_1, X)^2} \cdot \sqrt{1 - \text{Corr}(Y_2, X)^2}}.$$

Case 2: The slopes  $\beta_{11}$  and  $\beta_{21}$  have *different signs*. Then  $a_1 < 0$  and the correlations  $\text{Corr}(Y_1, X)$ ,  $\text{Corr}(Y_2, X)$  have different signs as well [see (11.17)]. In this case, the same equation holds, because  $R_{Y_1|X} \cdot R_{Y_2|X} = -\text{Corr}(Y_1, X) \cdot \text{Corr}(Y_2, X)$  [see Eqs. (11.18) and (11.22)] and, according to Corollary 7.23,  $\text{Corr}[E(Y_1|X), E(Y_2|X)] = -1$ .



## Chapter 12

# Linear Regression

In chapter 10, we introduced the general concepts of a conditional expectation and a regression, and in chapter 11 we treated the *residual* with respect to a conditional expectation, the concepts of a *conditional variance*, a *conditional covariance*, and a *partial correlation*. Now we turn to *parametrizations* of a conditional expectation. A parametrization serves to describe a conditional expectation with a few parameters (real numbers). Oftentimes these parameters have important meanings that differ between different parametrizations. We treat a *linear parametrization* of a conditional expectation, which is also called the *linear regression*, if it is uniquely defined. We start with the basic ideas, present the definitions, treat some examples, consider the relationship between a linear regression and a linear quasi-regression, and deal with uniqueness of a linear parametrization and the identification of the regression coefficients. Finally, we present a theorem on the invariance of regression coefficients and a theorem on the existence of a linear regression if the regressand and the regressors have a joint multivariate normal distribution.

### 12.1 Basic Ideas

Consider the random variables  $X: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  and  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ , and let  $Y$  be nonnegative or with finite expectation. Furthermore, let  $\mathcal{C} \subset \mathcal{A}$  be a  $\sigma$ -algebra and assume that  $X$  is  $\mathcal{C}$ -measurable. Now assume that there is a real-valued version  $E(Y|\mathcal{C})$  of the  $\mathcal{C}$ -conditional expectation of  $Y$  and coefficients  $\beta_0, \beta_1 \in \mathbb{R}$  such that

$$E(Y|\mathcal{C}) = \beta_0 + \beta_1 X. \quad (12.1)$$

Then we call the function  $g: \mathbb{R} \rightarrow \mathbb{R}$  defined by

$$g(x) := \beta_0 + \beta_1 x, \quad \forall x \in \mathbb{R}, \quad (12.2)$$

a *linear parametrization of  $E(Y|\mathcal{C})$  in  $X$* . This definition implies  $g(X) \in \mathcal{E}(Y|\mathcal{C})$ , where  $g(X)$  denotes the composition of  $X$  and  $g$ .

If  $\mathcal{C} = \sigma(X)$ , then a function  $g: \mathbb{R} \rightarrow \overline{\mathbb{R}}$  satisfying  $E(Y|\mathcal{C}) = E(Y|X) = g(X)$  always exists (see Rem. 10.3 and Cor. 10.23). However,  $g$  is not necessarily a linear function. Hence, even if  $\mathcal{C} = \sigma(X)$ , a *linear parametrization of  $E(Y|\mathcal{C}) = E(Y|X)$  in  $X$*  does *not necessarily exist*. Yet, if we assume that

- (a) Equations (12.1) and (12.2) hold, and
- (b) the variance of  $X$  is positive and finite,

then  $E(Y|X)$  and  $g$ , and therefore, the coefficients  $\beta_0$  and  $\beta_1$ , are uniquely defined (see section 12.5 and cf. Rem. 11.11). Under the assumptions (a) and (b), the function  $g$  is also called the *linear regression* of  $Y$  on  $X$  and the numbers  $\beta_0$  and  $\beta_1$  are called *regression coefficients*.

**Remark 12.1 (Composition of  $X$  and a Linear Function)** Because the conditional expectation  $E(Y|\mathcal{C})$  is a function with domain  $\Omega$ , strictly speaking, it is not a linear function itself. Assuming that Equation (12.1) holds and saying that  $E(Y|\mathcal{C})$  is a linear function of  $X$  we mean that  $E(Y|\mathcal{C})$  is the composition of the random variable  $X: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  and the linear function  $g: \mathbb{R} \rightarrow \mathbb{R}$  satisfying (12.2). This is why  $g$  and not  $E(Y|\mathcal{C})$  itself is called a linear parametrization and a linear regression if the assumptions (a) and (b) hold.  $\triangleleft$

**Remark 12.2 (Estimation)** Although estimation is beyond the scope of this book, it is worthwhile noting that estimation is one of the reasons why a parametrization is useful. The definition of a concrete version of a conditional expectation  $E(Y|X)$  requires that we know for all  $\omega \in \Omega$  which values  $E(Y|X)(\omega)$  are assigned to  $\omega$ . In empirical applications, these values are often *unknown*, i. e., we do not know which concrete number  $E(Y|X)(\omega)$  is assigned to a concrete  $\omega$ . In these cases estimating the values of the conditional expectation may be an issue. In particular, if Equations (12.1) and (12.2) hold and the variance of  $X$  is positive and finite, then the coefficients  $\beta_0$  and  $\beta_1$  — and with them  $E(Y|X) = \beta_0 + \beta_1 X$  — can be computed from estimable quantities such as the variance of  $X$  and the expectations and the covariance of  $X$  and  $Y$ . In this case, estimation of the values of  $E(Y|X)$  is relatively simple because the variance of  $X$  and the expectations and the covariance of  $X$  and  $Y$  can be estimated in a data sample.  $\triangleleft$

**Example 12.3 (Joe and Ann With Self-Selection – continued)** Table 12.1 (cf. also Table 11.2) shows nine random variables, the first five of which may be called *observable* (or *manifest*), whereas the last four are *unobservable* (or *latent*). The difference between the two kinds of random variables is that, in empirical applications, the values of the conditional expectations, the unobservable random variables, are unknown parameters that we might wish to estimate in a sample. These parameters can be computed from the joint distributions of the random variables involved. In this fictitious example, the information about the joint distribution of the random variables  $U$ ,  $X$ , and  $Y$  is contained in the second column of the table, whereas in empirical applications these parameters usually have to be estimated using a data sample. Examples in case are the conditional expectation values  $E(Y|U=Joe, X=0) = .70$  and  $E(Y|X=0) = .60$ . In contrast to the values of the conditional expectations, the values of the five observables are known for all eight possible outcomes  $\omega \in \Omega$  of the random experiment. For example, if  $\omega = (Joe, no, -)$ , then  $U(\omega) = Joe$ ,  $X(\omega) = 0$ , and  $Y(\omega) = 0$ , and these values are known, because the definitions of these observables do not involve unknown parameters that depend on the joint distribution of the random variables involved.

**Table 12.1.** Joe and Ann With Self-Selection: Conditional Expectations

Outcomes $\omega$			Observables					Conditional Expectations				
Unit	Treatment	Success	$P(\{\omega\})$	Person variable $U$	Indicator for Joe $1_{U=Joe}$	Indicator for Ann $1_{U=Ann}$	Treatment variable $X$	Outcome variable $Y$	$E(Y X, U)$	$E(Y X)$	$E(Y U)$	$P(X=1 U)$
(Joe, no, -)			.144	Joe	1	0	0	0	.70	.60	.704	.04
(Joe, no, +)			.336	Joe	1	0	0	1	.70	.60	.704	.04
(Joe, yes, -)			.004	Joe	1	0	1	0	.80	.42	.704	.04
(Joe, yes, +)			.016	Joe	1	0	1	1	.80	.42	.704	.04
(Ann, no, -)			.096	Ann	0	1	0	0	.20	.60	.352	.76
(Ann, no, +)			.024	Ann	0	1	0	1	.20	.60	.352	.76
(Ann, yes, -)			.228	Ann	0	1	1	0	.40	.42	.352	.76
(Ann, yes, +)			.152	Ann	0	1	1	1	.40	.42	.352	.76

In this example, we may consider, for instance, the conditional expectations

$$E(Y|X) = .60 - .18 \cdot X, \tag{12.3}$$

$$E(Y|U) = .352 + .352 \cdot 1_{U=Joe}, \tag{12.4}$$

and

$$E(Y|X, U) = .20 + .20 \cdot X + .50 \cdot 1_{U=Joe} - .10 \cdot X \cdot 1_{U=Joe}. \tag{12.5}$$

The computation of the parameters in Equations (12.3) and (12.5) is illustrated in Examples 12.16 and 12.24.  $\triangleleft$

## 12.2 Assumptions and Definitions

In this section we often refer to the following assumptions and the following notation.

### Notation and Assumptions 12.4

$Y: (\Omega, \mathcal{A}, P) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  and  $X := (X_1, \dots, X_n): (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}^n, \mathcal{B}_n)$  are random variables, where  $Y$  is nonnegative or has a finite expectation  $E(Y)$ . Furthermore,  $\mathcal{C} \subset \mathcal{A}$  is a  $\sigma$ -algebra and  $X$  is  $\mathcal{C}$ -measurable.

Referring to these assumptions we define a linear parametrization as follows:

**Definition 12.5 (Linear Parametrization)**

Let the assumptions 12.4 hold and let  $\beta_0, \beta_1, \dots, \beta_n \in \mathbb{R}$ . If there is a real-valued version  $E(Y|\mathcal{C}) \in \mathcal{E}(Y|\mathcal{C})$  such that

$$E(Y|\mathcal{C}) = \beta_0 + \sum_{i=1}^n \beta_i X_i, \quad (12.6)$$

then the function  $g: \mathbb{R}^n \rightarrow \mathbb{R}$  defined by

$$g(x) := \beta_0 + \sum_{i=1}^n \beta_i x_i, \quad \forall x = (x_1, \dots, x_n) \in \mathbb{R}^n, \quad (12.7)$$

is called a linear parametrization of  $E(Y|\mathcal{C})$  in  $X$ .

Note that, even if it exists, then a linear parametrization of  $E(Y|\mathcal{C})$  is not uniquely defined unless additional assumptions hold (see Example 12.9). Uniqueness of a linear parametrization is treated in Corollary 12.31.

**Remark 12.6 (Another Notation)** If  $\sigma(X) = \mathcal{C}$ , then Equation (12.6) is equivalent to

$$E(Y|X) = \beta_0 + \sum_{i=1}^n \beta_i X_i. \quad (12.8)$$

<

**Remark 12.7 (X-Conditional Mean Independence)** Equation (12.6) implies that  $E(Y|\mathcal{C}) \in \mathcal{E}(Y|X)$  which in turn implies that  $Y$  is  $X$ -conditionally mean independent from  $\mathcal{C}$  (see Def. 10.45). If  $\sigma(X) \neq \mathcal{C}$ , then this conditional mean independence does not necessarily hold. <

**Remark 12.8 (Other Versions and Other Factorizations)** Note that a linear parametrization  $g$  of  $E(Y|\mathcal{C})$  in  $X$  is a factorization of  $E(Y|X)$ , i. e.,  $E(Y|X) = g(X)$  is the composition of  $X$  and  $g$  (see section 10.4). Also note that there may be other factorizations  $g^*$  of  $E(Y|X)$  and versions in  $V^* \in \mathcal{E}(Y|\mathcal{C})$  that do not satisfy Equations (12.6) and (12.7), respectively. However, according to Theorem 10.9 (ii), Equation (12.6) implies

$$V^* \stackrel{p}{=} \beta_0 + \sum_{i=1}^n \beta_i X_i, \quad \forall V^* \in \mathcal{E}(Y|\mathcal{C}). \quad (12.9)$$

Furthermore, if  $g, g^*$  are factorizations of versions  $V, V^* \in \mathcal{E}(Y|X)$ , respectively, then

$$g^* \stackrel{p_X}{=} g \quad (12.10)$$

(see Cor. 10.29).

<

**Example 12.9 (Constant Regressor)** Suppose that  $E(Y|\mathcal{C}) = g(X) = \beta_0 + \beta_1 X$  and  $X$  is a constant function, i. e., there an  $\alpha \in \mathbb{R}$  such that, for all  $\omega \in \Omega$ ,  $X(\omega) = \alpha$ . Then  $\text{Var}(X) = 0$  and  $g, g^*: \mathbb{R} \rightarrow \mathbb{R}$  defined by

$$\forall x \in \mathbb{R}: g(x) = \beta_0 + \beta_1 x \quad \text{and} \quad g^*(x) = (\beta_0 - \gamma\alpha) + (\beta_1 + \gamma)x, \quad 0 \neq \gamma \in \mathbb{R},$$

are two linear parametrizations in  $X$  that differ from each other but satisfy  $g(X) = g^*(X) = E(Y|\mathcal{C})$ . This example shows that a linear parametrization of  $E(Y|\mathcal{C})$  in  $X$  is not uniquely defined unless additional assumptions hold.  $\triangleleft$

**Remark 12.10 (Linear Parametrizations in Different Random Variables)** Consider Equation (12.6), which involves a linear parametrization  $g$  of  $E(Y|\mathcal{C})$  in  $X$ . If  $Z = (Z_1, \dots, Z_m): \Omega \rightarrow \mathbb{R}^m$  is an  $m$ -variate random variable on  $(\Omega, \mathcal{A}, P)$ , and  $\sigma(Z) \subset \mathcal{C}$ , then for one and the same version  $E(Y|\mathcal{C}) \in \mathcal{E}(Y|\mathcal{C})$ , there may also be a linear parametrization  $f: \mathbb{R}^m \rightarrow \mathbb{R}$  of  $E(Y|\mathcal{C})$  in  $Z$  with coefficients  $\gamma_0, \gamma_1, \dots, \gamma_m \in \mathbb{R}$  satisfying

$$f(z) = \gamma_0 + \sum_{i=1}^m \gamma_i z_i, \quad \forall z = (z_1, \dots, z_m) \in \mathbb{R}^m. \quad (12.11)$$

In other words, one and the same version  $E(Y|\mathcal{C}) \in \mathcal{E}(Y|\mathcal{C})$  may have several parametrizations such as  $f$  and  $g$  that are linear in  $Z = (Z_1, \dots, Z_m)$  and  $X = (X_1, \dots, X_n)$ , respectively. Note that the regression coefficients  $\gamma_0, \gamma_1, \dots, \gamma_m \in \mathbb{R}$  and  $\beta_0, \beta_1, \dots, \beta_n \in \mathbb{R}$  pertaining to the two parametrizations  $f$  and  $g$  may differ from each other.  $\triangleleft$

**Remark 12.11 (Conditional Expectation Values)** If  $g$  is a linear parametrization of  $E(Y|X)$  in  $X$  satisfying Equation (12.7), then, according to Definition 10.33,

$$\begin{aligned} E(Y|X=x) &= E(Y|X_1=x_1, \dots, X_n=x_n) \\ &= g(x) = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n, \quad \forall x = (x_1, \dots, x_n) \in \mathbb{R}^n. \end{aligned} \quad (12.12)$$

Note that another factorization  $g^*$  of  $E(Y|X)$  might yield another conditional expectation value  $E(Y|X=x)$  if  $P(X=x) = 0$ .  $\triangleleft$

**Remark 12.12 ( $P_X$ -Equivalence of Different Parametrizations)** If  $g, g^*$  are factorizations of versions  $V, V^* \in \mathcal{E}(Y|X)$ , then, according to Equation (12.10),

$$g(x) = g^*(x), \quad \text{for } P_X\text{-a.a. } x \in \mathbb{R}^n, \quad (12.13)$$

[see Eq. (10.26)].  $\triangleleft$

## 12.3 Examples

**Example 12.13 (Univariate Real-Valued  $X$ )** If  $X: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  and

$$E(Y|X) = \beta_0 + \beta_1 X, \quad (12.14)$$

then the function  $g: \mathbb{R} \rightarrow \mathbb{R}$  defined by

$$g(x) = \beta_0 + \beta_1 x, \quad x \in \mathbb{R}, \quad (12.15)$$

is a linear parametrization of  $E(Y|X)$  in  $X$ . If  $0 < \text{Var}(X) < \infty$ , then  $g$  is uniquely defined and also called the *simple linear regression* of  $Y$  on  $X$  (see Cor. 12.31 and Rem. 12.34).  $\triangleleft$

**Example 12.14 (Intercept and Slope)** If  $E(Y|X) = \beta_0 + \beta_1 X$ , then

$$\beta_0 = E(Y|X=0). \quad (12.16)$$

Furthermore, if  $x_1, x_2 \in \mathbb{R}$  and  $x_2 > x_1$ , then

$$\beta_1 = \frac{1}{x_2 - x_1} \cdot [E(Y|X=x_2) - E(Y|X=x_1)] \quad (12.17)$$

[see Exercise 7-2 and Eq. (12.12)]. Equation (12.17) yields

$$\beta_1 = E(Y|X=x_2) - E(Y|X=x_1), \quad \text{if } x_2 - x_1 = 1. \quad (12.18)$$

This justifies calling  $\beta_0$  the *intercept* and  $\beta_1$  the *slope* of  $E(Y|X)$ , respectively (see Fig. 7.3). Note that these equations also apply if  $P(X=0) = P(X=x_1) = P(X=x_2) = 0$ . They even apply if  $0, x_1, x_2 \notin X(\Omega)$ .  $\triangleleft$

**Example 12.15 (Dichotomous Regressor)** If  $X$  is dichotomous with values 0 and 1 (see Example 5.10), then there is always a version  $E(Y|X) \in \mathcal{E}(Y|X)$  such that

$$E(Y|X) = \beta_0 + \beta_1 X \quad (12.19)$$

with

$$\beta_0 = E(Y|X=0), \quad (12.20)$$

and

$$\beta_1 = E(Y|X=1) - E(Y|X=0) \quad (12.21)$$

(for a proof see Th. 12.37).  $\triangleleft$

**Example 12.16 (Joe and Ann With Self-Selection – continued)** In Table 12.1,  $X$  is dichotomous. According to Example 12.15, this implies that there is a linear parametrization of  $E(Y|X)$  in  $X$  satisfying Equation (12.19). In this example,

$$E(Y|X) = .60 - .18 \cdot X, \quad (12.22)$$

and the function  $g: \mathbb{R} \rightarrow \mathbb{R}$  defined by  $g(x) = .60 - .18 \cdot x$ ,  $x \in \mathbb{R}$ , is a linear parametrization of  $E(Y|X)$  in  $X$ .

The intercept is most easily obtained via Equation (12.20):

$$\begin{aligned} \beta_0 &= E(Y|X=0) = P(Y=1|X=0) \\ &= \frac{P(Y=1, X=0)}{P(X=0)} = \frac{.336 + .024}{.144 + .336 + .096 + .024} = .60. \end{aligned} \quad (12.23)$$

The slope is obtained via Equation (12.21):

$$\begin{aligned}\beta_1 &= E(Y|X=1) - E(Y|X=0) = P(Y=1|X=1) - P(Y=1|X=0) \\ &= \frac{P(Y=1, X=1)}{P(X=1)} - P(Y=1|X=0) \\ &= \frac{.016 + .152}{.004 + .016 + .228 + .152} - .60 = .42 - .60 = -.18.\end{aligned}\tag{12.24}$$

◁

**Example 12.17 (Dichotomous Regressor – continued)** Continue Example 12.15 and define the random variable  $Z: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  by  $Z := 2X - 1$ . Then  $Z$  is dichotomous with values  $-1$  and  $1$ , and

$$\forall \omega \in \Omega: X(\omega) = 0 \Leftrightarrow Z(\omega) = -1 \quad \text{and} \quad X(\omega) = 1 \Leftrightarrow Z(\omega) = 1.$$

Note that  $\sigma(Z) = \sigma(X)$  holds for the  $\sigma$ -algebras generated by  $X$  and  $Z$ . Because  $X = \frac{1}{2}(Z+1)$ ,

$$E(Y|X) = \beta_0 + \beta_1 X = \beta_0 + \frac{\beta_1}{2} \cdot (Z+1) = \beta_0 + \frac{\beta_1}{2} + \frac{\beta_1}{2} \cdot Z = E(Y|Z).\tag{12.25}$$

the function  $g^*: \mathbb{R} \rightarrow \mathbb{R}$  defined by  $g^*(z) = \alpha_0 + \alpha_1 z$ ,  $z \in \mathbb{R}$ , is a linear parametrization of  $E(Y|X)$  in  $Z$ , where

$$\begin{aligned}\alpha_0 &= \beta_0 + \frac{\beta_1}{2} = E(Y|X=0) + \frac{E(Y|X=1) - E(Y|X=0)}{2} \\ &= \frac{E(Y|X=1) + E(Y|X=0)}{2}\end{aligned}\tag{12.26}$$

and

$$\alpha_1 = \frac{\beta_1}{2} = \frac{E(Y|X=1) - E(Y|X=0)}{2}.\tag{12.27}$$

Note that

$$\begin{aligned}\{X=1\} &= \left\{ \frac{1}{2}(Z+1) = 1 \right\} = \{Z=1\}, \\ \{X=0\} &= \left\{ \frac{1}{2}(Z+1) = 0 \right\} = \{Z=-1\}.\end{aligned}\tag{12.28}$$

Because  $X$  and  $Z$  are dichotomous with positive probabilities for both of their values, Equation (12.28) and  $E(Y|X=x) = E(Y|\{X=x\})$  [see Eq. (9.6)] imply

$$\alpha_0 = \frac{E(Y|Z=1) + E(Y|Z=-1)}{2},\tag{12.29}$$

$$\alpha_1 = \frac{E(Y|Z=1) - E(Y|Z=-1)}{2}.\tag{12.30}$$

Comparing Equations (12.26) and (12.27) to Equations (12.20) and (12.21) shows that the meaning of the regression coefficients depends on the choice of the random variable – here  $X$  or  $Z$  – with respect to which we consider a linear parametrization. ◁

**Example 12.18 (Joe and Ann With Self-Selection–continued)** Using the results of Example 12.16 as well as Equations (12.25) to (12.27) yields

$$E(Y|X) = E(Y|Z) = \frac{.42 + .60}{2} + \frac{.42 - .60}{2} = .51 - .09 \cdot Z. \quad (12.31)$$

Note again that  $E(Y|X)$  and  $E(Y|Z)$  are only different notations for a version of  $\mathcal{E}(Y|\mathcal{C})$ , where  $\mathcal{C} = \sigma(X) = \sigma(Z)$  and that  $g: \mathbb{R} \rightarrow \mathbb{R}$  defined by  $g(x) = .60 - .18 \cdot x$ ,  $x \in \mathbb{R}$ , and  $g^*: \mathbb{R} \rightarrow \mathbb{R}$  defined by  $g^*(z) = .51 - .09 \cdot z$ ,  $z \in \mathbb{R}$ , are two different linear parametrizations of one and the same version  $E(Y|\mathcal{C}) \in \mathcal{E}(Y|\mathcal{C})$ , one is linear in  $X$ , the other one linear in  $Z$ .  $\triangleleft$

**Example 12.19 (Quadratic Function)** Let  $X_1: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  be a real-valued random variable, let  $X_2 := X_1^2$ ,  $X := (X_1, X_2)$ , and assume that there is a version  $E(Y|X) \in \mathcal{E}(Y|X)$  with

$$E(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2. \quad (12.32)$$

Then the function  $g: \mathbb{R}^2 \rightarrow \mathbb{R}$  defined by

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2, \quad x = (x_1, x_2) \in \mathbb{R}^2, \quad (12.33)$$

is a linear parametrization of  $E(Y|X)$  in  $X = (X_1, X_1^2)$ .  $\triangleleft$

**Example 12.20 (Logarithmic Function)** Consider  $Z: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$ , a real-valued and positive random variable, define  $X := \ln Z$ , and assume that there is a version  $E(Y|X) \in \mathcal{E}(Y|X)$  with

$$E(Y|X) = \beta_0 + \beta_1 \ln Z = \beta_0 + \beta_1 X. \quad (12.34)$$

Then the function  $g: \mathbb{R} \rightarrow \mathbb{R}$  defined by

$$g(x) = \beta_0 + \beta_1 x, \quad x \in \mathbb{R}, \quad (12.35)$$

is a linear parametrization of  $E(Y|X)$  in  $X = \ln Z$ . Note that  $\mathcal{E}(Y|Z) = \mathcal{E}(Y|X)$ , because  $\sigma(X) = \sigma(Z)$  (see Exercise 12-1).  $\triangleleft$

**Example 12.21 (Two Regressors)** If  $X_i: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$ ,  $i = 1, 2$ , are univariate real-valued random variables,  $X := (X_1, X_2)$ , and there is a version  $E(Y|X) \in \mathcal{E}(Y|X)$  with

$$E(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2, \quad (12.36)$$

then the function  $g: \mathbb{R}^2 \rightarrow \mathbb{R}$  defined by

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2, \quad x = (x_1, x_2) \in \mathbb{R}^2, \quad (12.37)$$

is a linear parametrization of  $E(Y|X)$  in  $X = (X_1, X_2)$ .  $\triangleleft$

**Example 12.22 (Two Regressors and Their Product)** Let  $X_i: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$ ,  $i = 1, 2$ , be univariate real-valued random variables, define  $X_3 := X_1 \cdot X_2$  and  $X := (X_1, X_2, X_3)$ , and assume that there is a version  $E(Y|X) \in \mathcal{E}(Y|X)$  with

$$E(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 \cdot X_2. \tag{12.38}$$

Then the function  $g: \mathbb{R}^3 \rightarrow \mathbb{R}$  defined by

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3, \quad \forall x = (x_1, x_2, x_3) \in \mathbb{R}^3, \tag{12.39}$$

is a linear parametrization of  $E(Y|X)$  in  $X = (X_1, X_2, X_1 \cdot X_2)$ . ◁

**Remark 12.23 (Dichotomous Random Variables)** Note that such a linear parametrization of  $E(Y|X)$  always exists, if  $X_1$  and  $X_2$  are dichotomous with values 0 and 1 (see Example 5.10 and Exercise 12-2). ◁

**Example 12.24 (Joe and Ann With Self-Selection – continued)** Consider the random variables  $X$  and  $1_{U=Joe}$  specified in Table 12.1 and define  $Z := (Z_1, Z_2, Z_3) := (X, 1_{U=Joe}, X \cdot 1_{U=Joe})$ . Then, according to Example 12.22 and Remark 12.23,

$$E(Y|Z) = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot 1_{U=Joe} + \beta_3 \cdot X \cdot 1_{U=Joe}, \tag{12.40}$$

and the function  $g: \mathbb{R}^3 \rightarrow \mathbb{R}$  defined by

$$g(z) = .20 + .20 \cdot z_1 + .50 \cdot z_2 - .10 \cdot z_3, \quad \forall z = (z_1, z_2, z_3) \in \mathbb{R}^3,$$

is a linear parametrization of  $E(Y|Z)$  in  $Z = (X, 1_{U=Joe}, X \cdot 1_{U=Joe})$ .

The coefficients  $\beta_0$  to  $\beta_3$  in Equation (12.40) can also be obtained as follows (which is analogous to Exercise 12-2). Table 12.1 and Equation (12.40) yield:

$$\begin{aligned} \beta_0 &= E(Y|Z_1=0, Z_2=0, Z_3=0) = P(Y=1|X=0, U=Ann) = .20, \\ \beta_0 + \beta_1 &= E(Y|Z_1=1, Z_2=0, Z_3=0) = P(Y=1|X=1, U=Ann) = .40, \\ \beta_0 + \beta_2 &= E(Y|Z_1=0, Z_2=1, Z_3=0) = P(Y=1|X=0, U=Joe) = .70, \\ \beta_0 + \beta_1 + \beta_2 + \beta_3 &= E(Y|Z_1=1, Z_2=1, Z_3=1) = P(Y=1|X=1, U=Joe) = .80. \end{aligned}$$

Solving these equations for the four coefficients and inserting them into Equation (12.40) yields

$$E(Y|Z) = .20 + .20 \cdot X + .50 \cdot 1_{U=Joe} - .10 \cdot X \cdot 1_{U=Joe}. \tag{12.41}$$

In this example,  $\sigma(Z) = \sigma(X, U)$  and  $Z(\Omega) = \{0, 1\}^3$ . According to Remark 10.12, this implies  $E(Y|Z) = E(Y|X, U)$  and that the function  $g$  is also a linear parametrization of  $E(Y|X, U)$  in  $Z = (X, 1_{U=Joe}, X \cdot 1_{U=Joe})$ . ◁

**Remark 12.25 (Generalizing the Examples)** Generalizing the Examples 12.19 to 12.24, let  $Z: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Z, \mathcal{A}'_Z)$  be a (univariate or multivariate) random variable. For  $i = 1, \dots, n$ , let  $h_i: (\Omega'_Z, \mathcal{A}'_Z) \rightarrow (\mathbb{R}, \mathcal{B})$  be measurable functions, define  $X_i := h_i(Z)$ , and  $X := (X_1, \dots, X_n)$ . If there is a version  $E(Y|X) \in \mathcal{E}(Y|X)$  such that

$$E(Y|X) = \beta_0 + \sum_{i=1}^n \beta_i h_i(Z), \quad (12.42)$$

then  $g: \mathbb{R}^n \rightarrow \mathbb{R}$  defined by

$$g(x) = \beta_0 + \sum_{i=1}^n \beta_i x_i, \quad \forall x = (x_1, \dots, x_n) \in \mathbb{R}^n, \quad (12.43)$$

is a linear parametrization of  $E(Y|X)$  in  $X = (h_1(Z), \dots, h_n(Z))$ . Remember that  $\sigma(X) \subset \sigma(Z)$ , but note that  $\sigma(X) = \sigma(Z)$  does not necessarily hold. If  $\sigma(X) \neq \sigma(Z)$ , then there is not necessarily a version  $E(Y|Z) \in \mathcal{E}(Y|Z)$  with  $E(Y|Z) = E(Y|X)$ . However, if we assume that there is a version  $E(Y|Z) \in \mathcal{E}(Y|Z)$  with  $E(Y|Z) = E(Y|X)$ , then

$$E(Y|Z) = \beta_0 + \sum_{i=1}^n \beta_i h_i(Z). \quad (12.44)$$

In this case, the function  $g$  is also a linear parametrization of  $E(Y|Z)$  in  $X = (h_1(Z), \dots, h_n(Z))$ .  $\triangleleft$

## 12.4 Linear Quasi-Regression

In the following corollary  $Q_{lin}(Y|X_1, \dots, X_n)$  denotes the function that has been introduced in Definition 7.28. This corollary immediately follows from Theorem 10.26 and Definition 12.5.

### Corollary 12.26 (Linear Regression and Linear Quasi-Regression)

Let the assumptions 12.4 hold and suppose that  $E(Y^2), E(X_i^2) < \infty, i = 1, \dots, n$ . If there is a version  $E(Y|\mathcal{C}) \in \mathcal{E}(Y|\mathcal{C})$  with  $E(Y|\mathcal{C}) = \beta_0 + \sum_{i=1}^n \beta_i X_i$ , where  $\beta_0, \beta_1, \dots, \beta_n \in \mathbb{R}$ , then

$$E(Y|\mathcal{C}) = Q_{lin}(Y|X_1, \dots, X_n) = \beta_0 + \sum_{i=1}^n \beta_i X_i. \quad (12.45)$$

Hence, if  $E(Y|\mathcal{C}) = \beta_0 + \sum_{i=1}^n \beta_i X_i$ , then  $E(Y|\mathcal{C})$  and  $Q_{lin}(Y|X_1, \dots, X_n) = f(X)$  are identical, where  $f(X)$  is the composition of  $X$  and the linear quasi-regression  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  (see Def. 7.28).

**Remark 12.27 (Dichotomous Regressor)** If  $X$  is dichotomous with values 0 and 1 (see Example 5.10), then  $E(Y|X) = Q_{lin}(Y|X)$  (see Example 12.15 and Th. 12.37). If  $P(X=x_i) > 0$  for at least three different  $x_i \in \mathbb{R}$ , then it is not necessarily true that  $Q_{lin}(Y|X) \in \mathcal{E}(Y|X)$ . If  $Q_{lin}(Y|X) \notin \mathcal{E}(Y|X)$ , then there are no  $\beta_0, \beta_1 \in \mathbb{R}$  such that the function  $g$  defined by  $g(x) = \beta_0 + \beta_1 x, x \in \mathbb{R}$ , is a linear parametrization in  $X$  of a version  $E(Y|X) \in \mathcal{E}(Y|X)$  (see Example 12.41).  $\triangleleft$

**Remark 12.28 (Unbounded Regressor, Dichotomous Regressand)** Suppose that  $X$  and  $Y$  are real-valued random variables on  $(\Omega, \mathcal{A}, P)$ , and that  $Y$  is dichotomous with values 0 and 1. Because  $P(0 \leq Y \leq 1) = 1$ , Rules (ii) and (iii) of Box 10.3 imply

$$0 \leq E(Y|X) = P(Y=1|X) \leq 1. \tag{12.46}$$

Suppose that the regressor  $X$  is *not*  $P$ -almost surely bounded, i. e., suppose

$$\forall c \in \mathbb{R}, c > 0: P(X < -c) + P(X > c) > 0. \tag{12.47}$$

Then there is no linear parametrization of  $E(Y|X)$  in  $X$  with slope  $\beta_1 \neq 0$  (see Exercise 12-3 and cf. ch. 13). Note that the premise (12.47) holds, e. g., if  $X$  has a normal distribution.  $\triangleleft$

**Example 12.29 (No Treatment for Joe – continued)** In the example presented in Table 9.2 (p. 289),

$$E(Y|X) = \beta_0 + \beta_1 X = .6 - .2 \cdot X \tag{12.48}$$

(see Example 12.15) and

$$g(x) = \beta_0 + \beta_1 \cdot x = .6 - .2 \cdot x, \quad \forall x \in \mathbb{R}, \tag{12.49}$$

defines a linear parametrization  $g: \mathbb{R} \rightarrow \mathbb{R}$  of  $E(Y|X)$  in  $X$ . Hence, according to Remark 12.11, we may define

$$E(Y|X=x) = g(x) = .6 - .2 \cdot x, \quad \forall x \in \mathbb{R}, \tag{12.50}$$

(see Def. 10.33). For  $x=0$ , Equation (12.50) yields  $E(Y|X=0) = .6$ , and it yields  $E(Y|X=1) = .4$  for  $x=1$ . Note that the definition of the conditional expectation values  $E(Y|X=x)$  for  $x \in \mathbb{R} \setminus \{0, 1\}$  via Equation (12.50) is arbitrary, because  $P_X(X \in \mathbb{R} \setminus \{0, 1\}) = 0$ . Using any other factorization of  $E(Y|X)$  for the definition of the conditional expectation values  $E(Y|X=x)$  for  $x \in \mathbb{R} \setminus \{0, 1\}$  would do as well.

Remark 12.27 and Definition 7.28 imply that the function  $MSE: \mathbb{R}^2 \rightarrow [0, \infty[$  defined by

$$MSE(a_0, a_1) = E([Y - (a_0 + a_1 X)]^2), \quad \forall (a_0, a_1) \in \mathbb{R}^2, \tag{12.51}$$

has its minimum for  $(a_0, a_1) = (.6, -.2)$ . Hence, in this example,

$$E(Y|X) = Q_{lin}(Y|X) = .6 - .2 \cdot X,$$

and the function  $g: \mathbb{R} \rightarrow \mathbb{R}$  defined by Equation (12.49) is a linear parametrization of  $E(Y|X)$  in  $X$ .  $\triangleleft$

## 12.5 Uniqueness and Identification of Regression Coefficients

In Example 12.9 we showed that a linear parametrization is not uniquely defined unless additional assumptions hold. Such assumptions are specified in Corollary 12.31, which is uses the following notation and general assumptions:

**Notation and Assumptions 12.30**

Let the assumptions 12.4 hold. Furthermore,  $\mathbf{x} := [X_1, \dots, X_n]'$  is the column vector of  $X = (X_1, \dots, X_n)$ ,  $\boldsymbol{\mu} := [E(X_1), \dots, E(X_n)]'$  the column vector of the expectations of  $X_1, \dots, X_n$ , and  $\boldsymbol{\beta} := [\beta_1, \dots, \beta_n]'$  a column vector of  $n$  real numbers.

Assuming finite second moments of  $Y$  and  $X_1, \dots, X_n$ ,

$$\boldsymbol{\Sigma}_{xx} = E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})'] = \begin{bmatrix} \sigma_{X_1}^2 & \sigma_{X_1 X_2} & \cdots & \sigma_{X_1 X_n} \\ \sigma_{X_2 X_1} & \sigma_{X_2}^2 & \cdots & \sigma_{X_2 X_n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{X_n X_1} & \sigma_{X_n X_2} & \cdots & \sigma_{X_n}^2 \end{bmatrix}$$

denotes the variance-covariance matrix of  $X = (X_1, \dots, X_n)$  (see section 7.4.3). Furthermore,

$$\boldsymbol{\Sigma}_{yx} = [\sigma_{YX_1}, \dots, \sigma_{YX_n}]$$

denotes the row vector of the covariances  $Cov(Y, X_i) = \sigma_{YX_i}$ ,  $i = 1, \dots, n$ , and  $\boldsymbol{\Sigma}_{xy} := \boldsymbol{\Sigma}_{yx}'$  the column vector of these covariances. Remember that the notation  $X = (X_1, \dots, X_n)$  refers to an  $n$ -variate random variable, whereas  $\mathbf{x} = [X_1, \dots, X_n]$  denotes the row vector of the random variables  $X_1, \dots, X_n$ .

The following corollary immediately follows from Theorem 7.30 and Corollary 12.26. It shows how to compute (identify) the regression coefficients of a linear parametrization of  $E(Y|\mathcal{C})$  in  $X$  and it specifies sufficient conditions under which such a linear parametrization of  $E(Y|\mathcal{C})$  is uniquely defined.

**Corollary 12.31 (Identification of Parameters)**

Let the assumptions 12.30 hold. If there is a version  $E(Y|\mathcal{C}) \in \mathcal{E}(Y|\mathcal{C})$  with

$$E(Y|\mathcal{C}) = \beta_0 + \boldsymbol{\beta}'\mathbf{x} = \beta_0 + \sum_{i=1}^n \beta_i X_i, \quad (12.52)$$

then

$$\beta_0 = E(Y) - \boldsymbol{\beta}'\boldsymbol{\mu}. \quad (12.53)$$

If, in addition,  $Y$  and  $X_1, \dots, X_n$  have finite second moments and the inverse  $\boldsymbol{\Sigma}_{xx}^{-1}$  exists, then

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}_{xx}^{-1} \boldsymbol{\Sigma}_{xy}, \quad (12.54)$$

and the linear parametrization  $g: \mathbb{R}^n \rightarrow \mathbb{R}$  of  $E(Y|\mathcal{C})$  in  $X$  with

$$g(x) = \beta_0 + \sum_{i=1}^n \beta_i x_i, \quad \forall x = (x_1, \dots, x_n) \in \mathbb{R}^n, \quad (12.55)$$

is uniquely defined.

$$\begin{array}{c}
 \Omega \xrightarrow{X} \mathbb{R}^n \xrightarrow{g \text{ (linear regression)}} \mathbb{R} \\
 \underbrace{\hspace{10em}} \\
 E(Y|X) = g(X)
 \end{array}$$

**Figure 12.1.**  $E(Y|X)$  as the composition of  $X$  and the linear regression  $g$

## 12.6 Linear Regression

As already mentioned in the Preface, much empirical research uses some kind of regression in order to investigate how the expectation of one random variable depends on the values of one or more other random variables. In Definition 10.25 we introduced the concept of a regression as a special case of a factorization of a conditional expectation in which the regressor  $X$  is numerical. In Definition 12.5 we defined the concept of a parametrization of  $E(Y|\mathcal{C})$  that is linear in  $X$ . Such parametrizations of  $E(Y|\mathcal{C})$  are not necessarily uniquely defined. If there is a parametrization  $g$  of  $E(Y|\mathcal{C})$  that is linear in  $X$  and satisfies Equation (12.55), then we call it *the linear regression of  $Y$  on  $X$* . According to Corollary 12.31, the linear regression is uniquely defined.

### Definition 12.32 (Linear Regression)

Let the assumptions 12.30 hold and suppose that there is a version  $E(Y|\mathcal{C}) \in \mathcal{E}(Y|\mathcal{C})$  such that Equation (12.52) holds. Furthermore, assume that  $Y$  and  $X_1, \dots, X_n$  have finite second moments and that the inverse  $\Sigma_{xx}^{-1}$  exists. Then the function  $g: \mathbb{R}^n \rightarrow \mathbb{R}$  defined by Equation (12.55) is called *the linear regression of  $Y$  on  $X$  (with respect to  $P$ )*.

Figure 12.1 shows the conditional expectation  $E(Y|X)$  as the composition of  $X$  and the linear regression  $g$ . Hence, while the conditional expectation  $E(Y|X)$  is a function with domain  $\Omega$  and codomain  $\mathbb{R}$ , the linear regression  $g$  is a function with domain  $\mathbb{R}^n$  and codomain  $\mathbb{R}$ .

**Remark 12.33 (Simple and Multiple Linear Regression)** If  $n \geq 2$ , then the linear regression of  $Y$  on  $X$  is also called the *multiple linear regression* of  $Y$  on  $X$ . The coefficients  $\beta_0, \beta_1, \dots, \beta_n$  are called *regression coefficients*, and  $\beta_0$  the *intercept*. If  $n = 1$ , then a linear regression of  $Y$  on  $X$  is also called a *simple linear regression* of  $Y$  on  $X$  with *slope*  $\beta_1$ .  $\triangleleft$

**Remark 12.34 (Simple Regression as a Special Case)** If  $n = 1$  and we define  $X := X_1$ , then Equation (12.52) can be written

$$E(Y|\mathcal{C}) = \beta_0 + \beta_1 X. \quad (12.56)$$

If Equation (12.56) holds, then there is a version  $E(Y|X) \in \mathcal{E}(Y|X)$  such that

$$E(Y|\mathcal{C}) = E(Y|X) = \beta_0 + \beta_1 X = Q_{lin}(Y|X). \quad (12.57)$$

Therefore, we obtain the same results for the regression coefficients that have already been described in Theorem 7.14. In particular, Equation (12.53) yields

$$\beta_0 = E(Y) - \beta_1 E(X), \quad (12.58)$$

and (12.54) implies

$$\beta_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}. \quad (12.59)$$

If  $n = 1$  and  $X = X_1$ , then  $\Sigma_{xx}^{-1} = [\frac{1}{\text{Var}(X)}]$  and the existence of the inverse  $\Sigma_{xx}^{-1}$  is equivalent to  $\text{Var}(X) > 0$  (see Th. 7.14 and Rem. 7.25).  $\triangleleft$

In the following theorem, we consider the special case that  $Z = (Y, X_1, \dots, X_n)$  has an  $(n + 1)$ -variate normal distribution (see Def. 8.36).

### Theorem 12.35 (Linear Parametrization and Normal Distribution)

Let the assumptions 12.30 hold and let  $Z = (Y, X_1, \dots, X_n)$  be an  $(n + 1)$ -variate real-valued random variable on  $(\Omega, \mathcal{A}, P)$  with  $Z \sim \mathcal{N}_{\mu_z, \Sigma_{zz}}$ . Furthermore, assume that the inverse  $\Sigma_{xx}^{-1}$  exists. Then there is a version  $E(Y|X) \in \mathcal{E}(Y|X)$  with

$$E(Y|X) = \beta_0 + \beta' \mathbf{x}$$

such that Equations (12.53) and (12.54) hold for  $\beta_0$  and  $\beta$ , respectively.

Existence is proved by Rao (1973, pp. 523, Eq. 8a.2.16) and Corollary 12.31 implies that Equations (12.53) and (12.54) hold for  $\beta_0$  and  $\beta = [\beta_1, \dots, \beta_n]'$ .

## 12.7 Parametrizations of a Discrete Conditional Expectation

Now we consider two parametrizations of a discrete conditional expectation  $E(Y|Z)$ . In both cases, there are only a finite number of values of  $Z$ . In the first case, the possible values of  $Z$  are denoted  $z_1, \dots, z_n$ . In the second case, the notation is changed for didactic reasons.

### Theorem 12.36 (Means as Coefficients)

Let the assumptions 12.4 hold, assume that  $Z: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Z, \mathcal{A}'_Z)$  is a discrete random variable, and  $P(Z \in \{z_1, \dots, z_n\}) = 1$  with  $P(Z = z_i) > 0$ , for all  $i = 1, \dots, n$ . Furthermore, define  $X_i := 1_{Z=z_i}$ ,  $i = 1, \dots, n$ , and  $X := (X_1, \dots, X_n)$ . Then there is a version  $E(Y|Z) \in \mathcal{E}(Y|Z)$  with

$$E(Y|Z) = \sum_{i=1}^n \beta_i \cdot 1_{Z=z_i} = \sum_{i=1}^n \beta_i \cdot X_i, \quad (12.60)$$

where

$$\beta_i = E(Y|Z=z_i), \quad \forall i = 1, \dots, n. \quad (12.61)$$

The function  $g: \mathbb{R}^n \rightarrow \mathbb{R}$  defined by

$$g(x) = 0 + \sum_{i=1}^n \beta_i \cdot x_i, \quad \forall x = (x_1, \dots, x_n) \in \mathbb{R}^n, \quad (12.62)$$

is a linear parametrization of  $E(Y|Z)$  in  $X = (1_{Z=z_1}, \dots, 1_{Z=z_n})$ . If  $Z(\Omega) = \{z_1, \dots, z_n\}$ , then  $V = V^*$  for all  $V, V^* \in \mathcal{E}(Y|Z)$ .

(Proof p. 375)

Hence, in this parametrization the coefficients of Equation (12.7) are

$$\beta_0 = 0, \quad \beta_1 = E(Y|Z=z_1), \quad \dots, \quad \beta_n = E(Y|Z=z_n), \quad (12.63)$$

i. e., in this parametrization, the coefficients  $\beta_i$ ,  $i = 1, \dots, n$ , are the  $(Z=z_i)$ -conditional expectation values of  $Y$ .

In the following theorem we present another linear parametrization of  $E(Y|Z)$ , generalizing Example 12.15. For convenience, the possible values of  $Z$  are now denoted  $z_0, z_1, \dots, z_n$ . Aside from this change in the notation, the assumptions in Theorems 12.36 and 12.37 are identical.

**Theorem 12.37 (Differences Between Means as Coefficients)**

Let the assumptions 12.4 hold, assume that  $Z: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Z, \mathcal{A}'_Z)$  is discrete,  $P(Z \in \{z_0, z_1, \dots, z_n\}) = 1$  with  $P(Z=z_i) > 0$ , for all  $i = 0, 1, \dots, n$ . Furthermore, define  $X_i := 1_{Z=z_i}$ ,  $i = 1, \dots, n$ , and  $X := (X_1, \dots, X_n)$ . Then there is a version  $E(Y|Z) \in \mathcal{E}(Y|Z)$  with

$$E(Y|Z) = \beta_0 + \sum_{i=1}^n \beta_i \cdot 1_{Z=z_i} = \beta_0 + \sum_{i=1}^n \beta_i \cdot X_i, \quad (12.64)$$

where

$$\beta_0 = E(Y|Z=z_0) \quad (12.65)$$

and

$$\beta_i = E(Y|Z=z_i) - E(Y|Z=z_0), \quad \forall i = 1, \dots, n. \quad (12.66)$$

The function  $g: \mathbb{R}^n \rightarrow \mathbb{R}$  defined by

$$g(x) = \beta_0 + \sum_{i=1}^n \beta_i \cdot x_i, \quad \forall x = (x_1, \dots, x_n) \in \mathbb{R}^n, \quad (12.67)$$

is a linear parametrization of  $E(Y|Z)$  in  $X = (1_{Z=z_1}, \dots, 1_{Z=z_n})$ . If  $Z(\Omega) = \{z_0, z_1, \dots, z_n\}$ , then  $V = V^*$  for all  $V, V^* \in \mathcal{E}(Y|Z)$ .

(Proof p. 376)

Hence, in contrast to Equation (12.63), in this parametrization the coefficients are

$$\begin{aligned} \beta_0 &= E(Y|Z=z_0), \\ \beta_1 &= E(Y|Z=z_1) - E(Y|Z=z_0), \\ &\vdots \\ \beta_n &= E(Y|Z=z_n) - E(Y|Z=z_0). \end{aligned} \quad (12.68)$$

Now we present a lemma on the covariance matrix of the indicators  $1_{Z=z_i}$  for the values  $z_0, z_1, \dots, z_n$  of a discrete random variable  $Z$ . In particular this lemma helps to prove that the covariance matrix of the indicators  $1_{Z=z_1}, \dots, 1_{Z=z_n}$  is regular. This implies that the inverse of this covariance matrix exists so that Corollary 12.31 can be applied.

**Lemma 12.38 (Covariance Matrix of Indicators)**

Let  $Z: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Z, \mathcal{A}'_Z)$  be discrete with  $P(Z \in \{z_0, z_1, \dots, z_n\}) = 1$  and  $p_i := P(Z=z_i) > 0$ , for all  $i = 0, 1, \dots, n$ . Then:

- (i) The second moments of the indicators  $1_{Z=z_1}, \dots, 1_{Z=z_n}$  are finite, and for all  $i, j = 0, 1, \dots, n$ ,

$$\sigma_{ij} := \text{Cov}(1_{Z=z_i}, 1_{Z=z_j}) = \begin{cases} p_i \cdot (1 - p_i), & \text{if } i = j \\ -p_i \cdot p_j, & \text{if } i \neq j. \end{cases} \quad (12.69)$$

- (ii) For all  $i = 1, \dots, n$ ,

$$\sigma_{ii} > \sum_{j=1, j \neq i}^n |\sigma_{ij}|. \quad (12.70)$$

(Proof p. 376)

**Remark 12.39 (Strict Diagonal Dominance)** Proposition (ii) of Lemma 12.38 implies that the covariance matrix of the indicators  $1_{Z=z_1}, \dots, 1_{Z=z_n}$  is *strictly diagonally dominant*, i. e., it satisfies

$$|\sigma_{ii}| > \sum_{j=1, j \neq i}^n |\sigma_{ij}|, \quad \forall i = 1, \dots, n. \quad (12.71)$$

◁

**Remark 12.40 (Regularity of the Covariance Matrix of Indicators)** According to Corollary 5.6.17 of Horn and Johnson (1991, p. 302), (12.71) implies that the covariance matrix of the indicators  $1_{Z=z_1}, \dots, 1_{Z=z_n}$  is regular. In contrast, the covariance matrix of the indicators  $1_{Z=z_0}, 1_{Z=z_1}, \dots, 1_{Z=z_n}$  is *not* strictly diagonally dominant and it is *not regular*. The reason is that  $1_{Z=z_0} = 1 - \sum_{i=1}^n 1_{Z=z_i}$ , i. e.,  $1_{Z=z_0}$  is a linear combination of  $1_{Z=z_1}, \dots, 1_{Z=z_n}$ , which implies that the covariance matrix of  $1_{Z=z_0}, 1_{Z=z_1}, \dots, 1_{Z=z_n}$  is not regular (see Exercise 12-4).  $\triangleleft$

**Example 12.41 (Tom, Jim, and Kate – continued)** Table 5.1 displays an example in which the treatment variable  $X$  has three values. The conditional expectation  $E(Y|X) = P(Y=1|X)$  has three values as well, namely the following conditional probabilities:

$$P(Y=1|X=0) = \frac{(10 + 15 + 8) / 99}{(10 + 10 + 5 + 15 + 12 + 8) / 99} = .55,$$

$$P(Y=1|X=1) = \frac{(6 + 5 + 3) / 99}{(2 + 6 + 3 + 5 + 5 + 3) / 99} = .58\bar{3},$$

$$P(Y=1|X=2) = \frac{(4 + 3 + 1) / 99}{(1 + 4 + 2 + 3 + 4 + 1) / 99} = .5\bar{3}.$$

There are several linear parametrizations of the conditional expectation  $E(Y|X)$ . For example, we can use the linear parametrization in  $(X, X^2)$  specified in Equation (12.33), which yields

$$\begin{aligned} P(Y=1|X=0) &= \beta_0, \\ P(Y=1|X=1) &= \beta_0 + \beta_1 + \beta_2, \\ P(Y=1|X=2) &= \beta_0 + 2 \cdot \beta_1 + 4 \cdot \beta_2. \end{aligned}$$

Solving this equation system results in

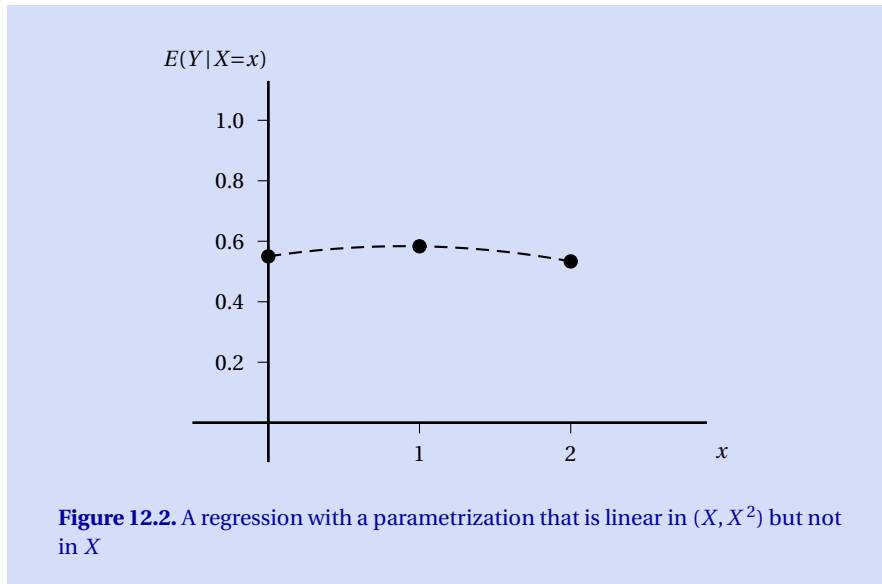
$$\begin{aligned} \beta_0 &= P(Y=1|X=0) = .55, \\ \beta_1 &= \frac{1}{2} \left[ -P(Y=1|X=2) + P(Y=1|X=1) - 3 \cdot P(Y=1|X=0) \right] = -.80, \\ \beta_2 &= \frac{1}{2} \left[ P(Y=1|X=2) - 2 \cdot P(Y=1|X=1) + P(Y=1|X=0) \right] = -.041\bar{6}. \end{aligned}$$

In this example,  $\beta_2 \neq 0$  and there is no linear parametrization of  $E(Y|X)$  in  $X$ . This is illustrated in Figure 12.2 (see Exercise 12-5). Note that  $X$  and the conditional expectation  $E(Y|X) = P(Y=1|X)$  have only three different values.

We could also use the linear parametrization specified in Equation (12.62). In this case, according to (12.63), the coefficients are

$$\beta_0 = 0, \quad \beta_1 = P(Y=1|X=0) = .55, \quad \beta_2 = P(Y=1|X=1) = .58\bar{3}, \quad \beta_3 = P(Y=1|X=2) = .5\bar{3}.$$

If we use the linear parametrization specified in Equation (12.67), then, according to (12.68), the coefficients are



$$\beta_0 = P(Y=1|X=0) = .55,$$

$$\beta_1 = P(Y=1|X=1) - P(Y=1|X=0) = .03,$$

$$\beta_2 = P(Y=1|X=2) - P(Y=1|X=0) = -.01\bar{6}.$$

&lt;

**Remark 12.42 ( $n$ -Valued Regressor)** If  $X$  is a discrete random variable with  $n$  different values, then there is a version  $P(Y=1|X)$  with a linear parametrization in  $(X, X^2, \dots, X^{n-1})$ , i. e.,  $P(Y=1|X)$  can be written as a polynomial in  $X$  of degree  $n-1$  (see Süli & Mayers, 2003, Theorem 6.1, Lagrange's interpolation theorem, p. 181).

&lt;

## 12.8 Invariance of Regression Coefficients

**Remark 12.43 (Simple vs. Multiple Regression)** Let the assumptions 12.4 hold with  $n = 2$  and let the second moments of  $Y, X_1$ , and  $X_2$  be finite. Assume that the inverse of the covariance matrix of  $(X_1, X_2)$  exists and that there is a version  $E(Y|X_1, X_2) \in \mathcal{E}(Y|X_1, X_2)$  such that

$$E(Y|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2. \quad (12.72)$$

Furthermore, assume that there is a version  $E(Y|X_1) \in \mathcal{E}(Y|X_1)$  with

$$E(Y|X_1) = \alpha_0 + \alpha_1 X_1. \quad (12.73)$$

Then  $\alpha_1 = \beta_1$  does not necessarily hold.  $\triangleleft$

In the following theorem we formulate a sufficient condition for  $\alpha_1 = \beta_1$  and generalize it to the  $n$ -variate case, where  $n \geq 2$ . That is, instead of  $E(Y|X_1)$ , we consider  $E(Y|X_1, \dots, X_m)$  and replace  $E(Y|X_1, X_2)$  by  $E(Y|X_1, \dots, X_n)$ , presuming  $m < n$ .

**Theorem 12.44 (Invariance of Regression Coefficients)**

Let the assumptions 12.4 hold, let  $m < n$ , and suppose that there is a version  $E(Y|X_1, \dots, X_n) \in \mathcal{E}(Y|X_1, \dots, X_n)$  such that

$$E(Y|X_1, \dots, X_n) = \beta_0 + \sum_{i=1}^m \beta_i X_i + \sum_{i=m+1}^n \beta_i X_i. \quad (12.74)$$

If

$$\forall i = m+1, \dots, n: \left( \beta_i = 0 \quad \text{or} \quad E(X_i | X_1, \dots, X_m) \stackrel{p}{=} E(X_i) \right), \quad (12.75)$$

then there is a version  $E(Y|X_1, \dots, X_m) \in \mathcal{E}(Y|X_1, \dots, X_m)$  such that

$$E(Y|X_1, \dots, X_m) = \alpha_0 + \sum_{i=1}^m \alpha_i X_i \quad (12.76)$$

with

$$\alpha_0 = \beta_0 + \sum_{i=m+1}^n \beta_i E(X_i) \quad (12.77)$$

and

$$\alpha_i = \beta_i, \quad \forall i = 1, \dots, m. \quad (12.78)$$

(Proof p. 377)

Note that

$$X_1, \dots, X_m \perp\!\!\!\perp X_i, \quad \forall i = m+1, \dots, n$$

is a sufficient condition for

$$E(X_i | X_1, \dots, X_m) \stackrel{p}{=} E(X_i), \quad \forall i = m+1, \dots, n$$

[see Eq. (12.75) and Box 10.2 (vi)].

## 12.9 Proofs

### ***Proof of Theorem 12.36***

The existence of a version  $E(Y|Z)$  satisfying (12.60) and (12.61) immediately follows from Definition 9.14 and Remark 10.11. Note that

$$\sigma(X) = \sigma(1_{Z=z_1}, \dots, 1_{Z=z_n}) = \sigma(\{\{Z=z_i\}: i=1, \dots, n\}) \subset \sigma(Z),$$

where  $X := (X_1, \dots, X_n) = (1_{Z=z_1}, \dots, 1_{Z=z_n})$ . This implies that the function  $g$  defined by Equation (12.62) is a linear parametrization of  $E(Y|Z)$  in  $X = (1_{Z=z_1}, \dots, 1_{Z=z_n})$  with coefficients

$$\beta_0 = 0, \quad \beta_1 = E(Y|Z=z_1), \quad \dots, \quad \beta_n = E(Y|Z=z_n)$$

(see Def. 12.5). Uniqueness is an immediate implication of Remark 10.12, provided that  $Z(\Omega) = \{z_1, \dots, z_n\}$ .

### **Proof of Theorem 12.37**

Let  $h: \Omega'_Z \rightarrow \mathbb{R}$  be a function such that  $h(Z) \in \mathcal{E}(Y|Z)$ . Then

$$h(Z) \stackrel{p}{=} \sum_{i=0}^n h(z_i) \cdot 1_{Z=z_i} \quad [(5.34)]$$

$$\stackrel{p}{=} \sum_{i=0}^n E(Y|Z=z_i) \cdot 1_{Z=z_i} \quad [(10.27)]$$

$$\begin{aligned} &\stackrel{p}{=} E(Y|Z=z_0) \cdot 1_{Z=z_0} + \sum_{i=1}^n E(Y|Z=z_i) \cdot 1_{Z=z_i} \\ &\stackrel{p}{=} E(Y|Z=z_0) - \sum_{i=1}^n E(Y|Z=z_0) \cdot 1_{Z=z_i} + \sum_{i=1}^n E(Y|Z=z_i) \cdot 1_{Z=z_i} \quad [(5.33)] \end{aligned}$$

$$\stackrel{p}{=} E(Y|Z=z_0) + \sum_{i=1}^n [E(Y|Z=z_i) - E(Y|Z=z_0)] \cdot 1_{Z=z_i}$$

$$\stackrel{p}{=} \beta_0 + \sum_{i=1}^n \beta_i \cdot 1_{Z=z_i},$$

where  $\beta_0 := E(Y|Z=z_0)$  and  $\beta_i := E(Y|Z=z_i) - E(Y|Z=z_0)$ , for  $i=1, \dots, n$ . Because the function on the right-hand side of the last equation is  $Z$ -measurable (see Rem. 2.17), it is an element of  $\mathcal{E}(Y|Z)$ , i. e.,

$$\beta_0 + \sum_{i=1}^n \beta_i \cdot 1_{Z=z_i} \in \mathcal{E}(Y|Z).$$

This proves equations (12.64), (12.65), and (12.66). Note that

$$\sigma(X) = \sigma(\{\{Z=z_i\}: i=1, \dots, n\}) \subset \sigma(\{\{Z=z_i\}: i=0, 1, \dots, n\}) \subset \sigma(Z),$$

where  $X := (X_1, \dots, X_n) = (1_{Z=z_1}, \dots, 1_{Z=z_n})$ . This implies that  $g$  [see Eq. (12.67)] is a linear parametrization of  $E(Y|Z)$  in  $X = (1_{Z=z_1}, \dots, 1_{Z=z_n})$  (see Def. 12.5). If  $Z(\Omega) = \{z_0, z_1, \dots, z_n\}$ , then Remark 10.12 implies that  $E(Y|Z)$  is the only version in  $\mathcal{E}(Y|Z)$ .

### **Proof of Lemma 12.38**

(i). For  $i=j$ , Equation (12.69) immediately follows from Equation (6.29) for the event  $A = \{Z=z_i\}$ . For  $i \neq j$ , Equation (12.69) follows from Equation (7.14) and the fact that  $P(Z=z_i, Z=z_j) = 0$  if  $i \neq j$ . Equation (12.69) also shows that the variances and covariances of the indicators  $1_{Z=z_1}, \dots, 1_{Z=z_n}$  are finite, which implies that their second moments are finite as well [see Box 7.1 (i)].

(ii). This proposition can be derived as follows: For all  $i = 1, \dots, n$ ,

$$\begin{aligned}
 \sigma_{ii} &= p_i \cdot (1 - p_i) && [(12.69) \text{ for } i = j] \\
 &= p_i \cdot \sum_{j=0, j \neq i}^n p_j && \left[ 1 - p_i = \sum_{j=0, j \neq i}^n p_j \right] \\
 &= \sum_{j=1, j \neq i}^n p_i p_j + p_i \cdot p_0 \\
 &> \sum_{j=1, j \neq i}^n p_i p_j && [p_i > 0, \forall i = 0, 1, \dots, n, p_i \cdot p_0 > 0] \\
 &= \sum_{j=1, j \neq i}^n |-p_i p_j| && [p_i p_j = |-p_i p_j|] \\
 &= \sum_{j=1, j \neq i}^n |\sigma_{ij}|. && [(12.69) \text{ for } i \neq j]
 \end{aligned}$$

### Proof of Theorem 12.44

$$\begin{aligned}
 &E(Y|X_1, \dots, X_m) \\
 \stackrel{\text{p}}{=} &E(E(Y|X_1, \dots, X_n)|X_1, \dots, X_m) && [m < n, \text{ Box 10.2 (v)}] \\
 \stackrel{\text{p}}{=} &E\left(\beta_0 + \sum_{i=1}^m \beta_i X_i + \sum_{i=m+1}^n \beta_i X_i \mid X_1, \dots, X_m\right) && [(12.74)] \\
 \stackrel{\text{p}}{=} &\beta_0 + E\left(\sum_{i=1}^m \beta_i X_i \mid X_1, \dots, X_m\right) + E\left(\sum_{i=m+1}^n \beta_i X_i \mid X_1, \dots, X_m\right) && [\text{Box 10.2 (xvi), (i)}] \\
 \stackrel{\text{p}}{=} &\beta_0 + \sum_{i=1}^m \beta_i X_i + \sum_{i=m+1}^n \beta_i E(X_i | X_1, \dots, X_m) && [\text{Box 10.2 (xiv), (xvi)}] \\
 \stackrel{\text{p}}{=} &\beta_0 + \sum_{i=1}^m \beta_i X_i + \sum_{i=m+1}^n \beta_i E(X_i) && [(12.75)] \\
 \stackrel{\text{p}}{=} &\alpha_0 + \sum_{i=1}^m \beta_i X_i. && [(12.77)]
 \end{aligned}$$

## 12.10 Exercises

▷ **Exercise 12-1** Consider Example 12.20 and show that  $\sigma(X) = \sigma(Z)$ .

▷ **Exercise 12-2** Let  $X$  and  $Z$  be dichotomous random variables on  $(\Omega, \mathcal{A}, P)$  with values  $x_1, x_2$ , and  $z_1, z_2$ , respectively (see Example 5.10), and  $k: \{0, 1\}^2 \rightarrow \mathbb{R}$  a (not necessarily measurable) function. Show that there are  $\beta_0, \dots, \beta_3 \in \mathbb{R}$  such that

$$k(X, Z) \stackrel{p}{=} \beta_0 + \beta_1 \cdot 1_{X=x_2} + \beta_2 \cdot 1_{Z=z_2} + \beta_3 \cdot 1_{X=x_2} \cdot 1_{Z=z_2},$$

and determine these coefficients in terms of the values of  $k(X, Z)$ . [Special cases are  $k(X, Z) = E(Y|X, Z)$  (see Example 12.22) and  $k(X, Z) = \text{logit}[P(Y=1|X, Z)]$  (see Example 13.24 with  $Z = U$ ).]

▷ **Exercise 12-3** Show that under the assumptions made in Remark 12.28, there is no linear parametrization  $g$  of  $E(Y|X)$  in  $X$  with  $g(x) = \beta_0 + \beta_1 x$ ,  $x \in \mathbb{R}$ , and slope  $\beta_1 \neq 0$ .

▷ **Exercise 12-4** Assume  $X_0, X_1, \dots, X_n: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  are random variables with finite second moments. Show: If  $X_0 = \mathbf{a}'\mathbf{x} + b$ , where  $\mathbf{x} := [X_1, \dots, X_n]'$ ,  $\mathbf{a}' := [a_1, \dots, a_n] \in \mathbb{R}^n$ , and  $b \in \mathbb{R}$ , then the variance-covariance matrix of  $X_0, X_1, \dots, X_n$  is singular.

▷ **Exercise 12-5** Compute the parameters of the quadratic function displayed in Figure 12.2 using the three conditional probabilities computed in Example 12.41.

## Solutions

▷ **Solution 12-1** Because the function  $\ln: ]0, \infty[ \rightarrow \mathbb{R}$  is continuous, it is  $(\mathcal{B}|_{]0, \infty[}, \mathcal{B})$ -measurable (see Klenke, 2013, Th. 1.88, p. 36). Hence, for  $X = \ln Z$ , Lemma 2.52 implies  $\sigma(X) \subset \sigma(Z)$ . Furthermore, the exponential function  $\exp: \mathbb{R} \rightarrow \mathbb{R}$  is continuous as well and therefore it is  $(\mathcal{B}, \mathcal{B})$ -measurable (see again Klenke, 2013, Th. 1.88, p. 36). Because  $Z = \exp(X)$ , Lemma 2.52 implies  $\sigma(Z) \subset \sigma(X)$ , and this yields  $\sigma(Z) = \sigma(X)$ .

▷ **Solution 12-2** Note that  $1_{X=x_2} = (1 - 1_{X=x_1})$ ,  $1_{Z=z_2} = (1 - 1_{Z=z_1})$ , and  $1_{X=x_i, Z=z_j} = 1_{X=x_i} \cdot 1_{Z=z_j}$ ,  $i, j = 1, 2$  [see Eq. (1.33)]. Hence,

$$\begin{aligned} k(X, Z) &\stackrel{p}{=} k(x_1, z_1) \cdot 1_{X=x_1} \cdot 1_{Z=z_1} + k(x_2, z_1) \cdot 1_{X=x_2} \cdot 1_{Z=z_1} + \\ &\quad k(x_1, z_2) \cdot 1_{X=x_1} \cdot 1_{Z=z_2} + k(x_2, z_2) \cdot 1_{X=x_2} \cdot 1_{Z=z_2} && \text{[Cor. 5.63]} \\ &= k(x_1, z_1) \cdot (1 - 1_{X=x_2}) \cdot (1 - 1_{Z=z_2}) + k(x_2, z_1) \cdot 1_{X=x_2} \cdot (1 - 1_{Z=z_2}) + \\ &\quad k(x_1, z_2) \cdot (1 - 1_{X=x_2}) \cdot 1_{Z=z_2} + k(x_2, z_2) \cdot 1_{X=x_2} \cdot 1_{Z=z_2} \\ &= k(x_1, z_1) + (k(x_2, z_1) - k(x_1, z_1)) \cdot 1_{X=x_2} + (k(x_1, z_2) - k(x_1, z_1)) \cdot 1_{Z=z_2} + \\ &\quad (k(x_2, z_2) - k(x_2, z_1) - k(x_1, z_2) + k(x_1, z_1)) \cdot 1_{X=x_2} \cdot 1_{Z=z_2}, \end{aligned}$$

where the last equation is obtained by multiplying out the parentheses and rearranging terms. Hence,

$$\begin{aligned} \beta_0 &= k(x_1, z_1) \\ \beta_1 &= k(x_2, z_1) - k(x_1, z_1) \\ \beta_2 &= k(x_1, z_2) - k(x_1, z_1) \\ \beta_3 &= k(x_2, z_2) - k(x_2, z_1) - k(x_1, z_2) + k(x_1, z_1). \end{aligned}$$

▷ **Solution 12-3** This proposition follows from

$$P[g(X) < 0] + P[g(X) > 1] = \begin{cases} P\left(X < -\frac{\beta_0}{\beta_1}\right) + P\left(X > \frac{1-\beta_0}{\beta_1}\right) > 0, & \text{if } \beta_1 > 0 \\ P\left(X > -\frac{\beta_0}{\beta_1}\right) + P\left(X < \frac{1-\beta_0}{\beta_1}\right) > 0, & \text{if } \beta_1 < 0, \end{cases}$$

because  $P[g(X) < 0] + P[g(X) > 1] > 0$  is a contradiction to (12.46).

▷ **Solution 12-4** Let  $\Sigma$  denote the variance-covariance matrix of  $X_0, X_1, \dots, X_n$ . Then

$$\begin{aligned} \Sigma &= \begin{bmatrix} \sigma_{X_0}^2 & \sigma_{X_0X_1} & \cdots & \sigma_{X_0X_n} \\ \sigma_{X_1X_0} & \sigma_{X_1}^2 & \cdots & \sigma_{X_1X_n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{X_nX_0} & \sigma_{X_nX_1} & \cdots & \sigma_{X_n}^2 \end{bmatrix} \\ &= E \left[ \begin{pmatrix} \mathbf{a}'\mathbf{x} - E(\mathbf{a}'\mathbf{x}) \\ \mathbf{x} - E(\mathbf{x}) \end{pmatrix} (\mathbf{a}'\mathbf{x} - E(\mathbf{a}'\mathbf{x}), \mathbf{x}' - E(\mathbf{x}')) \right] && [X_0 = \mathbf{a}'\mathbf{x} + b, (7.35)] \\ &= E \left[ \begin{matrix} \mathbf{a}'[\mathbf{x} - E(\mathbf{x})][\mathbf{x}' - E(\mathbf{x}')] \mathbf{a} & \mathbf{a}'[\mathbf{x} - E(\mathbf{x})][\mathbf{x}' - E(\mathbf{x}')] \\ [\mathbf{x} - E(\mathbf{x})][\mathbf{x}' - E(\mathbf{x}')] \mathbf{a} & [\mathbf{x} - E(\mathbf{x})][\mathbf{x}' - E(\mathbf{x}')] \end{matrix} \right] && [\text{Box 7.2 (iii), (v)}] \\ &= \begin{bmatrix} \mathbf{a}' E([\mathbf{x} - E(\mathbf{x})][\mathbf{x}' - E(\mathbf{x}')] \mathbf{a}) & \mathbf{a}' E([\mathbf{x} - E(\mathbf{x})][\mathbf{x}' - E(\mathbf{x}')] \mathbf{a}) \\ E([\mathbf{x} - E(\mathbf{x})][\mathbf{x}' - E(\mathbf{x}')] \mathbf{a}) & E([\mathbf{x} - E(\mathbf{x})][\mathbf{x}' - E(\mathbf{x}')] \mathbf{a}) \end{bmatrix}. && [(7.29)] \end{aligned}$$

The first row of this matrix is obtained by multiplying the lower two submatrices by  $\mathbf{a}'$  from the left. Hence, the first row of this variance-covariance matrix is a linear combination of its other rows. This implies that  $\Sigma$  is singular.

▷ **Solution 12-5** For  $x_1 = x$  and  $x_2 = x^2$ , the parametrization  $g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$  [see Eq. (12.33)] yields

$$.55 = \beta_0,$$

if  $x=0$ ,

$$.58\bar{3} = \beta_0 + \beta_1 + \beta_2,$$

if  $x=1$ , and

$$.53\bar{3} = \beta_0 + \beta_1 \cdot 2 + \beta_2 \cdot 4,$$

if  $x=2$ . Hence,  $\beta_0 = .55$  and solving the last two equations for the remaining two unknowns yields  $\beta_1 \approx .0750$  and  $\beta_2 \approx -.0416$ .



## Chapter 13

# Linear Logistic Regression

In chapter 12 we treated the notions of a linear parametrization of a conditional expectation and of a *linear regression*. In this chapter we turn to the *linear logit regression*, presuming that the regressand, say  $Y$ , is an indicator variable. In Remark 10.4 we noted that in this case a version  $E(Y|\mathcal{C})$  of the  $\mathcal{C}$ -conditional expectation of  $Y$  is also called a version of the  $\mathcal{C}$ -conditional probability of the event  $\{Y=1\}$  and that it is also denoted by  $P(Y=1|\mathcal{C})$ . As noted in Remark 12.28, if  $X$  is a  $\mathcal{C}$ -measurable real-valued random variable on  $(\Omega, \mathcal{A}, P)$  and  $X$  is  $P$ -almost surely unbounded, then there is no linear parametrization  $g$  in  $X$  of  $E(Y|\mathcal{C}) = P(Y=1|\mathcal{C}) = g(X)$  with a nonzero slope. However, in this case, there might be a linear logistic parametrization.

We begin with the logit transformation, define the logit of a  $\mathcal{C}$ -conditional probability  $P(Y=1|\mathcal{C})$ , a linear logistic parametrization and then present a theorem on uniqueness and the identification of the parameters. Finally, the concept of a linear logit regression is defined.

### 13.1 Logit Transformation of a Conditional Probability

The general assumptions and notation are as follows:

#### Notation and Assumptions 13.1

Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  be a dichotomous random variable with values 0 and 1, let  $\mathcal{C} \subset \mathcal{A}$  be a  $\sigma$ -algebra, and assume that there is a version  $P(Y=1|\mathcal{C}) \in \mathcal{P}(Y=1|\mathcal{C})$  with  $0 < P(Y=1|\mathcal{C}) < 1$ .

**Remark 13.2 (Necessary Condition)** If  $\{Y=1\} \in \mathcal{C}$  or  $\{Y=0\} \in \mathcal{C}$ , then there is no version  $P(Y=1|\mathcal{C}) \in \mathcal{P}(Y=1|\mathcal{C})$  with  $0 < P(Y=1|\mathcal{C}) < 1$ . Because  $\{Y=1\}, \{Y=0\} \in \sigma(Y)$ , there is no version  $P(Y=1|\mathcal{C}) \in \mathcal{P}(Y=1|\mathcal{C})$  with  $0 < P(Y=1|\mathcal{C}) < 1$  if  $\sigma(Y) \subset \mathcal{C}$  (see Exercise 13-1).  $\triangleleft$

**Remark 13.3 (The Logit Transformation)** Each version  $P(Y=1|\mathcal{C})$  of the conditional probability is  $P$ -almost surely bounded, because

$$P(\{\omega \in \Omega: 0 \leq P(Y=1|\mathcal{C})(\omega) \leq 1\}) = 1 \quad (13.1)$$

[see Box 10.3 (ii), (iii)]. It can be transformed using the logit transformation, i. e., using the function  $\text{logit}: ]0, 1[ \rightarrow \mathbb{R}$  defined by

$$\text{logit}(p) := \ln\left(\frac{p}{1-p}\right) = \ln(p) - \ln(1-p), \quad \forall p \in ]0, 1[, \quad (13.2)$$

where  $\ln$  denotes the natural logarithm. If  $p$  represents a probability, then  $\ln\left(\frac{p}{1-p}\right)$  is also called the *log-odds* of  $p$ . Figure 13.1 shows the graph of such a logit transformation. In the context of generalized linear models [see, e. g., Agresti, 2015; McCullagh & Nelder, 1989], this function is an example of a *link function*. Some algebra yields

$$p = \frac{\exp[\text{logit}(p)]}{1 + \exp[\text{logit}(p)]}, \quad \forall p \in ]0, 1[, \quad (13.3)$$

(see Exercise 13-2). ◀

**Remark 13.4 (Logistic Function)** The function  $h: \mathbb{R} \rightarrow ]0, 1[$  specified by

$$h(x) = \frac{\exp(x)}{1 + \exp(x)}, \quad \forall x \in \mathbb{R}, \quad (13.4)$$

is called the *logistic function*. Equation (13.3) implies that the logistic function is the inverse of the logit function. ◀

**Definition 13.5 (Logit of  $P(Y=1|\mathcal{C})$ )**

Let the assumptions 13.1 hold. Then

$$\text{logit}[P(Y=1|\mathcal{C})] := \ln\left(\frac{P(Y=1|\mathcal{C})}{1 - P(Y=1|\mathcal{C})}\right), \quad (13.5)$$

is called the *logit of  $P(Y=1|\mathcal{C})$* .

**Remark 13.6 (One-to-One Transformation)** Note that  $\text{logit}[P(Y=1|\mathcal{C})]$  denotes the composition of  $P(Y=1|\mathcal{C})$  and the function  $\text{logit}$  defined by Equation (13.2). Hence,  $\text{logit}[P(Y=1|\mathcal{C})]$  is a random variable on  $(\Omega, \mathcal{A}, P)$ . Also note that  $P(Y=1|\mathcal{C})$  and its logit contain the same information, i. e., the  $\sigma$ -algebras they generate are identical (see Lemma 13.7). While  $P(Y=1|\mathcal{C})$  informs us about the  $\mathcal{C}$ -conditional probability of the event  $\{Y=1\}$  on the *probability scale* with values between 0 and 1, a logit of  $P(Y=1|\mathcal{C})$  informs us about this conditional probability on the *log-odds scale* with values between  $-\infty$  and  $\infty$ . Applying (13.3) yields

$$P(Y=1|\mathcal{C}) = \frac{\exp(\text{logit}[P(Y=1|\mathcal{C})])}{1 + \exp(\text{logit}[P(Y=1|\mathcal{C})])}. \quad (13.6)$$

Hence,  $P(Y=1|\mathcal{C})$  and  $\text{logit}[P(Y=1|\mathcal{C})]$  are one-to-one transformations of each other. ◀

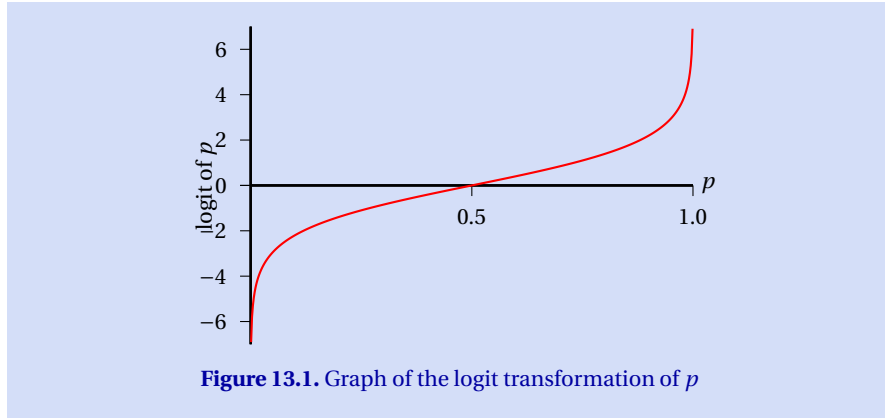


Figure 13.1. Graph of the logit transformation of  $p$

**Lemma 13.7 ( $\sigma$ -Algebra Generated by  $P(Y=1|\mathcal{C})$  and its Logit)**

Let the assumptions 13.1 hold. Then

$$\sigma[P(Y=1|\mathcal{C})] = \sigma(\text{logit}[P(Y=1|\mathcal{C})]). \tag{13.7}$$

(Proof p. 394)

**Remark 13.8 (Motivation for Considering the Logit of  $P(Y=1|\mathcal{C})$ )** Suppose that  $X$  is a real-valued  $\mathcal{C}$ -measurable random variable on  $(\Omega, \mathcal{A}, P)$  and that  $X$  is not  $P$ -almost surely bounded [see (12.47)]. Then it is still possible to assume that there is a version  $P(Y=1|\mathcal{C}) \in \mathcal{P}(Y=1|\mathcal{C})$  and numbers  $\lambda_0, \lambda_1 \in \mathbb{R}$  such that

$$\text{logit}[P(Y=1|\mathcal{C})] = \lambda_0 + \lambda_1 X. \tag{13.8}$$

In contrast, assuming  $P(Y=1|\mathcal{C}) = \lambda_0 + \lambda_1 X$  for real numbers  $\lambda_0, \lambda_1, \lambda_1 \neq 0$ , would be contradictory if  $X$  is not  $P$ -almost surely bounded.  $\triangleleft$

**13.2 Linear Logistic Parametrization**

**Notation and Assumptions 13.9**

Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  be a dichotomous random variable with values 0 and 1, let  $\mathcal{C} \subset \mathcal{A}$  be a  $\sigma$ -algebra, and assume that there is a version  $P(Y=1|\mathcal{C}) \in \mathcal{P}(Y=1|\mathcal{C})$  with  $0 < P(Y=1|\mathcal{C}) < 1$ . Furthermore, let  $X_i: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B}), i = 1, \dots, n$ , be real-valued random variables, define  $X := (X_1, \dots, X_n): (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}^n, \mathcal{B}_n)$ , and assume that  $X$  is  $\mathcal{C}$ -measurable.

Using this notation and these assumptions a linear logistic parametrization of a conditional probability  $P(Y=1|\mathcal{C})$  is now defined as follows:

**Definition 13.10 (Linear Logistic Parametrization)**

Let the assumptions 13.9 hold. If there are  $\lambda_0, \lambda_1, \dots, \lambda_n \in \mathbb{R}$  and a version  $P(Y=1|\mathcal{C}) \in \mathcal{P}(Y=1|\mathcal{C})$  such that

$$P(Y=1|\mathcal{C}) = \frac{\exp(\lambda_0 + \sum_{i=1}^n \lambda_i X_i)}{1 + \exp(\lambda_0 + \sum_{i=1}^n \lambda_i X_i)}, \quad (13.9)$$

then the function  $g: \mathbb{R}^n \rightarrow [0, 1]$  satisfying

$$g(x) = \frac{\exp(\lambda_0 + \sum_{i=1}^n \lambda_i x_i)}{1 + \exp(\lambda_0 + \sum_{i=1}^n \lambda_i x_i)}, \quad \forall x \in \mathbb{R}^n, \quad (13.10)$$

is called a linear logistic parametrization of  $P(Y=1|\mathcal{C})$  in  $X$ .

If  $\mathcal{C} = \sigma(X)$ , then Equation (13.9) is equivalent to

$$P(Y=1|X) = \frac{\exp(\lambda_0 + \sum_{i=1}^n \lambda_i X_i)}{1 + \exp(\lambda_0 + \sum_{i=1}^n \lambda_i X_i)}. \quad (13.11)$$

**Remark 13.11 (Univariate Real-Valued  $X$ )** If Equation (13.9) holds for  $n = 1$ , then there is a version  $P(Y=1|X) \in \mathcal{P}(Y=1|X)$  such that

$$P(Y=1|\mathcal{C}) = P(Y=1|X) = \frac{\exp(\lambda_0 + \lambda_1 X)}{1 + \exp(\lambda_0 + \lambda_1 X)}. \quad (13.12)$$

In Example 17.82 we present a sufficient condition of Equation (13.12) related to the normal distribution.  $\triangleleft$

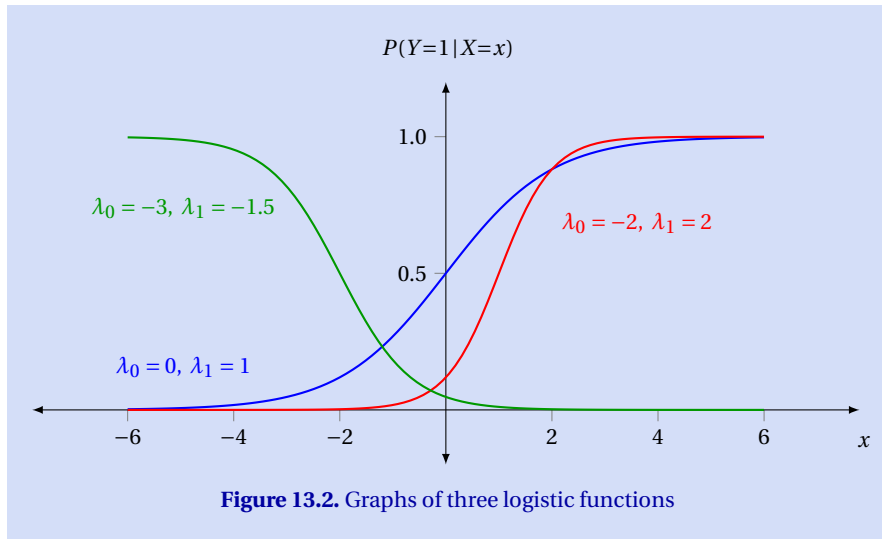
**Remark 13.12 (Conditional Probabilities)** If  $g$  is a linear logistic parametrization of  $P(Y=1|X)$  in  $X$  satisfying Equation (13.10), then, according to Definition 10.33, we can define

$$\begin{aligned} P(Y=1|X=x) &= P(Y=1|X_1=x_1, \dots, X_n=x_n) \\ &:= g(x) = \frac{\exp(\lambda_0 + \sum_{i=1}^n \lambda_i x_i)}{1 + \exp(\lambda_0 + \sum_{i=1}^n \lambda_i x_i)}, \quad \forall x \in \mathbb{R}^n. \end{aligned} \quad (13.13)$$

This definition is convenient, but note that another factorization  $g^*$  of  $P(Y=1|X)$  might yield other conditional probabilities  $P(Y=1|X=x)$  for values  $x$  of  $X$  with  $P(X=x) = 0$ . However, according to Equation (12.10), if  $g, g^*$  are factorizations of two versions  $V, V^* \in \mathcal{P}(Y=1|X)$ , then  $g(x) = g^*(x)$ , for  $P_X$ -almost all  $x \in \mathbb{R}^n$  [see Eq. (12.13)].  $\triangleleft$

**Remark 13.13 (Meaning of Coefficients)** Figure 13.2 displays the graphs of logistic transformations in which the logits of  $P(Y=1|X)$  are linear functions  $\lambda_0 + \lambda_1 X$ . As is easily seen,

$$P\left(Y=1 \mid X = -\frac{\lambda_0}{\lambda_1}\right) = \frac{\exp(0)}{1 + \exp(0)} = \frac{1}{1+1} = \frac{1}{2}.$$



This equation shows that  $x = -\frac{\lambda_0}{\lambda_1}$  is the point on the  $x$ -axis at which the conditional probability  $P(Y=1|X=x)$  is .5. Furthermore, the derivative of the linear parametrization  $g$  with respect to  $x$  is

$$\frac{d}{dx} g(x) = \frac{d}{dx} \frac{\exp(\lambda_0 + \lambda_1 x)}{1 + \exp(\lambda_0 + \lambda_1 x)} = \frac{\lambda_1 \exp(\lambda_0 + \lambda_1 x)}{(1 + \exp(\lambda_0 + \lambda_1 x))^2}, \quad (13.14)$$

Hence, the derivative (i. e., the slope) of  $g$  at  $x = -\frac{\lambda_0}{\lambda_1}$  is  $\frac{\lambda_1}{4}$  (see Exercise 13-4). <

### 13.3 A Parametrization of a Discrete Conditional Probability

In Theorem 12.37 we already considered a parametrization of a discrete conditional expectation  $E(Y|Z)$  in which the parameters  $\beta_1, \dots, \beta_n$  are the differences  $E(Y|Z=z_i) - E(Y|Z=z_0)$ ,  $i = 1, \dots, n$ . If  $Y$  is dichotomous, then there is also a logistic parametrization of the conditional probability  $E(Y|Z) = P(Y=1|Z)$ .

**Theorem 13.14 (Existence of the Logit Effects)**

Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  and  $Z: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Z, \mathcal{A}'_Z)$  be random variables, where  $Y$  is dichotomous with values 0 and 1, and  $Z$  is discrete with  $P(Z \in \{z_0, z_1, \dots, z_n\}) = 1$  and assume  $0 < P(Z=z_i) < 1$  and  $0 < P(Y=1|Z=z_i) < 1$ , for all  $i = 0, 1, \dots, n$ . Then  $P(Y=1|Z)$  is uniquely defined and there are coefficients  $\beta_0, \beta_1, \dots, \beta_n, \lambda_0, \lambda_1, \dots, \lambda_n \in \mathbb{R}$  such that

$$P(Y=1|Z) = \beta_0 + \sum_{i=1}^n \beta_i \cdot 1_{Z=z_i} \quad (13.15)$$

$$= \frac{\exp[\lambda_0 + \sum_{i=1}^n \lambda_i \cdot 1_{Z=z_i}]}{1 + \exp[\lambda_0 + \sum_{i=1}^n \lambda_i \cdot 1_{X=x_i}]} \quad (13.16)$$

with

$$\beta_0 = P(Y=1|Z=z_0) \quad (13.17)$$

$$= \frac{\exp(\lambda_0)}{1 + \exp(\lambda_0)} \quad (13.18)$$

and

$$\beta_i = P(Y=1|Z=z_0) - P(Y=1|Z=z_i) \quad (13.19)$$

$$= \frac{\exp(\lambda_0 + \lambda_i)}{1 + \exp(\lambda_0 + \lambda_i)} - \frac{\exp(\lambda_0)}{1 + \exp(\lambda_0)}. \quad (13.20)$$

(Proof p. 395)

**Remark 13.15 (Log Odds)** In terms of conditional probabilities, the logit intercept can be written

$$\lambda_0 = \ln \left[ \frac{P(Y=1|Z=z_0)}{1 - P(Y=1|Z=z_0)} \right]. \quad (13.21)$$

Hence,  $\lambda_0$  is the log odds of  $P(Y=1|Z=z_0)$ . Similarly,

$$\lambda_0 + \lambda_i = \ln \left[ \frac{P(Y=1|Z=z_i)}{1 - P(Y=1|Z=z_i)} \right], \quad \forall i = 1, \dots, n, \quad (13.22)$$

[see Eqs. (13.17) to (13.20) and (13.22)]. This equation shows that  $\lambda_0 + \lambda_i$  is the log odds of  $P(Y=1|Z=z_i)$ .  $\triangleleft$

**Remark 13.16 (Log Odds Ratio)** Equations (13.21) and (13.22) immediately imply

$$\lambda_i = \ln \left[ \frac{P(Y=1|Z=z_i)}{1 - P(Y=1|Z=z_i)} \right] - \ln \left[ \frac{P(Y=1|Z=z_0)}{1 - P(Y=1|Z=z_0)} \right] \quad (13.23)$$

$$= \ln \left[ \left( \frac{P(Y=1|Z=z_i)}{1 - P(Y=1|Z=z_i)} \right) / \left( \frac{P(Y=1|Z=z_0)}{1 - P(Y=1|Z=z_0)} \right) \right] \quad (13.24)$$

$$= \ln \left[ \frac{P(Y=1|Z=z_i) \cdot (1 - P(Y=1|Z=z_0))}{(1 - P(Y=1|Z=z_i)) \cdot P(Y=1|Z=z_0)} \right], \quad \forall i = 1, \dots, n,$$

for the logit effect  $\lambda_i$ . Hence,  $\lambda_i$ ,  $i = 1, \dots, n$ , is the difference between the log odds of  $P(Y=1|Z=z_i)$  and  $P(Y=1|Z=z_0)$ , respectively [see Eq. (13.23)]. Equation (13.24) shows that  $\lambda_i$  is the log odds ratio of  $P(Y=1|Z=z_i)$  and  $P(Y=1|Z=z_0)$ .  $\triangleleft$

**Remark 13.17 (Odds Ratio)** The exponential function of  $\lambda_i$  is

$$\begin{aligned}\exp(\lambda_i) &= \left( \frac{P(Y=1|Z=z_i)}{1-P(Y=1|Z=z_i)} \right) / \left( \frac{P(Y=1|Z=z_0)}{1-P(Y=1|Z=z_0)} \right) \\ &= \frac{P(Y=1|Z=z_i) \cdot (1-P(Y=1|Z=z_0))}{(1-P(Y=1|Z=z_i)) \cdot P(Y=1|Z=z_0)}, \quad \forall i = 1, \dots, n.\end{aligned}\quad (13.25)$$

This equation shows that the number  $\exp(\lambda_i)$  is the odds ratio of  $P(Y=1|Z=z_i)$  and  $P(Y=1|Z=z_0)$ .  $\triangleleft$

**Remark 13.18 (Risk Ratio)** Another closely related parameter is

$$\kappa_i := \frac{P(Y=1|Z=z_i)}{P(Y=1|Z=z_0)}, \quad \forall i = 1, \dots, n. \quad (13.26)$$

This parameter is called the *risk ratio* of  $P(Y=1|Z=z_i)$  and  $P(Y=1|Z=z_0)$ .  $\triangleleft$

**Remark 13.19 (Four Kinds of Effect Parameters)** Hence, under the assumptions of Theorem 13.14, we may consider four different kinds of effect parameters:  $\beta_i$ ,  $\lambda_i$ ,  $\exp(\lambda_i)$ , and  $\kappa_i$ . They all quantify the effect of  $x_i$  compared to  $x_0$  on  $Y$ , each one on a different scale.  $\triangleleft$

### 13.4 Identification of Coefficients of a Linear Logistic Parametrization

The following theorem specifies sufficient conditions under which a linear logit parametrization of  $P(Y=1|X)$  is uniquely defined.

#### Theorem 13.20 (Identification of Coefficients and Uniqueness)

Let the assumptions 13.9 hold and let  $\lambda_0, \lambda_1, \dots, \lambda_n \in \mathbb{R}$ . Furthermore, assume:

- (a) there is a version  $P(Y=1|\mathcal{C}) \in \mathcal{P}(Y=1|\mathcal{C})$  such that Equation (13.9) holds,
- (b)  $X_1, \dots, X_n$  have finite second moments, and
- (c) the inverse  $\Sigma_{xx}^{-1}$  of the covariance matrix of  $X = (X_1, \dots, X_n)$  exists.

Then, using

$$L := \text{logit}[P(Y=1|\mathcal{C})] = \lambda_0 + \boldsymbol{\lambda}'\mathbf{x}, \quad (13.27)$$

the following two equations hold:

$$\lambda_0 = E(L) - \boldsymbol{\lambda}'\boldsymbol{\mu}, \quad (13.28)$$

$$\boldsymbol{\lambda} = \Sigma_{xx}^{-1} \Sigma_{xl}, \quad (13.29)$$

where  $\boldsymbol{\mu} := [E(X_1), \dots, E(X_n)]'$  and  $\Sigma_{xl}$  denotes the column vector of the covariances  $\text{Cov}(X_i, L)$ . The coefficients  $\lambda_0$  and  $\boldsymbol{\lambda}$  are uniquely determined and

hence, the linear logistic parametrization  $g: \mathbb{R}^n \rightarrow [0, 1]$  of  $P(Y=1|\mathcal{C})$  satisfying

$$g(x) = \frac{\exp(\lambda_0 + \boldsymbol{\lambda}'\mathbf{x})}{1 + \exp(\lambda_0 + \boldsymbol{\lambda}'\mathbf{x})}, \quad \forall x \in \mathbb{R}^n, \quad (13.30)$$

is uniquely defined.

(Proof p. 395)

**Remark 13.21 (Identification vs. Estimation)** Note that, for a version  $P(Y=1|\mathcal{C}) \in \mathcal{P}(Y=1|\mathcal{C})$  satisfying assumptions 13.1, the logit of  $P(Y=1|\mathcal{C})$  is uniquely defined. Because the expectation vector  $\boldsymbol{\mu} = [E(X_1), \dots, E(X_n)]$ , the covariance matrix  $\boldsymbol{\Sigma}_{xx}$ , and the covariance vector  $\boldsymbol{\Sigma}_{xl}$  in Equation (13.29) are also uniquely defined, we can conclude that the coefficients of the linear logistic parametrization are uniquely defined (or ‘identified’) as well. Estimation in the logistic case is more difficult as compared to the linear regression, because, in contrast to  $Y$  (see Rem. 12.2), the random variable  $L = \text{logit}[P(Y=1|\mathcal{C})]$  is nonobservable. For methods of estimation, see, e. g., Agresti (2015) or McCullagh and Nelder (1989).  $\triangleleft$

### 13.5 Linear Logistic Regression and Linear Logit Regression

In Definition 13.10 we defined the concept of a linear logistic parametrization of  $P(Y=1|\mathcal{C})$  in  $X$ . Such parametrizations of  $P(Y=1|\mathcal{C})$  are not necessarily uniquely defined. If there is a linear logistic parametrization  $g$  of  $P(Y=1|\mathcal{C})$  in  $X$  that satisfies Equation (13.30), then we call it *the linear logistic regression of  $Y$  on  $X$* . According to Theorem 13.20, the linear logistic regression is uniquely defined.

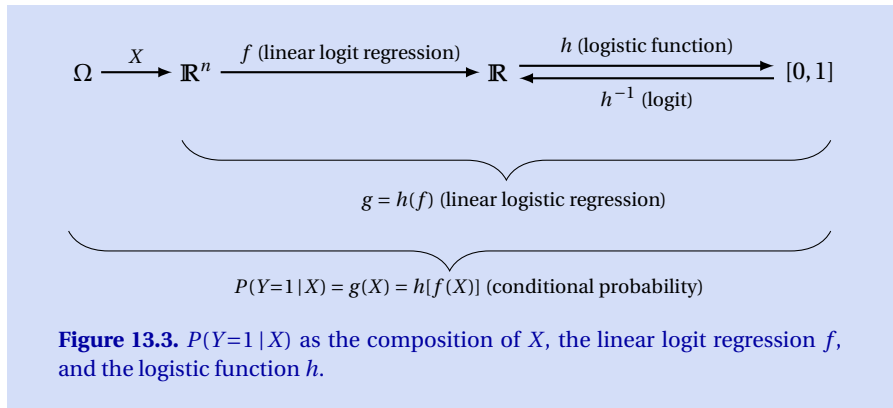
#### Definition 13.22 (Linear Logistic and Linear Logit Regression)

Let the assumptions 13.9 hold and suppose there are an  $P(Y=1|\mathcal{C}) \in \mathcal{P}(Y=1|\mathcal{C})$  and  $\lambda_0, \lambda_1, \dots, \lambda_n \in \mathbb{R}$  such that Equation (13.9) holds. Furthermore, assume that  $Y$  and  $X_1, \dots, X_n$  have finite second moments and that the inverse  $\boldsymbol{\Sigma}_{xx}^{-1}$  of the covariance matrix of  $X = (X_1, \dots, X_n)$  exists. Then the function  $g: \mathbb{R}^n \rightarrow [0, 1]$  defined by Equation (13.30) is called *the linear logistic regression or the linear inverse logit regression* and the function  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  defined by

$$f(x) = \lambda_0 + \boldsymbol{\lambda}'\mathbf{x}, \quad \forall x \in \mathbb{R}^n, \quad (13.31)$$

*the linear logit regression of  $Y$  on  $X$ .*

Figure 13.3 shows  $P(Y=1|X)$  as the composition of the functions introduced above. According to this figure,  $P(Y=1|X)$  is the composition of  $X$  and the linear logistic regression  $g$ , which itself is the composition of the linear logit regression



$f$  and the logistic function  $h$  [see Eq. (13.4)] that transforms a logit into a probability.

**Remark 13.23 (Simple and Multiple Linear Logistic Regression)** If  $n \geq 2$ , then a linear logistic regression is also called a *multiple linear logistic regression*. If  $n = 1$ , then it is also called a *simple linear logistic regression*.  $\triangleleft$

**Example 13.24 (Joe and Ann With Random Assignment – continued)** Table 9.1 shows the random variables  $U$ ,  $X$ , and  $Y$  as well as the conditional expectations  $E(Y|X, U)$  and  $E(Y|X)$ . Because  $X$  is dichotomous with values 0 and 1, the conditional probability  $E(Y|X) = P(Y=1|X)$  can always be written as a linear function of  $X$ . Applying the equations of Example 12.15) yields

$$P(Y=1|X) = .45 + .15 \cdot X \tag{13.32}$$

(see also Example 12.16).

Alternatively, we can also write the *logit* of  $E(Y|X) = P(Y=1|X)$  as a linear function of  $X$ . This yields

$$P(Y=1|X) = \frac{\exp(\alpha_0 + \alpha_1 \cdot X)}{1 + \exp(\alpha_0 + \alpha_1 \cdot X)} \approx \frac{\exp(-.201 + .606 \cdot X)}{1 + \exp(-.201 + .606 \cdot X)} \tag{13.33}$$

(see Exercise 13-5). Inserting the two values of  $X$ , this equation yields the probabilities

$$P(Y=1|X)(\omega) = P(Y=1|X=0) = .45, \quad \text{for } \omega \in \{X=0\}$$

and

$$P(Y=1|X)(\omega) = P(Y=1|X=1) = .60, \quad \text{for } \omega \in \{X=1\},$$

which is consistent with Equation (13.32). The function  $g: \mathbb{R} \rightarrow [0, 1]$  defined by

$$g(x) \approx \frac{\exp(-.201 + .606 \cdot x)}{1 + \exp(-.201 + .606 \cdot x)}, \quad \forall x \in \mathbb{R},$$

is the linear logistic regression of  $Y$  on  $X$ .  $\triangleleft$

**Remark 13.25 (Linear vs. Linear Logistic Regression)** In this particular example, there is no compelling reason to prefer the logistic linear regression over the (ordinary) linear regression. However, this does not apply any more if  $X$  is unbounded (see Rem. 13.8).

Furthermore, note that the standard computer programs (and the underlying statistical models) for linear regressions assume equality (homogeneity) of the conditional variances  $\text{Var}(Y|X=x)$  for different values  $x$  of  $X$ . In contrast, computer programs (and the underlying statistical models) for the analysis of linear logistic regressions allow for heterogeneous ( $X=x$ )-conditional variances of  $Y$ .

Remember that

$$\text{Var}(Y|X=x) = P(Y=1|X=x) \cdot [1 - P(Y=1|X=x)] \quad \text{for } P_X\text{-a.a. } x \in \Omega'_X \quad (13.34)$$

(see Eq. (11.49)). Equation (13.34) shows that  $\text{Var}(Y|X=x)$  depends on  $P(Y=1|X=x)$ . In Example 13.24, the ( $X=0$ )-conditional variance is

$$\text{Var}(Y|X=0) = P(Y=1|X=0) \cdot [1 - P(Y=1|X=0)] = .45 \cdot (1 - .45) = .2475,$$

whereas the ( $X=1$ )-conditional variance is

$$\text{Var}(Y|X=1) = P(Y=1|X=1) \cdot [1 - P(Y=1|X=1)] = .60 \cdot (1 - .60) = .24.$$

◁

**Example 13.26 (Joe and Ann With Random Assignment – continued)** According to Remark 12.23, the logit of  $E(Y|X, U) = P(Y=1|X, U)$  can be written as a linear function in  $X, 1_{U=Ann}, X \cdot 1_{U=Ann}$ . Hence, we can write

$$\begin{aligned} P(Y=1|X, U) &= \frac{\exp(\lambda_0 + \lambda_1 \cdot X + \lambda_2 \cdot 1_{U=Ann} + \lambda_3 \cdot X \cdot 1_{U=Ann})}{1 + \exp(\lambda_0 + \lambda_1 \cdot X + \lambda_2 \cdot 1_{U=Ann} + \lambda_3 \cdot X \cdot 1_{U=Ann})} \\ &\approx \frac{\exp(.847 + .539 \cdot X - 2.234 \cdot 1_{U=Ann} + .442 \cdot X \cdot 1_{U=Ann})}{1 + \exp(.847 + .539 \cdot X - 2.234 \cdot 1_{U=Ann} + .442 \cdot X \cdot 1_{U=Ann})} \end{aligned}$$

(see Exercise 13-6) Inserting the two values of  $X$  and the two values of  $1_{U=Ann}$ , this equation yields the four probabilities  $P(Y=1|X=x, U=u)$  listed in Table 9.1. The function  $g: \mathbb{R}^3 \rightarrow [0, 1]$  defined by

$$g(x) \approx \frac{\exp(.847 + .539 \cdot x_1 - 2.234 \cdot x_2 + .442 \cdot x_3)}{1 + \exp(.847 + .539 \cdot x_1 - 2.234 \cdot x_2 + .442 \cdot x_3)}, \quad \forall x \in \mathbb{R}^3,$$

is the linear logistic regression of  $Y$  on  $X = (X_1, X_2, X_3) = (X, 1_{U=Ann}, X \cdot 1_{U=Ann})$ .

Rearranging the equation for  $P(Y=1|X, U)$  yields

$$\begin{aligned} P(Y=1|X, U) &\stackrel{P}{=} \frac{\exp((\lambda_0 + \lambda_2 \cdot 1_{U=Ann}) + (\lambda_1 + \lambda_3 \cdot 1_{U=Ann}) \cdot X)}{1 + \exp((\lambda_0 + \lambda_2 \cdot 1_{U=Ann}) + (\lambda_1 + \lambda_3 \cdot 1_{U=Ann}) \cdot X)} \\ &\approx \frac{\exp((.847 - 2.234 \cdot 1_{U=Ann}) + (.539 + .442 \cdot 1_{U=Ann}) \cdot X)}{1 + \exp((.847 - 2.234 \cdot 1_{U=Ann}) + (.539 + .442 \cdot 1_{U=Ann}) \cdot X)}, \end{aligned}$$

showing that the logit is  $f_0(U) + f_1(U) \cdot X$  with

$$f_0(U) = \lambda_0 + \lambda_2 \cdot 1_{U=Ann} \approx .847 - 2.234 \cdot 1_{U=Ann}$$

and

$$f_1(U) = \lambda_1 + \lambda_3 \cdot 1_{U=Ann} = .539 + .442 \cdot 1_{U=Ann}.$$

The function  $f_0(U)$  is called the *logit intercept function* and  $f_1(U)$  the *logit effect function*. Note that

$$\begin{aligned} \alpha_1 &\approx .606 \\ &\neq E[f_1(U)] \approx .539 + .442 \cdot E(1_{U=Ann}) \approx .7599. \end{aligned}$$

Hence, although  $X$  and  $U$  are independent (see Example 5.37), the slope  $\alpha_1$  of the logit in the logistic parametrization of  $E(Y|X) = P(Y=1|X)$  is *not* equal to the expectation of the logit effect function  $f_1(U)$  of the logit in the logistic parametrization of  $E(Y|X, U) = P(Y=1|X, U)$ .

From a methodological point of view this means that randomized assignment of a unit to one of two treatment conditions—which creates independence of a treatment variable  $X$  and the person variable  $U$ —does not imply that the slope  $\alpha_1$  of the logit in the logistic parametrization of  $E(Y|X) = P(Y=1|X)$  can be interpreted as an average effect of treatment variable on  $Y$  [or on the logit of the linear logit parametrization of  $E(Y|X)$ ]. In examples in which  $f_1(U) = \lambda_1$  is a constant, this implies that  $\alpha_1 = \lambda_1$  does not follow from independence of  $X$  and  $U$ . In contrast, compare the corresponding invariance property formulated in Theorem 12.44 for a linear parametrization. <

**Remark 13.27 (Normal Distribution and Linear Logistic Regression)** Let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  be a continuous random variable, let  $A \in \mathcal{A}$  with  $0 < P(A) < 1$ , and let  $Y = 1_A$ . Furthermore, assume that  $X$  has a normal distribution with respect to each of the two conditional-probability measures  $P^{Y=y}$ ,  $y = 0, 1$ , and that  $Var(X|Y=0) = Var(X|Y=1)$ . Then  $P(Y=1|X)$  has a linear logistic parametrization (for a proof see Examples 17.80 to 17.82). <

**Example 13.28 (Joe and Ann With Latent Abilities)** Table 13.1 describes a random experiment that consists of sampling a person from the set  $\Omega_U := \{Joe, Ann\}$  of persons and observing whether (+) or not (−) problem 1, 2, and 3, respectively, is solved. The probabilities of the elementary events are displayed in the second column, which can be used to compute the values of the conditional probabilities  $P(Y_i=1|U)$  for Joe and for Ann, or more precisely, the values  $P(Y_i=1|U)(\omega)$  if  $\omega \in \{U=Joe\}$  and if  $\omega \in \{U=Ann\}$ . The last three columns present the logit transformations of these conditional probabilities, i. e.,

$$\text{logit}_i := \text{logit}[P(Y_i=1|U)] = \ln\left(\frac{P(Y_i=1|U)}{1 - P(Y_i=1|U)}\right), \quad \forall i = 1, \dots, m, \quad (13.35)$$

(see Def. 13.5), where  $m = 3$ . The probability space  $(\Omega, \mathcal{A}, P)$  is completely specified by  $\Omega := \{\omega_1, \dots, \omega_{16}\}$ ,  $\mathcal{A} := \mathcal{P}(\Omega)$ , and the sixteen probabilities  $P(\{\omega_i\})$  of the

**Table 13.1.** Joe and Ann With a Latent Ability Variable

Outcomes $\omega$				Observables				Conditional probabilities and their logits						
Unit	Problem 1	Problem 2	Problem 3	$P(\{\omega\})$	$U$	$Y_1$	$Y_2$	$Y_3$	$P(Y_1=1 U)$	$P(Y_2=1 U)$	$P(Y_3=1 U)$	$\text{logit}_1$	$\text{logit}_2$	$\text{logit}_3$
$\omega_1 = (\text{Joe}, -, -, -)$														
$\omega_2 = (\text{Joe}, -, -, +)$				0.0181	Joe	0	0	1	0.5000	0.7311	0.2689	0	1	-1
$\omega_3 = (\text{Joe}, -, +, -)$				0.1336	Joe	0	1	0	0.5000	0.7311	0.2689	0	1	-1
$\omega_4 = (\text{Joe}, +, -, -)$				0.0492	Joe	1	0	0	0.5000	0.7311	0.2689	0	1	-1
$\omega_5 = (\text{Joe}, -, +, +)$				0.0492	Joe	0	1	1	0.5000	0.7311	0.2689	0	1	-1
$\omega_6 = (\text{Joe}, +, -, +)$				0.0181	Joe	1	0	1	0.5000	0.7311	0.2689	0	1	-1
$\omega_7 = (\text{Joe}, +, +, -)$				0.1336	Joe	1	1	0	0.5000	0.7311	0.2689	0	1	-1
$\omega_8 = (\text{Joe}, +, +, +)$				0.0492	Joe	1	1	1	0.5000	0.7311	0.2689	0	1	-1
$\omega_9 = (\text{Ann}, -, -, -)$				0.0080	Ann	0	0	0	0.7311	0.8808	0.5000	1	2	0
$\omega_{10} = (\text{Ann}, -, -, +)$				0.0080	Ann	0	0	1	0.7311	0.8808	0.5000	1	2	0
$\omega_{11} = (\text{Ann}, -, +, -)$				0.0592	Ann	0	1	0	0.7311	0.8808	0.5000	1	2	0
$\omega_{12} = (\text{Ann}, +, -, -)$				0.0218	Ann	1	0	0	0.7311	0.8808	0.5000	1	2	0
$\omega_{13} = (\text{Ann}, -, +, +)$				0.0592	Ann	0	1	1	0.7311	0.8808	0.5000	1	2	0
$\omega_{14} = (\text{Ann}, +, -, +)$				0.0218	Ann	1	0	1	0.7311	0.8808	0.5000	1	2	0
$\omega_{15} = (\text{Ann}, +, +, -)$				0.1610	Ann	1	1	0	0.7311	0.8808	0.5000	1	2	0
$\omega_{16} = (\text{Ann}, +, +, +)$				0.1610	Ann	1	1	1	0.7311	0.8808	0.5000	1	2	0

Note: Logits are exact, whereas the probabilities are rounded.

elementary events  $\omega_i \in \Omega$ . Also note that the conditional probabilities  $P(Y_i=1|U)$  and their logits are completely determined by the sixteen probabilities  $P(\{\omega_i\})$ .

The probability measure  $P$  is such that it satisfies the following two conditions for  $m = 3$ , the number of items considered in Table 13.1:

$$\forall i, j \in \{1, \dots, m\} \exists \beta_{ij} \in \mathbb{R}: \text{logit}_i - \text{logit}_j = \beta_{ij}, \quad (13.36)$$

$$P(Y_i=1|U, Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_m) = P(Y_i=1|U), \quad \forall i = 1, \dots, m. \quad (13.37)$$

While (13.36) is easily checked by inspecting the last three columns of Table 13.1, we delay showing that (13.37) holds to Example 16.51. If (13.36) and (13.37) hold, then we say that  $Y_1$ ,  $Y_2$ , and  $Y_3$  satisfy the (two assumptions of the) Rasch model (cf. Rasch, 1960/1980). Assumption (13.36) is called *Rasch homogeneity*. According to this assumption, the logits of each pair of items are translations of each other. Assumption (13.37) postulates *U*-conditional mean independence of each item  $Y_i$  from the other items  $Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_m$  [see Def. 10.45 (ii) and Rem. 10.46 (ii)]. In chapter 16 it is shown that this assumption is equivalent to *U*-conditional independence of the random variables  $Y_i$  (see Example 16.51 for more details).

Now define

$$\xi := \text{logit}_1 \quad \text{and} \quad \beta_i := \beta_{1i}, \quad \forall i = 1, \dots, m. \quad (13.38)$$

In applications in which the  $Y_i$  indicate whether or not a problem is solved, the random variable  $\xi$  is called the *latent ability variable*, and  $\beta_i$  the *difficulty parameter of problem  $i$* , respectively. Inserting the definitions of  $\xi$  and  $\beta_i$  into (13.36) we receive

$$\text{logit}_i = \xi - \beta_i, \quad \forall i = 1, \dots, m. \quad (13.39)$$

Furthermore, inserting Equation (13.39) into (13.35) and solving it for  $P(Y_i=1|U)$  yields

$$P(Y_i=1|U) = \frac{\exp(\xi - \beta_i)}{1 + \exp(\xi - \beta_i)} \quad (13.40)$$

$$= P(Y_i=1|\xi), \quad \forall i = 1, \dots, m \quad (13.41)$$

[see Eq. (13.6) and Exercise 13-3]. According to Equation (13.41), the logits of the conditional probabilities  $P(Y_i=1|\xi)$  and  $P(Y_j=1|\xi)$  are translations of each other. Figure 13.4 shows the graphs of the factorizations  $g: \mathbb{R} \rightarrow [0, 1]$  of the conditional probabilities  $P(Y_i=1|\xi) = g(\xi)$ , i. e., the graphs of the regressions of  $Y_i$  on  $\xi$  (see Def. 10.25). This figure illustrates that these graphs are translations (parallel to the  $\xi$  axis) of each other.

The difficulty parameters  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$  can be computed as follows: Using Equations (13.36), (13.38), and the logits displayed in Table 13.1, we receive

$$\begin{aligned} \beta_1 &= \beta_{11} = \text{logit}_1 - \text{logit}_1 = 0, \\ \beta_2 &= \beta_{12} = \text{logit}_1 - \text{logit}_2 = -1, \\ \beta_3 &= \beta_{13} = \text{logit}_1 - \text{logit}_3 = 1. \end{aligned} \quad (13.42)$$

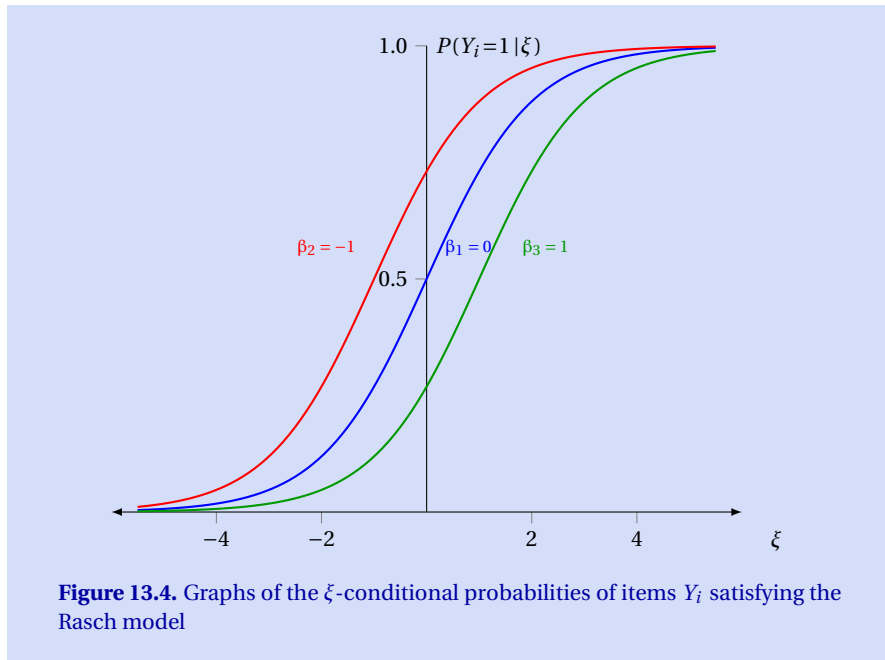
Note that the definitions of the latent variable  $\xi$  and the difficulty parameters  $\beta_i$  is arbitrary to some degree. Defining  $\xi^* := \xi + a$  and  $\beta_i^* := \beta_i + a$  for any  $a \in \mathbb{R}$  would do as well, i. e.,  $\xi^*$  and  $\beta_i^*$  also satisfy Equation (13.40).

Also note that the definition of  $\xi$  and Corollary 2.53 imply that there is a function  $f: \Omega_U \rightarrow \mathbb{R}$  such that  $\xi$  is the composition  $f \circ U = f(U)$  of  $U$  and  $f$ . The function  $f$  assigns to each  $u \in \Omega_U$  its value  $f(u)$ , where

$$\begin{aligned} \forall u \in \Omega_U \forall \omega \in \{U=u\}: \quad f(u) &= f[U(\omega)] = \xi(\omega) && [(2.25)] \\ &= \text{logit}[P(Y_1=1|U)(\omega)] && [(13.35), (13.38)] \\ &= \text{logit}[P(Y_1=1|U=u)]. && [(10.9)] \end{aligned} \quad (13.43)$$

Hence, the values of  $\xi$  are attributes of the persons in the set  $\Omega_U$ . This justifies calling such a value the *ability of person  $u$* , if the items  $Y_i$  indicate whether or not problem  $i$  is solved.

Finally, note that the values of the latent variable  $\xi$  are unknown in empirical applications. However, in Example 16.51 we treat some details on estimating the



value  $f(u)$  from the response pattern  $(y_1, \dots, y_m)$  of the items  $Y_1, \dots, Y_m$  that is observed for a person  $u \in \Omega_U$  if the random experiment is actually conducted, i. e., if we actually sample a person from a set of persons and observe, for  $i = 1, \dots, m$ , whether or not problem  $i$  is solved.  $\triangleleft$

### 13.6 Proofs

#### Proof of Lemma 13.7

Let  $V: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  be a version in  $\mathcal{P}(Y=1|\mathcal{C})$  with values in  $]0, 1[$ . Furthermore, let  $g: ]0, 1[ \rightarrow \mathbb{R}$  be a continuous and strictly monotone function with  $g(]0, 1[) = \mathbb{R}$ . Then the inverse function  $g^{-1}: \mathbb{R} \rightarrow ]0, 1[$  exists (see Ellis & Gulick, 2006, section 7.1) and it is continuous (Ellis & Gulick, 2006, Th. 7.4, p. 436) and strictly monotone with  $g^{-1}(\mathbb{R}) = ]0, 1[$ . Hence,

$$\begin{aligned}
 & \sigma[g(V)] \\
 &= (g \circ V)^{-1}(\mathcal{B}) && [(2.14)] \\
 &= V^{-1} \circ g^{-1}(\mathcal{B}) && [(2.29)] \\
 &= V^{-1} \circ g^{-1}[\sigma(\{]-\infty, b]: b \in \mathbb{R}\})] && [(1.19)] \\
 &= \sigma[V^{-1} \circ g^{-1}(\{]-\infty, b]: b \in \mathbb{R}\})] && [(2.12)] \\
 &= \sigma[V^{-1}(\{]-\infty, g^{-1}(b)]: b \in \mathbb{R}\})] && [\text{monotonicity, continuity}]
 \end{aligned}$$

$$\begin{aligned}
&= \sigma [V^{-1}(\{1 - \infty, c\} : c \in ]0, 1[\})] && \text{[domain of } g\text{]} \\
&= \sigma [V^{-1}(\{1 - \infty, b\} : b \in \mathbf{R}\})] && [b \leq 0: V^{-1}(1 - \infty, b) = \emptyset, b \geq 1: V^{-1}(1 - \infty, b) = \mathbf{R}] \\
&= V^{-1}[\sigma(\{1 - \infty, b\} : b \in \mathbf{R}\})] && \text{[(2.12)]} \\
&= V^{-1}(\mathcal{B}) && \text{[(1.19)]} \\
&= \sigma(V). && \text{[(2.14)]}
\end{aligned}$$

The result  $\sigma[g(V)] = \sigma(V)$  can now be applied to  $g = \text{logit}$  [see Eq. (13.2)].

### **Proof of Theorem 13.14**

By definition,  $P(Y=1|Z) = E(1_{Y=1}|Z)$ . Hence, the existence of coefficients  $\beta_0, \beta_1, \dots, \beta_n$  and a version  $P(Y=1|Z) \in \mathcal{P}(Y=1|Z)$  satisfying Equations (13.15), (13.17), and (13.19) has already been proved in Theorem 12.37. In order to show that there are  $\lambda_0, \lambda_1, \dots, \lambda_n$  satisfying Equation (13.16) we define

$$\lambda_0 := \text{logit}[P(Y=1|Z=z_0)], \quad (13.44)$$

[see Eq. (13.2)] and

$$\lambda_i := \text{logit}[P(Y=1|Z=z_i)] - \text{logit}[P(Y=1|Z=z_0)], \quad \forall i = 1, \dots, n. \quad (13.45)$$

These definitions and Equation (13.3) then yield

$$P(Y=1|Z=z_0) = \frac{\exp(\text{logit}[P(Y=1|Z=z_0)])}{1 + \exp(\text{logit}[P(Y=1|Z=z_0)])} = \frac{\exp(\lambda_0)}{1 + \exp(\lambda_0)},$$

and

$$P(Y=1|Z=z_i) = \frac{\exp(\text{logit}[P(Y=1|Z=z_i)])}{1 + \exp(\text{logit}[P(Y=1|Z=z_i)])} = \frac{\exp(\lambda_0 + \lambda_i)}{1 + \exp(\lambda_0 + \lambda_i)}, \quad \forall i = 1, \dots, n.$$

Hence, Equation (13.17) implies

$$\beta_0 = P(Y=1|Z=z_0) = \frac{\exp(\lambda_0)}{1 + \exp(\lambda_0)},$$

and Equation (13.19) yields

$$\begin{aligned}
\beta_i &= P(Y=1|Z=z_i) - P(Y=1|Z=z_0) \\
&= \frac{\exp(\lambda_0 + \lambda_i)}{1 + \exp(\lambda_0 + \lambda_i)} - \frac{\exp(\lambda_0)}{1 + \exp(\lambda_0)}, \quad \forall i = 1, \dots, n.
\end{aligned}$$

### **Proof of Theorem 13.20**

Denote  $\mathbf{x} := [X_1, \dots, X_n]'$ ,  $\boldsymbol{\mu} := [E(X_1), \dots, E(X_n)]'$ ,  $\boldsymbol{\lambda} = [\lambda_1, \dots, \lambda_n]'$ , as well as

$$L := \text{logit}[P(Y=1|\mathcal{C})] = \lambda_0 + \boldsymbol{\lambda}' \mathbf{x}$$

[see Eqs. (13.27) and (13.9)]. Taking the expectation on both sides, using the definition of  $\boldsymbol{\mu}$ , and rearranging yields

$$\lambda_0 = E(L) - \boldsymbol{\lambda}' \boldsymbol{\mu}.$$

Furthermore, consider the  $n$ -dimensional covariance vector

$$\begin{aligned} \boldsymbol{\Sigma}_{xI} &= \boldsymbol{\Sigma}_{x, \lambda_0 + \boldsymbol{\lambda}' x} && [(13.27)] \\ &= \boldsymbol{\Sigma}_{xx} \boldsymbol{\lambda}. && [\text{Box 7.3 (ii), (iii)}] \end{aligned}$$

Multiplying both sides by  $\boldsymbol{\Sigma}_{xx}^{-1}$  yields

$$\boldsymbol{\lambda} = \boldsymbol{\Sigma}_{xx}^{-1} \boldsymbol{\Sigma}_{xI}.$$

This equation also shows that the vector  $\boldsymbol{\lambda}$  is uniquely defined, and this implies that  $\lambda_0$  is uniquely defined as well. Uniqueness of  $\lambda_0$  and  $\boldsymbol{\lambda}$  implies that the linear logit parametrization  $g$  satisfying Equation (13.30) is uniquely defined as well.

### 13.7 Exercises

- ▷ **Exercise 13-1** Prove the propositions of Remark 13.2.
- ▷ **Exercise 13-2** Prove Equation (13.3).
- ▷ **Exercise 13-3** die Gleichung mit  $\zeta$ .
- ▷ **Exercise 13-4** Calculate the derivative (13.14).
- ▷ **Exercise 13-5** Consider Example 13.24 and compute the coefficients of the linear logit parametrization of  $E(Y|X)$ .
- ▷ **Exercise 13-6** Consider Example 13.24 and compute the coefficients of the linear logit parametrization of  $E(Y|X, U)$ .

### Solutions

▷ **Solution 13-1** Note that  $1_{Y=0} \cdot Y = 1_{Y=0} \cdot 0 = 0$  [see Eq. (2.49)], Hence, if  $\{Y=0\} \in \mathcal{C}$ , then Definition 10.2 (b) yields

$$\int 1_{Y=0} \cdot P(Y=1|\mathcal{C}) dP = \int 1_{Y=0} \cdot Y dP = \int 0 dP = 0. \quad (13.46)$$

Let  $P(Y=1|\mathcal{C}) \in \mathcal{P}(Y=1|\mathcal{C})$  be a version with  $P(Y=1|\mathcal{C}) \geq 0$  [for its existence see Box 10.3 (v)]. Then Theorem 3.43 yields  $1_{Y=0} \cdot P(Y=1|\mathcal{C}) \stackrel{p}{=} 0$ , i. e.,  $P(A) = 0$  for  $A := \{\omega \in \Omega: 1_{Y=0}(\omega) \cdot P(Y=1|\mathcal{C})(\omega) \neq 0\}$ . This implies  $P(A^c) = 1$  and

$$\{Y=0\} \cap A^c \subset \{P(Y=1|\mathcal{C}) = 0\} \cap A^c,$$

which in turn yields

$$\begin{aligned} &P(P(Y=1|\mathcal{C}) = 0) \\ &= P(\{P(Y=1|\mathcal{C}) = 0\} \cap A^c) && [\text{Box 4.1 (viii)}] \end{aligned}$$

$$\begin{aligned}
&\geq P(\{Y=0\} \cap A^c) && \text{[Box 4.1 (v)]} \\
&= P(Y=0) && \text{[Box 4.1 (viii)]} \\
&> 0. && \text{[Y dichotomous, Example 5.10]}
\end{aligned}$$

Hence, we have shown that  $\{Y=0\} \in \mathcal{C}$  implies  $P(P(Y=1|\mathcal{C}) = 0) > 0$ , which in turn implies that there is no version  $P(Y=1|\mathcal{C}) > 0$ , i. e., there is no version  $P(Y=1|\mathcal{C})$  with  $\{\omega \in \Omega: P(Y=1|\mathcal{C})(\omega) \leq 0\} = \emptyset$  [see Rem. 2.77 and (10.12)]. Now,

$$\begin{aligned}
\{Y=1\} \in \mathcal{C} &\Rightarrow \int 1_{Y=1} \cdot Y \, dP = \int 1_{Y=1} \cdot P(Y=1|\mathcal{C}) \, dP && \text{[Def. 10.2 (b)]} \\
&\Rightarrow \int (1 - 1_{Y=0}) \cdot Y \, dP = \int (1 - 1_{Y=0}) \cdot P(Y=1|\mathcal{C}) \, dP && \text{[Y dichotomous, (5.33)]} \\
&\Rightarrow E(Y) - \int 1_{Y=0} \cdot Y \, dP = E(Y) - \int 1_{Y=0} \cdot P(Y=1|\mathcal{C}) \, dP && \text{[(3.33), (6.1)]} \\
&\Rightarrow \int 1_{Y=0} \cdot Y \, dP = \int 1_{Y=0} \cdot P(Y=1|\mathcal{C}) \, dP. && \text{[E(Y) is finite]}
\end{aligned}$$

The last equation is identical to (13.46). Hence, we can also conclude that there is no version  $P(Y=1|\mathcal{C}) > 0$  if  $\{Y=1\} \in \mathcal{C}$ .

▷ **Solution 13-2**

$$\exp[\text{logit}(p)] = \exp\left[\ln\left(\frac{p}{1-p}\right)\right] = \frac{p}{1-p}.$$

Hence,

$$p = \frac{\frac{p}{1-p}}{\frac{1}{1-p}} = \frac{\frac{p}{1-p}}{\frac{1-p+p}{1-p}} = \frac{\frac{p}{1-p}}{\frac{1-p}{1-p} + \frac{p}{1-p}} = \frac{\frac{p}{1-p}}{1 + \frac{p}{1-p}} = \frac{\exp[\text{logit}(p)]}{1 + \exp[\text{logit}(p)]}.$$

▷ **Solution 13-3**

▷ **Solution 13-4** The chain rule and the quotient rule of differential calculus yield

$$\begin{aligned}
\frac{d}{dx} g(x) &= \frac{d}{dx} \frac{\exp(\lambda_0 + \lambda_1 x)}{1 + \exp(\lambda_0 + \lambda_1 x)} \\
&= \frac{\lambda_1 \exp(\lambda_0 + \lambda_1 x) (1 + \exp(\lambda_0 + \lambda_1 x)) - \exp(\lambda_0 + \lambda_1 x) \cdot \lambda_1 \exp(\lambda_0 + \lambda_1 x)}{(1 + \exp(\lambda_0 + \lambda_1 x))^2} \\
&= \frac{\lambda_1 \exp(\lambda_0 + \lambda_1 x) (1 + \exp(\lambda_0 + \lambda_1 x) - \exp(\lambda_0 + \lambda_1 x))}{(1 + \exp(\lambda_0 + \lambda_1 x))^2} \\
&= \frac{\lambda_1 \exp(\lambda_0 + \lambda_1 x)}{(1 + \exp(\lambda_0 + \lambda_1 x))^2}.
\end{aligned}$$

▷ **Solution 13-5** Inserting the value  $x = 0$  in the equation

$$\text{logit}[P(Y=1|X=x)] = \ln\left(\frac{P(Y=1|X=x)}{1 - P(Y=1|X=x)}\right) = \alpha_0 + \alpha_1 x$$

[see Eqs. (13.2) and (13.13)] yields

$$\text{logit}[P(Y=1|X=0)] = \ln\left(\frac{.45}{1-.45}\right) \approx -.201 \approx \alpha_0,$$

and inserting the value  $x = 1$  yields

$$\text{logit}[P(Y=1|X=1)] = \ln\left(\frac{.6}{1-.6}\right) \approx .406 \approx \alpha_0 + \alpha_1.$$

[In R these values are obtained by `qlogis(.45)` and `qlogis(.60)`, respectively.] Solving the last equation yields  $\alpha_1 \approx .406 - (-.201) = .606$ .

▷ **Solution 13-6** For  $x_1 = 0$ ,  $x_2 = 0$ , and  $x_3 = 0$ , the equation

$$\text{logit}[P(Y=1|X=0, U=Ann)] = \ln\left(\frac{P(Y=1|X=0, U=Ann)}{1-P(Y=1|X=0, U=Ann)}\right) = \lambda_0 + \lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3$$

[see Eqs. (13.2) and (13.13)] yields

$$\ln\left(\frac{P(Y=1|X=0, U=Joe)}{1-P(Y=1|X=0, U=Joe)}\right) = \ln\left(\frac{.7}{1-.7}\right) \approx -.847 = \lambda_0,$$

for  $x_1 = 1$ ,  $x_2 = 0$ , and  $x_3 = 0$  it yields

$$\ln\left(\frac{P(Y=1|X=1, U=Joe)}{1-P(Y=1|X=1, U=Joe)}\right) = \ln\left(\frac{.8}{1-.8}\right) \approx 1.386 \approx \lambda_0 + \lambda_1,$$

for  $x_1 = 0$ ,  $x_2 = 1$ , and  $x_3 = 0$  it yields

$$\ln\left(\frac{P(Y=1|X=0, U=Ann)}{1-P(Y=1|X=0, U=Ann)}\right) = \ln\left(\frac{.2}{1-.2}\right) \approx -1.386 \approx \lambda_0 + \lambda_2,$$

and for  $x_1 = 1$ ,  $x_2 = 1$ , and  $x_3 = 1$  it yields

$$\ln\left(\frac{P(Y=1|X=1, U=Ann)}{1-P(Y=1|X=1, U=Ann)}\right) = \ln\left(\frac{.4}{1-.4}\right) \approx -.406 \approx \lambda_0 + \lambda_1 + \lambda_2 + \lambda_3.$$

[In R these values are obtained by `qlogis(.7)` to `qlogis(.4)`.] Solving the second equation yields  $\lambda_1 \approx .539$ , solving the third equation yields  $\lambda_2 \approx -2.234$ , and solving the last one yields  $\lambda_3 \approx .442$ .

## Chapter 14

# Conditional Expectation With Respect to a Conditional-Probability Measure

In chapter 10 we treated the concept of a  $\mathcal{C}$ -conditional expectation  $E(Y|\mathcal{C})$  with respect to a probability measure  $P$  on a measurable space  $(\Omega, \mathcal{A})$ . In this chapter we introduce the concept of a  $\mathcal{C}$ -conditional expectation  $E^B(Y|\mathcal{C})$  of  $Y$  with respect to the *conditional-probability measure*  $P^B$  on  $(\Omega, \mathcal{A})$  (see Def. 4.29). A special case with  $\mathcal{C} = \sigma(X)$  is the  $X$ -conditional expectation of  $Y$  with respect to  $P^B$ , which is also denoted by  $E^B(Y|X)$ . If  $B = \{Z=z\}$  is the event that a random variable  $Z$  on  $(\Omega, \mathcal{A}, P)$  takes on the value  $z$  and  $P(Z=z) > 0$ , then we use the notation  $E^{Z=z}(Y|X)$  and call it a version of the  $X$ -conditional expectation of  $Y$  with respect to  $P^{Z=z}$ .

In empirical applications the conditional expectation  $E^{Z=z}(Y|X)$  can be used to describe how the conditional expectation values of  $Y$  depend on the values  $x$  of  $X$  given that  $Z$  takes on the value  $z$ . The dependency of  $Y$  on  $X$  described by  $E^{Z=z}(Y|X)$  may not only differ for different values  $z_1$  and  $z_2$  of  $Z$ , but it may also differ from the dependency described by the  $X$ -conditional expectation  $E(Y|X)$  of  $Y$  with respect to  $P$ . If, for instance,  $X$  denotes a treatment variable and  $Z = \text{sex}$  with values  $m$  (male) and  $f$  (female), then  $E^{Z=m}(Y|X)$  and  $E^{Z=f}(Y|X)$  refer to the  $X$ -conditional expectation of  $Y$  for males and females, respectively. In a data sample, these are the conditional expectations estimated using only the  $y$ -values and  $x$ -values obtained within the male and female subsamples, respectively. In contrast,  $E^{X=x}(Y|Z)$  refers to the  $Z$ -conditional expectation of  $Y$  given treatment  $x$ , and this is the conditional expectation estimated in the analysis of experimental or quasi-experimental data using only the  $y$ -values and  $z$ -values obtained in treatment condition  $x$ . If the treatment variable  $X$  is dichotomous with values 0 (control) and 1 (treatment), then  $g_1(Z) := E^{X=1}(Y|Z) - E^{X=0}(Y|Z)$  and  $g_1$  is the  $Z$ -conditional effect function of  $X$ . The values  $g_1(z)$  are the effects of  $X$  on  $Y$  given the value  $z$  of  $Z$ .

### 14.1 Introductory Examples

**Example 14.1 (Joe and Ann With Random Assignment – continued)** Table 14.1 displays the random variables  $U$ ,  $X$ ,  $Y$ , and, among other things, the conditional expectations  $E(Y|X)$  and  $E(Y|X, U)$ , the values of which have already been computed in Example 9.21. Here, the conditional expectation  $E(Y|X, U)$  is uniquely defined. According to Remark 12.23 it has a linear parametrization in

**Table 14.1.** Joe and Ann With Random Assignment: Conditional Expectations With Respect to  $P^{X=x}$

Outcomes $\omega$			Observables			Conditional expectations							
Unit	Treatment	Success	Person variable $U$	Treatment variable $X$	Outcome variable $Y$	$E(Y X, U)$	$E(Y X)$	$P(X=1 U)$	$E^{X=0}(Y U)$	$E^{X=1}(Y U)$	$g_1(U)$	$P^{X=0}(\{\omega\})$	$P^{X=1}(\{\omega\})$
(Joe, no, -)			Joe	0	0	.7	.45	.4	.7	.8	.1	.15	0
(Joe, no, +)			Joe	0	1	.7	.45	.4	.7	.8	.1	.35	0
(Joe, yes, -)			Joe	1	0	.8	.6	.4	.7	.8	.1	0	.1
(Joe, yes, +)			Joe	1	1	.8	.6	.4	.7	.8	.1	0	.4
(Ann, no, -)			Ann	0	0	.2	.45	.4	.2	.4	.2	.4	0
(Ann, no, +)			Ann	0	1	.2	.45	.4	.2	.4	.2	.1	0
(Ann, yes, -)			Ann	1	0	.4	.6	.4	.2	.4	.2	0	.3
(Ann, yes, +)			Ann	1	1	.4	.6	.4	.2	.4	.2	0	.2

$(1_{U=Joe}, X, 1_{U=Joe} \cdot X)$ . The coefficient of this linear parametrization can be computed analogously to Example 12.24, which yields

$$\begin{aligned}
 E(Y|X, U) &= .2 + .5 \cdot 1_{U=Joe} + .2 \cdot X - .1 \cdot 1_{U=Joe} \cdot X \\
 &= (.2 + .5 \cdot 1_{U=Joe}) + (.2 - .1 \cdot 1_{U=Joe}) \cdot X \\
 &= g_0(U) + g_1(U) \cdot X.
 \end{aligned}
 \tag{14.1}$$

In this equation,  $g_0$  is the  $U$ -conditional *intercept function* that assigns the person-specific intercept to each value  $u$  of the person variable  $U$ , and  $g_1$  is the  $U$ -conditional *effect function* that assigns the person-specific effect of  $X$  on  $Y$  to each value  $u$  of  $U$ . The function  $g_1(U)$  is a random variable on  $(\Omega, \mathcal{A}, P)$ . It is the composition of the person variable  $U$  and the effect function  $g_1$ . In this example, the treatment effect for  $u = Joe$  is  $g_1(Joe) = .2 - .1 \cdot 1 = .1$ , and for  $u = Ann$  it is  $g_1(Ann) = .2 - .1 \cdot 0 = .2$ . Hence,

$$g_1(U)(\omega) = .2 - .1 \cdot 1_{U=Joe}(\omega) = \begin{cases} .1, & \text{if } \omega \in \{U=Joe\}, \\ .2, & \text{if } \omega \in \{U=Ann\}. \end{cases}$$

In Theorem 15.3 we show that Equation (14.1) implies

$$g_0(U) = E^{X=0}(Y|U)
 \tag{14.2}$$

and

$$g_1(U) = E^{X=1}(Y|U) - E^{X=0}(Y|U), \quad (14.3)$$

where  $E^{X=x}(Y|U)$ ,  $x = 0, 1$ , denotes the  $U$ -conditional expectation of  $Y$  with respect to the conditional-probability measure  $P^{X=x}$ .

Furthermore, because  $E(1_{U=Joe}) = P(U=Joe) = .5$  [see Eq. (6.4)], the expectation of the effect function is

$$E_U(g_1) = E[g_1(U)] = E(.2 - .1 \cdot 1_{U=Joe}) = .2 - .1 \cdot E(1_{U=Joe}) = .2 - .1 \cdot .5 = .15 \quad (14.4)$$

[see Eq. (6.13)]. For simplicity, this expectation is also called the *average treatment effect*. In this example,  $E[g_1(U)]$  is also the slope of  $X$  in the equation

$$E(Y|X) = .45 + .15 \cdot X, \quad (14.5)$$

which can be computed applying the Equations of Example 12.15 (see also Example 12.16).

From a methodological point of view, note that in this example the function  $g_1$  and its expectation have a causal interpretation as a  $U$ -conditional effect function and the average effect of the treatment. Furthermore, the slope of  $X$  in Equation (14.5) is also identical to the average causal effect of the treatment variable  $X$ . As shown in Corollary 15.18, this follows from independence of  $X$  and  $U$  [see the column headed  $P(X=1|U)$  in Table 14.1, Example 5.37, and Rem. 10.59]. In empirical applications, in which draw one person and then assign him or her to one of two treatment conditions (see Table 14.1), independence of  $X$  and  $U$  is created by assigning the drawn person to treatment  $x$  with identical probabilities for all persons such that  $P(X=1|U) = P(X=1)$  (see Rem. 10.50).  $\triangleleft$

**Example 14.2 (Joe and Ann With Self-Selection – continued)** Table 14.2 shows another example with Joe and Ann. Some of the conditional expectations displayed have already been presented in Table 11.2. In Table 14.2, the treatment probability *does* depend on the person drawn, i. e.  $P(X=1|U) \neq P(X=1)$  [see the column headed  $P(X=1|U)$ ]. Therefore, according to Remark 10.50,  $X$  and  $U$  are not independent. In applied statistics, such a dependence between  $X$  and  $U$  is often created by ‘self-selection’ into one of the treatment conditions.

Inspecting the columns headed by  $E(Y|X, U)$  in Tables 14.1 and 14.2 shows that these columns are identical. Therefore, Equation (14.1) still holds for  $E(Y|X, U)$ , and this implies that the conditional treatment effect for  $u = Joe$  is again  $g_1(Joe) = .1$ , and for  $u = Ann$  it is  $g_1(Ann) = .2$ . Furthermore, because  $P(U=Joe) = .5$  still holds, the average treatment effect is again  $E[g_1(U)] = .15$ . [see also Eq. (14.4)]. In contrast to Example 14.1,  $E[g_1(U)]$  is *not identical* to the slope of  $X$  in the equation

$$E(Y|X) = .60 - .18 \cdot X, \quad (14.6)$$

which can be computed applying the Equations of Example 12.15 (see also Example 12.16). Now the slope  $-.18$  has no causal interpretation, because all individual

**Table 14.2.** Joe and Ann With Self-Selection: Conditional Expectations With Respect to  $P^{X=x}$

Unit Treatment Success	Outcomes $\omega$	Observables			Conditional expectations							
		$P(\{\omega\})$	Person variable $U$	Treatment variable $X$	Outcome variable $Y$	$E(Y X, U)$	$E(Y X)$	$P(X=1 U)$	$E^{X=0}(Y U)$	$E^{X=1}(Y U)$	$g_1(U)$	$P^{X=0}(\{\omega\})$
(Joe, no, -)	.144	Joe	0	0	.7	.60	.04	.7	.8	.1	.24	0
(Joe, no, +)	.336	Joe	0	1	.7	.60	.04	.7	.8	.1	.56	0
(Joe, yes, -)	.004	Joe	1	0	.8	.42	.04	.7	.8	.1	0	.01
(Joe, yes, +)	.016	Joe	1	1	.8	.42	.04	.7	.8	.1	0	.04
(Ann, no, -)	.096	Ann	0	0	.2	.60	.76	.2	.4	.2	.16	0
(Ann, no, +)	.024	Ann	0	1	.2	.60	.76	.2	.4	.2	.04	0
(Ann, yes, -)	.228	Ann	1	0	.4	.42	.76	.2	.4	.2	0	.57
(Ann, yes, +)	.152	Ann	1	1	.4	.42	.76	.2	.4	.2	0	.38

treatment effects are positive (.1 for Joe, .2 for Ann), whereas this slope is negative. Hence, this slope would be extremely misleading if used for the evaluation of the treatment. Note that the phenomenon that  $E[g_1(U)]$  and the slope of  $X$  in Equation (14.6) are *not identical* can only occur if  $X$  and  $U$  are not independent (see Theorem 15.14).

Hence, while the function  $g_1$  and its expectation  $E_U(g_1) = E[g_1(U)]$  can still be causally interpreted as a  $U$ -conditional effect function and the average effect of the treatment, respectively, the slope of  $X$  in Equation (14.6) does not have a causal meaning. ◁

**Example 14.3 (No Treatment for Joe – continued)** Table 14.3 displays a third example with Joe and Ann. Note that the values of the versions  $E(Y|X)$  and  $E(Y|X, U)$  in this table are specified for all  $\omega \in \Omega$  (see Examples 9.22, 9.23, and Exercise 14-1). However, in this example, there are infinitely many versions of the conditional expectation  $E(Y|X, U)$ . In Table 9.2 we already specified a version  $V \in \mathcal{E}(Y|X, U)$  with

$$V(\omega) = 0, \quad \text{if } \omega \in \{(Joe, yes, -), (Joe, yes, +)\},$$

and in Example 10.19 we noted that assigning any other real number instead would do as well. For instance, assigning

$$V^*(\omega) = 9, \quad \text{if } \omega \in \{(Joe, yes, -), (Joe, yes, +)\},$$

**Table 14.3.** No Treatment for Joe: Conditional Expectations With Respect to  $P^{X=x}$

Unit Treatment Success	Outcomes $\omega$ $P(\{\omega\})$	Observables			Conditional expectations						
		Person variable $U$	Treatment variable $X$	Outcome variable $Y$	$E(Y X, U)$	$E(Y X)$	$P(X=1 U)$	$E^{X=0}(Y U)$	$E^{X=1}(Y U)$	$P^{X=0}(\{\omega\})$ (rounded)	$P^{X=1}(\{\omega\})$
(Joe, no, -)	.152	Joe	0	0	.696	.6	0	.696	9	.245	0
(Joe, no, +)	.348	Joe	0	1	.696	.6	0	.696	9	.561	0
(Joe, yes, -)	0	Joe	1	0	9	.4	0	.696	9	0	0
(Joe, yes, +)	0	Joe	1	1	9	.4	0	.696	9	0	0
(Ann, no, -)	.096	Ann	0	0	.2	.6	.76	.2	.4	.155	0
(Ann, no, +)	.024	Ann	0	1	.2	.6	.76	.2	.4	.039	0
(Ann, yes, -)	.228	Ann	1	0	.4	.4	.76	.2	.4	0	.6
(Ann, yes, +)	.152	Ann	1	1	.4	.4	.76	.2	.4	0	.4

we define a second version  $V^* \in \mathcal{E}(Y|X, U)$ , provided, of course, that the other values are assigned as in Table 9.2. The version  $V^* = E(Y|X, U)$  is displayed in Table 14.3. In Example 10.19 we also noted that two versions  $V$  and  $V^*$  of  $E(Y|X, U)$  are  $P$ -equivalent.

For the version  $V$  (see the column headed  $E(Y|X, U)$  in Table 9.2) we can write

$$\begin{aligned}
 V &= .2 + .496 \cdot 1_{U=Joe} + .2 \cdot X - .896 \cdot 1_{U=Joe} \cdot X \\
 &= (.2 + .496 \cdot 1_{U=Joe}) + (.2 - .896 \cdot 1_{U=Joe}) \cdot X \\
 &= g_0(U) + g_1(U) \cdot X.
 \end{aligned}
 \tag{14.7}$$

(The coefficients of this equation are obtained analogously as in Example 12.24.) In this equation,  $g_0(U) := .2 + .496 \cdot 1_{U=Joe}$  holds for the  $U$ -conditional *intercept function* and

$$g_1(U) := .2 - .896 \cdot 1_{U=Joe}$$

holds for the  $U$ -conditional *effect function*  $g_1$ . For the version  $V$ , the value of  $g_1$  for  $u = Joe$  is  $g_1(Joe) = .2 - .896 \cdot 1 = -.696$ , and for  $u = Ann$  it is  $g_1(Ann) = .2 - .896 \cdot 0 = .2$ .

Furthermore, the version  $V^*$  satisfies

$$\begin{aligned}
 V^* &= .2 + .496 \cdot 1_{U=Joe} + .2 \cdot X + 8.104 \cdot 1_{U=Joe} \cdot X \\
 &= (.2 + .496 \cdot 1_{U=Joe}) + (.2 + 8.104 \cdot 1_{U=Joe}) \cdot X \\
 &= g_0^*(U) + g_1^*(U) \cdot X.
 \end{aligned}
 \tag{14.8}$$

In contrast to the two different versions of  $E(Y|X, U)$ , different versions of the  $U$ -conditional effect function are not necessarily  $P_U$ -equivalent. For instance, if we consider the version  $V \in \mathcal{E}(Y|X, U)$  specified above, then the associated effect function  $g_1$  has the two values

$$g_1(\text{Joe}) = .2 - .896 \cdot 1 = -.696 \quad \text{and} \quad g_1(\text{Ann}) = .2 - .896 \cdot 0 = .2,$$

whereas considering the version  $V^* \in \mathcal{E}(Y|X, U)$  specified above, the associated effect function  $g_1^*(U)$  has the two values

$$g_1^*(\text{Joe}) = .2 + 8.104 \cdot 1 = 8.304 \quad \text{and} \quad g_1^*(\text{Ann}) = .2 + 8.104 \cdot 0 = .2.$$

Consequently, the expectations of different versions of the effect function also differ from each other. Because  $P(U=\text{Joe}) = .152 + .348 + 0 + 0 = .5$  and  $P(U=\text{Ann}) = .096 + .024 + .228 + .152 = .5$ , the expectation of  $g_1$  is

$$E_U(g_1) = E[g_1(U)] = -.696 \cdot P(U=\text{Joe}) + .2 \cdot P(U=\text{Ann}) = -.696 \cdot .5 + .2 \cdot .5 = -.248,$$

whereas the expectation of  $g_1^*$  is

$$E_U(g_1^*) = E[g_1^*(U)] = 8.304 \cdot P(U=\text{Joe}) + .2 \cdot P(U=\text{Ann}) = 8.304 \cdot .5 + .2 \cdot .5 = 4.252.$$

For both computations we used Equation (6.15). ◁

**Remark 14.4 (Methodological Conclusions)** Examples 14.1 to 14.3 show that the notions effect function, conditional effects, and average effects are crucial for the evaluation of treatments, interventions, and expositions. This does not only apply to  $U$ -conditional effects, but also to effects conditioning on other variables, say  $Z$ , such as  $Z = \text{gender}$ ,  $Z = \text{severity of symptoms}$ ,  $Z = \text{educational status}$ , etc. Our examples with a dichotomous treatment variable  $X$  with values 0 and 1 show: Although a conditional expectation  $E(Y|X, U)$  is uniquely defined up to  $P$ -equivalence, this neither guarantees that the effect function is uniquely defined up to  $P_U$ -equivalence nor that the expectations of  $g_1$  and  $g_1^*$  are identical for different versions  $g_1, g_1^*$  of the effect function. This suggests that we need to learn more about the effect function and the conditional expectations  $E^{X=0}(Y|U)$  and  $E^{X=1}(Y|U)$  [see Eq. (14.3)]. In more general terms, we need to learn more about a conditional expectation  $E^B(Y|Z)$  with respect to a conditional-probability measure  $P^B$ . ◁

## 14.2 Assumptions and Definitions

In section 4.2 we considered a probability space  $(\Omega, \mathcal{A}, P)$  and an event  $B \in \mathcal{A}$  with  $P(B) > 0$ . According to Theorem 4.28, the function  $P^B: \mathcal{A} \rightarrow [0, 1]$  defined by

$$P^B(A) = P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad \forall A \in \mathcal{A}, \quad (14.9)$$

is a probability measure on  $(\Omega, \mathcal{A})$ , called the *B*-conditional-probability measure. We also noted that  $(\Omega, \mathcal{A}, P^B)$  is a probability space sharing the measurable space  $(\Omega, \mathcal{A})$  with the original probability space  $(\Omega, \mathcal{A}, P)$ .

In section 9.1 we used the conditional-probability measure  $P^B$  in order to introduce the *B*-conditional expectation

$$E(Y|B) = E^B(Y) = \int Y dP^B, \quad (14.10)$$

assuming  $P(B) > 0$  [see Eq. (9.2)] and that the expectation of  $Y$  with respect to  $P^B$  exists (see Def. 6.1), i. e., that  $Y$  is quasi-integrable with respect to  $P^B$  (see Def. 3.28). Hence,  $E(Y|B)$  denotes the *B*-conditional expectation of  $Y$  and, by definition, it is identical to the expectation  $E^B(Y)$  of  $Y$  with respect to  $P^B$ . If  $Z$  is a random variable on  $(\Omega, \mathcal{A}, P)$  and  $B = \{Z=z\} = \{\omega \in \Omega : Z(\omega) = z\}$  with  $P(Z=z) > 0$ , then we use the notation  $E^{Z=z}(Y)$  and

$$E(Y|Z=z) = E^{Z=z}(Y) \quad (14.11)$$

instead of  $E^B(Y)$  as well as  $P^{Z=z}$  instead of  $P^B$ . Note, however, that  $E(Y|Z=z)$  is also defined if  $P(Z=z) = 0$ . [For the definition see Eq. (10.27) and for uniqueness see Rem. 10.28.]

In this section, we often refer to the following assumptions and notation:

#### Notation and Assumptions 14.5

Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  be a random variable,  $B \in \mathcal{A}$  with  $P(B) > 0$ ,  $\mathcal{C} \subset \mathcal{A}$  a  $\sigma$ -algebra, and  $P^B$  defined by Equation (14.9). Furthermore, assume that  $Y$  is nonnegative or such that  $E^B(Y)$  is finite.

**Remark 14.6 (Finite Expectation With Respect to  $P^B$ )** If  $Y$  is a random variable with finite expectation  $E(Y)$ , then  $E^B(Y)$  is finite, too (see Exercise 14-2). In contrast, finiteness of  $E^B(Y)$  does not imply that  $E(Y)$  is finite.  $\triangleleft$

In chapter 10 we defined a version of a  $\mathcal{C}$ -conditional expectation  $E(Y|\mathcal{C})$  with respect to the measure  $P$  (see Def. 10.2). For convenience, we repeat this definition of a  $\mathcal{C}$ -conditional expectation, but now it is for a conditional-probability measure  $P^B$  defined by Equation (14.9).

Reading the following definition, note that  $V, X, Y$  are random variables on the probability space  $(\Omega, \mathcal{A}, P)$  if and only if they are random variables on  $(\Omega, \mathcal{A}, P^B)$ , provided, of course, that  $P(B) > 0$  so that  $P^B$  is defined (see Exercise 14-3). Also remember that  $\sigma(V) = V^{-1}(\bar{\mathcal{B}})$  denotes the  $\sigma$ -algebra generated by the random variable  $V: (\Omega, \mathcal{A}, P) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  (see section 2.3.2).

#### Definition 14.7 ( $\mathcal{C}$ -Conditional Expectation With Respect to $P^B$ )

Let the assumptions 14.5 hold. A random variable  $V: (\Omega, \mathcal{A}, P) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  is called a version of the  $\mathcal{C}$ -conditional expectation of  $Y$  with respect to  $P^B$ , if the following two conditions hold:

(a)  $\sigma(V) \subset \mathcal{C}$ .  
 (b)  $E^B(1_C \cdot V) = E^B(1_C \cdot Y), \quad \forall C \in \mathcal{C}$ .  
 If  $V$  satisfies (a) and (b), then we also use the notation  $E^B(Y|\mathcal{C}) := V$ .

According to Equation (6.1), condition (b) of this definition is equivalent to

$$\int 1_C \cdot V dP^B = \int 1_C \cdot Y dP^B, \quad \forall C \in \mathcal{C}. \tag{14.12}$$

This equation shows more clearly how the measure  $P^B$  is involved in the definition of  $E^B(Y|\mathcal{C})$ .

**Remark 14.8 (Existence)** The only difference between Definitions 10.2 and 14.7 is that, instead of referring to  $P$ , now we refer to the conditional-probability measure  $P^B$  defined by Equation (14.9). Hence, Theorem 10.9 (i) implies that, under the assumptions of Definition 14.7, there exists at least one version  $E^B(Y|\mathcal{C})$ . Furthermore, Theorem 10.9 (ii) implies that two versions  $V$  and  $V^*$  of the  $\mathcal{C}$ -conditional expectation of  $Y$  with respect to  $P^B$  are  $P^B$ -equivalent. In other words,  $E^B(Y|\mathcal{C})$  is  $P^B$ -unique (see Rem. 10.13).  $\triangleleft$

**Remark 14.9 (Uniqueness)** Theorem 10.9 (ii) implies that two versions  $V$  and  $V^*$  of the  $\mathcal{C}$ -conditional expectation of  $Y$  with respect to  $P^B$  are  $P^B$ -equivalent. In other words,  $E^B(Y|\mathcal{C})$  is  $P^B$ -unique (see Rem. 10.13).  $\triangleleft$

**Remark 14.10 (Set of all Versions of the  $\mathcal{C}$ -Conditional  $P^B$ -Expectation)** The notation  $E^B(Y|\mathcal{C})$  refers to a *version* of the  $\mathcal{C}$ -conditional expectation of  $Y$  with respect to  $P^B$ . In contrast, *the set* of all random variables  $V$  on  $(\Omega, \mathcal{A}, P)$  that satisfy conditions (a) and (b) of Definition 14.7 is denoted by  $\mathcal{E}^B(Y|\mathcal{C})$ . Hence, we can write  $E^B(Y|\mathcal{C}) \in \mathcal{E}^B(Y|\mathcal{C})$  (see also Exercise 14-4).  $\triangleleft$

**Remark 14.11 ( $E^B(Y|\mathcal{C})$  is not Necessarily  $P$ -Unique)** Although  $E^B(Y|\mathcal{C})$  is  $P^B$ -unique, it is *not necessarily  $P$ -unique* (see Rem. 10.13). In section 14.6 we present necessary and sufficient conditions for  $P$ -uniqueness of  $E^B(Y|\mathcal{C})$ , a property that has important implications (see, e. g., section 14.6.4 and Box 14.1).  $\triangleleft$

**Remark 14.12 (Properties of  $E^B(Y|\mathcal{C})$ )** Because  $E^B(Y|\mathcal{C})$  is a conditional expectation, the properties that have been treated in detail in chapters 10 and 11 analogously also apply to  $E^B(Y|\mathcal{C})$ . We simply have to exchange the probability measure  $P$  by  $P^B$ , the expectation  $E(\cdot)$  by  $E^B(\cdot)$ , the variance  $Var(\cdot)$  by  $Var^B(\cdot)$ , and the covariance  $Cov(\cdot, \cdot)$  by  $Cov^B(\cdot, \cdot)$ .  $\triangleleft$

**Remark 14.13 (Existence of a Real-Valued Version  $E^B(Y|\mathcal{C})$ )** Box 10.1 (x) immediately yields: If the assumptions 14.5 hold and  $E^B(Y)$  is finite, then there is a *real-valued* version  $V \in \mathcal{E}^B(Y|\mathcal{C})$ .  $\triangleleft$

**Remark 14.14 ( $\mathcal{C}$ -Conditional Probability With Respect to  $P^B$ )** Let the assumptions 14.5 hold and let  $A \in \mathcal{A}$ . Then we call

$$P^B(A|\mathcal{C}) := E^B(1_A|\mathcal{C}) \quad (14.13)$$

a version of the  $\mathcal{C}$ -conditional probability of  $A$  with respect to  $P^B$ . Correspondingly,  $\mathcal{P}^B(A|\mathcal{C})$  denotes the set of all these versions  $P^B(A|\mathcal{C})$ .  $\triangleleft$

Now we adapt notation and terminology of a  $\mathcal{C}$ -conditional expectation with respect to a conditional-probability measure to the case in which the  $\sigma$ -algebra  $\mathcal{C}$  is generated by a random variable.

**Remark 14.15 ( $X$ -Conditional Expectation With Respect to  $P^B$ )** Let the assumptions 14.5 hold and assume that  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  is a random variable. Then we define:

$$E^B(Y|X) := E^B(Y | \sigma(X)), \quad (14.14)$$

and call it a version of the  $X$ -conditional expectation of  $Y$  with respect to  $P^B$  or a version of the  $X$ -conditional  $P^B$ -expectation of  $Y$ .  $\triangleleft$

Correspondingly, we use  $\mathcal{E}^B(Y|X)$  to denote the set of all versions of the  $X$ -conditional expectation of  $Y$  with respect to  $P^B$ .

**Remark 14.16 ( $X$ -Conditional Probability of an Event With Respect to  $P^B$ )** If  $A \in \mathcal{A}$ , then

$$P^B(A|X) := E^B(1_A|X) \quad (14.15)$$

is called a version of the  $X$ -conditional expectation of  $A$  with respect to  $P^B$ , and  $\mathcal{P}^B(A|X)$  denotes the set of all these versions.  $\triangleleft$

If  $B$  is the event  $\{Z=z\} = \{\omega \in \Omega: Z(\omega) = z\}$  that a random variable  $Z$  takes on the value  $z$ , then we adapt the notation and the terminology correspondingly.

#### Notation and Assumptions 14.17

Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  and  $Z: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Z, \mathcal{A}'_Z)$  be random variables and  $\mathcal{C} \subset \mathcal{A}$  a  $\sigma$ -algebra. Furthermore, let  $z \in \Omega'_Z$  with  $P(Z=z) > 0$  and assume that  $Y$  is nonnegative or such that  $E^{Z=z}(Y)$  is finite.

**Remark 14.18 (Conditional Expectation With Respect to  $P^{Z=z}$ )** Let the assumptions 14.17 hold. Then we use the notation

$$E^{Z=z}(Y|\mathcal{C}) := E^{\{Z=z\}}(Y|\mathcal{C}) \quad (14.16)$$

and call it a version of the  $\mathcal{C}$ -conditional expectation of  $Y$  with respect to  $P^{Z=z}$ . The measure  $P^{Z=z}$  is defined by Equation (14.9) with  $B = \{Z=z\}$ . Correspondingly,  $\mathcal{E}^{Z=z}(Y|\mathcal{C})$  denotes the set of all versions of the  $\mathcal{C}$ -conditional expectation of  $Y$  with respect to  $P^{Z=z}$ .  $\triangleleft$

**Remark 14.19 ( $\mathcal{C}$ -Conditional Probability With Respect to  $P^{Z=z}$ )** Correspondingly, for  $A \in \mathcal{A}$  we define  $P^{Z=z}(A|\mathcal{C}) := E^{Z=z}(1_A|\mathcal{C})$ , a version of the  $\mathcal{C}$ -conditional probability of the event  $A$  with respect to the measure  $P^{Z=z}$ , and we use  $\mathcal{P}^{Z=z}(A|\mathcal{C})$  to denote the family of all versions of the  $\mathcal{C}$ -conditional probability of the event  $A$  with respect to the measure  $P^{Z=z}$ .  $\triangleleft$

In the next definition we additionally consider a random variable  $X$  and use it such that  $\sigma(X)$  takes the role of the  $\sigma$ -algebra  $\mathcal{C}$ .

**Notation and Assumptions 14.20**

Let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$ ,  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \overline{\mathcal{B}})$ , and  $Z: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Z, \mathcal{A}'_Z)$  be random variables. Furthermore, let  $z \in \Omega'_Z$  with  $P(Z=z) > 0$ , and assume that  $Y$  is nonnegative or such that  $E^{Z=z}(Y)$  is finite.

Under these assumptions we define an  $X$ -conditional expectation with respect to a  $(Z=z)$ -conditional probability measure as follows:

**Remark 14.21 ( $X$ -Conditional Expectation With Respect to  $P^{Z=z}$ )** Let the assumptions 14.20 hold. Then

$$E^{Z=z}(Y|X) := E^{(Z=z)}(Y|\sigma(X)), \tag{14.17}$$

is called a *version of the  $X$ -conditional expectation of  $Y$  with respect to  $P^{Z=z}$* .  $\triangleleft$

Again note the distinction between a version  $E^{Z=z}(Y|X)$  and  $\mathcal{E}^{Z=z}(Y|X)$ , the family of all versions of the  $X$ -conditional expectation of  $Y$  with respect to  $P^{Z=z}$ . Of course, what has been said in Remark 14.12 about the properties of a  $\mathcal{C}$ -conditional expectation with respect to  $P^B$  applies to  $E^{Z=z}(Y|X)$  as well.

**Remark 14.22 ( $X$ -Conditional Probability With Respect to  $P^{Z=z}$ )** Correspondingly, for  $A \in \mathcal{A}$  we define  $P^{Z=z}(A|X) := E^{Z=z}(1_A|X)$ , a version of the  $X$ -conditional probability of the event  $A$  with respect to the measure  $P^{Z=z}$ , and we use  $\mathcal{P}^{Z=z}(A|X)$  to denote the family of all these versions.  $\triangleleft$

**Example 14.23 (Joe and Ann With Self-Selection—continued)** We continue Example 14.2 illustrating how to compute  $E^{X=0}(Y|U)$  and  $E^{X=1}(Y|U)$ , which, in this example, are both uniquely defined. First we compute the probabilities of the elementary events with respect to the measures  $P^{X=0}$  and  $P^{X=1}$ , and then specify  $E^{X=0}(Y|U)$  and  $E^{X=1}(Y|U)$ .

Enumerating the eight elementary events  $\{\omega_1\}$  to  $\{\omega_8\}$  from top to bottom of the first column of Table 14.2, the probabilities of these elementary events with respect to  $P^{X=0}$  can be computed as follows:

$$\begin{aligned} P^{X=0}(\{\omega_1\}) &= P^{X=0}[\{(Joe, no, -)\}] = \frac{P[\{(Joe, no, -)\} \cap \{X=0\}]}{P(X=0)} \\ &= \frac{P[\{(Joe, no, -)\}]}{P(X=0)} = \frac{.144}{.144 + .336 + .096 + .024} = .24. \end{aligned}$$

For the elementary event  $\{\omega_2\}$  we obtain

$$\begin{aligned} P^{X=0}(\{\omega_2\}) &= P^{X=0}[\{(Joe, no, +)\}] = \frac{P[\{(Joe, no, +)\} \cap \{X=0\}]}{P(X=0)} \\ &= \frac{P[\{(Joe, no, +)\}]}{P(X=0)} = \frac{.336}{.144 + .336 + .096 + .024} = .56. \end{aligned}$$

For  $\{\omega_5\}$  we obtain

$$\begin{aligned} P^{X=0}(\{\omega_5\}) &= P^{X=0}[\{(Ann, no, -)\}] = \frac{P[\{(Ann, no, -)\} \cap \{X=0\}]}{P(X=0)} \\ &= \frac{P[\{(Ann, no, -)\}]}{P(X=0)} = \frac{.096}{.144 + .336 + .096 + .024} = .16, \end{aligned}$$

and for  $\{\omega_6\}$ ,

$$\begin{aligned} P^{X=0}(\{\omega_6\}) &= P^{X=0}[\{(Ann, no, +)\}] = \frac{P[\{(Ann, no, +)\} \cap \{X=0\}]}{P(X=0)} \\ &= \frac{P[\{(Ann, no, +)\}]}{P(X=0)} = \frac{.024}{.144 + .336 + .096 + .024} = .04. \end{aligned}$$

The probabilities of the other four elementary events with respect to  $P^{X=0}$  are 0 (see also the last but one column of Table 14.2). The probabilities of the eight elementary events with respect to  $P^{X=1}$  are computed analogously (see the last column of Table 14.2).

Now we specify the  $U$ -conditional expectation of  $Y$  with respect to  $P^{X=0}$ . Because  $Y$  is an indicator variable with values 0 and 1, the conditional expectation  $E^{X=0}(Y|U)$  can also be denoted by  $P^{X=0}(Y=1|U)$  [see Eq. (14.15)]. It has two different values, one for  $U(\omega) = Joe$  and one for  $U(\omega) = Ann$ . These values can be computed as follows:

$$P^{X=0}(Y=1|U=Joe) = \frac{P^{X=0}(Y=1, U=Joe)}{P^{X=0}(U=Joe)} = \frac{.56}{.24 + .56} = .7.$$

and

$$P^{X=0}(Y=1|U=Ann) = \frac{P^{X=0}(Y=1, U=Ann)}{P^{X=0}(U=Ann)} = \frac{.04}{.04 + .16} = .2.$$

The results of the corresponding computations for  $P^{X=1}(Y=1|U)$  are displayed in the last but two column of Table 14.2).

Note that, in this example, there is only one single element of  $\mathcal{E}^{X=0}(Y|U)$  and one single element of  $\mathcal{E}^{X=1}(Y|U)$ , i. e., the conditional expectations  $E^{X=0}(Y|U)$  and  $E^{X=1}(Y|U)$  are uniquely defined. In contrast, in Example 14.24 there is one single element of  $\mathcal{E}^{X=0}(Y|U)$ , but an infinite number of different elements of  $\mathcal{E}^{X=1}(Y|U)$ .  $\triangleleft$

**Example 14.24 (No Treatment for Joe – continued)** We continue Example 14.3. In this example,  $E^{X=0}(Y|U)$  is uniquely defined (see Exercise 14-7). In contrast, this is not true for  $E^{X=1}(Y|U)$ . The probabilities of the elementary events with

respect to the measures  $P^{X=x}$  are computed analogously to Example 14.23. The results are displayed in the last two columns of Table 14.3.

Now we specify a version of the  $U$ -conditional expectation of  $Y$  with respect to the measure  $P^{X=1}$ . Because  $Y$  is an indicator variable with values 0 and 1, a version of this conditional expectation can also be denoted by  $P^{X=1}(Y=1|U)$  [see Eq. (14.15)]. It has two different values, one for  $U(\omega) = Joe$  and one for  $U(\omega) = Ann$ . The latter is

$$P^{X=1}(Y=1|U=Ann) = \frac{P^{X=1}(Y=1, U=Ann)}{P^{X=1}(U=Ann)} = \frac{.4}{.6 + .4} = .4.$$

In contrast, the fraction  $P^{X=1}(Y=1, U=Joe)/P^{X=1}(U=Joe)$  is not defined, because  $P^{X=1}(U=Joe) = 0$ . Nevertheless, a value  $P^{X=1}(Y=1|U=Joe)$  of a factorization of  $P^{X=1}(Y=1|U)$  is defined [see Def. 10.33 and also Eq. (14.34)]. In this case, we can choose any real number as the value of  $P^{X=1}(Y=1|U)$  for  $\omega \in \{U=Joe\} = \{\omega_1, \dots, \omega_4\}$ . For example,

$$V_1(\omega) = \begin{cases} 9, & \text{if } U(\omega) = Joe, \\ .4, & \text{if } U(\omega) = Ann, \end{cases}$$

[see Eq. (10.31) and also Eq. (14.41)] defines a first element of  $\mathcal{P}^{X=1}(Y=1|U)$  (see Table 14.3), and

$$V_1^*(\omega) = \begin{cases} 0, & \text{if } U(\omega) = Joe, \\ .4, & \text{if } U(\omega) = Ann, \end{cases}$$

is a second element of the set  $\mathcal{P}^{X=1}(Y=1|U)$ . Obviously,  $V_1$  and  $V_1^*$  are  $P^{X=1}$ -equivalent, because  $P^{X=1}(A_1) = 0$ , where

$$\begin{aligned} A_1 &:= \{\omega \in \Omega: V_1(\omega) \neq V_1^*(\omega)\} = \{\omega_1, \dots, \omega_4\} \\ &= \{(Joe, no, -), (Joe, no, +), (Joe, yes, -), (Joe, yes, +)\}. \end{aligned}$$

This probability can be computed by

$$P^{X=1}(A_1) = \sum_{\omega \in A_1} P^{X=1}(\{\omega\}) = 0 + 0 + 0 + 0 = 0$$

(see the last column of Table 14.3). In contrast,

$$P(A_1) = \sum_{\omega \in A_1} P(\{\omega\}) = .152 + .348 + 0 + 0 = .5.$$

Hence, the two versions  $V_1$  and  $V_1^*$  of the  $U$ -conditional  $P^{X=1}$ -expectation of  $Y$  are  $P^{X=1}$ -equivalent, i. e., they are equivalent with respect to the measure  $P^{X=1}$ . However, the two versions are *not equivalent* with respect to the measure  $P$ . (This issue will be treated in more detail in section 14.6.)

Note that the values of a version  $E^{X=x}(Y|U)$  of the conditional expectation are defined for *all* elements  $\omega \in \Omega$  and that these conditional expectations are random variables on all three probability spaces  $(\Omega, \mathcal{A}, P)$ ,  $(\Omega, \mathcal{A}, P^{X=0})$ , and  $(\Omega, \mathcal{A}, P^{X=1})$ , because they share the same measurable space  $(\Omega, \mathcal{A})$  (see Def. 5.1). Furthermore, the values of such a version  $E^{X=x}(Y|U)$  only depend on the drawn person. This illustrates that they are measurable with respect to  $U$  [see Def. 14.7 (b)].  $\triangleleft$

### 14.3 Properties

**Remark 14.25 (Rules of Computation)** The rules of computation for conditional expectations  $E(Y|\mathcal{C})$  with respect to a probability measure  $P$  analogously hold for conditional expectation  $E^B(Y|\mathcal{C})$  with respect to a conditional-probability measure  $P^B$ . For example, according to Rule (iv) of Box 10.1,

$$E^B[E^B(Y|\mathcal{C})] = E^B(Y). \quad (14.18)$$

Similarly, according to Rule (iii) of Box 10.1,

$$E^B(\alpha \cdot Y|\mathcal{C}) \stackrel{P^B}{=} \alpha \cdot E^B(Y|\mathcal{C}), \quad \alpha \in \mathbb{R}. \quad (14.19)$$

We simply have to exchange the notation  $E(\cdot)$ , which refers to the measure  $P$ , by  $E^B(\cdot)$  referring to the conditional-probability measure  $P^B$  and, of course, exchange  $P$  by  $P^B$ .  $\triangleleft$

In the following theorem we extend Equation (14.19) showing how to deal with a  $\mathcal{C}$ -conditional expectation with respect to  $P^{Z=z}$  of  $f(Z) \cdot Y$ .

**Theorem 14.26 (Regressand  $f(Z) \cdot Y$ )**

Let the assumptions 14.17 hold. If  $f: (\Omega'_Z, \mathcal{A}'_Z) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  is a measurable function and  $f(z) \in \mathbb{R}$ , then

$$E^{Z=z}[f(Z) \cdot Y|\mathcal{C}] \stackrel{P^{Z=z}}{=} f(z) \cdot E^{Z=z}(Y|\mathcal{C}). \quad (14.20)$$

(Proof p. 426)

**Remark 14.27 (Two Special Cases)** For the constant  $Y = 1$ , Rule (i) of Box 10.1 yields  $E^{Z=z}(Y|\mathcal{C}) \stackrel{P^{Z=z}}{=} 1$ . Therefore, Equation (14.20) implies

$$E^{Z=z}[f(Z)|\mathcal{C}] \stackrel{P^{Z=z}}{=} f(z) \quad (14.21)$$

Another special case of Equation (14.20) is

$$E^{Z=z}[f(Z)] = f(z), \quad (14.22)$$

which follows from Remark 10.5 and (14.21) for  $\mathcal{C} = \{\Omega, \emptyset\}$ .  $\triangleleft$

**Remark 14.28 (Two Probability Spaces)** There are also some properties of a conditional expectation  $E^B(Y|\mathcal{C})$  that are related to the fact that two probability spaces,  $(\Omega, \mathcal{A}, P)$  and  $(\Omega, \mathcal{A}, P^B)$ , are involved. By definition, a version of the conditional expectation  $E^B(Y|\mathcal{C})$  with respect to  $P^B$  is a random variable on the probability space  $(\Omega, \mathcal{A}, P^B)$ . Therefore, it is also a random variable on  $(\Omega, \mathcal{A}, P)$ . However, different elements of  $\mathcal{E}^B(Y|\mathcal{C})$  are not necessarily  $P$ -equivalent; they are necessarily equivalent only with respect to  $P^B$ . Hence, if  $V, V^* \in \mathcal{E}^B(Y|\mathcal{C})$ , then the expectations  $E(V)$  and  $E(V^*)$  with respect to  $P$  may differ from each other, whereas  $E^B(V)$  and  $E^B(V^*)$  are necessarily identical. These issues are treated in detail in section 14.6.2.  $\triangleleft$

### 14.4 Partial Conditional Expectation

Now we introduce the concept of a *partial conditional expectation* using a factorization of a version  $g(X, Z) = E(Y|X, Z) \in \mathcal{E}(Y|X, Z)$ . We show how this concept is related to a conditional expectation with respect to a conditional probability measure. In Definition 14.29 we refer to the functions  $g_z: \Omega'_X \rightarrow \bar{\mathbb{R}}$  that, for all  $z \in \Omega'_Z$ , are defined by

$$g_z(x) = g(x, z), \quad \forall x \in \Omega'_X. \quad (14.23)$$

Referring to the concept of an  $(X=x, Z=z)$ -conditional expectation value introduced in Definition 10.33, we can write

$$g_z(x) = g(x, z) = E(Y|X=x, Z=z), \quad \forall (x, z) \in \Omega'_X \times \Omega'_Z. \quad (14.24)$$

Note that in Equations (14.23) to (14.25) we do not assume  $P(Z=z) > 0$ .

**Definition 14.29 (Partial Conditional Expectation)**

Let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$ ,  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\bar{\mathbb{R}}, \mathcal{B})$ , and  $Z: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Z, \mathcal{A}'_Z)$  be random variables and assume that  $Y$  is nonnegative or with finite expectation  $E(Y)$ . Furthermore, let  $g(X, Z) = E(Y|X, Z) \in \mathcal{E}(Y|X, Z)$  and, for  $z \in \Omega'_Z$ , let the function  $g_z$ , be defined by Equation (14.23). Then the function  $E(Y|X, Z=z): \Omega \rightarrow \bar{\mathbb{R}}$  defined by

$$E(Y|X, Z=z) := g_z(X) \quad (14.25)$$

is called a *version of the partial  $(X, Z=z)$ -conditional expectation of  $Y$  (with respect to  $P$ )*.

To emphasize, for each  $z \in \Omega'_Z$ , the function  $E(Y|X, Z=z)$  denotes the composition of  $X$  and  $g_z$ . Hence, for each  $z \in \Omega'_Z$  it is a random variable on  $(\Omega, \mathcal{A}, P)$  that is  $X$ -measurable (see Lemma 2.52). In Theorem 14.33 we show that  $E(Y|X, Z=z)$  is a version of the conditional expectation of  $Y$  on  $X$  with respect to  $P^{Z=z}$ , provided that  $P(Z=z) > 0$ .

**Remark 14.30 (Partial Conditional Probability)** If  $A \in \mathcal{A}$ , then we also use the notation  $P(A|X, Z=z) := E(1_A|X, Z=z)$  and call it the partial  $(X, Z=z)$ -conditional probability of the event  $A$  (with respect to  $P$ ). Furthermore, if  $Y$  is dichotomous with values 0 and 1, then we also use the notation  $P(Y=1|X, Z=z) := E(Y|X, Z=z)$  and call it the partial  $(X, Z=z)$ -conditional probability of the event  $\{Y=1\}$ , or simply of  $Y=1$  (with respect to  $P$ ).  $\triangleleft$

**Remark 14.31 (Factorization and Partial Conditional Expectation)** If the assumptions of Definition 14.29 hold, then

$$1_{Z=z} \cdot g(X, Z) = 1_{Z=z} \cdot g_z(X). \quad (14.26)$$

(For a proof see Exercise 14-5).  $\triangleleft$

**Remark 14.32 (Discrete  $Z$ )** Under the assumptions of Definition 14.29, suppose that  $Z: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Z, \mathcal{A}'_Z)$  is discrete and  $Z(\Omega) \subset \Omega'_Z$  is finite or countable with  $\{z\} \in \mathcal{A}'_Z$  for all  $z \in Z(\Omega)$ . Then

$$E(Y|X, Z) = \sum_{z \in Z(\Omega)} E(Y|X, Z=z) \cdot 1_{Z=z} \quad (14.27)$$

holds for the specific version  $E(Y|X, Z)$  that is used in Definition 14.29 (see Exercise 14-6). Furthermore,

$$V \stackrel{P}{=} \sum_{z \in Z(\Omega)} E(Y|X, Z=z) \cdot 1_{Z=z}, \quad \forall V \in \mathcal{E}(Y|X, Z). \quad (14.28)$$

◁

According to the following theorem, the partial conditional expectation  $E(Y|X, Z=z)$  is also a version of the  $X$ -conditional expectation of  $Y$  with respect to  $P^{Z=z}$ , provided that  $z \in \Omega'_Z$  with  $P(Z=z) > 0$ .

**Theorem 14.33 (Relationship Between  $E(Y|X, Z=z)$  and  $E^{Z=z}(Y|X)$ )**

Let the assumptions of Definition 14.29 hold and suppose that  $z \in \Omega'_Z$  with  $P(Z=z) > 0$ . Then

$$E(Y|X, Z=z) \in \mathcal{E}^{Z=z}(Y|X), \quad (14.29)$$

and therefore

$$E(Y|X, Z=z) \stackrel{P^{Z=z}}{=} E^{Z=z}(Y|X), \quad \forall E^{Z=z}(Y|X) \in \mathcal{E}^{Z=z}(Y|X). \quad (14.30)$$

(Proof p. 426)

**Remark 14.34 (An Immediate Implication)** If the assumptions of Theorem 14.33 hold and if  $Z$  is discrete with  $P(Z \in \Omega'_0) = 1$  and  $P(Z=z) > 0$  for all  $z \in \Omega'_0$  (see Def. 5.56), then

$$V \stackrel{P}{=} \sum_{z \in \Omega'_0} E^{Z=z}(Y|X) \cdot 1_{Z=z}, \quad \forall V \in \mathcal{E}(Y|X, Z). \quad (14.31)$$

◁

## 14.5 Factorization

Let the assumptions 14.5 hold and let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  be a random variable. Because  $E^B(Y|X)$  is measurable with respect to  $X$ , Lemma 2.52 implies that there is a measurable function  $g_B: (\Omega'_X, \mathcal{A}'_X) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  such that

$$E^B(Y|X) = g_B(X) \quad (14.32)$$

is the composition of  $X$  and  $g_B$ . A function  $g_B$  satisfying Equation (14.32) is called a *factorization* of the version  $E^B(Y|X) \in \mathcal{E}^B(Y|X)$  (see sections 10.4.1 and 10.4.4).

**14.5.1 Conditional Expectation Value With Respect to  $P^B$**

The values of a factorization  $g_B$  of  $E^B(Y|X)$  are called  $(X=x)$ -conditional expectation values with respect to  $P^B$  and they are denoted by

$$E^B(Y|X=x) := g_B(x), \quad \forall x \in \Omega'_X, \tag{14.33}$$

(see Def. 10.33). Instead of  $E^B(1_A|X=x)$  we also use the notation  $P^B(A|X=x)$ , provided that  $A \in \mathcal{A}$ .

Correspondingly, under the assumptions 14.5,

$$E^{Z=z}(Y|X=x) := g_{Z=z}(x), \quad \forall x \in \Omega'_X \tag{14.34}$$

and  $P^{Z=z}(A|X=x) := E^{Z=z}(1_A|X=x)$  for  $A \in \mathcal{A}$ . Note that  $g_{Z=z}$  is not necessarily identical to the function  $g_z$  defined by Equation (14.23).

**Remark 14.35 (Relationship Between the Functions  $g_z$  and  $g_{Z=z}$ )** Let  $g(X, Z) \in \mathcal{E}(Y|X, Z)$  and  $g_{Z=z}(X) \in \mathcal{E}^{Z=z}(Y|X)$ . Then the relationship between the function  $g_z$  defined by Equation (14.23) and the function  $g_{Z=z}$  defined by Equation (14.34) is as follows:

- (i)  $g_z \stackrel{P^{Z=z}}{=} g_{Z=z}$  if  $P(Z=z) > 0$
- (ii)  $g_z = g_{Z=z}$  if  $z \in \Omega'_Z$  with  $P(Z=z, X=x) > 0$  for all  $x \in \Omega'_X$

(see Exercise 14-8). Note that  $g_{Z=z}$  is only defined if  $P(Z=z) > 0$ . ◁

**Remark 14.36 (Relationship Between Factorizations)** Suppose that the assumptions 14.20 hold, where  $Y$  is nonnegative or with finite expectation  $E(Y)$ . Then Theorem 14.33 implies

$$\begin{aligned} E^{Z=z}(Y|X=x) &= E(Y|X=x, Z=z) \\ &= E(Y|X=x, 1_{Z=z}=1), \quad \text{for } P_X^{Z=z}\text{-a.a. } x \in \Omega'_X, \end{aligned} \tag{14.35}$$

where  $E^{Z=z}(Y|X=x)$  and  $E(Y|X=x, Z=z)$  are the conditional expectation values defined by (14.34) and (10.27), respectively (see Exercise 14-9). Note that  $B = \{1_B = 1\}$ . Therefore, according to (14.33), for  $B \in \mathcal{A}$  with  $P(B) > 0$ ,  $Z = 1_B$ , and  $z = 1$ , the first of these two equations yields

$$E^B(Y|X=x) = E(Y|X=x, 1_B=1), \quad \text{for } P_X^B\text{-a.a. } x \in \Omega'_X. \tag{14.36}$$

◁

**Remark 14.37 (A Sufficient Condition for Uniqueness)** Applying Remark 2.71 to Equation (14.35) yields

$$E^{Z=z}(Y|X=x) = E(Y|X=x, Z=z) = E^{X=x, Z=z}(Y) \tag{14.37}$$

for all  $x \in \Omega'_X$  for which  $P^{Z=z}(X=x) > 0$ , or equivalently, for which  $P(X=x, Z=z) > 0$ , where  $E^{X=x, Z=z}(Y)$  denotes the expectation of  $Y$  with respect to the conditional-probability measure  $P^B$  with  $B = \{X=x\} \cap \{Z=z\}$ . Hence,  $E^{Z=z}(Y|X=x)$  is uniquely defined if  $P(X=x, Z=z) > 0$ . ◁

### 14.5.2 Uniqueness of Factorizations

**Remark 14.38 (Uniqueness of Factorizations)** For a fixed version  $E^B(Y|X)$ , the factorization  $g_B$  of  $E^B(Y|X)$  is uniquely defined, provided that  $\Omega'_X$  is identical to the image  $X(\Omega) = \{\omega \in \Omega: X(\omega) \in \Omega'_X\}$ . If  $\Omega'_X \neq X(\Omega)$ , then there can be different factorizations of a single version  $E^B(Y|X)$  (see Rem. 10.28).

If  $g_B$  and  $g_B^*$  are factorizations of two versions  $V$  and  $V^*$  of  $\mathcal{E}^B(Y|X)$ , respectively, then, according to Corollary 10.29,

$$g_B \stackrel{P_X^B}{=} g_B^*, \quad (14.38)$$

i. e.,  $g_B$  and  $g_B^*$  are  $P_X^B$ -equivalent, where  $P_X^B$  denotes the probability measure on  $(\Omega'_X, \mathcal{A}'_X)$  defined by

$$P_X^B(A') = P^B(X \in A'), \quad \forall A' \in \mathcal{A}'_X. \quad (14.39)$$

Hence, because  $P^B(X \in A')$  is just another notation for  $P^B[X^{-1}(A')]$ ,  $P_X^B$  is the image measure of  $P^B$  under  $X$  (see Def. 5.3 and Rem. 5.5).  $\triangleleft$

**Remark 14.39 (Values of the  $X$ -Conditional Expectation With Respect to  $P^B$ )** If  $P(B) > 0$  and  $E^B(Y|X) \in \mathcal{E}^B(Y|X)$ , then, for all  $x \in \Omega'_X$ ,

$$E^B(Y|X)(\omega) = E^B(Y|X=x), \quad \forall \omega \in \{X=x\} \quad (14.40)$$

(see Rem. 10.37). This also implies that the value of  $E^B(Y|X)$  is constant on all sets  $\{X=x\}$ ,  $x \in \Omega'_X$ . Correspondingly, if  $P(Z=z) > 0$ , then, for all  $x \in \Omega'_X$ ,

$$E^{Z=z}(Y|X)(\omega) = E^{Z=z}(Y|X=x), \quad \forall \omega \in \{X=x\}. \quad (14.41)$$

In other words, whenever the random variable  $X$  takes on the value  $x$ , and this is the case if  $\omega \in \{X=x\}$ , then the random variable  $E^{Z=z}(Y|X)$  takes on the value  $E^{Z=z}(Y|X=x)$ . Note that Equation (14.41) also holds if  $\Omega$  is finite or countable and some  $\omega \in \{X=x\}$  have probability  $P^{Z=z}(\{\omega\}) = 0$ . [As an example, consider the values of  $E^{X=1}(Y|U)$  for  $\omega \in \{(Joe, yes, -), (Joe, yes, +)\}$  in Example 14.24.]  $\triangleleft$

## 14.6 Uniqueness

### 14.6.1 A Necessary and Sufficient Condition of Uniqueness

In Theorem 10.17 we presented a necessary and sufficient condition for uniqueness of a conditional expectation. In the following corollary we translate this result to a conditional expectation with respect to  $P^B$ .

**Corollary 14.40 (Uniqueness of  $E^B(Y|\mathcal{C})$ )**

Let the assumptions 14.5 hold, let  $\mathcal{E}$  be a finite or countable partition of  $\Omega$ , and assume  $\mathcal{C} = \sigma(\mathcal{E})$ . Then  $V = V^*$  for all  $V, V^* \in \mathcal{E}^B(Y|\mathcal{C})$  if and only if

$$P^B(A) > 0, \quad \forall A \in \mathcal{E}. \quad (14.42)$$

**Remark 14.41 (Values of  $E^B(Y|\mathcal{C})$ )** In other words, under the assumptions of Corollary 14.40, the conditional expectation  $E^B(Y|\mathcal{C})$  is uniquely defined if and only if (14.42) holds. Furthermore, if (14.42) holds, then

$$E^B(Y|\mathcal{C}) = \sum_{A \in \mathcal{E}} E^B(Y|A) \cdot 1_A \quad (14.43)$$

[see Eq. (10.14)] and

$$\forall A \in \mathcal{E}: \quad E^B(Y|\mathcal{C})(\omega) = E^B(Y|A), \quad \text{if } \omega \in A \quad (14.44)$$

[see Eq. (10.15)]. The last equation shows that  $E^B(Y|\mathcal{C})$  describes how the conditional expectation values  $E^B(Y|A)$  depend on the events  $A \in \mathcal{E}$ .  $\triangleleft$

The corresponding result for the  $X$ -conditional expectation  $E^B(Y|X)$  of  $Y$  with respect to  $P^B$  is as follows:

**Corollary 14.42 (Uniqueness of  $E^B(Y|X)$ )**

Let the assumptions 14.5 hold. Furthermore, let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  be a random variable such that  $\Omega'_X$  is finite or countable,  $X(\Omega) = \Omega'_X$ , and  $\mathcal{A}'_X = \mathcal{P}(\Omega'_X)$ . Then  $V = V^*$  for all  $V, V^* \in \mathcal{E}^B(Y|X)$  if and only if

$$P^B(X=x) > 0, \quad \forall x \in X(\Omega). \quad (14.45)$$

**Remark 14.43 (Values of  $E^B(Y|X)$ )** Hence, under the assumptions of Corollary 14.42,  $E^B(Y|X)$  is uniquely defined if and only if (14.45) holds. And, if (14.45) holds, then

$$E^B(Y|X) = \sum_{x \in X(\Omega)} E^B(Y|X=x) \cdot 1_{X=x} \quad (14.46)$$

and

$$\forall x \in X(\Omega): \quad E^B(Y|X)(\omega) = E^B(Y|X=x), \quad \text{if } \omega \in \{X=x\}. \quad (14.47)$$

This equation shows that the conditional expectation  $E^B(Y|X)$  describes how the conditional expectation values  $E^B(Y|X=x)$  depend on the values  $x$  of  $X$ .  $\triangleleft$

**Example 14.44 (Joe and Ann)** Applying Corollary 14.42 to the introductory Examples 14.1 and 14.2 implies that the conditional expectations  $E^{X=0}(Y|U)$  and  $E^{X=1}(Y|U)$  are uniquely defined. Furthermore, according to Remark 14.37, in these two examples, the conditional expectation values  $E^{X=x}(Y|U=u)$  are also uniquely defined, which follows from the fact that  $P^{X=x}(U=u) > 0$  for all pairs  $(x, u)$  of values of  $X$  and  $U$ . In contrast, in Example 14.3, only  $E^{X=0}(Y|U)$  is uniquely defined, but  $E^{X=1}(Y|U)$  is not. In fact, in this example,  $E^{X=1}(Y|U)$  is even not  $P$ -unique, although it is  $P^{X=1}$ -unique. This issue is dealt with in section 14.6.2.  $\triangleleft$

### 14.6.2 Uniqueness w.r.t. $P$ and Other Probability Measures

A conditional expectation  $E^B(Y|\mathcal{C})$  is always  $P^B$ -unique (see Rem. 14.9). However, we may also ask if  $E^B(Y|\mathcal{C})$  is  $Q$ -unique, where  $Q$  is *any* probability measure on  $(\Omega, \mathcal{A})$ . This includes  $Q = P$ , but also  $Q = P^C$ , where  $C \in \mathcal{A}$  and  $C \neq B$  (see Remarks 5.17 and 5.20). If  $E^B(Y|\mathcal{C})$  is  $Q$ -unique, then two versions  $V, V^* \in \mathcal{E}^B(Y|\mathcal{C})$  do not only have identical distributions, expectations, variances, etc. with respect to  $P^B$ , but also with respect to  $Q$ , i. e.,  $Q_V = Q_{V^*}$  (see Cors. 5.24 and 6.17). The following example shows why this is of interest.

**Example 14.45 (Pre-Post Design)** Suppose that  $X$  is an indicator variable with values 0 (control) and 1 (treatment),  $Y$  represents *life satisfaction after treatment*, and  $Z$  *life satisfaction before treatment*. Then  $P$ -uniqueness of  $E^{X=x}(Y|Z)$  is crucial if we consider

$$E(Y|X, Z) \stackrel{P}{=} g_0(Z) + g_1(Z) \cdot X$$

with

$$g_1(Z) \stackrel{P}{=} E^{X=1}(Y|Z) - E^{X=0}(Y|Z),$$

or the expectation

$$E[g_1(Z)] = E[E^{X=1}(Y|Z)] - E[E^{X=0}(Y|Z)].$$

Furthermore, for  $x = 0$  and  $x = 1$ , we may also consider the  $(X=x)$ -conditional expectation values of  $g_1(Z)$

$$E[g_1(Z)|X=x] = E[E^{X=1}(Y|Z)|X=x] - E[E^{X=0}(Y|Z)|X=x],$$

the *average effect of  $X$  on  $Y$*  given  $x = 0$  (control) given  $x = 1$  (treatment), respectively. Considering  $E[g_1(Z)|X=x]$ , where  $x = 0$  or  $x = 1$ , it is crucial that  $E^{X=0}(Y|Z)$  and  $E^{X=1}(Y|Z)$  are unique with respect to the measure  $P^{X=x}$  for the following reason: For  $x = 0$  or  $x = 1$ , if  $P(X=x) > 0$  and  $E^{X=1}(Y|Z)$  is  $P^{X=x}$ -unique, then the conditional expectation value  $E[E^{X=1}(Y|Z)|X=x]$  is identical for different versions  $E^{X=1}(Y|Z) \in \mathcal{E}^{X=1}(Y|Z)$ . Correspondingly, for  $x = 0$  or  $x = 1$ , if  $P(X=x) > 0$  and  $E^{X=0}(Y|Z)$  is  $P^{X=x}$ -unique, then  $E[E^{X=0}(Y|Z)|X=x]$  is identical for different versions  $E^{X=0}(Y|Z) \in \mathcal{E}^{X=0}(Y|Z)$  (see Exercise 14-10).  $\triangleleft$

**14.6.3 Necessary and Sufficient Conditions of P-Uniqueness**

Now we present conditions that are equivalent to  $P$ -uniqueness of  $E^B(Y|\mathcal{C})$ . Note that, in this theorem, we do not refer to the expectation  $E^B(V)$  of  $V$  with respect to the measure  $P^B$ , but to the expectation  $E(V)$  of  $V$  with respect to  $P$ .

**Theorem 14.46 (Conditions Equivalent to  $P$ -Uniqueness of  $E^B(Y|\mathcal{C})$ )**  
 Let the assumptions 14.5 hold. Then the following propositions are equivalent to each other:

- (a)  $E^B(Y|\mathcal{C})$  is  $P$ -unique.
- (b)  $P \ll_{\mathcal{C}} P^B$ .
- (c)  $P(B|\mathcal{C}) \underset{P}{>} 0$ .

If there is a version  $V \in \mathcal{E}^B(Y|\mathcal{C})$  such that  $E(V)$  is finite, then (a) to (c) are also equivalent to

- (d)  $\forall V, V^* \in \mathcal{E}^B(Y|\mathcal{C}): E(V) = E(V^*)$ .

(Proof p. 427)

**Remark 14.47 (Sufficient Conditions for Finiteness of  $E[E^B(Y|\mathcal{C})]$ )** Remember, the expectation of a random variable  $Y$  exists if  $\int Y^+ dP$  or  $\int Y^- dP$  are finite (see Def. 3.28). Hence, the expectation  $E[E^B(Y|\mathcal{C})]$  of any version  $E^B(Y|\mathcal{C}) \in \mathcal{E}^B(Y|\mathcal{C})$  exists and is finite, e. g., if one of the following conditions holds:

- (a)  $\mathcal{C}$  is a finite set and  $E^B(Y)$  is finite (see Exercise 14-11),
- (b)  $E^B(Y|\mathcal{C})$  has only a finite number of real values (see Rem. 6.5),
- (c)  $E^B(Y|\mathcal{C})$  is  $P$ -almost surely bounded on both sides, i. e.,  
 $\exists \alpha \in \mathbb{R}: -\alpha \leq E^B(Y|\mathcal{C}) \leq \alpha$  [see Eq. (3.50)], or
- (d)  $Y$  is  $P$ -almost surely bounded on both sides, i. e.,  
 $\exists \alpha \in \mathbb{R}: -\alpha \leq Y \leq \alpha$  [ $E^B(\cdot) = E^{1B=1}(\cdot)$ , and see Eq. (14.30), and Box 10.3 (v) (vii), and (c)].

A special case of (d) is  $0 \leq Y \leq \alpha$ , for  $0 < \alpha \in \mathbb{R}$ . Another one is  $Y = 1_A$ , if  $A \in \mathcal{A}$ .  
 $\triangleleft$

In the following corollary we translate Theorem 14.46 to the special case of an  $X$ -conditional expectation  $E^B(Y|X)$  with respect to  $P^B$  and apply Lemma 5.29.

**Corollary 14.48 ( $P$ -Uniqueness of  $E^B(Y|X)$ )**  
 Let the assumptions 14.5 hold and let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  be a random variable. Then the following propositions are equivalent to each other:

- (a)  $E^B(Y|X)$  is  $P$ -unique.
- (b)  $P \ll_{\sigma(X)} P^B$ .

(c)  $P(B|X) \stackrel{P}{>} 0$ .

(d)  $P_X \ll_{\mathcal{A}'_X} P_X^B$ .

If there is a version  $V \in \mathcal{E}^B(Y|X)$  such that  $E(V)$  is finite, then (a) to (d) are also equivalent to

(e)  $\forall V, V^* \in \mathcal{E}^B(Y|X): E(V) = E(V^*)$ .

**Remark 14.49 (Absolute Continuity if X is Discrete)** Under the assumptions of Corollary 14.48: If  $X$  is discrete and  $\{x\} \in \mathcal{A}'_X$  for all  $x \in \Omega'_X$ , then  $P_X \ll_{\mathcal{A}'_X} P_X^B$  [see Cor. 14.48 (d)] is equivalent to

$$\forall x \in \Omega'_X: P^B(X=x) = 0 \Rightarrow P(X=x) = 0 \tag{14.48}$$

(see Exercise 14-12). ◁

**Example 14.50 (Joe and Ann With Random Assignment – continued)** The last two columns of Table 14.1 display the conditional-probability measures  $P^{X=0}$  and  $P^{X=1}$ . The values of  $P^{X=1} = P^B$  with  $B = \{X=1\}$  were already computed in Example 4.34. The last two columns of Table 14.1 show that  $P^{X=x}(U=u) > 0$  for all pairs of values of  $X$  and  $U$ . This implies that the conditional expectations  $E^{X=x}(Y|U)$  are uniquely defined for both values of  $X$  (see Cor. 14.42), which in turn implies that they are  $P$ -unique. Furthermore, according to Equation (14.35) the expectations

$$E[E^{X=x}(Y|U)] = \sum_u E(Y|X=x, U=u) \cdot P(U=u), \quad x = 0, 1, \tag{14.49}$$

are finite. According to Remark 14.47 (b), this follows from the fact that the conditional expectation values  $E(Y|X=x, U=u) = E^{X=x}(Y|U=u)$  are finite. Finally,  $E(V) = E(V^*) = E[E^{X=x}(Y|U)]$ , for all  $V, V^* \in \mathcal{E}^{X=x}(Y|U)$  [see Cor. 14.48 (e)]. ◁

**Example 14.51 (No Treatment for Joe – continued)** Continuing Example 14.24, consider the event

$$\{X=1\} = \{(Joe, yes, -), (Joe, yes, -), (Ann, yes, -), (Ann, yes, +)\},$$

that the *drawn person is treated* and the event

$$\{U=Joe\} = \{(Joe, no, -), (Joe, no, +), (Joe, yes, -), (Joe, yes, +)\},$$

that *Joe is drawn*. In this example, we already computed the  $(X=1)$ -conditional probability  $P^{X=1}(U=Joe) = 0$  and the (unconditional) probability  $P(U=Joe) = .50$ . Because  $\{U=Joe\} \in \sigma(U)$ , in this example, it is not true that  $P^{X=1}(A) = 0$  implies  $P(A) = 0$  for all  $A \in \sigma(U)$ . Therefore,  $P \ll_{\sigma(U)} P^{X=1}$  does not hold (see Def. 3.70). Hence, Corollary 14.48 implies that the conditional expectation  $E^{X=1}(Y|U)$  is not  $P$ -unique (see also Examples 14.3 and 14.24).

In Table 14.3 (p. 403), the values of  $E^{X=1}(Y|U)$  are not uniquely defined for all four  $\omega \in \{U=Joe\}$ . Instead of 9, we could have assigned *any* real number to these

four possible outcomes  $\omega$ , because  $P^{X=1}(U=Joe) = 0$ . Nevertheless,  $E^{X=1}(Y|U)$  is  $P^{X=1}$ -unique. However, because  $E^{X=1}(Y|U)$  is not  $P$ -unique, in this example,  $E(V_1) = E(V_1^*)$  does *not* hold for all  $V_1, V_1^* \in \mathcal{E}^{X=1}(Y|U)$  (see Th. 14.46). This has already been illustrated in Example 14.3.  $\triangleleft$

**14.6.4 Properties Related to  $P$ -Uniqueness**

Box 14.1 summarizes some important properties related to  $P$ -uniqueness (for proofs see Exercise 14-13), some of which have already been treated and illustrated in section 14.6.3. In the following remarks we comment some of the implications of  $P$ -uniqueness.

**Remark 14.52 (Implications of  $P$ -Uniqueness)** Suppose that  $Y$  is nonnegative or with finite expectation  $E(Y)$ , that  $E^B(Y|\mathcal{C})$  is  $P$ -unique, and  $C \in \mathcal{A}$  with  $P(C) > 0$ . Then, according to property (v) of Box 14.1,  $E^B(Y|\mathcal{C})$  is also  $P^C$ -unique, and according to property (vi) of Box 14.1, the distributions  $P_V^C$  with respect to  $P^C$  of all versions  $V \in \mathcal{E}^B(Y|\mathcal{C})$  are identical (see Cor. 5.24). This implies, e. g., that the expectation  $E^C[E^B(Y|\mathcal{C})]$  of  $E^B(Y|\mathcal{C})$  with respect to the conditional-probability measure  $P^C$  is identical for all versions  $V \in \mathcal{E}^B(Y|\mathcal{C})$  [see Box 14.1 (vii)]. The same applies to its variance  $Var^C[E^B(Y|\mathcal{C})]$  [see Rem. 6.27 and Box 6.2 (v)] as well as to its covariance  $Cov^C[E^B(Y|\mathcal{C}), W]$  with another random variable  $W: (\Omega, \mathcal{A}, P) \rightarrow (\bar{\mathbb{R}}, \mathcal{B})$  [see Box 7.1 (x)], provided that this variance and this covariance with respect to  $P^C$  exist.  $\triangleleft$

**Remark 14.53 (The Special Case  $C = \Omega$ )** A special case is  $C = \Omega$ . Hence, under  $P$ -uniqueness of  $E^B(Y|\mathcal{C})$ , the following equations hold for all  $V, V^* \in \mathcal{E}^B(Y|\mathcal{C})$ :

$$P_V = P_{V^*}, \tag{14.50}$$

$$E(V) = E(V^*), \tag{14.51}$$

$$Var(V) = Var(V^*), \tag{14.52}$$

$$Cov(V, W) = Cov(V^*, W), \tag{14.53}$$

provided that these expectations, variances, and covariances with respect to the measure  $P$  exist.  $\triangleleft$

**Remark 14.54 (Another Implication of  $P$ -Uniqueness)** Let the assumptions 14.5 hold. Furthermore, let  $\mathcal{D} \subset \mathcal{A}$  be a  $\sigma$ -algebra,  $\mathcal{C} \subset \mathcal{D}$ ,  $C \in \mathcal{A}$  with  $P(C) > 0$ , and suppose that  $E^B(Y|\mathcal{C})$  is  $P^C$ -unique. Then

$$E^C[E^B(Y|\mathcal{C})|\mathcal{D}] \stackrel{P^C}{=} E^B(Y|\mathcal{C}), \tag{14.54}$$

which follows from Rule (vii) of Box 10.1, because, by definition,  $E^B(Y|\mathcal{C})$  is  $\mathcal{C}$ -measurable, and because we assume  $\mathcal{C} \subset \mathcal{D}$ . For the special case  $C = \Omega$ , this yields: If  $\mathcal{C} \subset \mathcal{D}$  and  $E^B(Y|\mathcal{C})$  is  $P$ -unique, then

$$E[E^B(Y|\mathcal{C})|\mathcal{D}] \stackrel{P}{=} E^B(Y|\mathcal{C}). \tag{14.55}$$

$\triangleleft$

**Box 14.1** *P*-Uniqueness of  $E^B(Y|\mathcal{C})$ 

Let  $(\Omega, \mathcal{A}, P)$  be a probability space, let  $\mathcal{C}, \mathcal{D} \subset \mathcal{A}$  be  $\sigma$ -algebras, let  $B, C \in \mathcal{A}$  with  $P(B), P(C) > 0$ , and let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  be a random variable that is nonnegative or with finite expectation  $E(Y)$ . Then:

$$E^B(Y|\mathcal{C}) \text{ is } P\text{-unique} \Leftrightarrow \forall V, V^* \in \mathcal{E}^B(Y|\mathcal{C}): V \stackrel{P}{=} V^* \quad (\text{i})$$

$$E^B(Y|\mathcal{C}) \text{ is } P\text{-unique} \Leftrightarrow P \ll_{\mathcal{C}} P^B \quad (\text{ii})$$

$$E^B(Y|\mathcal{C}) \text{ is } P\text{-unique} \Leftrightarrow P(B|\mathcal{C}) \stackrel{P}{>} 0 \quad (\text{iii})$$

$$E^B(Y|\mathcal{C}) \text{ is } P\text{-unique} \Rightarrow E^B(Y|\mathcal{D}) \text{ is } P\text{-unique, if } \mathcal{D} \subset \mathcal{C} \quad (\text{iv})$$

$$E^B(Y|\mathcal{C}) \text{ is } P\text{-unique} \Rightarrow E^B(Y|\mathcal{C}) \text{ is } P^C\text{-unique} \quad (\text{v})$$

$$E^B(Y|\mathcal{C}) \text{ is } P\text{-unique} \Rightarrow \forall V, V^* \in \mathcal{E}^B(Y|\mathcal{C}): P_V^C = P_{V^*}^C \quad (\text{vi})$$

$$E^B(Y|\mathcal{C}) \text{ is } P\text{-unique} \Rightarrow \forall V, V^* \in \mathcal{E}^B(Y|\mathcal{C}): E^C(V) = E^C(V^*). \quad (\text{vii})$$

If  $E^B(Y|\mathcal{C})$  or  $E^C(Y|\mathcal{C})$  is real-valued and  $\alpha, \beta \in \mathbb{R}$ , then

$$E^B(Y|\mathcal{C}), E^C(Y|\mathcal{C}) \text{ are } P\text{-unique} \Rightarrow \alpha E^B(Y|\mathcal{C}) + \beta E^C(Y|\mathcal{C}) \text{ is } P\text{-unique.} \quad (\text{viii})$$

If  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  is a random variable and  $\Omega'_X$  finite or countable, then:

$$\left( \forall x \in \Omega'_X: P^B(X=x) > 0 \right) \Rightarrow E^B(Y|X) \text{ is } P\text{-unique.} \quad (\text{ix})$$

**Remark 14.55 (Expectation of a Linear Combination)** An implication of Box 14.1 (viii) and Box 6.1 (vii) is:

$$E[\alpha \cdot E^B(Y|\mathcal{C}) + \beta \cdot E^C(Y|\mathcal{C})] = \alpha \cdot E[E^B(Y|\mathcal{C})] + \beta \cdot E[E^C(Y|\mathcal{C})], \quad (14.56)$$

provided that  $E^B(Y|\mathcal{C})$  and  $E^C(Y|\mathcal{C})$  are  $P$ -unique, that  $E^B(Y|\mathcal{C})$  or  $E^C(Y|\mathcal{C})$  is real-valued, and the expectation  $E[E^B(Y|\mathcal{C})]$  or  $E[E^C(Y|\mathcal{C})]$  is finite. Under these assumptions, for all real-valued versions  $V_B \in \mathcal{E}^B(Y|\mathcal{C})$  and  $V_C \in \mathcal{E}^C(Y|\mathcal{C})$ ,

$$E(\alpha \cdot V_B + \beta \cdot V_C) = \alpha \cdot E(V_B) + \beta \cdot E(V_C).$$

◁

**Example 14.56 (Joe and Ann With Self-Selection – continued)** Consider again Table 14.2 (p. 402). In this example, the function  $g_1(U) = E^{X=1}(Y|U) - E^{X=0}(Y|U)$  is a uniquely defined random variable on  $(\Omega, \mathcal{A}, P)$ , because  $E^{X=0}(Y|U)$  as well as  $E^{X=1}(Y|U)$  are uniquely defined, i. e., each of the sets  $\mathcal{E}^{X=0}(Y|U)$  and  $\mathcal{E}^{X=1}(Y|U)$  has only one single element. This implies that the average treatment effect

$$E[g_1(U)] = E[E^{X=1}(Y|U) - E^{X=0}(Y|U)]$$

is uniquely defined as well. Note that the assumptions of Box 14.1 (viii) are less restrictive, because they allow that each of the sets  $\mathcal{E}^{X=0}(Y|U)$  and  $\mathcal{E}^{X=1}(Y|U)$  has

more than one element. The requirement of Box 14.1 (viii) is not uniqueness but only  $P$ -uniqueness.

◁

**Example 14.57 (No Treatment for Joe – continued)** In the example presented in Table 14.3 (p. 403), the set  $\mathcal{E}^{X=0}(Y|U)$  has only one single element. However,  $\mathcal{E}^{X=1}(Y|U)$  has infinitely many elements: Replacing the value 9 by any other real number yields a new element of  $\mathcal{E}^{X=1}(Y|U)$ . More important, it is *not true* that all these elements are pairwise  $P$ -equivalent. Hence, in this example, the function  $g_1(U) = E^{X=1}(Y|U) - E^{X=0}(Y|U)$  is *not  $P$ -unique* (see Example 14.51). In other words, it *not true* that all elements of the set

$$\{V_1 - V_0 : V_0 \in \mathcal{E}^{X=0}(Y|U), V_1 \in \mathcal{E}^{X=1}(Y|U)\}$$

are pairwise  $P$ -equivalent. Therefore, there can be versions  $V_0, V_0^* \in \mathcal{E}^{X=0}(Y|U)$  and  $V_1, V_1^* \in \mathcal{E}^{X=1}(Y|U)$  such that

$$E(V_1 - V_0) \neq E(V_1^* - V_0^*).$$

This means that there is no uniquely defined average treatment effect  $E[g_1(U)]$  in this example.

◁

In the following corollary we extend Theorem 14.33, adding another assumption. Remember, if assumptions of Definition 14.29 hold and  $P(Z=z) > 0$ , then according to Theorem 14.33,

$$E^{Z=z}(Y|X) \stackrel{p_{Z=z}}{=} E(Y|X, Z=z), \tag{14.57}$$

referring to the partial conditional expectation  $E(Y|X, Z=z)$  [see Def. 14.29].

**Corollary 14.58 (Implications of  $P^C$ -Uniqueness of  $E^{Z=z}(Y|X)$ )**

Let the assumptions of Definition 14.29 hold, suppose  $P(Z=z) > 0$ , and let  $C \in \mathcal{A}$  with  $P(C) > 0$ . If  $E^{Z=z}(Y|X)$  is  $P^C$ -unique, then

$$E^{Z=z}(Y|X) \stackrel{P^C}{=} E(Y|X, Z=z), \tag{14.58}$$

and

$$E^{Z=z}(Y|X=x) = E(Y|X=x, Z=z), \text{ for } P_X^C\text{-a.a. } x \in \Omega'_X. \tag{14.59}$$

(Proof p. 428)

**Remark 14.59 (Implications of  $P$ -Uniqueness of  $E^{Z=z}(Y|X)$ )** For  $C = \Omega$ , this corollary yields: If  $E^{Z=z}(Y|X)$  is  $P$ -unique, then

$$E^{Z=z}(Y|X) \stackrel{P}{=} E(Y|X, Z=z), \tag{14.60}$$

and

$$E^{Z=z}(Y|X=x) = E(Y|X=x, Z=z), \text{ for } P_X\text{-a.a. } x \in \Omega'_X. \tag{14.61}$$

◁

Now we consider the family of factorizations  $g_B$  of  $E^B(Y|X)$ , which are defined by Equation (14.32). Because each element of  $\mathcal{E}^B(Y|X)$  has at least one factorization  $g_B$ , there is a family of factorizations, which are random variables on the probability space  $(\Omega'_X, \mathcal{A}'_X, P_X)$ .

The next corollary immediately follows from Corollary 5.25 (i) if  $P^C$  takes the role of  $P$  and  $P_X^C$  the role of  $P_X$ .

**Corollary 14.60 ( $P^C$ -Uniqueness and  $P_X^C$ -Uniqueness)**

Let the assumptions 14.5 hold, let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  be a random variable, let  $V = g_B(X)$  and  $V^* = g_B^*(X)$  be two elements of  $\mathcal{E}^B(Y|X)$ , and let  $C \in \mathcal{A}$  with  $P(C) > 0$ . Then

$$V \stackrel{P^C}{=} V^* \Leftrightarrow g_B \stackrel{P_X^C}{=} g_B^*. \quad (14.62)$$

**Remark 14.61 ( $P$ -Uniqueness and  $P_X$ -Uniqueness)** For  $C = \Omega$ , Corollary 14.60 yields: If  $V = g_B(X)$  and  $V^* = g_B^*(X)$  are two elements of  $\mathcal{E}^B(Y|X)$ , then

$$g_B(X) \stackrel{P}{=} g_B^*(X) \Leftrightarrow g_B \stackrel{P_X}{=} g_B^*. \quad (14.63)$$

Note that both sides of (14.63) are not only equivalent to each other but also to

$$g_B(x) = g_B^*(x), \quad \text{for } P_X\text{-a.a. } x \in \Omega'_X. \quad (14.64)$$

<

**Remark 14.62 (Some Formulas for the Expectation of  $E^B(Y|X)$ )** Suppose that the assumptions 14.5 hold, where  $Y$  is nonnegative or with finite expectation  $E(Y)$ , and let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  be a random variable. If  $E^B(Y|X)$  is  $P$ -unique, then, according to Equations (6.13) and (14.36),

$$E[E^B(Y|X)] = \int E^B(Y|X=x) P_X(dx) = \int E(Y|X=x, 1_B=1) P_X(dx). \quad (14.65)$$

Furthermore, if  $X$  is discrete (see Def. 5.56) and  $P_X(\Omega'_0) = 1$ , then

$$E[E^B(Y|X)] = \int E^B(Y|X=x) P_X(dx) \quad [(6.13)]$$

$$= \sum_{x \in \Omega'_0} E^B(Y|X=x) \cdot P(X=x) \quad [(6.15)] \quad (14.66)$$

$$= \sum_{x \in \Omega'_0} E(Y|X=x, 1_B=1) \cdot P(X=x). \quad [(14.36)]$$

<

**Remark 14.63 (Some Formulas for the Expectation of  $E^{Z=z}(Y|X)$ )** Correspondingly, let the assumptions 14.20 hold, where  $Y$  is nonnegative or with finite expectation  $E(Y)$ . If  $E^{Z=z}(Y|X)$  is  $P$ -unique, then for  $B = \{Z=z\}$ , Equations (14.65) and (14.61) yield

$$E[E^{Z=z}(Y|X)] = \int E^{Z=z}(Y|X=x) P_X(dx) = \int E(Y|X=x, Z=z) P_X(dx), \quad (14.67)$$

and, if  $X$  is discrete (see Def. 5.56) and  $P_X(\Omega'_0) = 1$ , then Equations (14.66) and (14.61) yield

$$\begin{aligned} E[E^{Z=z}(Y|X)] &= \int E^{Z=z}(Y|X=x) P_X(dx) \\ &= \sum_{x \in \Omega'_0} E(Y|X=x, Z=z) \cdot P(X=x). \end{aligned} \quad (14.68)$$

◁

### 14.7 Conditional Mean Independence With Respect to $P^{Z=z}$

According to the following theorem, a numerical random variable  $Y$  on  $(\Omega, \mathcal{A}, P)$  that is nonnegative or with finite expectation  $E(Y)$  is  $\mathcal{C}$ -conditionally mean independent from  $Z$  with respect to  $P^{Z=z}$ . For simplicity, we use the following notation

$$\sigma(\mathcal{C}, \mathcal{D}) := \sigma(\mathcal{C} \cup \mathcal{D}), \quad \sigma(\mathcal{C}, Z) := \sigma(\mathcal{C} \cup \sigma(Z)), \quad (14.69)$$

$$E^{Z=z}(Y|\mathcal{C}, \mathcal{D}) := E^{Z=z}(Y|\sigma(\mathcal{C} \cup \mathcal{D})), \quad (14.70)$$

$$E^{Z=z}(Y|\mathcal{C}, Z) := E^{Z=z}(Y|\sigma(\mathcal{C} \cup \sigma(Z))). \quad (14.71)$$

#### Theorem 14.64 (Conditional Mean Independence)

Let the assumptions 14.17 hold, where  $Y$  is nonnegative or with finite expectation  $E(Y)$ . Furthermore, let  $Z: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Z, \mathcal{A}'_Z)$  be a random variable, and suppose that  $z \in \Omega'_Z$  with  $P(Z=z) > 0$ . Then

$$E^{Z=z}(Y|\mathcal{C}), E(Y|\mathcal{C}, Z) \in \mathcal{E}^{Z=z}(Y|\mathcal{C}, Z),$$

which implies

$$E^{Z=z}(Y|\mathcal{C}, Z) \stackrel{p^{Z=z}}{=} E^{Z=z}(Y|\mathcal{C}) \stackrel{p^{Z=z}}{=} E(Y|\mathcal{C}, Z). \quad (14.72)$$

(Proof p. 428)

**Remark 14.65 (A Caveat)** If  $\sigma(\mathcal{C}, Z) \neq \mathcal{C}$ , then  $E^{Z=z}(Y|\mathcal{C}, Z)$  and  $E(Y|\mathcal{C}, Z)$  are not necessarily elements of  $\mathcal{E}^{Z=z}(Y|\mathcal{C})$  [see Def. 14.7 (a)]. Nevertheless,

$$E^{Z=z}(Y|\mathcal{C}) \stackrel{p^{Z=z}}{=} E(Y|\mathcal{C}, Z).$$

According to Box 6.1 (viii) and Box 10.2 (iv), this implies

$$E^{Z=z}(Y) = E^{Z=z}[E^{Z=z}(Y|\mathcal{C})] = E^{Z=z}[E(Y|\mathcal{C}, Z)], \quad (14.73)$$

and, for  $\mathcal{C} = \sigma(X)$ , using Equation (14.14),

$$E^{Z=z}(Y) = E^{Z=z}[E^{Z=z}(Y|X)] = E^{Z=z}[E(Y|X, Z)]. \quad (14.74)$$

Furthermore, using (14.30), this equation yields

$$E^{Z=z}(Y) = E^{Z=z}[E^{Z=z}(Y|X)] = E^{Z=z}[E(Y|X, Z=z)] \quad (14.75)$$

Note that we still presume that assumptions of Theorem 14.64 hold.  $\triangleleft$

**Remark 14.66 (Two Implications Concerning Mean Independence)** Let the assumptions 14.20 hold. Then

$$E^{Z=z}(Y|\mathcal{C}, X) \underset{P^{Z=z}}{=} E^{Z=z}(Y|\mathcal{C}), \quad \text{if } \sigma(X) \subset \sigma(Z) \quad (14.76)$$

(see Exercise 14-14). Furthermore, considering a  $\sigma$ -algebra  $\mathcal{D}$  and assuming that  $Y$  is nonnegative or with finite expectation  $E(Y)$ ,

$$E^{Z=z}[E(Y|\mathcal{C}, Z)|\mathcal{D}] \underset{P^{Z=z}}{=} E^{Z=z}(Y|\mathcal{D}), \quad \text{if } \mathcal{D} \subset \sigma(\mathcal{C}, Z) \quad (14.77)$$

(see Exercise 14-15).  $\triangleleft$

In the following theorem we generalize the propositions of Remark 14.63.

**Theorem 14.67 (Expectation of  $E^{Z=z}(Y|X, W)$  With Respect to  $P^{W=w}$ )**

Let the assumptions 14.20 hold, where  $Y$  is nonnegative or with finite expectation  $E(Y)$ , let  $W: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_W, \mathcal{A}'_W)$  be a random variable, let  $w \in \Omega'_W$  with  $P(W=w) > 0$ , and assume that  $E^{Z=z}(Y|X, W)$  is  $P^{W=w}$ -unique. Then

$$\begin{aligned} E^{W=w}[E^{Z=z}(Y|X, W)] &= \int E^{Z=z}(Y|X=x, W=w) P_X^{W=w}(dx) \\ &= \int E(Y|X=x, W=w, Z=z) P_X^{W=w}(dx). \end{aligned} \quad (14.78)$$

(Proof p. 430)

**Remark 14.68 (Discrete  $X$ )** If the assumptions of Theorem 14.67 hold,  $X$  is discrete (see Def. 5.56), and  $P_X(\Omega'_0) = 1$ , then Equations (14.78), (6.15), (14.59), and (9.13) yield

$$\begin{aligned} E^{W=w}[E^{Z=z}(Y|X, W)] &= \int E^{Z=z}(Y|X=x, W=w) P_X^{W=w}(dx) \\ &= \sum_{x \in \Omega'_0} E(Y|X=x, W=w, Z=z) \cdot P^{W=w}(X=x) \\ &= \sum_{x \in \Omega'_0} E(Y|X=x, W=w, Z=z) \cdot P(X=x|W=w). \end{aligned} \quad (14.79)$$

$\triangleleft$

In the following theorem we study an implication of conditional mean independence [see Def. 10.45 (ii)] on conditional expectations with respect to  $P^{Z=z}$ .

**Theorem 14.69 (An Implication of Conditional Mean Independence)**

Let the assumptions 14.17 hold, where  $Y$  is nonnegative or with finite expectation  $E(Y)$ , and let  $\mathcal{C}, \mathcal{D} \subset \mathcal{A}$  be  $\sigma$ -algebras. Then

$$E(Y|\mathcal{C}, \mathcal{D}, Z) \stackrel{p}{=} E(Y|\mathcal{C}, Z) \Rightarrow E^{Z=z}(Y|\mathcal{C}, \mathcal{D}) \stackrel{p^{Z=z}}{=} E^{Z=z}(Y|\mathcal{C}). \quad (14.80)$$

*(Proof p. 430)*

**Remark 14.70 (An Implication of Conditional Mean Independence)** Let the assumptions of Theorem 14.69 hold. If  $X$  and  $W$  are two random variables on the probability space  $(\Omega, \mathcal{A}, P)$ , then

$$E(Y|X, W, Z) \stackrel{p}{=} E(Y|X, Z) \Rightarrow E^{Z=z}(Y|X, W) \stackrel{p^{Z=z}}{=} E^{Z=z}(Y|X). \quad (14.81)$$

Furthermore, for  $X = \alpha$ ,  $\alpha \in \Omega'_X$ , we can conclude that  $E(Y|W, Z) \stackrel{p}{=} E(Y|Z)$  implies

$$E^{Z=z}(Y|W) \stackrel{p^{Z=z}}{=} E^{Z=z}(Y). \quad (14.82)$$

◁

**14.8 Proofs**

**Proof of Theorem 14.26**

Because  $f(Z) \stackrel{p^{Z=z}}{=} f(z)$  (see Rem. 9.1),

$$\begin{aligned} E^{Z=z}[f(Z) \cdot Y|\mathcal{C}] &\stackrel{p^{Z=z}}{=} E^{Z=z}[f(z) \cdot Y|\mathcal{C}] && \text{[Box 10.1 (ix)]} \\ &\stackrel{p^{Z=z}}{=} f(z) \cdot E^{Z=z}(Y|\mathcal{C}). && \text{[Box 10.1 (iii)]} \end{aligned}$$

**Proof of Theorem 14.33**

Let the assumptions of Definition 14.29 hold. Because  $g: \Omega'_X \times \Omega'_Z \rightarrow \bar{\mathbf{R}}$  is  $(\mathcal{A}'_X \otimes \mathcal{A}'_Z, \bar{\mathcal{B}})$ -measurable, the function  $g_z$  defined in (14.23) is  $(\mathcal{A}'_X, \bar{\mathcal{B}})$  measurable (see Bauer, 2001, Lemma 23.5. p. 138). Hence, condition (a) of Definition 14.7 holds. Furthermore, for all  $C \in \sigma(X)$ , we can conclude  $C \cap \{Z=z\} \in \sigma(X, Z)$ , and therefore, for all  $C \in \sigma(X)$ ,

$$\begin{aligned}
\int 1_C E(Y|X, Z=z) dP^{Z=z} &= \frac{1}{P(Z=z)} \int 1_C \cdot 1_{Z=z} \cdot E(Y|X, Z=z) dP && [(9.16), (9.11)] \\
&= \frac{1}{P(Z=z)} \int 1_C \cdot 1_{Z=z} \cdot g_z(X) dP && [(14.25)] \\
&= \frac{1}{P(Z=z)} \int 1_C \cdot 1_{Z=z} \cdot g(X, Z) dP && [(14.26)] \\
&= \frac{1}{P(Z=z)} \int 1_C \cdot 1_{Z=z} \cdot E(Y|X, Z) dP && [\text{Ass. of Def. 14.29}] \\
&= \frac{1}{P(Z=z)} \int 1_C \cdot 1_{Z=z} \cdot Y dP && [\text{Def. 10.2 (b)}] \\
&= \int 1_C \cdot Y dP^{Z=z}. && [(9.16)]
\end{aligned}$$

This shows that condition (b) of Definition 14.7 holds as well.

### **Proof of Theorem 14.46**

(b)  $\Rightarrow$  (a). Remember that (b) is defined by

$$\forall C \in \mathcal{C}: P^B(C) = 0 \Rightarrow P(C) = 0.$$

If  $V, V^* \in \mathcal{E}^B(Y|\mathcal{C})$ , then  $\{V \neq V^*\} \in \mathcal{C}$  [see Rem. 2.67 (c)],  $P^B(\{V \neq V^*\}) = 0$ , and (b) implies  $P(\{V \neq V^*\}) = 0$ .

(a)  $\Rightarrow$  (b). This proposition is proved by contraposition, i. e., we show  $\neg$  (b)  $\Rightarrow$   $\neg$  (a). Assume that there is an  $A \in \mathcal{C}$  with  $P^B(A) = 0$  and  $P(A) > 0$ , and let  $V \in \mathcal{E}^B(Y|\mathcal{C})$  with  $V(\omega) = 0$  for all  $\omega \in A$ . [Note that if  $V' \in \mathcal{E}^B(Y|\mathcal{C})$  and  $A \in \mathcal{C}$  with  $P^B(A) = 0$ , then  $V := V' \cdot 1_{A^c} \in \mathcal{E}^B(Y|\mathcal{C})$ .] Then  $V^* = V + 1_A$  is also  $\mathcal{C}$ -measurable and, for all  $C \in \mathcal{C}$ ,

$$\begin{aligned}
\int 1_C V^* dP^B &= \int 1_C \cdot (V + 1_A) dP^B \\
&= \int 1_C V dP^B + \int 1_C 1_A dP^B && [(3.33)] && (14.83) \\
&= \int 1_C V dP^B, && [\text{Lemma 3.45}]
\end{aligned}$$

because  $\int 1_C \cdot 1_A dP^B = P^B(C \cap A) = 0$ . Hence, according to Definition 14.7,  $V^* = V + 1_A \in \mathcal{E}^B(Y|\mathcal{C})$ . However,  $P(\{V \neq V^*\}) = P(A) > 0$ .

(c)  $\Rightarrow$  (b). Let  $P(B|\mathcal{C}) \stackrel{p}{\geq} 0$ . This implies  $P(A) = 0$ , where

$$A = \{\omega \in \Omega: P(B|\mathcal{C})(\omega) \leq 0\},$$

and, according to Remark 2.67 (a),  $A \in \mathcal{C}$ . However, if  $P(A) = 0$ , then, according to Rule (ix) of Box 4.1,

$$P(A \cup C) = P(C), \quad \forall C \in \mathcal{C}. \quad (14.84)$$

Now, let  $C \in \mathcal{C}$  with  $P^B(C) = 0$ . This yields

$$\begin{aligned}
\int P(B|\mathcal{C}) \cdot 1_C dP &= \int 1_B \cdot 1_C dP && [\text{Def. 10.2 (b)}] \\
&= P(B \cap C) = P^B(C) \cdot P(B) = 0. && [(1.33), (3.9), (4.14)]
\end{aligned}$$

Because  $P(B|\mathcal{C}) \cdot 1_C \stackrel{p}{\geq} 0$  [see Box 10.3 (ii)], Lemma 3.44 yields  $P(B|\mathcal{C}) \cdot 1_C \stackrel{p}{=} 0$ , which is equivalent to

$$\begin{aligned}
 1 &= P(\{\omega \in \Omega: P(B|\mathcal{C})(\omega) \cdot 1_C(\omega) = 0\}) \\
 &= P(\{\omega \in \Omega: \omega \in A \text{ or } \omega \in C^c\}) = P(A \cup C^c) \\
 &= P(C^c). \tag{14.84}
 \end{aligned}$$

Hence,  $P(C) = 0$ , which shows that  $P(B|\mathcal{C}) \geq_P 0$  implies (b).

(b)  $\Rightarrow$  (c). This proposition is proved by contraposition, i. e., we show  $\neg(c) \Rightarrow \neg(b)$ . Again, let  $A = \{\omega \in \Omega: P(B|\mathcal{C})(\omega) \leq 0\}$  and assume  $P(A) > 0$ . Now,  $P(B|\mathcal{C})(\omega) \leq 0$  for all  $\omega \in A$  implies  $P(B|\mathcal{C}) \cdot 1_A \leq 0$ . Therefore,

$$\begin{aligned}
 0 &\leq E[P(B|\mathcal{C}) \cdot 1_A] && [(3.50)] \\
 &= E[P(B \cdot 1_A|\mathcal{C})] && [A \in \mathcal{C}, \text{ Box 10.1 (xiv)}] \\
 &= E(1_B \cdot 1_A) && [10.1 (iv)] \\
 &= E(1_{A \cap B}) && [(1.33)]. \\
 &= P(A \cap B). && [(6.4)]
 \end{aligned}$$

Because  $P(B) > 0$ , the equation  $P(A \cap B) = P^B(A) \cdot P(B) = 0$  implies  $P^B(A) = 0$  [see (4.14)]. Hence,  $A \in \mathcal{C}$  with  $P^B(A) = P(A \cap B)/P(B) = 0$  and  $P(A) > 0$ .

(b)  $\Rightarrow$  (d). As has been shown above, (b) is equivalent to  $P$ -uniqueness of  $E^B(Y|\mathcal{C})$ , and according to Box 6.1 (ix),  $P$ -uniqueness of  $E^B(Y|\mathcal{C})$  implies (d).

(d)  $\Rightarrow$  (b). This proposition is proved by contraposition, i. e., we show  $\neg(b) \Rightarrow \neg(d)$ . Assume that there is an  $A \in \mathcal{C}$  with  $P^B(A) = 0$  and  $P(A) > 0$ , and let  $V \in \mathcal{E}^B(Y|\mathcal{C})$  be real-valued [see Box 10.1 (x)] with a finite expectation  $E(V)$ . Then  $V^* = V + 1_A$  is also  $\mathcal{C}$ -measurable and, for all  $C \in \mathcal{C}$ ,

$$\int 1_C V^* dP^B = \int 1_C V dP^B. \tag{14.83}$$

Therefore, according to Definition 14.7,  $V^* = V + 1_A \in \mathcal{E}^B(Y|\mathcal{C})$  and  $P(\{V \neq V^*\}) = P(A) > 0$ . Now,

$$\begin{aligned}
 E(V^*) &= \int V^* dP && [(6.1)] \\
 &= \int (V + 1_A) dP && [V^* := V + 1_A] \\
 &= \int V dP + \int 1_A dP && [(3.33)] \\
 &= E(V) + P(A). && [(6.1), (6.4)]
 \end{aligned}$$

Because  $E(V)$  is finite and  $P(A) > 0$  it follows that  $E(V) \neq E(V^*)$ .

**Proof of Corollary 14.58**

If  $E^{Z=z}(Y|X)$  is  $P^C$ -unique, then all pairs of elements of  $\mathcal{E}^{Z=z}(Y|X)$  are  $P^C$ -equivalent. According to Theorem 14.33,  $E(Y|X, Z=z)$  is an element of  $\mathcal{E}^{Z=z}(Y|X)$ , which implies Equation (14.58). Finally, according to Corollary 5.25, Equation (14.59) is equivalent to (14.58).

**Proof of Theorem 14.64**

First, we show that  $E(Y|\mathcal{C}, Z) \in \mathcal{E}^{Z=z}(Y|\mathcal{C}, Z)$ . By Definition 10.2 (a),  $\sigma[E(Y|\mathcal{C}, Z)] \subset \sigma(\mathcal{C}, Z)$ , which implies that condition (a) of Definition 14.7 holds. In order to show condition (b) of Definition 14.7 note that, for all  $C \in \sigma(\mathcal{C}, Z)$ ,  $\{Z=z\}, \{Z=z\} \cap C \in \sigma(\mathcal{C}, Z)$  and

$1_{\{Z=z\} \cap C} = 1_{Z=z} 1_C$ . Hence, for all  $C \in \sigma(\mathcal{C}, Z)$ ,

$$\begin{aligned} E^{Z=z}[1_C E(Y|\mathcal{C}, Z)] &= \frac{1}{P(Z=z)} E[1_{Z=z} 1_C \cdot E(Y|\mathcal{C}, Z)] && [(9.11)] \\ &= \frac{1}{P(Z=z)} E(1_{Z=z} 1_C \cdot Y) && [\text{Def. 10.2 (b)}] \\ &= E^{Z=z}(1_C \cdot Y). && [(9.11)] \end{aligned}$$

Now we show that  $E^{Z=z}(Y|\mathcal{C}) \in \mathcal{E}^{Z=z}(Y|\mathcal{C}, Z)$ . By Definition 14.7 (a) and the definition of  $\sigma(\mathcal{C}, Z)$ ,

$$\sigma[E^{Z=z}(Y|\mathcal{C})] \subset \mathcal{C} \subset \sigma(\mathcal{C}, Z),$$

and hence,  $\sigma[E^{Z=z}(Y|\mathcal{C}, Z)] \subset \sigma(\mathcal{C}, Z)$ .

Now we show that condition (b) of Definition 14.7 holds. Note that  $\sigma(Z) \upharpoonright_{\{Z=z\}} = \{\emptyset, \{Z=z\}\}$  (see the definition of the trace of a set system in Example 1.10). Therefore,

$$\begin{aligned} \mathcal{C} \cup \sigma(Z) \upharpoonright_{\{Z=z\}} &= (\mathcal{C} \upharpoonright_{\{Z=z\}}) \cup (\sigma(Z) \upharpoonright_{\{Z=z\}}) && [\text{Example 1.10}] \\ &= (\mathcal{C} \upharpoonright_{\{Z=z\}}) \cup \{\emptyset, \{Z=z\}\} && [(14.85)] \\ &= \mathcal{C} \upharpoonright_{\{Z=z\}}. && [\emptyset \in \mathcal{C}, \Omega \cup \{Z=z\} \in \mathcal{C} \upharpoonright_{\{Z=z\}}] \end{aligned}$$

Now

$$\begin{aligned} \sigma(\mathcal{C}, Z) \upharpoonright_{\{Z=z\}} &= \sigma[(\mathcal{C} \cup \sigma(Z)) \upharpoonright_{\{Z=z\}}] && [(14.69)] \\ &= \sigma(\mathcal{C} \upharpoonright_{\{Z=z\}}) && [(14.85)] \\ &= \mathcal{C} \upharpoonright_{\{Z=z\}}. && [(1.15)] \end{aligned}$$

Hence, for all  $C \in \sigma(\mathcal{C}, Z)$ , there is an  $A_C \in \mathcal{C}$  such that  $\{Z=z\} \cap C = \{Z=z\} \cap A_C$ , which implies

$$1_{Z=z} 1_C = 1_{Z=z} 1_{A_C}. \quad (14.86)$$

Therefore, for all  $C \in \sigma(\mathcal{C}, Z)$ ,

$$\begin{aligned} E^{Z=z}[1_C \cdot E^{Z=z}(Y|\mathcal{C})] &= \frac{1}{P(Z=z)} E[1_{Z=z} 1_C \cdot E^{Z=z}(Y|\mathcal{C})] && [(9.11)] \\ &= \frac{1}{P(Z=z)} E[1_{Z=z} 1_{A_C} \cdot E^{Z=z}(Y|\mathcal{C})] && [(14.86)] \\ &= E^{Z=z}[1_{A_C} \cdot E^{Z=z}(Y|\mathcal{C})] && [(9.11)] \\ &= E^{Z=z}(1_{A_C} \cdot Y) && [\text{Def. 14.7 (b)}] \\ &= \frac{1}{P(Z=z)} E(1_{Z=z} 1_{A_C} \cdot Y) && [(9.11)] \\ &= \frac{1}{P(Z=z)} E(1_{Z=z} 1_C \cdot Y) && [(14.86)] \\ &= E^{Z=z}(1_C \cdot Y). && [(9.11)] \end{aligned}$$

**Proof of Theorem 14.67**

Let  $g_{Z=z}(X, W) = E^{Z=z}(Y | X, W)$  [see (14.34)]. Then

$$g_{Z=z}(X, W) \cdot 1_{W=w} = g_{Z=z}(X, w) \cdot 1_{W=w}. \quad (14.87)$$

Furthermore,

$$E^{W=w}[E^{Z=z}(Y | X, W)] = \frac{1}{P(W=w)} \int E^{Z=z}(Y | X, W) \cdot 1_{W=w} dP \quad [(9.11)]$$

$$= \frac{1}{P(W=w)} \int E^{Z=z}(Y | X, W=w) \cdot 1_{W=w} dP \quad [(14.87)]$$

$$= \int E^{Z=z}(Y | X, W=w) dP^{W=w} \quad [(9.11)]$$

$$= \int E^{Z=z}(Y | X=x, W=w) P_X^{W=w}(dx), \quad [(6.13)]$$

which proves the first equation of (14.78).

Now let  $g(X, W, Z) = E(Y | X, W, Z)$ . Then, for  $g_z(X, W)$  [see (14.23)],

$$g_z(X, W) \cdot 1_{W=w} = g_z(X, w) \cdot 1_{W=w}. \quad (14.88)$$

Furthermore, because  $E^{Z=z}(Y | X, W)$  is  $P^{W=w}$ -unique,

$$E^{W=w}[E^{Z=z}(Y | X, W)] = E^{W=w}[E(Y | X, W, Z=z)] \quad [(14.58), \text{Box 6.1 (ix)}]$$

$$= \frac{1}{P(W=w)} \int E(Y | X, W, Z=z) \cdot 1_{W=w} dP \quad [(9.11)]$$

$$= \frac{1}{P(W=w)} \int E(Y | X, W=w, Z=z) \cdot 1_{W=w} dP \quad [(14.88)]$$

$$= \int E(Y | X, W=w, Z=z) dP^{W=w} \quad [(9.11)]$$

$$= \int E(Y | X=x, W=w, Z=z) P_X^{W=w}(dx). \quad [(6.13)]$$

This proves the second equation of (14.78).

**Proof of Theorem 14.69**

Note that  $E^{Z=z}(Y | \mathcal{C})$  is  $\mathcal{C}$ -measurable and therefore also  $\sigma(\mathcal{C}, \mathcal{D})$ -measurable [see Eq. (14.69)].

Furthermore,

$$E^{Z=z}(Y | \mathcal{C}, \mathcal{D}) \stackrel{p^{Z=z}}{=} E(Y | \mathcal{C}, \mathcal{D}, Z) \quad [\text{Eqs. (14.72)}]$$

$$\stackrel{p^{Z=z}}{=} E(Y | \mathcal{C}, Z) \quad [E(Y | \mathcal{C}, \mathcal{D}, Z) \stackrel{p}{=} E(Y | \mathcal{C}, Z), \text{Cor. 5.22}]$$

$$\stackrel{p^{Z=z}}{=} E^{Z=z}(Y | \mathcal{C}). \quad [\text{Eqs. (14.72)}]$$

## 14.9 Exercises

- ▷ **Exercise 14-1** Consider Table 14.3. Why do the values of the conditional expectation  $E(Y|X, U)$  have to be identical for  $\omega_3 = (\text{Joe}, \text{yes}, -)$  and  $\omega_4 = (\text{Joe}, \text{yes}, +)$ ?
- ▷ **Exercise 14-2** Show that if  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  is a random variable with finite expectation  $E(Y)$ , then  $E^B(Y)$  is finite.
- ▷ **Exercise 14-3** Let  $(\Omega, \mathcal{A}, P), (\Omega, \mathcal{A}, Q)$  be probability spaces, and consider the random variable  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$ . Show that  $X$  is a random variable on  $(\Omega, \mathcal{A}, P)$  if and only if it is also a random variable on  $(\Omega, \mathcal{A}, Q)$ .
- ▷ **Exercise 14-4** Show that the assumptions 14.5 imply  $\mathcal{E}^B(1_B \cdot Y|\mathcal{C}) = \mathcal{E}^B(Y|\mathcal{C})$ .
- ▷ **Exercise 14-5** Prove Equation (14.26).
- ▷ **Exercise 14-6** Prove Equation (14.27).
- ▷ **Exercise 14-7** Compute the values of the conditional expectation  $E^{X=0}(Y|U)$  in the example presented in Table 14.3 (p. 403).
- ▷ **Exercise 14-8** Prove propositions (i) and (ii) of Remark 14.35.
- ▷ **Exercise 14-9** Show that Theorem 14.33 implies Equation (14.35).
- ▷ **Exercise 14-10** Prove: For all  $x \in X(\Omega) = \{0, 1\}$ : If  $P(X=x) > 0$  and  $E^{X=0}(Y|Z)$  is  $P^{X=x}$ -unique, then  $E(V_0|X=x) = E(V_0^*|X=x)$  for all versions  $V_0, V_0^* \in \mathcal{E}^{X=0}(Y|Z)$ .
- ▷ **Exercise 14-11** Show that the expectation  $E(V)$  of a version  $V \in \mathcal{E}^B(Y|\mathcal{C})$  exists and is finite if  $\mathcal{C}$  is a finite set and  $E^B(Y)$  is finite.
- ▷ **Exercise 14-12** Prove Remark 14.49 for  $X$  being discrete.
- ▷ **Exercise 14-13** Prove the propositions summarized in Box 14.1.
- ▷ **Exercise 14-14** Prove Equation (14.76).
- ▷ **Exercise 14-15** Prove Equation (14.77).

## Solutions

- ▷ **Solution 14-1** If  $E(Y|X, U)(\omega_3) \neq E(Y|X, U)(\omega_4)$ , then  $E(Y|X, U)$  would not be measurable with respect to  $(X, U)$ . Measurability of a random variable  $V$  with respect to  $(X, U)$  requires that  $V$  takes on only one single value for all  $\omega \in (X, U)^{-1}(\{(1, \text{Joe})\})$  (see Cor. 2.53).
- ▷ **Solution 14-2** According to Equation (9.11),

$$E^B(Y) = \frac{1}{P(B)} E(1_B Y).$$

If  $Y$  has a finite expectation with respect to  $P$ , then  $\int Y^+ dP < \infty$  and  $\int Y^- dP < \infty$ . Because  $0 \leq 1_B Y^+ \leq Y^+$  and  $0 \leq 1_B Y^- \leq Y^-$ , Equation (3.24) yields  $\int 1_B Y^+ dP < \infty$  and  $\int 1_B Y^- dP < \infty$ , and therefore  $-\infty < \int 1_B Y dP < \infty$ . Therefore, if  $-\infty < E(Y) < \infty$ , then  $-\infty < E(1_B Y) < \infty$ .

▷ **Solution 14-3** Definition 5.1 of a random variable  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  only requires that there is a probability measure, here denoted  $P$ , on a measurable space  $(\Omega, \mathcal{A})$  and that  $X$  is measurable with respect to  $\mathcal{A}$ , i. e.,  $X^{-1}(A') \in \mathcal{A}$ , for all  $A' \in \mathcal{A}'_X$ . Hence, if  $X$  is a random variable on  $(\Omega, \mathcal{A}, P)$ , then it is also a random variable on  $(\Omega, \mathcal{A}, Q)$ , whenever  $P$  and  $Q$  are probability measures on  $(\Omega, \mathcal{A})$ .

▷ **Solution 14-4** If  $V \in \mathcal{E}^B(1_B \cdot Y | \mathcal{C})$ , then, according to Definition 14.7 (a), it is measurable with respect to  $\mathcal{C}$ . Furthermore, for all  $C \in \mathcal{C}$ ,

$$\begin{aligned} E^B(1_C \cdot V) &= E^B(1_C \cdot 1_B \cdot Y) && \text{[Def. 14.7 (b)]} \\ &= \frac{1}{P(B)} \cdot E(1_C \cdot 1_B \cdot Y) && \text{[(9.7)]} \\ &= \frac{1}{P(B)} \cdot E(1_C \cdot 1_B \cdot Y) && [1_B \cdot 1_B = 1_B] \\ &= E^B(1_C \cdot Y). && \text{[(9.7)]} \end{aligned}$$

Hence,  $V \in \mathcal{E}^B(Y | \mathcal{C})$ . Vice versa, if  $V^* \in \mathcal{E}^B(Y | \mathcal{C})$ , then it is measurable with respect to  $\mathcal{C}$  and, for all  $C \in \mathcal{C}$ ,

$$\begin{aligned} E^B(1_C \cdot V^*) &= E^B(1_C \cdot Y) && \text{[Def. 14.7 (b)]} \\ &= \frac{1}{P(B)} \cdot E(1_C \cdot 1_B \cdot Y) && \text{[(9.7)]} \\ &= \frac{1}{P(B)} \cdot E(1_C \cdot 1_B \cdot 1_B \cdot Y) && [1_B \cdot 1_B = 1_B] \\ &= E^B(1_C \cdot 1_B \cdot Y). && \text{[(9.7)]} \end{aligned}$$

Hence,  $V^* \in \mathcal{E}^B(1_B \cdot Y | \mathcal{C})$ .

▷ **Solution 14-5** For all  $\omega \in \Omega$ ,

$$\begin{aligned} 1_{Z=z}(\omega) \cdot g[X(\omega), Z(\omega)] &= \begin{cases} 0, & \text{if } 1_{Z=z}(\omega) = 0 \\ g_z[X(\omega)], & \text{if } 1_{Z=z}(\omega) = 1 \end{cases} \\ &= \begin{cases} 1_{Z=z}(\omega) \cdot g_z[X(\omega)], & \text{if } 1_{Z=z}(\omega) = 0 \\ 1_{Z=z}(\omega) \cdot g_z[X(\omega)], & \text{if } 1_{Z=z}(\omega) = 1 \end{cases} \\ &= 1_{Z=z}(\omega) \cdot g_z[X(\omega)]. \end{aligned}$$

▷ **Solution 14-6**

$$\begin{aligned} \forall \omega \in \Omega: & \left( \sum_{z \in Z(\Omega)} E(Y | X, Z=z) \cdot 1_{Z=z} \right) (\omega) \\ &= \left( \sum_{z \in Z(\Omega)} g_z(X) \cdot 1_{Z=z} \right) (\omega) && \text{[(14.25)]} \\ &= \left( \sum_{z \in Z(\Omega)} g(X, Z) \cdot 1_{Z=z} \right) (\omega) && \text{[(14.26)]} \\ &= \sum_{z \in Z(\Omega)} g(X, Z)(\omega) \cdot 1_{Z=z}(\omega) && \text{[(2.31)]} \\ &= g(X, Z)(\omega) && \text{[(2.26)]} \\ &= E(Y | X, Z)(\omega). && \text{[Def. 14.29]} \end{aligned}$$

▷ **Solution 14-7** The values of the conditional expectation  $E^{X=0}(Y|U)$  are the two conditional expectation values  $E^{X=0}(Y|U=Joe)$  and  $E^{X=0}(Y|U=Ann)$  with respect to the measure  $P^{X=0}$ . Because  $E^{X=0}(Y|U=u) = P^{X=0}(Y=1|U=u)$ , they can be computed as follows:

$$P^{X=0}(Y=1|U=Joe) = \frac{P^{X=0}(Y=1, U=Joe)}{P^{X=0}(U=Joe)} \approx \frac{.561}{.561 + .245} \approx .696$$

and

$$P^{X=0}(Y=1|U=Ann) = \frac{P^{X=0}(Y=1, U=Ann)}{P^{X=0}(U=Ann)} \approx \frac{.039}{.039 + .155} \approx .2.$$

▷ **Solution 14-8** (i)

If  $g(X, Z) \in \mathcal{E}(Y|X, Z)$  and  $g_{Z=z}(X) \in \mathcal{E}^{Z=z}(Y|X=x)$ , then for all  $C \in \sigma(X)$ ,

$$\begin{aligned} & \int 1_C \cdot g_z(X) dP^{Z=z} \\ &= \frac{1}{P(Z=z)} \int 1_C \cdot 1_{Z=z} g_z(X) dP && [(9.11)] \\ &= \frac{1}{P(Z=z)} \int 1_C \cdot 1_{Z=z} g(X, Z) dP && [(14.26)] \\ &= \frac{1}{P(Z=z)} \int 1_C \cdot 1_{Z=z} Y dP && [C \cap \{Z=z\} \in \sigma(X, Z), \text{Def. 10.2 (b)}] \\ &= \int 1_C Y dP^{Z=z} && [(9.11)] \\ &= \int 1_C g_{Z=z}(X) dP^{Z=z}. && [\text{Def. 14.7 (b)}] \end{aligned}$$

Hence,  $g_z(X) \stackrel{P^{Z=z}}{=} g_{Z=z}(X)$  (see Th. 3.48) or, equivalently,  $g_z(x) = g_{Z=z}(x)$ , for  $P_X^{Z=z}$ -a.a.  $x \in \Omega'_X$  [see Cor. 5.25 (i) and (5.12)].

(ii) If  $P(X=x, Z=z) > 0$  for all  $x \in \Omega'_X$ , then  $P(Z=z) > 0$  [see Box 4.1 (v)] and

$$P_X^{Z=z}(\{x\}) = P^{Z=z}(X=x) = \frac{P(X=x, Z=z)}{P(Z=z)} > 0, \quad \forall x \in \Omega'_X.$$

Hence, if  $P(X=x, Z=z) > 0$ , then Remark 2.71 and  $g_z(X) \stackrel{P^{Z=z}}{=} g_{Z=z}(X)$  imply  $g_z(x) = g_{Z=z}(x)$  for all  $x \in \Omega'_X$ .

▷ **Solution 14-9**

$$\begin{aligned} E(Y|X, Z=z) & \stackrel{P^{Z=z}}{=} E^{Z=z}(Y|X) && [(14.30)] \\ & \stackrel{P^{Z=z}}{=} E^{1_{Z=z}=1}(Y|X) && [\{Z=z\} = \{1_{Z=z}=1\}] \\ & \stackrel{P^{Z=z}}{=} E(Y|X, 1_{Z=z}=1). && [(14.30)] \end{aligned}$$

For  $E(Y|X, Z) = g(X, Z) \in \mathcal{E}(Y|X, Z)$  and  $P_X^{Z=z}$ -a.a.  $x \in \Omega'_X$ ,

$$\begin{aligned} E(Y|X=x, Z=z) &= g(x, z) && [(10.27)] \\ &= g_z(x) && [(14.23)] \\ &= g_{Z=z}(x) && [\text{Rem. 14.35 (i)}] \\ &= E^{Z=z}(Y|X=x). && [(14.34)] \end{aligned}$$

▷ **Solution 14-10** This proposition follows from the definition of uniqueness of a conditional expectation with respect to a probability measure  $P^{X=x}$ , equivalence of two random variables with respect to a probability measure  $P^{X=x}$ , Corollaries 5.24 and 6.17, and Equations (9.5) and (9.6) with  $B = \{X=x\}$ .

▷ **Solution 14-11** If  $\mathcal{C} = \{A_1, \dots, A_n\}$  is finite, then there is a finite partition  $\{B_1, \dots, B_m\}$  of  $\Omega$  with  $\mathcal{C} = \sigma(\{B_1, \dots, B_m\})$  (see Rem. 1.21). Then, according to Lemma 2.19 and Definition 14.7 (b),

$$V = \sum_{j=1}^m \alpha_j 1_{B_j}, \quad \text{where } \begin{cases} \alpha_j = \int 1_{B_j} \cdot Y \, dP^B, & \text{if } P^B(B_j) > 0, \\ \text{any } \alpha_j \in \mathbb{R}, & \text{if } P^B(B_j) = 0, \end{cases}$$

is a version  $V \in \mathcal{E}^B(Y|\mathcal{C})$ . Hence, if  $E^B(Y) = \int Y \, dP^B$  is finite, then, for any such choice of the  $\alpha_j$ ,

$$E(V) = \sum_{j=1}^m \alpha_j \cdot P(B_j)$$

[see Eq. (6.3)] is finite as well.

▷ **Solution 14-12**

(d)  $\Rightarrow$  (14.48). By definition, condition (d) of Corollary 14.48 is equivalent to

$$\forall A' \in \mathcal{A}'_X: P_X^B(A') = 0 \Rightarrow P_X(A') = 0.$$

If  $X$  is discrete and  $\{x\} \in \mathcal{A}'_X$  for all  $x \in \Omega'_X$ , then we can choose  $A' = \{x\}$  for all  $x \in \Omega'_X$ , and this yields

$$\forall x \in \Omega'_X: P_X^B(\{x\}) = 0 \Rightarrow P_X(\{x\}) = 0,$$

which is (14.48) in a different notation [see Eqs. (5.2) and (5.4)].

(14.48)  $\Rightarrow$  (d). Assume that (14.48) holds and let  $A' \in \mathcal{A}'_X$  with  $P_X^B(A') = 0$ . Then

$$\begin{aligned} P_X^B(A') &= \sum_{\substack{x \in A' \\ P(X=x) > 0}} P_X^B(\{x\}) && \text{[Th. 4.28, Box 4.1 (i)]} \\ &= \sum_{\substack{x \in A' \\ P(X=x) > 0}} P^B(X=x) && \text{[(5.2), (5.4)]} \\ &= 0. \end{aligned}$$

Because a sum of nonnegative summands can only be 0 if and only if all summands are 0, the last equation implies  $P^B(X=x) = 0$  for all  $x \in A'$ . Now (14.48) yields  $P(X=x) = 0$  for all  $x \in A'$ . Therefore,

$$P(A') = \sum_{\substack{x \in A' \\ P(X=x) > 0}} P(\{x\}) = \sum_{\substack{x \in A' \\ P(X=x) > 0}} P(X=x) = 0$$

[see again Eqs. (5.2) and (5.4)].

- ▷ **Solution 14-13** (i). This is the definition of  $P$ -uniqueness of  $E^B(Y|\mathcal{C})$  (see Rem. 10.13).  
(ii), (iii). These propositions have been proved in Theorem 14.46.  
(iv). This proposition follows from Theorem 14.46, because  $P \ll_{\mathcal{C}} P^B$  is equivalent to

$$\forall C \in \mathcal{C}: P^B(C) = 0 \Rightarrow P(C) = 0,$$

and  $\mathcal{D} \subset \mathcal{C}$  implies that this implication also holds for all  $C \in \mathcal{D}$ , i. e.,  $P \ll_{\mathcal{D}} P^B$ .

(v). This proposition immediately follows from applying Corollary 5.22 to  $E^B(Y|\mathcal{C})$ .

(vi), (vii). According to (v),  $P$ -uniqueness of  $E^B(Y|\mathcal{C})$  implies  $V \stackrel{P}{=} V^*$ . Theorem 2.85 then implies  $P_V^C = P_{V^*}^C$ . If the two distributions are identical, then the corresponding expectations are identical as well (see Cor. 6.17).

(viii). This proposition immediately follows from Remark 2.76 (iii).

(ix). Note that  $P^B(X=x) > 0$ , for all  $x \in \Omega'_X$ , implies  $X(\Omega) = \Omega'_X$ . Therefore, and because uniqueness of  $E^B(Y|X)$  implies that  $E^B(Y|X)$  is  $P$ -unique, the proposition is an immediate implication of Corollary 14.42.

▷ **Solution 14-14** If  $X$  is measurable with respect to  $Z$ , then

$$\begin{aligned} E^{Z=z}(Y|\mathcal{C}, X) &\stackrel{P^{Z=z}}{=} E^{Z=z}[E^{Z=z}(Y|\mathcal{C}, Z)|\mathcal{C}, X] && \text{[Box 10.1 (v)]} \\ &\stackrel{P^{Z=z}}{=} E^{Z=z}[E^{Z=z}(Y|\mathcal{C})|\mathcal{C}, X] && \text{[(14.72), Box 10.1 (ix)]} \\ &\stackrel{P^{Z=z}}{=} E^{Z=z}(Y|\mathcal{C}). && \text{[Box 10.1 (vii)]} \end{aligned}$$

▷ **Solution 14-15** If  $\mathcal{D} \subset \sigma(\mathcal{C}, Z)$ , then

$$\begin{aligned} E^{Z=z}(Y|\mathcal{D}) &\stackrel{P^{Z=z}}{=} E^{Z=z}[E^{Z=z}(Y|\mathcal{C}, Z)|\mathcal{D}] && \text{[Box 10.1 (v)]} \\ &\stackrel{P^{Z=z}}{=} E^{Z=z}[E(Y|\mathcal{C}, Z)|\mathcal{D}]. && \text{[(14.72), Box 10.1 (ix)]} \end{aligned}$$



## Chapter 15

# Conditional Effect Functions of a Discrete Regressor

In chapter 14 we treated  $E^{X=x}(Y|Z)$ , the  $Z$ -conditional expectation of  $Y$  with respect to the  $(X=x)$ -conditional probability measure  $P^{X=x}$ . There we already noted that, if the values of  $X$  represent treatment conditions, then  $E^{X=x}(Y|Z)$  refers to the  $Z$ -conditional expectation of  $Y$  given treatment  $x$ . If  $X$  is dichotomous with values 0 and 1, then the values  $g_1(z)$  of the function  $g_1(Z) := E^{X=1}(Y|Z) - E^{X=0}(Y|Z)$  are the effects of  $X$  on  $Y$  given the value  $z$  of  $Z$ , and  $g_1$  is called the *Z*-conditional effect function of  $X$ .

From a methodological point of view, the values  $g_1(z)$  of the  $Z$ -conditional effect function of  $X$  are of interest for at least two reasons. *First*, the conditional effect  $g_1(z)$  describes the effect of  $X$  on  $Y$  for a fixed value  $z$  of  $Z$ . Thus the impact of  $Z$  on  $X$  and  $Y$  is controlled by keeping  $Z$  constant on one of its values. *Second*, the conditional effect  $g_1(z)$  is *more specific* and therefore more informative than the unconditional effect  $E(Y|X=1) - E(Y|X=0)$ . Knowing such conditional effects, we can choose individualized treatments.

In this chapter we introduce the concepts of conditional intercept functions and conditional effect functions and consider these functions for the parametrizations of the conditional expectation  $E(Y|X, Z)$  that have been treated in chapters 12 and 13.

### 15.1 Assumptions and Definitions

In section 14.1 we treated three examples that motivated introducing the conditional expectations  $E^{X=x}(Y|Z)$ . These examples, in which the person variable  $U$  takes the role of  $Z$ , also motivate the present chapter on conditional effect functions. In Examples 14.1 and 14.2, the conditional effect function  $g_1$  and each of its two values  $g_1(\text{Joe})$  and  $g_1(\text{Ann})$  are uniquely defined, whereas in Example 14.3 this is not the case. While in the latter example the value  $g_1(\text{Ann})$  is uniquely defined, the value  $g_1(\text{Joe})$  is not, and we can choose any real number as the value  $g_1(\text{Joe}) = g_1(U)(\omega)$  if  $U(\omega) = \text{Joe}$ , and still  $g_0(U)$  and  $g_1(U)$  satisfy

$$E(Y|X, U) \stackrel{\bar{p}}{=} g_0(U) + g_1(U) \cdot X, \quad (15.1)$$

for all versions  $E(Y|X, U) \in \mathcal{E}(Y|X, U)$ . In other words, the function  $g_1(U)$  specified in Example 14.3 is only one out of infinitely many versions of such a conditional effect function satisfying Equation (15.1), and even the expectations

$E[g_1(U)]$  are not necessarily identical. Therefore, we have to introduce an assumption that guarantees that the values of the functions  $g_0(U)$  and  $g_1(U)$  are not arbitrary and that the expectations  $E[g_0(U)]$  and  $E[g_1(U)]$  are uniquely defined. Instead of the person variable  $U$  used in the examples with Joe and Ann, now we choose  $Z$  as a random variable with respect to which we consider conditional intercept and effect functions.

Throughout this chapter we refer to the following assumptions and notation.

### Notation and Assumptions 15.1

Let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$ ,  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$ , and  $Z: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Z, \mathcal{A}'_Z)$  be random variables, where  $E(Y)$  is finite. Furthermore, assume that  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  is discrete with  $P(X \in \{x_0, x_1, \dots, x_n\}) = 1$  and  $P(X = x_i) > 0$ , for all  $i = 0, 1, \dots, n$ .

**Remark 15.2 (An Additional Assumption)** The following additional assumption is often used in this chapter:

$$P(X = x_i | Z) \gtrsim 0, \quad \forall i = 0, 1, \dots, n. \quad (15.2)$$

According to Corollary 14.48, this assumption is equivalent to each of the following conditions:

$$P\text{-uniqueness of the conditional expectations } E^{X=x_i}(Y|Z), \quad (15.3)$$

$$P \ll_{\sigma(Z)} P^{X=x_i}, \quad \forall i = 0, 1, \dots, n, \quad (\text{absolute continuity}) \quad (15.4)$$

$$E(V_i) = E(V_i^*), \quad \forall V_i, V_i^* \in \mathcal{E}^{X=x_i}(Y|Z), \quad \forall i = 0, 1, \dots, n. \quad (15.5)$$

Also remember,  $P$ -uniqueness of the conditional expectations  $E^{X=x_i}(Y|Z)$  implies that different versions of  $E^{X=x_i}(Y|Z)$  do not only have identical distributions with respect to the conditional-probability measure  $P^{X=x_i}$  [see Def. 4.29 and (9.4)] but also with respect to  $P$ .  $\triangleleft$

## 15.2 Conditional Intercept Function and Effect Functions

### Theorem 15.3 (Existence of Conditional Intercept and Effect Functions)

Let the assumptions 15.1 hold. Then there are an  $E(Y|X, Z) \in \mathcal{E}(Y|X, Z)$  and, for all  $i = 0, 1, \dots, n$ , real-valued  $E^{X=x_i}(Y|Z) \in \mathcal{E}^{X=x_i}(Y|Z)$  such that

$$E(Y|X, Z) = g_0(Z) + \sum_{i=1}^n g_i(Z) \cdot 1_{X=x_i}, \quad (15.6)$$

with

$$g_0(Z) := E^{X=x_0}(Y|Z), \quad (15.7)$$

and

$$g_i(Z) := E^{X=x_i}(Y|Z) - E^{X=x_0}(Y|Z), \quad \forall i = 1, \dots, n. \quad (15.8)$$

(Proof p. 452)

Referring to Equation (15.6), now we can define the conditional intercept function and the conditional effect functions as follows.

**Definition 15.4 (Conditional Intercept Function and Effect Functions)**

Let the assumptions 15.1 hold as well as (15.2). Then the function  $g_0: (\Omega'_Z, \mathcal{A}'_Z) \rightarrow (\mathbb{R}, \mathcal{B})$  is called the  $Z$ -conditional intercept function, and, for all  $i = 1, \dots, n$ , the function  $g_i: (\Omega'_Z, \mathcal{A}'_Z) \rightarrow (\mathbb{R}, \mathcal{B})$ , the  $Z$ -conditional effect function of  $x_i$  vs.  $x_0$  on  $Y$  pertaining to the version  $E(Y|X, Z) \in \mathcal{E}(Y|X, Z)$  in Equation (15.6).

**Remark 15.5 (The Functions  $g_i$  vs. the Functions  $g_i(Z)$ )** Note that the functions  $g_i(Z)$ ,  $i = 0, 1, \dots, n$ , denote the compositions of  $Z$  and  $g_i$ . Because  $Z$  is a random variable on  $(\Omega, \mathcal{A}, P)$ , the compositions  $g_i(Z)$  are random variables on  $(\Omega, \mathcal{A}, P)$  as well. In contrast, the functions  $g_i: (\Omega'_Z, \mathcal{A}'_Z) \rightarrow (\mathbb{R}, \mathcal{B})$ ,  $i = 0, 1, \dots, n$ , are not random variables on  $(\Omega, \mathcal{A}, P)$ . However, they are random variables on the probability space  $(\Omega'_Z, \mathcal{A}'_Z, P_Z)$ , where  $P_Z$  denotes the distribution of  $Z$ .  $\triangleleft$

**Remark 15.6 ( $P$ -Uniqueness of the Conditional Intercept and Effect Functions)**

Suppose that the assumptions of Definition 15.4 hold. Then, (15.3) and (2.36) imply that all measurable functions  $g_0^*, g_1^*, \dots, g_n^*: (\Omega'_Z, \mathcal{A}'_Z) \rightarrow (\mathbb{R}, \mathcal{B})$  with

$$\left( g_0^*(Z) + \sum_{i=1}^n g_i^*(Z) \cdot 1_{X=x_i} \right) \in \mathcal{E}(Y|X, Z)$$

are  $P_Z$ -unique, and according to Corollary 5.25 (i), this implies

$$g_i^*(Z) \stackrel{p}{=} g_i(Z), \quad \forall i = 0, 1, \dots, n. \quad (15.9)$$

Hence, under the assumptions mentioned above, the compositions of the conditional intercept and effect functions are  $P$ -unique. This in turn implies that the expectations and the variances of the conditional intercept and effect functions are uniquely defined, provided that they exist.  $\triangleleft$

**Remark 15.7 (Partial Conditional Expectation  $E(Y|X, Z=z)$ )** Let the assumptions of Definition 15.4 hold. Then Equation (15.6) implies

$$E(Y|X, Z=z) = g_0(z) + \sum_{i=1}^n g_i(z) \cdot 1_{X=x_i} \quad (15.10)$$

for the partial conditional expectation  $E(Y|X, Z=z)$  (see Def. 14.29). This equation justifies the terminology introduced in Definition 15.4.  $\triangleleft$

**Remark 15.8 (( $Z=z$ )-Conditional Intercept and Effects)** Let the assumptions of Definition 15.4 hold and assume  $z \in \Omega'_Z$  with  $P(Z=z) > 0$ . Then Equation (15.10) and (14.29) imply that there is an  $E^{Z=z}(Y|X) \in \mathcal{E}^{Z=z}(Y|X)$  with

$$E^{Z=z}(Y|X) = E(Y|X, Z=z) = g_0(z) + \sum_{i=1}^n g_i(z) \cdot 1_{X=x_i}. \quad (15.11)$$

If  $P(Z=z) > 0$ , then, according to Remark 2.71, the coefficients  $g_i(z)$ ,  $i = 0, 1, \dots, n$ , are uniquely determined. The number  $g_0(z)$  is called the ( $Z=z$ )-*conditional intercept* and  $g_i(z)$  the ( $Z=z$ )-*conditional effect* of  $x_i$  vs.  $x_0$  on  $Y$ , where  $i = 1, \dots, n$ . Equation (15.11) and Theorem 12.37 imply that  $g_0(z)$  is the intercept and  $g_i(z)$ ,  $i = 1, \dots, n$ , are the regression coefficients pertaining to a linear parametrization in  $1_{X=x_1}, \dots, 1_{X=x_n}$  of the  $X$ -conditional expectation of  $Y$  with respect to the measure  $P^{Z=z}$ .

Furthermore, if  $Y$  has finite second moments with respect to  $P^{Z=z}$  and the matrix of the covariances of the indicators  $1_{X=x_1}, \dots, 1_{X=x_n}$  with respect to  $P^{Z=z}$  is regular, then Corollary 12.31 can be applied. Note that regularity of the matrix of the covariances of indicators  $1_{X=x_1}, \dots, 1_{X=x_n}$  with respect to  $P^{Z=z}$  holds, if (15.2) and  $P(Z=z) > 0$ , because then  $P^{Z=z}(X=x_i) > 0$  for all  $i = 1, \dots, n$  (see Lemma 12.38 and Rem. 12.39).

Also note that the prerequisite of Corollary 12.31 that  $E(Y^2)$  is finite can be neglected because finiteness of  $E(Y)$  already implies that  $E(Y \cdot 1_{X=x_i})$  is finite [see Lemma 3.33 (ii)] and that the covariance vector of  $(1_{X=x_1}, \dots, 1_{X=x_n})$  and  $Y$  exists. Therefore, Equations (12.53) and (12.54) can be applied as well.  $\triangleleft$

**Remark 15.9 (Versions of  $E^{Z=z}(Y|X)$ )** Under the assumptions of Remark 15.8, Equation (15.11) and Remark 14.9 immediately imply

$$V_z \underset{P^{Z=z}}{=} g_0(z) + \sum_{i=1}^n g_i(z) \cdot 1_{X=x_i}, \quad \forall V_z \in \mathcal{E}^{Z=z}(Y|X). \quad (15.12)$$

$\triangleleft$

**Remark 15.10 (Partial Conditional Expectation  $E(Y|X=x_i, Z)$ )** Let the assumptions of Definition 15.4 hold and let  $E(Y|X=x_i, Z)$  denote the partial conditional expectation (see Def. 14.29 and Eq. 14.23). Then Equation (15.6) implies that there is a version  $E(Y|X=x_i, Z)$  such that

$$E(Y|X=x_i, Z) = g_0(Z) + g_i(Z), \quad \forall i = 1, \dots, n. \quad (15.13)$$

$\triangleleft$

**Remark 15.11 (Conditional Expectation  $E^{X=x_i}(Y|Z)$ )** If the assumptions of Definition 15.4 hold, then, for all  $i = 1, \dots, n$ , there is an  $E^{X=x_i}(Y|Z) \in \mathcal{E}^{X=x_i}(Y|Z)$  with

$$E^{X=x_i}(Y|Z) = E(Y|X=x_i, Z) = g_0(Z) + g_i(Z). \quad (15.14)$$

$\triangleleft$

**Remark 15.12 (Versions of  $E^{X=x_i}(Y|Z)$ )** Remark 14.9 and Equation (15.14) imply, for all  $i = 1, \dots, n$ ,

$$V_{x_i} \stackrel{P}{=} g_0(Z) + g_i(Z), \quad \forall V_{x_i} \in \mathcal{E}^{X=x_i}(Y|Z). \quad (15.15)$$

<

**Example 15.13 (No Treatment for Joe – continued)** In Example 14.3, the conditional expectation  $E^{X=1}(Y|Z)$  is not  $P$ -unique. Although the  $U$ -conditional intercept function  $g_0$  and the conditional effect  $g_1(Ann)$  are uniquely defined, the  $U$ -conditional effect function  $g_1$  is not defined, because  $P(X=1|U) \stackrel{P}{>} 0$  [see (15.2)] does not hold. In this example, there are (infinitely) many functions  $g_1$  satisfying Equation (15.1). This example emphasizes the importance of (15.2), which, in the definition of a conditional effect function (see Def. 15.4), is assumed to hold. <

### 15.3 Implications of Independence of $X$ and $Z$ for Regression Coefficients

In the following theorem we presume that the assumptions of Definition 15.4 hold, which implies that there is a version  $E(Y|X, Z)$  of the  $(X, Z)$ -conditional expectation of  $Y$  such that Equation (15.6) holds. Now we consider the implications for the conditional expectation  $E(Y|X)$  if we additionally assume

$$E[E^{X=x_i}(Y|Z) | X] \stackrel{P}{=} E[E^{X=x_i}(Y|Z)], \quad \forall i = 0, 1, \dots, n. \quad (15.16)$$

Note that this equation follows from independence of  $X$  and  $Z$  [see (14.32) with  $B = \{X=x_i\}$ , (14.17), Theorem 5.52, and Box 10.2 (vi)].

Remember, if the assumptions of Definition 15.4 hold, then, according to Theorem 12.37, there is a version  $E(Y|X) \in \mathcal{E}(Y|X)$  such that

$$E(Y|X) = \beta_0 + \sum_{i=1}^n \beta_i \cdot 1_{X=x_i} \quad (15.17)$$

with

$$\beta_0 = E(Y|X=x_0) \quad (15.18)$$

and

$$\beta_i = E(Y|X=x_i) - E(Y|X=x_0). \quad (15.19)$$

**Theorem 15.14 (Mean Independence and Average Effects)**

*Let the assumptions of Definition 15.4 hold and let  $g_i(Z)$ ,  $i = 1, \dots, n$ , be the functions defined in Theorem 15.3. If Equation (15.16) holds, then there is a version  $E(Y|X) \in \mathcal{E}(Y|X)$  satisfying Equations (15.17) to (15.19) with*

$$\beta_0 = E[g_0(Z)] \quad (15.20)$$

and

$$\beta_i = E[g_i(Z)], \quad \forall i = 1, \dots, n. \quad (15.21)$$

(Proof p. 453)

**Remark 15.15 (Uniqueness of Regression Coefficients)** According to Equations (15.20) and (15.21), the regression coefficients  $\beta_0, \beta_1, \dots, \beta_n$  are uniquely defined. The crucial assumption for uniqueness is  $P(X \in \{x_0, x_1, \dots, x_n\}) = 1$  and  $P(X = x_i) > 0$ , for all  $i = 0, 1, \dots, n$ .  $\triangleleft$

**Remark 15.16 (Independence of  $X$  and  $Z$ )** Independence of  $X$  and  $Z$  implies (15.16), which in turn implies

$$E[g_i(Z) | X] \stackrel{p}{=} E[g_i(Z)], \quad \forall i = 0, 1, \dots, n. \quad (15.22)$$

Note that Equation (15.22) also immediately follows from Theorem 5.52 and Box 10.2 (vi).  $\triangleleft$

**Remark 15.17 (The Role of Randomization)** From a methodological point of view it should be noted that independence of a treatment variable  $X$  and a variable  $Z$  can be created by randomized assignment of the observational unit (e. g., a person) to one of the treatment conditions, provided that  $Z$  represents a pretreatment variable. Randomized assignment creates independence of  $X$  and *all* pretreatment variables. Examples for such pretreatment variables are the person variable  $U$  (see Table 14.1 for a concrete example) as well as any function of  $U$  such as sex, race, and any other attribute of persons prior to treatment.

This means that we can ignore a pretreatment variable  $Z$  in a randomized experiment with treatment variable  $X$  and dichotomous response variable  $Y$  and still interpret the coefficient  $\beta_i$  as the average effect  $E[g_i(Z)]$  of  $x_i$  compared to  $x_0$  on  $Y$ , where  $i = 1, \dots, n$  [see Eq. 15.21].  $\triangleleft$

According to Corollary 15.18, independence of  $X$  and  $Z$  implies  $P$ -uniqueness of the functions  $g_i(Z)$ , provided that we presume  $P(X = x_i) > 0$  for all  $i = 0, 1, \dots, n$ .

**Corollary 15.18 (Independence of  $X$  and  $Z$  and  $P$ -Uniqueness)**

Let the assumptions 15.1 hold. If  $X \perp\!\!\!\perp_p Z$ , then  $E^{X=x_i}(Y|Z)$  and the functions  $g_i(Z)$  in Equations (15.7) and (15.8) are  $P$ -unique, for all  $i = 0, 1, \dots, n$ .

(Proof p. 453)

**Example 15.19 (Joe and Ann With Random Assignment—continued)** In Example 14.1 we already showed that the slope in

$$E(Y|X) = .45 + .15 \cdot X,$$

[see Eq. (14.5)] is identical to the expectation  $E_U(g_1) = E[g_1(U)] = .15$  of the  $U$ -conditional effect function  $g_1$ . Similarly, using Equation (14.1) and  $E(1_{U=Joe}) = .5$  (see Table 14.1),

$$E[g_0(U)] = E(.2 + .5 \cdot 1_{U=Joe}) = .2 + .5 \cdot E(1_{U=Joe}) = .2 + .5 \cdot .5 = .45,$$

which is identical to the intercept in the equation for  $E(Y|X)$ . This illustrates Theorem 15.14.  $\triangleleft$

**Example 15.20 (Joe and Ann With Self-Selection – continued)** In Example 14.2 we already showed that the slope in

$$E(Y|X) = .60 - .18 \cdot X,$$

[see Eq. (14.6)] is not identical to the expectation

$$E_U(g_1) = E[g_1(U)] = .15$$

of the  $U$ -conditional effect function  $g_1$ . In this example, neither independence of  $U$  and  $X$  nor  $E[E_U(g_1)|X] = E[g_1(U)]$  hold, and the slope  $-.18$  in the equation for  $E(Y|X)$  cannot be used for the evaluation of the treatment effect. This emphasizes the importance of assumption (15.16), which is the crucial assumption made in Theorem 15.14.  $\triangleleft$

## 15.4 Adjusted Conditional Effect Functions

**Remark 15.21 (Methodological Background)** In Definition 15.4 we introduced the  $Z$ -conditional intercept function  $g_0$  as well as the  $Z$ -conditional effect functions  $g_i$ ,  $i = 1, \dots, n$ , where  $Z = (Z_1, \dots, Z_m)$  can be an  $m$ -dimensional random variable consisting of  $m$  uni-dimensional random variables. Examples of such random variables are *pretest score* ( $Z_1$ ), *sex* ( $Z_2$ ), *educational status* ( $Z_3$ ), *body mass index* ( $Z_4$ ), and *blood type* ( $Z_5$ ). As mentioned before, conditioning on a (possibly multi-dimensional) random variable  $Z = (Z_1, \dots, Z_m)$  also serves to obtain more specific effects that are more informative than unconditional effects.

However, the  $(Z_1, \dots, Z_m)$ -conditional effects described by the effect functions  $g_i$  might be too fine-grained and one may wish to re-aggregate them. The most radical re-aggregation is to consider the (unconditional) expectation (the ‘average’) of the  $(Z_1, \dots, Z_m)$ -conditional effects.

Aside from the average effects, i. e., the (unconditional) expectations of the functions  $g_i(Z_1, \dots, Z_m)$ , we might also be interested in the conditional expectation values of the functions  $g_i(Z_1, \dots, Z_m)$  given  $Z_2 = \textit{male}$  and given  $Z_2 = \textit{female}$ , or in the conditional expectation values of the functions  $g_i(Z_1, \dots, Z_m)$  given various values of  $Z_1$  (*pretest*). This way of re-aggregation may be called *coarsening the effects*. Knowing such coarsened effects is very important, e. g., if, knowing the  $(Z_1, \dots, Z_m)$ -conditional effect functions (e. g., from a previous study), we want to conduct an as much as possible individualized treatment, but are not able to assess all components  $Z_1, \dots, Z_m$  of  $Z$ , but just  $Z_1$  or just  $Z_1$  and  $Z_2$ , for instance.

In Definition 15.22 we consider re-aggregating the  $Z$ -conditional intercept function and the  $Z$ -conditional effect functions to a  $W$ -conditional effect function that is adjusted for  $Z$ , where  $W$  is another random variable. If  $Z = (Z_1, \dots, Z_m)$

with  $m \geq 2$ , then  $W = Z_1$  and  $W = (Z_1, Z_2)$  are examples in case. In these two cases,  $W$  would be  $Z$ -measurable.

Note that re-aggregating the  $Z$ -conditional effect functions is not equivalent to ignoring  $Z$  and conditioning on  $W$  instead. More precisely, assume (a)  $Z = (Z_1, \dots, Z_m)$  and there are real-valued functions  $g_i$  such that

$$E(Y|X, Z) \stackrel{p}{=} g_0(Z) + \sum_{i=1}^n g_i(Z) \cdot 1_{X=x_i}, \quad (15.23)$$

(b)  $W = (Z_{i_1}, \dots, Z_{i_k})$ , for  $\{i_1, \dots, i_k\} \subset \{1, \dots, m\}$ , and (c) there are real-valued functions  $f_i$  such that

$$E(Y|X, W) \stackrel{p}{=} f_0(W) + \sum_{i=1}^n f_i(W) \cdot 1_{X=x_i}. \quad (15.24)$$

Then  $f_i(W) = E[g_i(Z)|W]$ , for  $i \in \{1, \dots, n\}$ , does not necessarily hold (see Example 15.24).

Considering  $E(Y|X, W)$  instead of  $E(Y|X, Z)$ , we might miss the purpose of controlling and adjusting for important confounders contained in  $Z$  but not in  $W$ . In contrast, using the conditional expectation  $E[g_i(Z)|W]$  (see Def. 15.22), we still control for  $Z$ , and with it we control for potential confounders contained in  $Z$ . Although, re-aggregating the  $Z$ -conditional effect functions in this way, i. e., considering  $E[g_i(Z)|W]$ , we obtain less informative and less individualized conditional effects, the purpose of controlling and adjusting for important confounders is still fulfilled.  $\triangleleft$

**Definition 15.22 (Adjusted Conditional Effect Function)**

Let the assumptions of Definition 15.4 hold, let  $E(Y|X, Z) \in \mathcal{E}(Y|X, Z)$  and  $g_i$ ,  $i = 1, \dots, n$ , be the functions satisfying Equation (15.6), and let  $W: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_W, \mathcal{A}'_W)$  be a random variable.

- (i) Then, for each  $i = 1, \dots, n$ , the conditional expectation  $E[g_i(Z)|W]$  is called a  $Z$ -adjusted  $W$ -conditional effect function of  $x_i$  vs.  $x_0$  on  $Y$ .
- (ii) Furthermore, for each  $i = 1, \dots, n$ , the expectation  $E[g_i(Z)]$  is called the  $Z$ -adjusted effect of  $x_i$  vs.  $x_0$  on  $Y$  or the average of the  $Z$ -conditional effects of  $x_i$  vs.  $x_0$  on  $Y$  or simply the average effect of  $x_i$  vs.  $x_0$ , if there is no ambiguity about  $Y$  and  $Z$ .

According to Remark 15.2, the conditional expectations  $E[g_i(Z)|W]$ ,  $i = 1, \dots, n$ , are  $P$ -unique and the expectations  $E[g_i(Z)]$  are uniquely determined. The values of  $E[g_i(Z)|W]$  are the  $(W=w)$ -conditional effects of  $x_i$  vs.  $x_0$  on  $Y$  that are adjusted for  $Z$ .

**Remark 15.23 ( $Z$ -Adjusted  $W$ -Conditional Expectation of  $Y$  given  $X=x_i$ )** Using the definition of the functions  $g_i(Z)$  [see Eq. (15.8)] and Rule (xv) of Box 10.2 yield

**Table 15.1.** Four Persons With Self-Selection to Treatment

Outcomes $\omega$			Observables				Conditional Expectations				
Unit	Treatment	Success	$P(\{\omega\})$	Person variable $U$	Sex $Z$	Treatment variable $X$	Outcome variable $Y$	$E(Y X, U)$	$E(Y X, Z)$	$E(Y X)$	$P(X=1 U)$
(Joe, no, -)			.0675	Joe	m	0	0	.7	.66	.63	.1
(Joe, no, +)			.1575	Joe	m	0	1	.7	.66	.63	.1
(Joe, yes, -)			.0050	Joe	m	1	0	.8	.44	.44	.1
(Joe, yes, +)			.0200	Joe	m	1	1	.8	.44	.44	.1
(Jim, no, -)			.0175	Jim	m	0	0	.3	.66	.63	.9
(Jim, no, +)			.0075	Jim	m	0	1	.3	.66	.63	.9
(Jim, yes, -)			.1350	Jim	m	1	0	.4	.44	.44	.9
(Jim, yes, +)			.0900	Jim	m	1	1	.4	.44	.44	.9
(Sue, no, -)			.0600	Sue	f	0	0	.7	.60	.63	.2
(Sue, no, +)			.1400	Sue	f	0	1	.7	.60	.63	.2
(Sue, yes, -)			.0200	Sue	f	1	0	.6	.44	.44	.2
(Sue, yes, +)			.0300	Sue	f	1	1	.6	.44	.44	.2
(Ann, no, -)			.0400	Ann	f	0	0	.2	.60	.63	.8
(Ann, no, +)			.0100	Ann	f	0	1	.2	.60	.63	.8
(Ann, yes, -)			.1200	Ann	f	1	0	.4	.44	.44	.8
(Ann, yes, +)			.0800	Ann	f	1	1	.4	.44	.44	.8

$$\begin{aligned}
 & E[g_i(Z) | W] \\
 &= E[E^{X=x_i}(Y|Z) - E^{X=x_0}(Y|Z) | W] \\
 &= E[E^{X=x_i}(Y|Z) | W] - E[E^{X=x_0}(Y|Z) | W], \quad \forall i = 1, \dots, n.
 \end{aligned}
 \tag{15.25}$$

A conditional expectation  $E[E^{X=x_i}(Y|Z) | W]$  is called a *Z-adjusted W-conditional expectation of Y given X = x<sub>i</sub>*, where  $i = 0, 1, \dots, n$ . ◁

**Example 15.24 (Four Persons With Self-Selection to Treatment)** In Table 15.1 we present a new example displaying the conditional expectations  $E(Y|X, U)$ ,  $E(Y|X, Z)$ ,  $E(Y|X)$ , and  $P(X=1|U)$  with dichotomous random variables  $X$ ,  $Y$ , and  $Z$ , where  $X$  indicates with its values 0 and 1 whether or not the person is treated,  $Y$  indicates with its values 0 and 1 whether or not the person is successful, the values of  $Z$  are *male* and *female*, and  $U$  is the person variable with values *Joe*, *Jim*, *Sue*, and *Ann*.

We use the example of Table 15.1 to show that re-aggregating the  $U$ -conditional effect function is not equivalent to ignoring  $U$  and conditioning on  $Z$  instead. In order to formulate this more precisely, consider the equations

$$E(Y|X, U) = g_0(U) + g_1(U) \cdot X \tag{15.26}$$

and

$$E(Y|X, Z) = f_0(Z) + f_1(Z) \cdot X. \quad (15.27)$$

We show that, in this example, the function  $f_1(Z)$  is not identical to the  $U$ -adjusted  $Z$ -conditional expectation  $E[g_1(U)|Z]$ .

Using the values  $E(Y|X=x, U=u)$  of the conditional expectation  $E(Y|X, U)$  displayed in Table 15.1 yields the following values of the functions  $g_i$  of Equation (15.26):

$$g_0(\text{Joe}) = .7, \quad g_0(\text{Jim}) = .3, \quad g_0(\text{Sue}) = .7, \quad g_0(\text{Ann}) = .4, \quad (15.28)$$

and

$$g_1(\text{Joe}) = .1, \quad g_1(\text{Jim}) = .1, \quad g_1(\text{Sue}) = -.1, \quad g_1(\text{Ann}) = .2 \quad (15.29)$$

Similarly, using the values  $E(Y|X=x, Z=z)$  of the conditional expectation  $E(Y|X, Z)$  displayed in Table 15.1 yields the following values of the functions  $f_i$  of Equation (15.27):

$$f_0(\text{male}) = .66, \quad f_0(\text{female}) = .60 \quad (15.30)$$

and

$$f_1(\text{male}) = .22, \quad f_1(\text{female}) = .16. \quad (15.31)$$

Hence, the ( $Z=\text{male}$ )-conditional effect of  $X$  on  $Y$  is .22 and the ( $Z=\text{female}$ )-conditional effect of  $X$  on  $Y$  is .16. In this example, these two numbers are misleading if used for the evaluation of the treatment effect, because the individual treatment effects for Joe and Jim are both .1, and for Sue and Ann they are  $-.1$  and .2, respectively.

In contrast, computing the two values of the  $U$ -adjusted  $Z$ -conditional expectation  $E[g_1(U)|Z]$  yields

$$E[g_1(U)|Z=\text{male}] = \sum_u g_1(u) \cdot P(U=u|Z=\text{male}) = .1 \cdot .5 + .1 \cdot .5 - .1 \cdot 0 + .2 \cdot 0 = .1,$$

and correspondingly,

$$E[g_1(U)|Z=\text{female}] = \sum_u g_1(u) \cdot P(U=u|Z=\text{female}) = .1 \cdot 0 + .1 \cdot 0 - .1 \cdot .5 + .2 \cdot .5 = .5.$$

Hence, the  $U$ -adjusted ( $Z=\text{male}$ )-conditional expectation value  $E[g_1(U)|Z=\text{male}] = .1$  is the average effect of the treatment for males, and the  $U$ -adjusted ( $Z=\text{female}$ )-conditional expectation value  $E[g_1(U)|Z=\text{female}] = .5$  is the average effect of the treatment for females. These effects are less informative than the person-specific effects  $g_1(u)$  [see (15.29)]. Nevertheless, in this example they still inform us about the (conditional) expectations of these person-specific effects for the two sexes. In other words, in this example, they are the average treatment effects of the males and of the females, respectively.  $\triangleleft$

## 15.5 Conditional Logit Effect Functions

In the previous sections of this chapter we studied the conditional intercept and conditional effect functions  $g_0, g_1, \dots, g_n$ . Now we consider the special case, in which  $Y$  is dichotomous with values 0 and 1 (see Example 5.10). In this case  $E(Y|X, Z)$  is also called a conditional probability and is denoted by  $P(Y=1|X, Z)$  (see Rem. 10.4). If  $Y$  is dichotomous, then aside from the conditional intercept and conditional effect functions  $g_0, g_1, \dots, g_n$ , there are also  $Z$ -conditional logit intercept and logit effect functions (see Example 13.24), denoted  $f_i$ .

### Theorem 15.25 (Existence of the Conditional Logit Effect Functions)

Let the assumptions 15.1 hold, let  $Y$  be dichotomous with values 0 and 1, and suppose there is a  $P(Y=1|X, Z) \in \mathcal{P}(Y=1|X, Z)$  with  $0 < P(Y=1|X, Z) < 1$ . Then there are a version  $P(Y=1|X, Z) \in \mathcal{P}(Y=1|X, Z)$ , measurable functions  $g_0, g_1, \dots, g_n: (\Omega'_Z, \mathcal{A}'_Z) \rightarrow (\mathbb{R}, \mathcal{B})$ , measurable functions  $f_0, f_1, \dots, f_n: (\Omega'_Z, \mathcal{A}'_Z) \rightarrow (\mathbb{R}, \mathcal{B})$ , and, for all  $i = 0, 1, \dots, n$ , a real-valued  $P^{X=x_i}(Y=1|Z) \in \mathcal{P}^{X=x_i}(Y=1|Z)$  such that

$$P(Y=1|X, Z) = g_0(Z) + \sum_{i=1}^n g_i(Z) \cdot 1_{X=x_i} \quad (15.32)$$

$$= \frac{\exp[f_0(Z) + \sum_{i=1}^n f_i(Z) \cdot 1_{X=x_i}]}{1 + \exp[f_0(Z) + \sum_{i=1}^n f_i(Z) \cdot 1_{X=x_i}]} \quad (15.33)$$

with

$$g_0(Z) := P^{X=x_0}(Y=1|Z) \quad (15.34)$$

$$= \frac{\exp[f_0(Z)]}{1 + \exp[f_0(Z)]} \quad (15.35)$$

and, for  $i = 1, \dots, n$ ,

$$g_i(Z) := P^{X=x_i}(Y=1|Z) - P^{X=x_0}(Y=1|Z) \quad (15.36)$$

$$= \frac{\exp[f_0(Z) + f_i(Z)]}{1 + \exp[f_0(Z) + f_i(Z)]} - \frac{\exp[f_0(Z)]}{1 + \exp[f_0(Z)]}. \quad (15.37)$$

(Proof p. 453)

**Remark 15.26 ( $P$ -Uniqueness of the Functions  $f_i(Z)$ )** Under the assumptions of Theorem 15.25 and the additional assumption (15.2), for all measurable functions  $f_0^*, f_1^*, \dots, f_n^*: (\Omega'_Z, \mathcal{A}'_Z) \rightarrow (\mathbb{R}, \mathcal{B})$  satisfying Equations (15.34) to (15.37),

$$f_i^*(Z) \stackrel{p}{=} f_i(Z), \quad \forall i = 0, 1, \dots, n, \quad (15.38)$$

(see Exercise 15-1). Hence,  $E[f_i^*(Z)] = E[f_i(Z)]$ ,  $i = 1, \dots, n$ , provided that these expectations exist.  $\triangleleft$

**Definition 15.27 (Conditional Logit Intercept and Effect Functions)**

Let the assumptions of Theorem 15.25 as well as (15.2) hold. Then the function  $f_0: (\Omega'_Z, \mathcal{A}'_Z) \rightarrow (\mathbb{R}, \mathcal{B})$  is called the  $Z$ -conditional logit intercept function, and, for all  $i = 1, \dots, n$ , the function  $f_i: (\Omega'_Z, \mathcal{A}'_Z) \rightarrow (\mathbb{R}, \mathcal{B})$ , the  $Z$ -conditional logit effect function of  $x_i$  vs.  $x_0$  on  $Y$  pertaining to the version  $P(Y=1|X, Z) \in \mathcal{P}(Y=1|X, Z)$  in Equation (15.33).

**Remark 15.28 (Partial Conditional Probability  $P(Y=1|X, Z=z)$ )** Let the assumptions of Theorem 15.25 hold. Then Equations (15.33) and (14.25) imply

$$P(Y=1|X, Z=z) = \frac{\exp(f_0(z) + \sum_{i=1}^n f_i(z) \cdot 1_{X=x_i})}{1 + \exp(f_0(z) + \sum_{i=1}^n f_i(z) \cdot 1_{X=x_i})} \quad (15.39)$$

for the partial conditional probability  $P(Y=1|X, Z=z)$  (see Rem. 14.30). This equation justifies the terminology introduced in Definition 15.27.  $\triangleleft$

**Remark 15.29 (( $Z=z$ )-Conditional Logit Intercept and Effects)** If the assumptions of Theorem 15.25 hold and  $z \in \Omega'_Z$  with  $P(Z=z) > 0$ , then Equation (15.39) and (14.29) imply that there is a  $P^{Z=z}(Y=1|X) \in \mathcal{P}^{Z=z}(Y=1|X)$  with

$$P^{Z=z}(Y=1|X) = P(Y=1|X, Z=z) = \frac{\exp(f_0(z) + \sum_{i=1}^n f_i(z) \cdot 1_{X=x_i})}{1 + \exp(f_0(z) + \sum_{i=1}^n f_i(z) \cdot 1_{X=x_i})}. \quad (15.40)$$

Equation (15.40) and Theorem 13.20 imply that  $f_0(z)$  is the intercept and  $f_i(z)$ ,  $i = 1, \dots, n$ , are the coefficients pertaining to a linear logistic parametrization of  $P^{Z=z}(Y=1|X)$ . According to Lemma 12.38 and Remark 12.39,  $P^{Z=z}(X=x_i) > 0$  for all  $i = 1, \dots, n$  implies that the matrix of the covariances of the random variables  $1_{X=x_1}, \dots, 1_{X=x_n}$  with respect to the measure  $P^{Z=z}$  is regular. Therefore, we can apply Theorem 13.20 for identifying the coefficients  $f_0(z)$  and  $f_i(z)$ ,  $i = 1, \dots, n$ .

Under the assumptions of Theorem 15.25, Remark 14.9 and Equation (15.40) imply

$$V_z \stackrel{=}{=}_{P^{Z=z}} \frac{\exp(f_0(z) + \sum_{i=1}^n f_i(z) \cdot 1_{X=x_i})}{1 + \exp(f_0(z) + \sum_{i=1}^n f_i(z) \cdot 1_{X=x_i})}, \quad \forall V_z \in \mathcal{P}^{Z=z}(Y=1|X). \quad (15.41)$$

$\triangleleft$

**Remark 15.30 (Conditional Probability  $P^{X=x_i}(Y=1|Z)$ )** Remark 14.9 and Equation (15.32) imply

$$V_0 \stackrel{=}{=}_{P^{X=x}} \frac{\exp(f_0(Z))}{1 + \exp(f_0(Z))}, \quad \forall V_0 \in \mathcal{P}^{X=0}(Y=1|Z), \quad (15.42)$$

and, for all  $i = 1, \dots, n$ ,

$$V_i \stackrel{=}{=}_{P^{X=x_i}} \frac{\exp(f_0(Z) + f_i(Z))}{1 + \exp(f_0(Z) + f_i(Z))}, \quad \forall V_i \in \mathcal{P}^{X=x_i}(Y=1|Z). \quad (15.43)$$

$\triangleleft$

**Remark 15.31 (Conditional Log Odds Functions)** In terms of conditional probabilities, the  $Z$ -conditional logit intercept function  $f_0$  satisfies

$$f_0(Z) = \ln \left[ \frac{P^{X=x_0}(Y=1|Z)}{1 - P^{X=x_0}(Y=1|Z)} \right] \quad (15.44)$$

(cf. Rem. 13.15). Hence,  $f_0$  may also be called the  $Z$ -conditional log odds function of  $P^{X=x_0}(Y=1|Z)$ . Similarly,

$$f_0(Z) + f_i(Z) = \ln \left[ \frac{P^{X=x_i}(Y=1|Z)}{1 - P^{X=x_i}(Y=1|Z)} \right], \quad i = 1, \dots, n \quad (15.45)$$

(see Exercise 15-2). The function  $f_0 + f_i: (\Omega'_Z, \mathcal{A}'_Z) \rightarrow (\mathbb{R}, \mathcal{B})$  satisfying  $(f_0 + f_i)(Z) = f_0(Z) + f_i(Z)$  is called the  $Z$ -conditional log odds function of  $P^{X=x_i}(Y=1|Z)$ .  $\triangleleft$

**Remark 15.32 (Conditional Log Odds Ratio Functions)** Equations (15.44) and (15.45) imply

$$f_i(Z) = \ln \left[ \frac{P^{X=x_i}(Y=1|Z)}{1 - P^{X=x_i}(Y=1|Z)} \right] - \ln \left[ \frac{P^{X=x_0}(Y=1|Z)}{1 - P^{X=x_0}(Y=1|Z)} \right] \quad (15.46)$$

$$= \ln \left[ \frac{P^{X=x_i}(Y=1|Z)}{1 - P^{X=x_i}(Y=1|Z)} \bigg/ \frac{P^{X=x_0}(Y=1|Z)}{1 - P^{X=x_0}(Y=1|Z)} \right], \quad i = 1, \dots, n, \quad (15.47)$$

for the  $Z$ -conditional logit effect functions  $f_i$  (cf. Rem. 13.16).

Referring to Equation (15.47),  $f_i$ ,  $i = 1, \dots, n$ , is also called the  $Z$ -conditional odds ratio function of  $P^{X=x_i}(Y=1|Z)$  and  $P^{X=x_0}(Y=1|Z)$ .  $\triangleleft$

**Remark 15.33 (Conditional Odds Ratio Functions)** The exponential function of  $f_i(Z)$  is

$$\exp[f_i(Z)] = \frac{P^{X=x_i}(Y=1|Z)}{1 - P^{X=x_i}(Y=1|Z)} \bigg/ \frac{P^{X=x_0}(Y=1|Z)}{1 - P^{X=x_0}(Y=1|Z)}, \quad i = 1, \dots, n. \quad (15.48)$$

The composite function  $\exp(f_i): (\Omega'_Z, \mathcal{A}'_Z) \rightarrow (\mathbb{R}, \mathcal{B})$  is called the  $Z$ -conditional odds ratio function of  $P^{X=x_i}(Y=1|Z)$  and  $P^{X=x_0}(Y=1|Z)$  (cf. Rem. 13.17).  $\triangleleft$

**Remark 15.34 (Conditional Risk Ratio Functions)** Another closely related function describing conditional effects of  $(X=x_i)$  compared to  $(X=x_0)$  is

$$k_i(Z) := \frac{P^{X=x_i}(Y=1|Z)}{P^{X=x_0}(Y=1|Z)}, \quad i = 1, \dots, n. \quad (15.49)$$

The function  $k_i: (\Omega'_Z, \mathcal{A}'_Z) \rightarrow (\mathbb{R}, \mathcal{B})$  satisfying (15.49) is called the  $Z$ -conditional risk ratio function of  $P^{X=x_i}(Y=1|Z)$  and  $P^{X=x_0}(Y=1|Z)$  (cf. Rem. 13.18).  $\triangleleft$

**Remark 15.35 (Four Kinds of Conditional Effect Functions)** Hence, under the assumptions of Theorem 15.25, we considered four kinds of different  $Z$ -conditional effect functions:  $g_i$ ,  $f_i$ ,  $\exp(f_i)$ , and  $k_i$ . They all describe  $Z$ -conditional effect functions of  $x_i$  compared to  $x_0$  on  $Y$  on different scales.  $\triangleleft$

## 15.6 Implications of Independence of $X$ and $Z$ for the Logit Regression Coefficients

In section 15.3 we already treated the implication of

$$E[g_i(Z)|X] = E[g_i(Z)], \quad \forall i = 0, 1, \dots, n, \quad (15.50)$$

[see Eq. (15.22)]. According to Theorem 15.14, this equation implies

$$E[g_i(Z)] = \beta_i, \quad \forall i = 0, 1, \dots, n, \quad (15.51)$$

where  $\beta_i$  are the parameters in the equation

$$P(Y=1|X) = \beta_0 + \sum_{i=1}^n \beta_i \cdot 1_{X=x_i}. \quad (15.52)$$

Note that, under the assumptions of Theorem 15.25, there are coefficients  $\beta_0, \beta_1, \dots, \beta_n \in \mathbb{R}$  such that Equation (15.52) holds. Remember, according to Remark 15.16, Equation (15.50) holds if  $X$  and  $Z$  are independent.

**Remark 15.36 (Mean Independence and Logit Effect Functions)** Note that the analog to Equation (15.51) *does not hold* for the expectation of the logit effect functions  $f_i$  specified in Equation (15.33). That is, although, under the assumptions of Theorem 15.25, there are always coefficients  $\alpha_0, \alpha_1, \dots, \alpha_n \in \mathbb{R}$  such that

$$P(Y=1|X) = \frac{\exp(\alpha_0 + \sum_{i=1}^n \alpha_i \cdot 1_{X=x_i})}{1 + \exp(\alpha_0 + \sum_{i=1}^n \alpha_i \cdot 1_{X=x_i})}, \quad (15.53)$$

neither

$$E[f_i(Z)|X] \stackrel{p}{=} E[f_i(Z)], \quad \forall i = 0, 1, \dots, n, \quad (15.54)$$

nor independence of  $X$  and  $Z$  imply that  $\alpha_i$  is identical to  $E[f_i(Z)]$  (see Example 15.38). In fact it is even possible that  $\alpha_i$  is negative and  $E[f_i(Z)]$  is positive (see Example 15.39).  $\triangleleft$

**Remark 15.37 (Implications for Methodology)** From a methodological point of view this means that randomized assignment of a unit to one of the treatment conditions—which creates independence of a treatment variable  $X$  and the person variable  $U$ —does not imply that the regression coefficients in the logistic parametrization of  $P(Y=1|X)$  can be interpreted as the expectation of the corresponding functions  $f_i(U)$ . As mentioned before, it is even possible that a coefficient  $\alpha_i$  is negative and  $E[f_i(Z)]$  is positive (see Example 15.39). In contrast, compare the corresponding invariance property formulated in Theorem 12.44 for a linear parametrization of a conditional expectation.  $\triangleleft$

**Example 15.38 (Joe and Ann With Random Assignment—continued)** Now we show by an example that independence of  $X$  and  $Z$  does *not imply* that the coefficients

$\alpha_i$  of Equation (15.53) are identical to  $E[f_i(Z)]$ , where the functions  $f_i$  are specified in Theorem 15.25. (In contrast, compare Equations (14.4) and (14.5) for the effect function  $g_1$ .)

In Examples 13.24 and 13.26 we computed

$$P(Y=1|X) = \frac{\exp(\alpha_0 + \alpha_1 \cdot X)}{1 + \exp(\alpha_0 + \alpha_1 \cdot X)} \approx \frac{\exp(-.201 + .606 \cdot X)}{1 + \exp(-.201 + .606 \cdot X)} \quad (15.55)$$

and

$$\begin{aligned} P(Y=1|X, U) &\stackrel{P}{=} \frac{\exp((\lambda_0 + \lambda_2 \cdot 1_{U=Ann}) + (\lambda_1 + \lambda_3 \cdot 1_{U=Ann}) \cdot X)}{1 + \exp((\lambda_0 + \lambda_2 \cdot 1_{U=Ann}) + (\lambda_1 + \lambda_3 \cdot 1_{U=Ann}) \cdot X)} \\ &\approx \frac{\exp((.847 - 2.234 \cdot 1_{U=Ann}) + (.539 + .442 \cdot 1_{U=Ann}) \cdot X)}{1 + \exp((.847 - 2.234 \cdot 1_{U=Ann}) + (.539 + .442 \cdot 1_{U=Ann}) \cdot X)}, \end{aligned} \quad (15.56)$$

showing that the logit of  $P(Y=1|X, U)$  is  $f_0(U) + f_1(U) \cdot X$  with logit intercept function  $f_0$  satisfying

$$f_0(U) = \lambda_0 + \lambda_2 \cdot 1_{U=Ann} \approx .847 - 2.234 \cdot 1_{U=Ann}$$

and logit effect function  $f_1$  satisfying

$$f_1(U) = \lambda_1 + \lambda_3 \cdot 1_{U=Ann} = .539 + .442 \cdot 1_{U=Ann}.$$

Note that

$$\alpha_1 \approx .606 \neq E[f_1(U)] \approx .539 + .442 \cdot E(1_{U=Ann}) \approx .760.$$

Hence, although  $X$  and  $U$  are independent, the logit effect  $\alpha_1$  of  $X$  in the logistic parametrization of  $P(Y=1|X)$  is *not equal* to the expectation of the function  $f_1(U)$  in the logistic parametrization of  $P(Y=1|X, U)$ .  $\triangleleft$

**Example 15.39 (Joe and Ann: Reversed Average Logit Effect)** Table 15.2 displays an example in which the coefficient  $\alpha_1$  in the equation

$$P(Y=1|X) = \frac{\exp(\alpha_0 + \alpha_1 \cdot X)}{1 + \exp(\alpha_0 + \alpha_1 \cdot X)} \quad (15.57)$$

is negative, whereas the expectation  $E[f_1(U)]$  of the function  $f_1(U)$  in the equation

$$P(Y=1|X, U) \stackrel{P}{=} \frac{\exp(f_0(U) + f_1(U) \cdot X)}{1 + \exp(f_0(U) + f_1(U) \cdot X)} \quad (15.58)$$

is positive, although  $X$  and  $U$  are independent. The coefficients of Equation (15.57) are  $\alpha_0 \approx 1.005$  and  $\alpha_1 \approx -.490$ , whereas the expectation of the conditional logit effect function is  $E_U(f_1) = E[f_1(U)] \approx .552$  (see Exercise 15-3).  $\triangleleft$

**Table 15.2.** Joe and Ann: Reversed Average Logit Effect

Outcomes $\omega$		Observables			Conditional expectations								
Unit	Treatment Success	$P(\omega)$	Person variable $U$	Treatment variable $X$	Outcome variable $Y$	$E(Y X, U)$	$E(Y X)$	$P(X=1 U)$	$E^{X=0}(Y U)$	$E^{X=1}(Y U)$	$\text{logit}[E(Y X, U)]$	$g_1(U) = \frac{E^{X=1}(Y U) - E^{X=0}(Y U)}{E^{X=1}(Y U) + E^{X=0}(Y U)}$	$f_1(U)$
(Joe, no, -)		.067	Joe	0	0	.732	.732	.5	.732	.984	1.005	.252	3.114
(Joe, no, +)		.183	Joe	0	1	.732	.732	.5	.732	.984	1.005	.252	3.114
(Joe, yes, -)		.004	Joe	1	0	.984	.626	.5	.732	.984	4.119	.252	3.114
(Joe, yes, +)		.246	Joe	1	1	.984	.626	.5	.732	.984	4.119	.252	3.114
(Ann, no, -)		.067	Ann	0	0	.732	.732	.5	.732	.268	1.005	-.464	-2.010
(Ann, no, +)		.183	Ann	0	1	.732	.732	.5	.732	.268	1.005	-.464	-2.010
(Ann, yes, -)		.183	Ann	1	0	.268	.626	.5	.732	.268	-1.005	-.464	-2.010
(Ann, yes, +)		.067	Ann	1	1	.268	.626	.5	.732	.268	-1.005	-.464	-2.010

**15.7 Proofs**

***Proof of Theorem 15.3***

For all  $i = 0, 1, \dots, n$ ,

$$E(Y) \text{ is finite} \Rightarrow E^{X=x_i}(Y) \text{ is finite} \quad [\text{Rem. 14.6}]$$

$$\Rightarrow \text{there is a real-valued } E^{X=x_i}(Y|Z) \in \mathcal{E}^{X=x_i}(Y|Z). \quad [\text{Box 10.2 (x)}]$$

According to Theorem 10.9 and Corollary 10.23, finiteness of  $E(Y)$  also implies that there is a  $g(X, Z) \in \mathcal{E}(Y|X, Z)$  such that, for real-valued versions  $E^{X=x_i}(Y|Z)$

$$g(X, Z) \stackrel{p}{=} \sum_{i=0}^n E^{X=x_i}(Y|Z) \cdot 1_{X=x_i} \quad [(14.31)]$$

$$\stackrel{p}{=} E^{X=x_0}(Y|Z) \cdot 1_{X=x_0} + \sum_{i=1}^n E^{X=x_i}(Y|Z) \cdot 1_{X=x_i}$$

$$\stackrel{p}{=} E^{X=x_0}(Y|Z) - \sum_{i=1}^n E^{X=x_0}(Y|Z) \cdot 1_{X=x_i} + \sum_{i=1}^n E^{X=x_i}(Y|Z) \cdot 1_{X=x_i} \quad [(5.33)]$$

$$\stackrel{p}{=} E^{X=x_0}(Y|Z) + \sum_{i=1}^n [E^{X=x_i}(Y|Z) - E^{X=x_0}(Y|Z)] \cdot 1_{X=x_i}.$$

Because the function on the right-hand side of the last equation is  $(X, Z)$ -measurable and  $P$ -equivalent to  $g(X, Z) \in \mathcal{E}(Y|X, Z)$  (see Th. 2.57), it is an element of  $\mathcal{E}(Y|X, Z)$ . Defining the specific version

$$E(Y|X, Z) := E^{X=x_0}(Y|Z) + \sum_{i=1}^n [E^{X=x_i}(Y|Z) - E^{X=x_0}(Y|Z)] \cdot 1_{X=x_i}$$

completes the proof.

**Proof of Theorem 15.14**

For all versions  $E(Y|X) \in \mathcal{E}(Y|X)$ ,

$$\begin{aligned}
 E(Y|X) &\stackrel{P}{=} E[E(Y|X, Z)|X] && \text{[Box 10.2 (v)]} \\
 &\stackrel{P}{=} E \left[ g_0(Z) + \sum_{i=1}^n g_i(Z) \cdot 1_{X=x_i} \middle| X \right] && \text{[(15.6)]} \\
 &\stackrel{P}{=} E[g_0(Z)|X] + \sum_{i=1}^n E[g_i(Z) \cdot 1_{X=x_i} | X] && \text{[Box 10.2 (xv)]} \\
 &\stackrel{P}{=} E[g_0(Z)|X] + \sum_{i=1}^n E[g_i(Z)|X] \cdot 1_{X=x_i} \quad [\sigma(1_{X=x_i}) \subset \sigma(X), (10.74), (2.36)] \\
 &\stackrel{P}{=} E[g_0(Z)] + \sum_{i=1}^n E[g_i(Z)] \cdot 1_{X=x_i}. && \text{[(15.22)]}
 \end{aligned}$$

According to Remark 2.18, the right-hand side of the last equation above is  $X$ -measurable. Because it is  $P$ -equivalent to all versions  $E(Y|X) \in \mathcal{E}(Y|X)$ , it is an element of  $\mathcal{E}(Y|X)$  [see (10.12)], and it satisfies Equations (15.17) to (15.21).

**Proof of Corollary 15.18**

According to Box 10.2 (vi) and Equations (6.5) and (10.4), independence of  $X$  and  $Z$  implies  $P(X=x_i|Z) \stackrel{P}{=} P(X=x_i)$ . Because  $P(X=x_i) > 0$ , we can conclude  $P(X=x_i|Z) \stackrel{P}{>} 0$ . Furthermore,  $P(X=x_i|Z) \stackrel{P}{>} 0$  is equivalent to  $P$ -uniqueness of  $E^{X=x_i}(Y|Z)$  [see Cor. 14.48 (a) and (c)]. According to Box 14.1 (viii), this implies  $P$ -uniqueness of the functions  $g_i(Z)$ ,  $i = 0, 1, \dots, n$ .

**Proof of Theorem 15.25**

By definition,  $P(Y=1|X, Z) = E(1_{Y=1}|X, Z)$ . Hence, the existence of measurable functions  $g_0, g_1, \dots, g_n: (\Omega'_Z, \mathcal{A}'_Z) \rightarrow (\mathbb{R}, \mathcal{B})$  satisfying Equations (15.32), (15.34), and (15.36) has already been proved in Theorem 15.3. In order to show that there are measurable functions  $f_0, f_1, \dots, f_n: (\Omega'_Z, \mathcal{A}'_Z) \rightarrow (\mathbb{R}, \mathcal{B})$  satisfying Equation (15.33) we define

$$l_i(Z) := \text{logit}[P^{X=x_i}(Y=1|Z)], \quad \forall i = 0, 1, \dots, n, \quad (15.59)$$

using the logit of  $P^{X=x_i}(Y=1|Z)$  defined by Equation (13.5). Furthermore, we define

$$f_0(Z) := l_0(Z), \quad (15.60)$$

and

$$f_i(Z) := l_i(Z) - l_0(Z), \quad \forall i = 1, \dots, n. \quad (15.61)$$

These definitions and Equation (13.6) then yield

$$P^{X=x_i}(Y=1|Z) = \frac{\exp(\text{logit}[P^{X=x_i}(Y=1|Z)])}{1 + \exp(\text{logit}[P^{X=x_i}(Y=1|Z)])} = \frac{\exp(l_i(Z))}{1 + \exp(l_i(Z))}, \quad \forall i = 0, 1, \dots, n.$$

Hence, for  $i = 0$ , Equation (15.34) implies

$$g_0 = P^{X=x_0}(Y=1|Z) = \frac{\exp(f_0(Z))}{1 + \exp(f_0(Z))},$$

which proves (15.35). Furthermore, Equation (15.36) yields

$$\begin{aligned} g_i(Z) &= P^{X=x_i}(Y=1|Z) - P^{X=x_0}(Y=1|Z) \\ &= \frac{\exp(f_0(Z) + f_i(Z))}{1 + \exp(f_0(Z) + f_i(Z))} - \frac{\exp(f_0(Z))}{1 + \exp(f_0(Z))}, \quad \forall i = 1, \dots, n, \end{aligned}$$

which proves (15.37).

Finally, for all  $\omega \in \Omega$ ,

$$\left( f_0(Z) + \sum_{i=1}^n f_i(Z) \cdot 1_{X=x_i} \right)(\omega) = \begin{cases} (f_0(Z) + f_j(Z))(\omega), & \text{if } X(\omega) = x_j, \quad j = 1, \dots, n \\ f_0(Z)(\omega), & \text{otherwise.} \end{cases}$$

For all  $\omega \in \{X=x_0\}$ ,

$$\begin{aligned} & \frac{\exp \left[ \left( f_0(Z) + \sum_{i=1}^n f_i(Z) \cdot 1_{X=x_i} \right)(\omega) \right]}{1 + \exp \left[ \left( f_0(Z) + \sum_{i=1}^n f_i(Z) \cdot 1_{X=x_i} \right)(\omega) \right]} \\ &= \frac{\exp [f_0(Z)(\omega)]}{1 + \exp [f_0(Z)(\omega)]} \quad [1_{X=x_i}(\omega) = 0, \quad \forall i = 1, \dots, n] \\ &= g_0(Z). \quad [(15.35)] \end{aligned}$$

Hence, for all  $j = 1, \dots, n$  and all  $\omega \in \{X=x_j\}$ ,

$$\begin{aligned} & \frac{\exp \left[ \left( f_0(Z) + \sum_{i=1}^n f_i(Z) \cdot 1_{X=x_i} \right)(\omega) \right]}{1 + \exp \left[ \left( f_0(Z) + \sum_{i=1}^n f_i(Z) \cdot 1_{X=x_i} \right)(\omega) \right]} \\ &= \frac{\exp [(f_0(Z) + f_j(Z))(\omega)]}{1 + \exp [(f_0(Z) + f_j(Z))(\omega)]} \\ &= [g_0(Z) + g_j(Z)](\omega) \quad [(15.36), (15.35)] \\ &= \left[ g_0(Z) + \sum_{i=1}^n g_i(Z) \cdot 1_{X=x_i} \right](\omega) \quad [1_{X=x_i}(\omega) = 1 \Leftrightarrow x_i = x_j] \\ &= P(Y=1|X, Z)(\omega). \quad [(15.32)] \end{aligned}$$

## 15.8 Exercises

▷ **Exercise 15-1** Prove Equation (15.38).

▷ **Exercise 15-2** Prove Equations (15.44) and (15.45).

▷ **Exercise 15-3** Using the results displayed in Table 15.2, compute the coefficients  $\alpha_0$  and  $\alpha_1$  of Equation (15.57) as well as the expectation of the conditional logit effect function  $f_1$ .

## Solutions

▷ **Solution 15-1** According to Theorem 14.46, assuming (15.2) implies that the conditional probabilities  $P^{X=x_i}(Y=1|Z)$  are  $P$ -unique, which in turn implies that the functions  $g_i$  are  $P$ -unique [see (2.36)]. If, for all  $i = 0, 1, \dots, n$ , the  $P^{X=x_i}(Y=1|Z)$  are  $P$ -unique, then the functions  $l_i(Z)$  and their differences [see Eqs. (15.60) and (15.61)] are  $P$ -unique as well [see (2.34) and Rem. 2.76].

▷ **Solution 15-2** The composition of  $Z$  and the conditional logit intercept function can be written

$$f_0(Z) = \text{logit}(P^{X=x_0}(Y=1|Z)) \quad [(15.35), (13.3)]$$

$$= \ln \left[ \frac{P^{X=x_0}(Y=1|Z)}{1 - P^{X=x_0}(Y=1|Z)} \right]. \quad [(13.2)]$$

Furthermore,

$$f_0(Z) + f_i(Z) = \text{logit}(P^{X=x_i}(Y=1|Z)) \quad [(15.37), (13.3)]$$

$$= \ln \left[ \frac{P^{X=x_i}(Y=1|Z)}{1 - P^{X=x_i}(Y=1|Z)} \right]. \quad [(13.2)]$$

▷ **Solution 15-3** Using Equations (13.6), (13.8), as well as the conditional probabilities  $P(Y=1|X=x)$  displayed in Table 15.2 yields

$$\alpha_0 = \text{logit}[P(Y=1|X=0)] \approx \ln(.732/(1-.732)) \approx 1.005,$$

$$\alpha_0 + \alpha_1 = \text{logit}[P(Y=1|X=1)] \approx \ln(.626/(1-.626)) \approx .515,$$

and  $\alpha_1 \approx .515 - 1.005 = -.490$ .

Similarly,

$$\lambda_0 = \text{logit}[P(Y=1|X=0, U=Joe)] \approx \ln(.732/(1-.732)) \approx 1.005,$$

$$\lambda_0 + \lambda_1 = \text{logit}[P(Y=1|X=1, U=Joe)] \approx \ln(.984/(1-.984)) \approx 4.119,$$

$$\lambda_0 + \lambda_2 = \text{logit}[P(Y=1|X=0, U=Ann)] \approx \ln(.732/(1-.732)) \approx 1.005,$$

$$\lambda_0 + \lambda_1 + \lambda_2 + \lambda_3 = \text{logit}[P(Y=1|X=1, U=Ann)] \approx \ln(.268/(1-.268)) \approx -1.005,$$

and this yields

$$\lambda_1 \approx 4.119 - 1.005 = 3.114,$$

$$\lambda_2 \approx 1.005 - 1.005 = 0,$$

$$\lambda_3 \approx -1.005 - 1.005 - 3.114 - 0 = -5.124.$$

Hence, the composition of  $U$  and the logit effect function  $f_1$  is

$$f_1(U) = \lambda_1 + \lambda_3 \cdot 1_{U=Ann} \approx 3.114 - 5.124 \cdot 1_{U=Ann},$$

and taking its expectation we receive

$$E_U(f_1) = E[f_1(U)] \quad [(6.13)]$$

$$\approx 3.114 - 5.124 \cdot .5 = .552. \quad [\text{Table 15.2, Box 6.1 (ii), (iii)}]$$



**Part IV**  
**Conditional Independence and**  
**Conditional Distribution**



# Chapter 16

## Conditional Independence

In section 4.3 we introduced independence of events and sets of events and in section 4.4 we treated conditional independence of events given an event that has a positive probability. There we noted that conditional independence given an event  $B$  is equivalent to independence with respect to the conditional-probability measure  $P^B$ . Furthermore, in section 5.4 we used these definitions in order to introduce independence of random variables. In this chapter we generalize these concepts and define conditional independence of events, of sets of events, and of random variables *given a  $\sigma$ -algebra or a random variable*. Further, we study the relationship between conditional independence and conditional mean independence (see section 10.6). For further implications of conditional independence see section 17.6.

### 16.1 Assumptions and Definitions

In this section we make the following assumptions and use the following notation.

#### Notation and Assumptions 16.1

Let  $Z: (\Omega, \mathcal{A}, P) \rightarrow (\Omega_Z^I, \mathcal{A}_Z^I)$  be a random variable, let  $\mathcal{C} \subset \mathcal{A}$  be a  $\sigma$ -algebra, and  $A, B \in \mathcal{A}$ .

Also remember,  $P(A|\mathcal{C}) := E(1_A|\mathcal{C})$ , where  $E(1_A|\mathcal{C})$  denotes the  $\mathcal{C}$ -conditional expectation of the indicator  $1_A$  of  $A \in \mathcal{A}$ , and  $P(A|Z) := E(1_A|Z)$  the  $Z$ -conditional probability of  $A$  (see Def. 10.2 and Rem. 10.4). Finally, see Remark 5.17 for the concept of  $P$ -equivalence and section 10.2 for propositions on  $P$ -equivalence of conditional expectations.

#### 16.1.1 Two Events

#### Definition 16.2 ( $\mathcal{C}$ -Conditional Independence of Two Events)

Let the assumptions 16.1 hold.

(i)  $A$  and  $B$  are called  $\mathcal{C}$ -conditionally  $P$ -independent, denoted  $A \perp\!\!\!\perp B | \mathcal{C}$ , if

$$P(A \cap B | \mathcal{C}) \stackrel{P}{=} P(A | \mathcal{C}) \cdot P(B | \mathcal{C}). \quad (16.1)$$

(ii)  $A$  and  $B$  are called  $Z$ -conditionally  $P$ -independent, denoted  $A \perp\!\!\!\perp B | Z$ , if

$$P(A \cap B | Z) \stackrel{P}{=} P(A | Z) \cdot P(B | Z). \quad (16.2)$$

A synonym for  $\mathcal{C}$ -conditional  $P$ -independence of  $A$  and  $B$  is *conditional independence of  $A$  and  $B$  given  $\mathcal{C}$  with respect to the probability measure  $P$* .

**Remark 16.3 (Symmetry)** Obviously,  $\mathcal{C}$ -conditional independence of  $A$  and  $B$  with respect to  $P$  is *symmetric* in the following sense:

$$A \perp\!\!\!\perp B | \mathcal{C} \Leftrightarrow B \perp\!\!\!\perp A | \mathcal{C}. \quad (16.3)$$

◁

Note that we did not exclude that the events  $A$  or  $B$  or both are elements of the  $\sigma$ -algebra  $\mathcal{C}$ . This case is considered in the following lemma.

**Lemma 16.4 (A Sufficient Condition)**

Let the assumptions 16.1 hold. Then

$$A \in \mathcal{C} \text{ or } B \in \mathcal{C} \Rightarrow A \perp\!\!\!\perp B | \mathcal{C} \quad (16.4)$$

and

$$A \in \sigma(Z) \text{ or } B \in \sigma(Z) \Rightarrow A \perp\!\!\!\perp B | Z. \quad (16.5)$$

(Proof p. 478)

**Remark 16.5 (An Immediate Implication)** If  $(\Omega, \mathcal{A}, P)$  is a probability space, then  $A \perp\!\!\!\perp B | \mathcal{A}$  for all  $A, B \in \mathcal{A}$ . ◁

**Corollary 16.6 (Two Formulations of  $Z$ -Conditional  $P$ -Independence)**

Let the assumptions 16.1 hold. Then the following propositions are equivalent to each other:

(i)  $A \perp\!\!\!\perp B | Z$ .

(ii)  $P(A \cap B | Z=z) = P(A | Z=z) \cdot P(B | Z=z)$ , for  $P_Z$ -a.a.  $z \in \Omega'_Z$ .

(Proof p. 479)

**Remark 16.7 (The Special Case  $P(Z=z) > 0$ )** Corollary 16.6 and Remark 2.71 imply: If  $A \perp\!\!\!\perp B \mid Z$  and  $P(Z=z) > 0$ , then

$$P(A \cap B \mid Z=z) = P(A \mid Z=z) \cdot P(B \mid Z=z). \quad (16.6)$$

According to Equation (10.30),  $P(A \mid Z=z) = P(A \mid \{Z=z\})$ , if  $P(Z=z) > 0$ . Therefore, Equation (16.6) is consistent with Equation (4.27).  $\triangleleft$

**Definition 16.8 (( $Z=z$ )-Conditional Independence of Two Events)**

Let the assumptions 16.1 hold and assume that  $z \in \Omega'_Z$  with  $P(Z=z) > 0$ . Then  $A$  and  $B$  are called ( $Z=z$ )-conditionally independent, denoted  $A \perp\!\!\!\perp B \mid Z=z$ , if Equation (16.6) holds.

**Remark 16.9 (Conditioning on a Constant)** Let the assumptions 16.1 hold. If  $\mathcal{C} = \{\Omega, \emptyset\}$ , then Remark 10.5 implies  $P(A \mid \mathcal{C}) = P(A)$ ,  $P(B \mid \mathcal{C}) = P(B)$ , and  $P(A \cap B \mid \mathcal{C}) = P(A \cap B)$ . Hence, in this case,  $\mathcal{C}$ -conditional independence and (unconditional) independence of two events are equivalent, i. e.,

$$\mathcal{C} = \{\Omega, \emptyset\} \Rightarrow (A \perp\!\!\!\perp B \mid \mathcal{C} \Leftrightarrow A \perp\!\!\!\perp B). \quad (16.7)$$

If  $Z = \alpha$ ,  $\alpha \in \Omega'_Z$ , i. e., if  $Z$  is a constant random variable, then  $\sigma(Z) = \{\Omega, \emptyset\}$ . Therefore, (16.7) implies

$$(\exists \alpha \in \Omega'_Z: Z = \alpha) \Rightarrow (A \perp\!\!\!\perp B \mid Z \Leftrightarrow A \perp\!\!\!\perp B). \quad (16.8)$$

A more general proposition is that independence and  $Z$ -conditional independence of two events are equivalent if  $Z \stackrel{P}{=} \alpha$ ,  $\alpha \in \Omega'_Z$ , i. e.,

$$(\exists \alpha \in \Omega'_Z: Z \stackrel{P}{=} \alpha) \Rightarrow (A \perp\!\!\!\perp B \mid Z \Leftrightarrow A \perp\!\!\!\perp B) \quad (16.9)$$

(see Exercise 16-1).  $\triangleleft$

### 16.1.2 Two Sets of Events

Using the concept of conditional independence of two events, we define  $\mathcal{C}$ -conditional  $P$ -independence of two sets of events. One or both of these sets of events can be a  $\sigma$ -algebra, but this is not required.

**Definition 16.10 ( $\mathcal{C}$ -Conditional Independence of Two Sets of Events)**

Let the assumptions 16.1 hold and let  $\mathcal{D}, \mathcal{E} \subset \mathcal{A}$ . Then:

- (i)  $\mathcal{D}$  and  $\mathcal{E}$  are called  $\mathcal{C}$ -conditionally  $P$ -independent, denoted  $\mathcal{D} \perp\!\!\!\perp \mathcal{E} \mid \mathcal{C}$ , if for all pairs  $(A, B) \in \mathcal{D} \times \mathcal{E}$ , the events  $A$  and  $B$  are  $\mathcal{C}$ -conditionally  $P$ -independent.

(ii)  $\mathcal{D}$  and  $\mathcal{E}$  are called *Z-conditionally P-independent*, denoted  $\mathcal{D} \perp\!\!\!\perp_P \mathcal{E} | Z$ , if for all pairs  $(A, B) \in \mathcal{D} \times \mathcal{E}$ , the events  $A$  and  $B$  are *Z-conditionally P-independent*.

**Remark 16.11 (Conditioning on a Constant)** Remark 16.9 implies

$$\mathcal{C} = \{\Omega, \emptyset\} \Rightarrow (\mathcal{D} \perp\!\!\!\perp_P \mathcal{E} | \mathcal{C} \Leftrightarrow \mathcal{D} \perp\!\!\!\perp_P \mathcal{E}). \quad (16.10)$$

Hence,  $\mathcal{D} \perp\!\!\!\perp_P \mathcal{E}$  (see Def. 4.40) is a special case of  $\mathcal{D} \perp\!\!\!\perp_P \mathcal{E} | \mathcal{C}$  with  $\mathcal{C} = \{\Omega, \emptyset\}$ . Similarly,

$$(\exists \alpha \in \Omega'_Z: Z = \alpha) \Rightarrow (\mathcal{D} \perp\!\!\!\perp_P \mathcal{E} | Z \Leftrightarrow \mathcal{D} \perp\!\!\!\perp_P \mathcal{E}). \quad (16.11)$$

Again, a more general proposition, which follows from (16.9), is that *independence* and *Z-conditional independence* of two sets of events are equivalent if  $Z = \bar{\alpha}$ ,  $\alpha \in \Omega'_Z$ , i. e.,

$$(\exists \alpha \in \Omega'_Z: Z = \bar{\alpha}) \Rightarrow (\mathcal{D} \perp\!\!\!\perp_P \mathcal{E} | Z \Leftrightarrow \mathcal{D} \perp\!\!\!\perp_P \mathcal{E}). \quad (16.12)$$

(see Exercise 16-2). ◁

According to the following lemma, two  $\sigma$ -algebras are conditionally independent if two  $\cap$ -stable set systems (see Def. 1.36) that generate these  $\sigma$ -algebras are conditionally independent.

**Lemma 16.12 ( $\cap$ -Stable Generators)**

Let the assumptions 16.1 hold, let  $\mathcal{D}, \mathcal{E} \subset \mathcal{A}$ , and suppose that  $\mathcal{D}$  and  $\mathcal{E}$  are  $\cap$ -stable. Then

$$\mathcal{D} \perp\!\!\!\perp_P \mathcal{E} | \mathcal{C} \Leftrightarrow \sigma(\mathcal{D}) \perp\!\!\!\perp_P \sigma(\mathcal{E}) | \mathcal{C}. \quad (16.13)$$

(Proof p. 479)

### 16.1.3 Two Random Variables

The sets of events  $\mathcal{D}$  and  $\mathcal{E}$  occurring in Definition 16.10 may also be two  $\sigma$ -algebras, such as  $\sigma(X)$  and  $\sigma(Y)$ , the  $\sigma$ -algebras generated by random variables  $X$  and  $Y$ , respectively (see Def. 2.26). This case will now be used to define *conditional P-independence of two random variables*.

**Definition 16.13 (Conditional P-Independence of Two Random Variables)**

Let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$ ,  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Y, \mathcal{A}'_Y)$ ,  $Z: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Z, \mathcal{A}'_Z)$  be random variables and  $\mathcal{C} \subset \mathcal{A}$  a  $\sigma$ -algebra. Then:

- (i)  $X$  and  $Y$  are called  $\mathcal{C}$ -conditionally  $P$ -independent, denoted  $X \perp\!\!\!\perp_P Y | \mathcal{C}$ , if  $\sigma(X)$  and  $\sigma(Y)$  are  $\mathcal{C}$ -conditionally  $P$ -independent.
- (ii)  $X$  and  $Y$  are called  $Z$ -conditionally  $P$ -independent, denoted  $X \perp\!\!\!\perp_P Y | Z$ , if  $\sigma(X)$  and  $\sigma(Y)$  are  $Z$ -conditionally  $P$ -independent.

**Remark 16.14 (Three Equivalent Notations)** Let  $\{X \in A'\} := \{\omega \in \Omega: X(\omega) \in A'\}$  and  $\{Y \in B'\} := \{\omega \in \Omega: Y(\omega) \in B'\}$  denote the events that  $X$  takes on a value in  $A'$  and  $Y$  takes on a value in  $B'$ , respectively. Then we use the following equivalent notations for  $\mathcal{C}$ -conditional independence of two random variables:

- (i)  $X \perp\!\!\!\perp_P Y | \mathcal{C}$   
(ii)  $\sigma(X) \perp\!\!\!\perp_P \sigma(Y) | \mathcal{C}$   
(iii)  $\{X \in A'\} \perp\!\!\!\perp_P \{Y \in B'\} | \mathcal{C}$ ,  $\forall (A', B') \in \mathcal{A}'_X \times \mathcal{A}'_Y$ .

Each of these notations is equivalent to

$$P(X \in A', Y \in B' | \mathcal{C}) \stackrel{P}{=} P(X \in A' | \mathcal{C}) \cdot P(Y \in B' | \mathcal{C}), \quad \forall (A', B') \in \mathcal{A}'_X \times \mathcal{A}'_Y. \quad (16.14)$$

Analogously,  $X \perp\!\!\!\perp_P Y | Z$  is equivalent to

$$P(X \in A', Y \in B' | Z) \stackrel{P}{=} P(X \in A' | Z) \cdot P(Y \in B' | Z), \quad \forall (A', B') \in \mathcal{A}'_X \times \mathcal{A}'_Y. \quad (16.15)$$

◁

**Remark 16.15 (An Implication for Random Variables)** The following proposition follows from Definition 16.13. If  $\mathcal{D}, \mathcal{E} \subset \mathcal{A}$ ,  $\sigma(X) \subset \mathcal{D}$ , and  $\sigma(Y) \subset \mathcal{E}$ , then

$$\mathcal{D} \perp\!\!\!\perp_P \mathcal{E} | \mathcal{C} \quad \Rightarrow \quad X \perp\!\!\!\perp_P Y | \mathcal{C}. \quad (16.16)$$

◁

## 16.2 Properties

Now we study some implications of conditional independence of two sets of events and of two random variables. According to Example 1.37, the set  $\{A\}$  is a  $\cap$ -stable generator of  $\sigma(\{A\}) = \{A, A^c, \Omega, \emptyset\}$ . Furthermore,  $\sigma(1_A) = \sigma(\{A\})$  (see Example 2.31). Therefore, Lemma 16.12, Definitions 16.2, 16.13, and  $1_{B^c} = 1 - 1_B$  imply the following corollary.

### Corollary 16.16 (Equivalent Propositions)

Let the assumptions 16.1 hold. Then the following propositions are equivalent to each other:

- (i)  $\{A\} \perp\!\!\!\perp_P \{B\} | \mathcal{C}$ .
- (ii)  $\sigma(\{A\}) \perp\!\!\!\perp_P \sigma(\{B\}) | \mathcal{C}$ .
- (iii)  $1_A \perp\!\!\!\perp_P 1_B | \mathcal{C}$ .
- (iv)  $A \perp\!\!\!\perp_P B | \mathcal{C}$ .
- (v)  $B \perp\!\!\!\perp_P A | \mathcal{C}$ .
- (vi)  $A \perp\!\!\!\perp_P B^c | \mathcal{C}$ .

Lemma 16.4 and Definition 16.13 imply the following corollary.

**Corollary 16.17 (Sub- $\sigma$ -Algebras)**

Let the assumptions 16.1 hold and let  $\mathcal{D}, \mathcal{E} \subset \mathcal{A}$ . Then

$$(\mathcal{D} \subset \mathcal{C} \vee \mathcal{E} \subset \mathcal{C}) \Rightarrow \mathcal{D} \perp\!\!\!\perp_P \mathcal{E} | \mathcal{C}. \quad (16.17)$$

Furthermore, if the assumptions of Definition 16.13 hold, then

$$(\sigma(X) \subset \sigma(Z) \vee \sigma(Y) \subset \sigma(Z)) \Rightarrow X \perp\!\!\!\perp_P Y | Z. \quad (16.18)$$

**Remark 16.18 (An Implication for Compositions)** Corollary 16.17 and Lemma 2.52 imply: If  $g(Z)$  is a composition of  $Z$  and an  $(\mathcal{A}'_Z, \overline{\mathcal{B}})$ -measurable function  $g: (\Omega'_Z, \mathcal{A}'_Z) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ , then  $X \perp\!\!\!\perp_P g(Z) | Z$ .  $\triangleleft$

Now we consider conditional independence of two discrete random variables. Remember, a random variable  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  is called *discrete* if there is a finite or countable set  $\Omega'_{X0} \subset \Omega'_X$  with  $P_X(\Omega'_{X0}) = 1$  and  $\{x\} \in \mathcal{A}'_X$  for all  $x \in \Omega'_{X0}$  (see Def. 5.56). If we consider two discrete random variables  $X, Y$  and adopt the corresponding notation for  $Y$ , then Equation (16.14) can be replaced by a simpler equation specified in the following corollary.

**Corollary 16.19 (Conditional Independence of Discrete Random Variables)**

Let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  and  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Y, \mathcal{A}'_Y)$  be discrete random variables and let  $\mathcal{C} \subset \mathcal{A}$  be a  $\sigma$ -algebra. Then  $X$  and  $Y$  are  $\mathcal{C}$ -conditionally  $P$ -independent, if and only if

$$P(X=x, Y=y | \mathcal{C}) \stackrel{=}{=} P(X=x | \mathcal{C}) \cdot P(Y=y | \mathcal{C}), \quad \forall (x, y) \in \Omega'_{X0} \times \Omega'_{Y0}. \quad (16.19)$$

(For a proof see the proof of the more general Corollary 16.47.)

**Box 16.1 Notation**

Let  $(\Omega, \mathcal{A}, P)$  be a probability space, let  $A, B \in \mathcal{A}$ , let  $\mathcal{C}, \mathcal{D}, \mathcal{E} \subset \mathcal{A}$ , where  $\mathcal{C}$  is a  $\sigma$ -algebra, and let  $X, Y, Z$  be random variables on  $(\Omega, \mathcal{A}, P)$ . Then the

symbol	denotes
$A \perp\!\!\!\perp B \mid \mathcal{C}$	$\mathcal{C}$ -conditional $P$ -independence of $A$ and $B$
$\mathcal{E} \perp\!\!\!\perp \mathcal{D} \mid \mathcal{C}$	$\mathcal{C}$ -conditional $P$ -independence of $\mathcal{E}$ and $\mathcal{D}$
$X \perp\!\!\!\perp Y \mid \mathcal{C}$	$\mathcal{C}$ -conditional $P$ -independence of $X$ and $Y$
$X \perp\!\!\!\perp \mathcal{D} \mid \mathcal{C}$	$\mathcal{C}$ -conditional $P$ -independence of $X$ and $\mathcal{D}$ , i. e., $\sigma(X) \perp\!\!\!\perp \mathcal{D} \mid \mathcal{C}$
$A \perp\!\!\!\perp B \mid Z$	$Z$ -conditional $P$ -independence of $A$ and $B$ , i. e., $A \perp\!\!\!\perp B \mid \sigma(Z)$
$\mathcal{E} \perp\!\!\!\perp \mathcal{D} \mid Z$	$Z$ -conditional $P$ -independence of $\mathcal{E}$ and $\mathcal{D}$ , i. e., $\mathcal{E} \perp\!\!\!\perp \mathcal{D} \mid \sigma(Z)$

Equation (16.19) only refers to all pairs  $(x, y) \in \Omega'_{X0} \times \Omega'_{Y0}$ , whereas (16.14) refers to all pairs  $(A', B')$  of elements in the  $\sigma$ -algebras  $\mathcal{A}'_X$  and  $\mathcal{A}'_Y$ , respectively.

**Remark 16.20 (Conditioning on a Random Variable)** If  $Z: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Z, \mathcal{A}'_Z)$ , then (16.19) can be written

$$P(X=x, Y=y \mid Z) \stackrel{\text{def}}{=} P(X=x \mid Z) \cdot P(Y=y \mid Z), \quad \forall (x, y) \in \Omega'_{X0} \times \Omega'_{Y0}. \quad (16.20)$$

◁

**Example 16.21 (Joe and Ann With no Individual Treatment Effects)** Table 16.1 displays a new example with Joe and Ann. As before, the random experiment consists of drawing Joe or Ann, each one with probability .50, observing whether  $(X=1)$  or not  $(X=0)$  the drawn person receives treatment, and whether  $(Y=1)$  or not  $(Y=0)$  a success criterion is reached. In this example,  $X$  and  $Y$  are  $U$ -conditionally independent with respect to  $P$ . In other words, conditioning on one of the two persons,  $X$  and  $Y$  are  $P$ -independent. Intuitively speaking, this means that there are no individual treatment effects.

According to Definition 16.13 we have to show that the two  $\sigma$ -algebras  $\sigma(X)$  and  $\sigma(Y)$  are  $U$ -conditionally  $P$ -independent, where  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  and  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Y, \mathcal{A}'_Y)$  with  $\Omega'_X = \Omega'_Y = \{0, 1\}$  and  $\mathcal{A}'_X = \mathcal{A}'_Y = \{\{0, 1\}, \emptyset, \{0\}, \{1\}\}$ . In other words, we have to show

$$P(X \in A', Y \in B' \mid U) = P(X \in A' \mid U) \cdot P(Y \in B' \mid U), \quad \forall (A', B') \in \mathcal{A}'_X \times \mathcal{A}'_Y. \quad (16.21)$$

According to Corollary 16.16 it suffices to consider  $A' = B' = \{1\}$ , using the notations  $P(X \in A' \mid U) = P(X=1 \mid U)$ ,  $P(Y \in B' \mid U) = P(Y=1 \mid U)$ , and  $P(X \in A', Y \in B' \mid U) =$

**Table 16.1.** Joe and Ann With no Individual Treatment Effects

Outcomes $\omega$			Observables			Conditional Probabilities					
Unit	Treatment	Success									
			$P(\{\omega\})$	Person variable $U$	Treatment variable $X$	Outcome variable $Y$	$P(Y=1 X, U) = E(Y X, U)$	$P(Y=1 X) = E(Y X)$	$P(Y=1 U) = E(Y U)$	$P(X=1 U) = E(X U)$	$P(X=1, Y=1 U)$
(Joe, no, -)			.04	Joe	0	0	.60	.76	.60	3/4	9/20
(Joe, no, +)			.06	Joe	0	1	.60	.76	.60	3/4	9/20
(Joe, yes, -)			.12	Joe	1	0	.60	.68	.60	3/4	9/20
(Joe, yes, +)			.18	Joe	1	1	.60	.68	.60	3/4	9/20
(Ann, no, -)			.08	Ann	0	0	.80	.76	.80	1/3	8/30
(Ann, no, +)			.32	Ann	0	1	.80	.76	.80	1/3	8/30
(Ann, yes, -)			.04	Ann	1	0	.80	.68	.80	1/3	8/30
(Ann, yes, +)			.16	Ann	1	1	.80	.68	.80	1/3	8/30

$P(X=1, Y=1|U)$ . Looking at the last three columns of conditional probabilities in Table 16.1 shows that, in this example,

$$P(X=1, Y=1|U) = P(X=1|U) \cdot P(Y=1|U). \tag{16.22}$$

◁

The following theorem adds another important property of conditional independence that is also useful in some proofs.

**Theorem 16.22 (A Property Equivalent to Conditional P-Independence)**

Let  $(\Omega, \mathcal{A}, P)$  be a probability space, let  $\mathcal{C}, \mathcal{D}, \mathcal{E} \subset \mathcal{A}$ , and let  $\mathcal{C}, \mathcal{D}$  be  $\sigma$ -algebras. Then

$$\mathcal{E} \perp\!\!\!\perp_P \mathcal{D} | \mathcal{C} \iff \forall A \in \mathcal{E}: P(A|\mathcal{C}, \mathcal{D}) \stackrel{P}{=} P(A|\mathcal{C}). \tag{16.23}$$

(Proof p. 480)

In the following remark we use the notation  $E(Y|\mathcal{C}, Z) := E(Y|\sigma[\mathcal{C} \cup \sigma(Z)])$  for a version of the  $\sigma[\mathcal{C} \cup \sigma(Z)]$ -conditional expectation of  $Y$  and  $P(A|\mathcal{C}, Z) := P(A|\sigma[\mathcal{C} \cup \sigma(Z)])$  for a version of the  $\sigma[\mathcal{C} \cup \sigma(Z)]$ -conditional probability of an event  $A \in \mathcal{A}$ .

**Remark 16.23 (Sufficient Condition for Conditional Independence of Events)**

An immediate implication of Theorem 16.22 is

$$\forall A, B \in \mathcal{A}: \left( P(A|1_B, \mathcal{C}) \stackrel{P}{=} P(A|\mathcal{C}) \Rightarrow A \perp\!\!\!\perp B | \mathcal{C} \right). \quad (16.24)$$

This proposition may also be written:

$$\forall A, B \in \mathcal{A}: \left( E(1_A|1_B, \mathcal{C}) \stackrel{P}{=} E(1_A|\mathcal{C}) \Rightarrow A \perp\!\!\!\perp B | \mathcal{C} \right). \quad (16.25)$$

◁

The following corollary follows from Theorem 16.22 for  $A := \{Y=y\}$ .

**Corollary 16.24 (An Implication of Conditional Independence)**

Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Y, \mathcal{A}'_Y)$  be a random variable with  $\{y\} \in \mathcal{A}'_Y$  for all  $y \in \Omega'_Y$ , and let  $\mathcal{C}, \mathcal{D} \subset \mathcal{A}$  be  $\sigma$ -algebras. Then

$$Y \perp\!\!\!\perp \mathcal{D} | \mathcal{C} \Rightarrow \forall y \in \Omega'_Y: P(Y=y|\mathcal{C}, \mathcal{D}) \stackrel{P}{=} P(Y=y|\mathcal{C}). \quad (16.26)$$

**Remark 16.25 (Rewriting the Corollary for Random Variables)** In addition to the assumptions of Corollary 16.24, let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  and  $Z: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Z, \mathcal{A}'_Z)$  be random variables. Then

$$Y \perp\!\!\!\perp X | Z \Rightarrow \forall y \in \Omega'_Y: P(Y=y|X, Z) \stackrel{P}{=} P(Y=y|Z). \quad (16.27)$$

◁

Now we consider a special case of Theorem 16.22 for a *discrete* random variable. This means that we consider a random variable  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Y, \mathcal{A}'_Y)$  such that there is a finite or countable set  $\Omega'_{Y0} \subset \Omega'_Y$  with  $P_Y(\Omega'_{Y0}) = 1$  and  $\{y\} \in \mathcal{A}'_Y$  for all  $y \in \Omega'_{Y0}$  (see Def. 5.56). In this theorem we also use the notation  $Y \perp\!\!\!\perp \mathcal{D} | \mathcal{C}$  introduced in Box 16.1.

**Theorem 16.26 (An Property Equivalent to of Conditional Independence)**

Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Y, \mathcal{A}'_Y)$  be a discrete random variable and let  $\mathcal{C}, \mathcal{D} \subset \mathcal{A}$  be  $\sigma$ -algebras. Then

$$Y \perp\!\!\!\perp \mathcal{D} | \mathcal{C} \Leftrightarrow \forall y \in \Omega'_{Y0}: P(Y=y|\mathcal{C}, \mathcal{D}) \stackrel{P}{=} P(Y=y|\mathcal{C}). \quad (16.28)$$

(Proof p. 480)

**Remark 16.27 (Rewriting the Theorem for Random Variables)** Suppose that  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$ ,  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Y, \mathcal{A}'_Y)$ , and  $Z: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Z, \mathcal{A}'_Z)$  are random variables, and that  $Y$  is discrete. Then for  $\mathcal{D} = \sigma(X)$  and  $\mathcal{C} = \sigma(Z)$ , Theorem 16.26 immediately yields:

$$Y \perp\!\!\!\perp X | Z \Leftrightarrow \forall y \in \Omega'_{Y0}: P(Y=y|X, Z) \stackrel{P}{=} P(Y=y|Z). \quad (16.29)$$

◁

Box 16.2 summarizes some important propositions on conditional independence of  $\sigma$ -algebras, and Box 16.3 translates these propositions to conditional independence of random variables. Proofs are provided in Exercise 16-3.

Considering a value  $z \in \Omega'_Z$  for which  $P(Z=z) > 0$ , some rules of Box 16.2 may also have implications for conditional independence with respect to a conditional-probability measure  $P^{Z=z}$ . For example, if  $X$  is a numerical  $Z$ -measurable function, then according to Lemma 2.52, there is a measurable function  $g: (\bar{\mathbb{R}}, \bar{\mathcal{B}}) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  such that  $X = g(Z)$ . This implies that  $X \stackrel{P^{Z=z}}{=} g(z)$  (see Rem. 9.1). Therefore, for  $\alpha = g(z)$ , Box 16.2 (xii) yields the following corollary.

**Corollary 16.28 (Measurability and Conditional Independence w.r.t.  $P^{Z=z}$ )**

Let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\bar{\mathbb{R}}, \bar{\mathcal{B}})$  and  $Z: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Z, \mathcal{A}'_Z)$  be random variables, let  $P(Z=z) > 0$ , and  $\mathcal{C}, \mathcal{E} \subset \mathcal{A}$ , where  $\mathcal{C}$  is a  $\sigma$ -algebra. Then

$$X \text{ is } Z\text{-measurable} \Rightarrow X \underset{P^{Z=z}}{\perp\!\!\!\perp} \mathcal{E} \mid \mathcal{C}. \quad (16.30)$$

**Remark 16.29 ( $Z$ -Measurability and Independence w.r.t.  $P^{Z=z}$ )** For  $\mathcal{C} = \{\Omega, \emptyset\}$ , Corollary 16.28 implies that  $\mathcal{E}$  and  $X$  are independent with respect to  $P^{Z=z}$ , i. e.,

$$X \text{ is } Z\text{-measurable} \Rightarrow X \underset{P^{Z=z}}{\perp\!\!\!\perp} \mathcal{E}. \quad (16.31)$$

This result is also an immediate implication of Lemmas 2.52 and 5.51.  $\triangleleft$

**Example 16.30 (Conditional Independence Treatment and Person Variables)** Table

16.2 displays some parameters of a new random experiment. In this experiment a person (unit) is drawn from a set of six persons with a sampling probability  $P(U=u) = 1/6$  for each person. There are four males ( $Z=m$ ) and two females ( $Z=f$ ). If a male person is drawn, then he receives treatment ( $X=1$ ) with probability  $P(X=1 \mid U=u) = 3/4$ , if a female person is drawn, then she gets treatment with probability  $P(X=1 \mid U=u) = 1/4$ . The table also displays the conditional expectation values  $E(Y \mid X=0, U=u)$  and  $E(Y \mid X=1, U=u)$  of a real-valued outcome variable  $Y$ .

In this example,

$$P(X=1 \mid Z, U) = P(X=1 \mid Z) = \frac{3}{4} \cdot 1_{Z=m} + \frac{1}{4} \cdot 1_{Z=f}.$$

However, if the first of these two equations holds, then  $P(X=0 \mid Z, U) = P(X=0 \mid Z)$  holds as well, because  $P(X=0 \mid Z) = 1 - P(X=1 \mid Z)$  and  $P(X=0 \mid Z, U) = 1 - P(X=1 \mid Z, U)$  (see again Exercise 16-4). Therefore,

$$P(X=x \mid Z, U) = P(X=x \mid Z), \quad x = 0, 1, \quad (16.32)$$

and, according to (16.29), the treatment variable  $X$  and the person variable  $U$  are conditionally independent given  $Z$  (sex).  $\triangleleft$

**Box 16.2 Conditional Independence of Set Systems**

Let  $(\Omega, \mathcal{A}, P)$  be a probability space,  $\mathcal{C}, \mathcal{D}, \mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3 \subset \mathcal{A}$  be  $\sigma$ -algebras, and  $\mathcal{E}, \mathcal{F}, \mathcal{G} \subset \mathcal{A}$ . Then:

$$\mathcal{E} \perp\!\!\!\perp \mathcal{F} \mid \mathcal{C} \Leftrightarrow \mathcal{F} \perp\!\!\!\perp \mathcal{E} \mid \mathcal{C} \quad (\text{i})$$

$$\mathcal{E} \perp\!\!\!\perp \sigma(\mathcal{F}, \mathcal{G}) \mid \mathcal{C} \Rightarrow \mathcal{E} \perp\!\!\!\perp \mathcal{F} \mid \mathcal{C} \wedge \mathcal{E} \perp\!\!\!\perp \mathcal{G} \mid \mathcal{C} \quad (\text{ii})$$

$$\perp\!\!\!\perp \mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3 \Rightarrow \mathcal{D}_1 \perp\!\!\!\perp \mathcal{D}_2 \mid \mathcal{D}_3 \quad (\text{iii})$$

$$\mathcal{F} \subset \mathcal{C} \Rightarrow \mathcal{E} \perp\!\!\!\perp \mathcal{F} \mid \mathcal{C} \quad (\text{iv})$$

$$\mathcal{F} = \{\Omega, \emptyset\} \Rightarrow \mathcal{E} \perp\!\!\!\perp \mathcal{F} \mid \mathcal{C} \quad (\text{v})$$

$$\mathcal{E} \perp\!\!\!\perp \mathcal{F} \mid \mathcal{C} \wedge \mathcal{G} \subset \mathcal{F} \Rightarrow \mathcal{E} \perp\!\!\!\perp \mathcal{G} \mid \mathcal{C} \quad (\text{vi})$$

$$\mathcal{D}_1 \perp\!\!\!\perp \mathcal{D}_2 \mid \mathcal{C} \Leftrightarrow \mathcal{D}_1 \perp\!\!\!\perp \sigma(\mathcal{C}, \mathcal{D}_2) \mid \mathcal{C} \quad (\text{vii})$$

$$\mathcal{E} \perp\!\!\!\perp \sigma(\mathcal{D}_1, \mathcal{D}_2) \mid \mathcal{C} \Leftrightarrow \mathcal{E} \perp\!\!\!\perp \mathcal{D}_1 \mid \mathcal{C} \wedge \mathcal{E} \perp\!\!\!\perp \mathcal{D}_2 \mid \sigma(\mathcal{C}, \mathcal{D}_1) \quad (\text{viii})$$

$$\mathcal{E} \perp\!\!\!\perp \sigma(\mathcal{D}_1, \mathcal{D}_2) \Leftrightarrow \mathcal{E} \perp\!\!\!\perp \mathcal{D}_1 \wedge \mathcal{E} \perp\!\!\!\perp \mathcal{D}_2 \mid \mathcal{D}_1. \quad (\text{ix})$$

Let  $\mathcal{F}_0 \subset \mathcal{F}$ ,  $\mathcal{G}_0 \subset \mathcal{G}$ , and  $\mathcal{F}_0, \mathcal{F}, \mathcal{G}_0, \mathcal{G} \subset \mathcal{A}$  be  $\sigma$ -algebras. Then:

$$\mathcal{F} \perp\!\!\!\perp \mathcal{G} \mid \mathcal{C} \Rightarrow \mathcal{F} \perp\!\!\!\perp \mathcal{G} \mid \sigma(\mathcal{C}, \mathcal{F}_0, \mathcal{G}_0). \quad (\text{x})$$

Let  $X_1, X_2, X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  be random variables. Then:

$$X_1 \stackrel{P}{=} X_2 \wedge X_1 \perp\!\!\!\perp \mathcal{E} \mid \mathcal{C} \Rightarrow X_2 \perp\!\!\!\perp \mathcal{E} \mid \mathcal{C} \quad (\text{xi})$$

$$X \stackrel{P}{=} \alpha, \alpha \in \Omega'_X \Rightarrow X \perp\!\!\!\perp \mathcal{E} \mid \mathcal{C}. \quad (\text{xii})$$

**Example 16.31 (Joe and Ann With no Individual Treatment Effects—continued)**

In Example 16.21 we presented Table 16.1 and showed that  $X$  and  $Y$  are  $U$ -conditionally independent. According to Theorem 16.26

$$P(Y=y|X, U) \stackrel{P}{=} P(Y=y|U), \quad \forall y = 0, 1,$$

is equivalent to  $U$ -conditional independence of  $X$  and  $Y$ . Comparing the columns headed by  $P(Y=1|X, U)$  and  $P(Y=1|U)$  in Table 16.1 confirms that this condition is satisfied for  $y=1$ . However, this implies that  $P(Y=0|X, U) \stackrel{P}{=} P(Y=0|U)$  holds as well (see Exercise 16-4).  $\triangleleft$

**Box 16.3 Conditional Independence of Random Variables**

Let  $W, X, X_1, X_2, X_3, Y,$  and  $Z$  be random variables on the probability space  $(\Omega, \mathcal{A}, P)$ . Then:

$$X \perp\!\!\!\perp_p Y | Z \Leftrightarrow Y \perp\!\!\!\perp_p X | Z \quad (\text{i})$$

$$X \perp\!\!\!\perp_p (W, Y) | Z \Rightarrow X \perp\!\!\!\perp_p W | Z \wedge X \perp\!\!\!\perp_p Y | Z \quad (\text{ii})$$

$$\perp\!\!\!\perp_{i=1}^3 X_i \Rightarrow X_1 \perp\!\!\!\perp_p X_2 | X_3 \quad (\text{iii})$$

$$\sigma(Y) \subset \sigma(Z) \Rightarrow X \perp\!\!\!\perp_p Y | Z \quad (\text{iv})$$

$$X = \alpha, \alpha \in \Omega'_X \Rightarrow X \perp\!\!\!\perp_p Y | Z \quad (\text{v})$$

$$X \perp\!\!\!\perp_p Y | Z \wedge \sigma(W) \subset \sigma(Y) \Rightarrow X \perp\!\!\!\perp_p W | Z \quad (\text{vi})$$

$$X \perp\!\!\!\perp_p Y | Z \Leftrightarrow X \perp\!\!\!\perp_p (Z, Y) | Z \quad (\text{vii})$$

$$X \perp\!\!\!\perp_p (W, Y) | Z \Leftrightarrow X \perp\!\!\!\perp_p Y | Z \wedge X \perp\!\!\!\perp_p W | (Z, Y) \quad (\text{viii})$$

$$X \perp\!\!\!\perp_p (W, Y) \Leftrightarrow X \perp\!\!\!\perp_p Y \wedge X \perp\!\!\!\perp_p W | Y. \quad (\text{ix})$$

If  $Y_0$  and  $W_0$  are random variables on  $(\Omega, \mathcal{A}, P)$  that are measurable with respect to  $Y$  and  $W$ , respectively, then

$$W \perp\!\!\!\perp_p Y | Z \Rightarrow W \perp\!\!\!\perp_p Y | (Z, Y_0, W_0) \quad (\text{x})$$

$$X_1 \stackrel{=}{\perp\!\!\!\perp}_p X_2 \wedge X_1 \perp\!\!\!\perp_p Y | Z \Rightarrow X_2 \perp\!\!\!\perp_p Y | Z \quad (\text{xi})$$

$$X \stackrel{=}{\perp\!\!\!\perp}_p \alpha, \alpha \in \Omega'_X \Rightarrow X \perp\!\!\!\perp_p Y | Z. \quad (\text{xii})$$

### 16.3 Conditional Independence and Conditional Mean Independence

Now we generalize the product rule for the expectations of independent random variables [see Box 6.1 (x) and Proposition (16.7)] using the notation  $X \perp\!\!\!\perp_p Y | \mathcal{C}$  introduced in Definition 16.13.

#### Theorem 16.32 (Product Rule Under Conditional $P$ -Independence)

Let  $X, Y: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  be random variables that are both nonnegative or both with finite expectations, and let  $\mathcal{C} \subset \mathcal{A}$  be a  $\sigma$ -algebra. Then

$$X \perp\!\!\!\perp_p Y | \mathcal{C} \Rightarrow E(X \cdot Y | \mathcal{C}) \stackrel{=}{\perp\!\!\!\perp}_p E(X | \mathcal{C}) \cdot E(Y | \mathcal{C}). \quad (16.33)$$

(Proof p. 481)

**Table 16.2.** Z-Conditional Independence of X and U

Persons $u$	Conditional expectation values					
	Person variable $U$	$P(U=u)$	Sex $Z$	$E(Y X=0, U=u)$	$E(Y X=1, U=u)$	$P(X=1 U=u)$
$u_1$		1/6	$m$	68	81	3/4
$u_2$		1/6	$m$	78	86	3/4
$u_3$		1/6	$m$	88	100	3/4
$u_4$		1/6	$m$	98	103	3/4
$u_5$		1/6	$f$	106	114	1/4
$u_6$		1/6	$f$	116	130	1/4

**Remark 16.33 (Conditioning on a Random Variable)** If  $Z: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Z, \mathcal{A}'_Z)$  is a random variable, then, under the assumptions of Theorem 16.32,

$$X \perp\!\!\!\perp Y | Z \Rightarrow E(X \cdot Y | Z) \stackrel{P}{=} E(X | Z) \cdot E(Y | Z). \tag{16.34}$$

Note that  $E(X \cdot Y | Z) \stackrel{P}{=} E(X | Z) \cdot E(Y | Z)$  is equivalent to  $Cov(X, Y | Z) \stackrel{P}{=} 0$ , provided that  $E(X^2)$  and  $E(Y^2)$  are finite [see Rule (i) of Box 11.2].  $\triangleleft$

In section 10.6 we introduced the concept of conditional mean independence. Now we consider its relationship to conditional independence.

**Theorem 16.34 (Conditional Mean Independence)**

Let  $(\Omega, \mathcal{A}, P)$  be a probability space, let  $\mathcal{C}, \mathcal{D}, \mathcal{E} \subset \mathcal{A}$  be  $\sigma$ -algebras, and let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  be a random variable that is nonnegative or has a finite expectation  $E(Y)$ . If  $\sigma(Y) \subset \mathcal{E}$ , then

$$\mathcal{E} \perp\!\!\!\perp \mathcal{D} | \mathcal{C} \Rightarrow E(Y | \mathcal{C}, \mathcal{D}) \stackrel{P}{=} E(Y | \mathcal{C}). \tag{16.35}$$

(Proof p. 482)

An immediate implication is

$$Y \perp\!\!\!\perp \mathcal{D} | \mathcal{C} \Rightarrow E(Y | \mathcal{C}, \mathcal{D}) \stackrel{P}{=} E(Y | \mathcal{C}). \tag{16.36}$$

**Remark 16.35 (Conditioning on Random Variables)** Another implication is: If  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$ ,  $Z: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Z, \mathcal{A}'_Z)$  are random variables and  $Y: (\Omega, \mathcal{A}, P) \rightarrow$

$(\overline{\mathbb{R}}, \overline{\mathcal{B}})$  is a random variable that is nonnegative or has a finite expectation  $E(Y)$ , then

$$Y \perp\!\!\!\perp_p X|Z \Rightarrow E(Y|X, Z) \stackrel{\overline{p}}{=} E(Y|Z). \quad (16.37)$$

◁

**Remark 16.36 (Implications of Independence)** For  $Z$  being a constant, the last remark yields

$$Y \perp\!\!\!\perp_p X \Rightarrow E(Y|X) \stackrel{\overline{p}}{=} E(Y) \quad (16.38)$$

[cf. Box 10.2 (vi)].

◁

While Theorem 16.34 deals with an implication of conditional independence on conditional mean independence, now we present some conditions that are equivalent to conditional independence.

**Theorem 16.37 (Characterizations of Conditional Independence)**

Let  $(\Omega, \mathcal{A}, P)$  be a probability space, let  $\mathcal{C}, \mathcal{D}, \mathcal{E} \subset \mathcal{A}$  be  $\sigma$ -algebras, and let  $W, Y: (\Omega, \mathcal{A}, P) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  be real-valued random variables. Then the following propositions are equivalent to each other:

- (i)  $\mathcal{E} \perp\!\!\!\perp_p \mathcal{D}|\mathcal{C}$ .
- (ii)  $E(Y|\mathcal{C}, \mathcal{D}) \stackrel{\overline{p}}{=} E(Y|\mathcal{C})$ , for all nonnegative  $Y$  with finite expectation and  $\sigma(Y) \subset \mathcal{E}$ .
- (iii)  $E(W \cdot Y|\mathcal{C}) \stackrel{\overline{p}}{=} E(W|\mathcal{C}) \cdot E(Y|\mathcal{C})$ , for all nonnegative  $Y$  with finite expectation and  $\sigma(Y) \subset \mathcal{E}$  and all nonnegative  $W$  with  $\sigma(W) \subset \mathcal{D}$ .
- (iv)  $E[E(Y|\mathcal{C})|\mathcal{D}] \stackrel{\overline{p}}{=} E(Y|\mathcal{D})$ , for all nonnegative  $Y$  with finite expectation and  $\sigma(Y) \subset \sigma(\mathcal{C}, \mathcal{E})$ .

(Proof p. 483)

In the next theorem we assume  $Z \perp\!\!\!\perp_p Y|\mathcal{C}$  and present an implication for the conditional expectation  $E^{Z=z}(Y|\mathcal{C})$  (see ch. 14).

**Theorem 16.38 (An Implication of  $\mathcal{C}$ -Conditional  $P$ -Independence)**

Let  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  and  $Z: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Z, \mathcal{A}'_Z)$  be random variables, let  $z \in \Omega'_Z$  with  $\{z\} \in \mathcal{A}'_Z$  and  $P(Z=z) > 0$ , and let  $Y$  be nonnegative or such that  $E^{Z=z}(Y) = \int Y dP^{Z=z}$  is finite. Furthermore, let  $\mathcal{C} \subset \mathcal{A}$  be a  $\sigma$ -algebra. If  $Z \perp\!\!\!\perp_p Y|\mathcal{C}$ , then  $\mathcal{E}(Y|\mathcal{C}) \subset \mathcal{E}^{Z=z}(Y|\mathcal{C})$ , and therefore,

$$E^{Z=z}(Y|\mathcal{C}) \stackrel{\overline{p^{Z=z}}}{=} E(Y|\mathcal{C}). \quad (16.39)$$

(Proof p. 483)

Note that, even if  $Z \perp\!\!\!\perp_p Y|\mathcal{C}$ , then  $E^{Z=z}(Y|\mathcal{C})$  is not necessarily an element of  $\mathcal{E}(Y|\mathcal{C})$ , because Equation (16.39) does not imply  $E^{Z=z}(Y|\mathcal{C}) \stackrel{\overline{p}}{=} E(Y|\mathcal{C})$ . For  $\mathcal{C} = \{\Omega, \emptyset\}$ , Theorem 16.38 implies the following corollary:

**Corollary 16.39 (An Implication of  $P$ -Independence)**

Let the assumptions of Theorem 16.38 hold. If  $Z \perp\!\!\!\perp_p Y$ , then

$$E^{Z=z}(Y) = E(Y). \quad (16.40)$$

**16.4 Families of Events**

Now we extend the concept of conditional independence to more than two events, more than two sets of events, and more than two random variables.

Three events  $A_1, A_2, A_3$  are called ' $\mathcal{C}$ -conditionally  $P$ -independent', if

$$P(A_i \cap A_j | \mathcal{C}) \stackrel{p}{=} P(A_i | \mathcal{C}) \cdot P(A_j | \mathcal{C}), \quad \forall i, j = 1, 2, 3, \quad i \neq j, \quad (16.41)$$

(pairwise conditional independence) and

$$P(A_1 \cap A_2 \cap A_3 | \mathcal{C}) \stackrel{p}{=} P(A_1 | \mathcal{C}) \cdot P(A_2 | \mathcal{C}) \cdot P(A_3 | \mathcal{C}) \quad (16.42)$$

(triple-wise conditional independence). We use the notation

$$\perp\!\!\!\perp_p A_1, A_2, A_3 | \mathcal{C}$$

for  $\mathcal{C}$ -conditional  $P$ -independence of  $A_1, A_2, A_3$ . Four events are conditionally  $P$ -independent if the corresponding product rule holds for all pairs, all triples, and for the quadruple. The general definition is as follows:

**Definition 16.40 (Conditional Independence of a Family of Events)**

Let  $(\Omega, \mathcal{A}, P)$  be a probability space,  $I$  be a nonempty set,  $A_i \in \mathcal{A}$ , for all  $i \in I$ , and  $\mathcal{C} \subset \mathcal{A}$  be a  $\sigma$ -algebra. Then

- (i)  $(A_i, i \in I)$  is called a family of  $\mathcal{C}$ -conditionally  $P$ -independent events, denoted  $\perp\!\!\!\perp_p (A_i, i \in I) | \mathcal{C}$ , if

$$P\left(\bigcap_{i \in J} A_i \mid \mathcal{C}\right) \stackrel{p}{=} \prod_{i \in J} P(A_i | \mathcal{C}), \quad \text{for all finite sets } J \subset I. \quad (16.43)$$

- (ii) Let  $Z: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Z, \mathcal{A}'_Z)$  be a random variable. Then  $(A_i, i \in I)$  is called a family of  $Z$ -conditionally independent events, denoted  $\perp\!\!\!\perp_p (A_i, i \in I) | Z$ , if (16.43) holds for  $\mathcal{C} = \sigma(Z)$ .

**16.5 Families of Set Systems**

Using the concept of conditional  $P$ -independence of families of events, we can now define *conditional  $P$ -independence of families of set systems* or of sets of

events. Three sets  $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$  of events are called  $\mathcal{C}$ -conditionally  $P$ -independent if, for all  $i, j = 1, 2, 3, i \neq j$ ,

$$P(A_i \cap A_j | \mathcal{C}) \stackrel{=}{=} P(A_i | \mathcal{C}) \cdot P(A_j | \mathcal{C}), \quad \forall (A_i, A_j) \in \mathcal{E}_i \times \mathcal{E}_j, \quad (16.44)$$

and

$$P(A_1 \cap A_2 \cap A_3 | \mathcal{C}) \stackrel{=}{=} P(A_1 | \mathcal{C}) \cdot P(A_2 | \mathcal{C}) \cdot P(A_3 | \mathcal{C}), \quad (16.45)$$

$$\forall (A_1, A_2, A_3) \in \mathcal{E}_1 \times \mathcal{E}_2 \times \mathcal{E}_3.$$

The general definition for a family  $(\mathcal{E}_i, i \in I)$  of sets of events is as follows:

**Definition 16.41 (Conditional Independence of a Family of Sets of Events)**

Let  $(\Omega, \mathcal{A}, P)$  be a probability space,  $I$  be a nonempty set, and  $\mathcal{E}_i \subset \mathcal{A}, i \in I$ .

- (i) Then  $(\mathcal{E}_i, i \in I)$  is called a family of  $\mathcal{C}$ -conditionally  $P$ -independent sets of events, denoted  $\perp\!\!\!\perp_P(\mathcal{E}_i, i \in I) | \mathcal{C}$ , if each family  $(A_i, i \in I)$  of events  $A_i \in \mathcal{E}_i, i \in I$ , is  $\mathcal{C}$ -conditionally  $P$ -independent.
- (ii) Furthermore, let  $Z: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Z, \mathcal{A}'_Z)$  be a random variable. Then  $(\mathcal{E}_i, i \in I)$  is called a family of  $Z$ -conditionally  $P$ -independent sets of events, denoted  $\perp\!\!\!\perp_P(\mathcal{E}_i, i \in I) | Z$ , if each family  $(A_i, i \in I)$  of events  $A_i \in \mathcal{E}_i, i \in I$ , is  $Z$ -conditionally  $P$ -independent.

**Remark 16.42 (Sub-Families)** This definition immediately implies:

$$\perp\!\!\!\perp_P(\mathcal{E}_i, i \in I) | \mathcal{C} \Leftrightarrow \forall J \subset I: \perp\!\!\!\perp_P(\mathcal{E}_i, i \in J) | \mathcal{C}. \quad (16.46)$$

◁

**Remark 16.43 (Smallest  $\sigma$ -Algebra)** If  $\mathcal{C} = \{\Omega, \emptyset\}$ , then

$$\perp\!\!\!\perp_P(\mathcal{E}_i, i \in I) | \mathcal{C} \Leftrightarrow \perp\!\!\!\perp_P(\mathcal{E}_i, i \in I) \quad (16.47)$$

[see Eq. (10.2) and Remark 10.5]. Hence,  $P$ -independence of families of events (see Def. 4.40) is a special case of  $\mathcal{C}$ -conditional  $P$ -independence of families of events. ◁

**Theorem 16.44 (Conditional Independence of  $\cap$ -Stable Set Systems)**

If  $(\Omega, \mathcal{A}, P)$  is a probability space,  $\mathcal{C} \subset \mathcal{A}$  a  $\sigma$ -algebra, and  $(\mathcal{E}_i, i \in I)$ , is a family of  $\cap$ -stable set systems on  $\mathcal{A}$ , then

$$\perp\!\!\!\perp_P(\mathcal{E}_i, i \in I) | \mathcal{C} \Leftrightarrow \perp\!\!\!\perp_P(\sigma(\mathcal{E}_i), i \in I) | \mathcal{C}. \quad (16.48)$$

(Proof p. 484)

## 16.6 Families of Random Variables

Now we turn to the concept of conditional  $P$ -independence of (families of) random variables. Three random variables  $X_1$ ,  $X_2$ , and  $X_3$  are called  $\mathcal{C}$ -conditionally  $P$ -independent if Equations (16.44) and (16.45) hold for  $\mathcal{E}_i = \sigma(X_i)$ . Note that  $\mathcal{C}$ -conditional  $P$ -independence of  $X_1$ ,  $X_2$ , and  $X_3$  implies that  $X_1$  and  $X_2$ ,  $X_1$  and  $X_3$ , as well as  $X_2$  and  $X_3$  are  $\mathcal{C}$ -conditionally  $P$ -independent. Of course, the same applies to  $Z$ -conditional  $P$ -independence.

### Definition 16.45 (Conditional Independence of Random Variables)

Let  $I$  be a nonempty set,  $X_i: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_i, \mathcal{A}'_i)$ ,  $i \in I$ , be random variables, and let  $\mathcal{C} \subset \mathcal{A}$  be a  $\sigma$ -algebra.

- (i) Then  $(X_i, i \in I)$  is called a family of  $\mathcal{C}$ -conditionally  $P$ -independent random variables, denoted  $\perp\!\!\!\perp (X_i, i \in I) | \mathcal{C}$ , if

$$\perp\!\!\!\perp (\sigma(X_i), i \in I) | \mathcal{C}. \quad (16.49)$$

- (ii) Let  $Z: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Z, \mathcal{A}'_Z)$  be a random variable. Then  $(X_i, i \in I)$  is called a family of  $Z$ -conditionally  $P$ -independent random variables, denoted  $\perp\!\!\!\perp (X_i, i \in I) | Z$ , if

$$\perp\!\!\!\perp (\sigma(X_i), i \in I) | Z. \quad (16.50)$$

**Remark 16.46 (Conditional Independence of Subfamilies)** If  $(X_i, i \in I)$  is a family of  $\mathcal{C}$ -conditionally  $P$ -independent random variables, i. e., if  $\perp\!\!\!\perp (X_i, i \in I) | \mathcal{C}$ , then  $\perp\!\!\!\perp (X_i, i \in J) | \mathcal{C}$ , for all  $J \subset I$ , because  $\Omega'_i \in \sigma(X_i)$ , for all  $i \in I \setminus J$ . For example, if  $X_1, X_2, X_3$  are  $\mathcal{C}$ -conditionally  $P$ -independent, then they are also pairwise  $\mathcal{C}$ -conditionally  $P$ -independent, because  $\Omega'_1 \in \sigma(X_1)$ ,  $\Omega'_2 \in \sigma(X_2)$ , and  $\Omega'_3 \in \sigma(X_3)$ , respectively (cf. Rem. 5.40).  $\triangleleft$

Now we consider conditional independence of  $n$  discrete random variables generalizing Corollary 16.19. Remember, a random variable  $X_i: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_i, \mathcal{A}'_i)$ , is called *discrete* if there is a finite or countable set  $\Omega'_{i0} \subset \Omega'_i$  with  $P_{X_i}(\Omega'_{i0}) = 1$  and  $\{x_i\} \in \mathcal{A}'_i$  for all  $x_i \in \Omega'_{i0}$ ,  $i = 1, \dots, n$  (see Def. 5.56).

### Corollary 16.47 (Conditional Independence of $n$ Discrete Random Variables)

Let  $X_i: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_i, \mathcal{A}'_i)$ ,  $i = 1, \dots, n$ , be discrete random variables and let  $\mathcal{C} \subset \mathcal{A}$  be a  $\sigma$ -algebra. Then  $\perp\!\!\!\perp (X_i, i = 1, \dots, n) | \mathcal{C}$  if and only if

$$P(X_1 = x_1, \dots, X_n = x_n | \mathcal{C}) \stackrel{=}{=} P(X_1 = x_1 | \mathcal{C}) \cdot \dots \cdot P(X_n = x_n | \mathcal{C}), \quad (16.51)$$

$$\forall (x_1, \dots, x_n) \in \Omega'_{10} \times \dots \times \Omega'_{n0}.$$

(Proof p. 484)

**Remark 16.48 (Conditioning on a Random Variable)** If  $Z: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_Z, \mathcal{A}'_Z)$ , then (16.51) can be written

$$P(X=x, Y=y|Z) \stackrel{\text{p}}{=} P(X=x|Z) \cdot P(Y=y|Z), \quad \forall (x, y) \in \Omega'_{X0} \times \Omega'_{Y0}. \quad (16.52)$$

◁

**Lemma 16.49 (Unions of Independent  $\sigma$ -Algebras)**

Under the assumptions of Definition 16.45, let  $J$  be a nonempty set, and let  $\{I_j, j \in J\}$  be a set of pairwise disjoint subsets of  $I$ . Then

$$\prod_{\text{p}} (\sigma(X_i, i \in I_j), j \in J) \mid \mathcal{C}. \quad (16.53)$$

(Proof p. 485)

If  $Y_1, \dots, Y_m: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  are discrete random variables and  $X$  is a random variable on  $(\Omega, \mathcal{A}, P)$ , then the following corollary presents a useful characterization of  $X$ -conditional independence of the  $Y_1, \dots, Y_m$ .

**Corollary 16.50 ( $X$ -Conditional Independence of Discrete Random Variables)**

Let  $X: (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_X, \mathcal{A}'_X)$  be a random variable, let  $(Y_1, \dots, Y_m): (\Omega, \mathcal{A}, P) \rightarrow (\Omega'_1 \times \dots \times \Omega'_m, \mathcal{A}'_1 \otimes \dots \otimes \mathcal{A}'_m)$  be a discrete multivariate random variable, and  $\Omega'_{i0} \subset \Omega'_i$  finite or countable with  $P(Y_i \in \Omega'_{i0}) = 1$ ,  $i = 1, \dots, m$ . Then the following propositions are equivalent to each other:

(i)  $\prod_{\text{p}} Y_1, \dots, Y_m \mid X$ .

(ii) For all  $i = 1, \dots, m$  and all  $y \in \Omega'_{i0}$ :

$$P(Y_i=y|X, Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_m) \stackrel{\text{p}}{=} P(Y_i=y|X).$$

(iii) For all  $i = 2, \dots, m$  and all  $y \in \Omega'_{i0}$ :

$$P(Y_i=y|X, Y_1, \dots, Y_{i-1}) \stackrel{\text{p}}{=} P(Y_i=y|X).$$

(iv) For all  $i = 1, \dots, m-1$  and all  $y \in \Omega'_{i0}$ :

$$P(Y_i=y|X, Y_{i+1}, \dots, Y_m) \stackrel{\text{p}}{=} P(Y_i=y|X).$$

(Proof p. 485)

**Example 16.51 (Joe and Ann With Latent Abilities – continued)** Table 13.1 describes an example in which three dichotomous variables  $Y_1, Y_2, Y_3$  satisfy the assumptions (13.36) and (13.37) that define a Rasch model. Because  $Y$  is dichotomous with values 0 and 1, the second of these assumptions

$$P(Y_i=1|U, Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_m) = P(Y_i=1|U), \quad \forall i = 1, \dots, m, \quad (16.54)$$

is equivalent to  $\prod_p(Y_1, \dots, Y_m) | U$  (see Cor. 16.50) and, because

$$P(Y_i=0 | U, Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_m) = 1 - P(Y_i=1 | U, Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_m),$$

also to

$$P(Y_i=0 | U, Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_m) = P(Y_i=0 | U), \quad \forall i = 1, \dots, m.$$

Using the probabilities displayed in Table 13.1, Equation (4.2), and Notation (5.4), Equation (16.54) can be verified for  $i = 1$  as follows:

$$\begin{aligned} P(Y_1=1 | U=Joe, Y_2=1, Y_3=1) &= \frac{P(Y_1=1, U=Joe, Y_2=1, Y_3=1)}{P(U=Joe, Y_2=1, Y_3=1)} \\ &= \frac{.0492}{.0492 + .0492} = .5 \\ &= P(Y_1=1 | U=Joe). \end{aligned}$$

Analogously, we obtain

$$P(Y_1=1 | U=u, Y_2=y_2, Y_3=y_3) = \frac{P(Y_1=1, U=u, Y_2=y_2, Y_3=y_3)}{P(U=Joe, Y_2=y_2, Y_3=y_3)} = P(Y_1=1 | U=u)$$

for all other combinations of values of  $U$ ,  $Y_2$ , and  $Y_3$ .

The corresponding equations for

$$P(Y_2=1 | U=Joe, Y_1, Y_3) \quad \text{and} \quad P(Y_3=1 | U=Joe, Y_1, Y_2),$$

which are also assumed with (16.54) can be checked analogously.

Assumptions (13.36), (13.37), and the definition of  $\xi$  [see (13.38)] also imply  $\xi$ -conditional independence of the random variables  $Y_1, \dots, Y_m$ , i. e., they imply

$$P(Y_i=1 | \xi, Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_m) = P(Y_i=1 | \xi), \quad \forall i = 1, \dots, m \quad (16.55)$$

(see Exercise 16-5), which is equivalent to

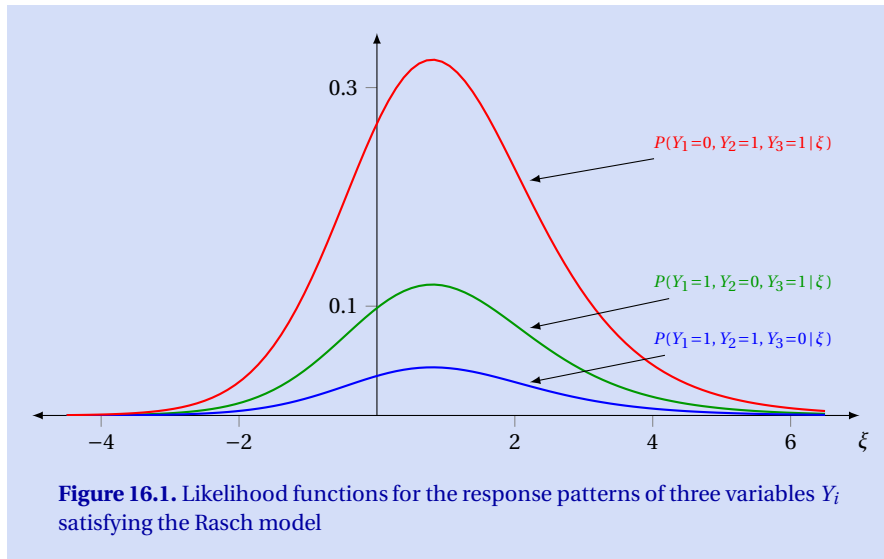
$$\forall J \subset \{1, \dots, m\}, \forall y_i \in \{0, 1\}: \quad P\left(\bigcap_{i \in J} \{Y_i = y_i\} \mid \xi\right) = \prod_{i \in J} P(Y_i = y_i | \xi) \quad (16.56)$$

(see Exercise 16-6).

Applying Equation (16.56) for  $J = \{1, 2, 3\}$  and Equation (13.40) to the conditional probabilities  $P(Y_i=1 | U) = P(Y_i=1 | \xi)$  displayed in Table 13.1 yields the functions

$$\begin{aligned} P(Y_1=0, Y_2=1, Y_3=1 | \xi) &= P(Y_1=0 | \xi) \cdot P(Y_2=1 | \xi) \cdot P(Y_3=1 | \xi) \\ &= \left(1 - \frac{\exp(\xi - \beta_1)}{1 + \exp(\xi - \beta_1)}\right) \cdot \frac{\exp(\xi - \beta_2)}{1 + \exp(\xi - \beta_2)} \cdot \frac{\exp(\xi - \beta_3)}{1 + \exp(\xi - \beta_3)} \end{aligned}$$

(see Fig. 16.1), where the difficulty parameters  $\beta_i$  have been computed in Equation (13.42). The conditional probabilities  $P(Y_1=1, Y_2=0, Y_3=1 | \xi)$  and  $P(Y_1=1, Y_2=1, Y_3=0 | \xi)$  are computed analogously.



**Figure 16.1.** Likelihood functions for the response patterns of three variables  $Y_i$  satisfying the Rasch model

These functions are called *likelihood functions*. The graphs of these functions in Figure 16.1 illustrate that the most likely ability parameter of a person solving exactly two items is the value of  $\xi$  at which the functions  $P(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3 | \xi)$  have their maximum.

The likelihood functions can also be used for the estimation of the ability of a person, for which the score pattern  $(0, 1, 1)$ ,  $(1, 0, 1)$ , or  $(1, 1, 0)$  is observed. The value of  $\xi$  (ability) at which these functions have their maximum, the *maximum likelihood estimate*, is the most likely ability score of such a person. For each value of  $\xi$  other than this maximum likelihood estimate, the probability of such a score pattern is smaller.  $\triangleleft$

## 16.7 Proofs

### *Proof of Lemma 16.4*

If  $B \in \mathcal{C}$ , then  $1_B$  is  $(\mathcal{C}, \mathcal{B})$ -measurable (see Example 2.12). Hence,

$$\begin{aligned}
 P(A \cap B | \mathcal{C}) &\stackrel{P}{=} E(1_{A \cap B} | \mathcal{C}) \\
 &\stackrel{P}{=} E(1_A \cdot 1_B | \mathcal{C}) && [(1.33)] \\
 &\stackrel{P}{=} E(1_A | \mathcal{C}) \cdot 1_B && [\text{Box 10.1 (xiv)}] \\
 &\stackrel{P}{=} E(1_A | \mathcal{C}) \cdot E(1_B | \mathcal{C}). && [\text{Box 10.1 (vii)}]
 \end{aligned}$$

If we assume  $A \in \mathcal{C}$ , the proof is analogous, exchanging the roles of  $A$  and  $B$ . Proposition (16.5) is an immediate implication of (16.4).

**Proof of Corollary 16.6**

$$\begin{aligned}
& A \perp\!\!\!\perp_P B \mid Z \\
\Leftrightarrow & P(A \cap B \mid Z) \stackrel{=}{=} P(A \mid Z) \cdot P(B \mid Z) && \text{[Def. 16.2 (ii)]} \\
\Leftrightarrow & P(A \cap B \mid Z=z) = P(A \mid Z=z) \cdot P(B \mid Z=z), \text{ for } P_Z\text{-a.a. } z \in \Omega'_Z. && \text{[Cor. 10.39 (i)]}
\end{aligned}$$

**Proof of Lemma 16.12**

- (i)  $\sigma(\mathcal{D}) \perp\!\!\!\perp_P \sigma(\mathcal{E}) \mid \mathcal{C} \Rightarrow \mathcal{D} \perp\!\!\!\perp_P \mathcal{E} \mid \mathcal{C}$  follows from Definition 16.10 (i), because  $\mathcal{D} \subset \sigma(\mathcal{D})$  and  $\mathcal{E} \subset \sigma(\mathcal{E})$ .
- (ii)  $\mathcal{D} \perp\!\!\!\perp_P \mathcal{E} \mid \mathcal{C} \Rightarrow \sigma(\mathcal{D}) \perp\!\!\!\perp_P \sigma(\mathcal{E}) \mid \mathcal{C}$ . Assume  $\mathcal{D} \perp\!\!\!\perp_P \mathcal{E} \mid \mathcal{C}$ , apply Equation (10.2), and define the set of events

$$\mathcal{F} := \{A \in \mathcal{A} : \forall B \in \mathcal{E} : E(1_A \cdot 1_B \mid \mathcal{C}) \stackrel{=}{=} E(1_A \mid \mathcal{C}) \cdot E(1_B \mid \mathcal{C})\}.$$

The assumption implies  $\mathcal{D} \subset \mathcal{F}$ . Now we show (a)  $\Omega \in \mathcal{F}$  and (b)  $\mathcal{F}$  is closed under complement and countable disjoint union, i. e.,  $\mathcal{F}$  is a Dynkin system (see Def. 1.40).

(a) Because  $1_\Omega = 1$ , for all  $B \in \mathcal{E}$ ,

$$\begin{aligned}
E(1_\Omega \cdot 1_B \mid \mathcal{C}) & \stackrel{=}{=} E(1 \cdot 1_B \mid \mathcal{C}) \\
& \stackrel{=}{=} 1 \cdot E(1_B \mid \mathcal{C}) && \text{[Box 10.1 (iii)]} \\
& \stackrel{=}{=} E(1 \mid \mathcal{C}) \cdot E(1_B \mid \mathcal{C}) && \text{[Box 10.1 (i)]} \\
& \stackrel{=}{=} E(1_\Omega \mid \mathcal{C}) \cdot E(1_B \mid \mathcal{C}).
\end{aligned}$$

Hence,  $\Omega \in \mathcal{F}$ .

(b) Assume that  $A \in \mathcal{F}$ . Then, for all  $B \in \mathcal{E}$ ,

$$\begin{aligned}
E(1_{A^c} \cdot 1_B \mid \mathcal{C}) & \stackrel{=}{=} E[(1 - 1_A) \cdot 1_B \mid \mathcal{C}] \\
& \stackrel{=}{=} E(1_B \mid \mathcal{C}) - E(1_A \cdot 1_B \mid \mathcal{C}) && \text{[Box 10.1 (xvi)]} \\
& \stackrel{=}{=} E(1_B \mid \mathcal{C}) - E(1_A \mid \mathcal{C}) \cdot E(1_B \mid \mathcal{C}) && \text{[def. of } \mathcal{F}] \\
& \stackrel{=}{=} [1 - E(1_A \mid \mathcal{C})] \cdot E(1_B \mid \mathcal{C}) \\
& \stackrel{=}{=} E(1 - 1_A \mid \mathcal{C}) \cdot E(1_B \mid \mathcal{C}) && \text{[Box 10.1 (i), (xvi)]} \\
& \stackrel{=}{=} E(1_{A^c} \mid \mathcal{C}) \cdot E(1_B \mid \mathcal{C}).
\end{aligned}$$

Hence,  $A \in \mathcal{F}$  implies  $A^c \in \mathcal{F}$ .

(c) Assume  $A_1, A_2, \dots \in \mathcal{F}$  and  $A_i \cap A_j = \emptyset$  for  $i \neq j$ . Then, for all  $B \in \mathcal{E}$ ,

$$\begin{aligned}
E\left(1_{\bigcup_{i=1}^{\infty} A_i} \cdot 1_B \mid \mathcal{C}\right) &\stackrel{=}{=} E\left[\left(\sum_{i=1}^{\infty} 1_{A_i}\right) \cdot 1_B \mid \mathcal{C}\right] && [(1.37)] \\
&\stackrel{=}{=} E\left(\sum_{i=1}^{\infty} (1_{A_i} \cdot 1_B) \mid \mathcal{C}\right) \\
&\stackrel{=}{=} \sum_{i=1}^{\infty} E(1_{A_i} \cdot 1_B \mid \mathcal{C}) && [\text{Th. 10.21 (ii)}] \\
&\stackrel{=}{=} \sum_{i=1}^{\infty} E(1_{A_i} \mid \mathcal{C}) \cdot E(1_B \mid \mathcal{C}) && [\text{def. of } \mathcal{F}] \\
&\stackrel{=}{=} E\left(\sum_{i=1}^{\infty} 1_{A_i} \mid \mathcal{C}\right) \cdot E(1_B \mid \mathcal{C}) && [\text{Th. 10.21 (ii)}] \\
&\stackrel{=}{=} E\left(1_{\bigcup_{i=1}^{\infty} A_i} \mid \mathcal{C}\right) \cdot E(1_B \mid \mathcal{C}). && [(1.37)]
\end{aligned}$$

Hence,  $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$ . This proves that  $\mathcal{F}$  is a Dynkin system (see Def. 1.40). Because  $\mathcal{D} \subset \mathcal{F}$ , we can conclude  $\delta(\mathcal{D}) \subset \mathcal{F}$ , where  $\delta(\mathcal{D})$  denotes the Dynkin system generated by  $\mathcal{D}$ , i. e., the smallest Dynkin system that contains  $\mathcal{D}$  (see Rem. 1.14). According to Theorem 1.41 (ii),  $\cap$ -stability of  $\mathcal{D}$  implies  $\delta(\mathcal{D}) = \sigma(\mathcal{D})$ . Hence,  $\sigma(\mathcal{D}) \subset \mathcal{F}$ .

Now we have shown:  $\mathcal{D} \perp\!\!\!\perp_P \mathcal{E} \mid \mathcal{C} \Rightarrow \sigma(\mathcal{D}) \perp\!\!\!\perp_P \mathcal{E} \mid \mathcal{C}$ . Because  $\sigma(\mathcal{D}) \perp\!\!\!\perp_P \mathcal{E} \mid \mathcal{C} \Leftrightarrow \mathcal{E} \perp\!\!\!\perp_P \sigma(\mathcal{D}) \mid \mathcal{C}$ , this also implies  $\sigma(\mathcal{D}) \perp\!\!\!\perp_P \mathcal{E} \mid \mathcal{C} \Rightarrow \sigma(\mathcal{D}) \perp\!\!\!\perp_P \sigma(\mathcal{E}) \mid \mathcal{C}$ .

### ***Proof of Theorem 16.22***

$$\begin{aligned}
&\mathcal{E} \perp\!\!\!\perp_P \mathcal{D} \mid \mathcal{C} \\
\Leftrightarrow E(1_A \cdot 1_B \mid \mathcal{C}) &\stackrel{=}{=} E(1_A \mid \mathcal{C}) \cdot E(1_B \mid \mathcal{C}), \quad \forall (A, B) \in \mathcal{E} \times \mathcal{D} && [(10.2), \text{Def. 16.10 (i)}] \\
\Leftrightarrow E(1_A \cdot W \mid \mathcal{C}) &\stackrel{=}{=} E(1_A \mid \mathcal{C}) \cdot E(W \mid \mathcal{C}), \quad \forall A \in \mathcal{E} \text{ and all nonnegative } \mathcal{D}\text{-measurable } W \\
\Leftrightarrow E(1_A \mid \mathcal{C}, \mathcal{D}) &\stackrel{=}{=} E(1_A \mid \mathcal{C}), \quad \forall A \in \mathcal{E} && [\text{Th. 10.51}] \\
\Leftrightarrow P(A \mid \mathcal{C}, \mathcal{D}) &\stackrel{=}{=} P(A \mid \mathcal{C}), \quad \forall A \in \mathcal{E}. && [(10.2)]
\end{aligned}$$

The equivalence of the second and third propositions is obtained using the proof of Theorem 16.32, with  $W$  taking the role of  $X$ , where  $W = \sum_{i=1}^{\infty} \alpha_i 1_{A_i}$  and  $Y = 1_A$ .

### ***Proof of Theorem 16.26***

$\Rightarrow$  This is an implication of Corollary 16.24.  
 $\Leftarrow$  For  $A'_Y \in \mathcal{A}'_Y$  define  $A'_{Y_0} := A'_Y \cap \Omega'_{Y_0}$ , which is finite or countable. Then,  $1_{Y \in A'_Y} \stackrel{=}{=} 1_{Y \in A'_{Y_0}}$ . Furthermore,

$$\begin{aligned}
P(Y=y|\mathcal{C},\mathcal{D}) &\stackrel{\bar{P}}{=} P(Y=y|\mathcal{C}), \quad \forall y \in \Omega'_{Y_0} \\
\Rightarrow E\left(\sum_{y \in A'_{Y_0}} 1_{Y=y} \middle| \mathcal{C}, \mathcal{D}\right) &\stackrel{\bar{P}}{=} E\left(\sum_{y \in A'_{Y_0}} 1_{Y=y} \middle| \mathcal{C}\right) \quad [(10.2), \text{Th. 10.21 (ii)}] \\
\Rightarrow E(1_{Y \in A'_{Y_0}} | \mathcal{C}, \mathcal{D}) &\stackrel{\bar{P}}{=} E(1_{Y \in A'_{Y_0}} | \mathcal{C}) \quad [(1.36) (1.37)] \\
\Rightarrow E(1_{Y \in A'_Y} | \mathcal{C}, \mathcal{D}) &\stackrel{\bar{P}}{=} E(1_{Y \in A'_Y} | \mathcal{C}) \quad [\text{Box 10.1 (ix)}] \\
\Rightarrow Y &\stackrel{\bar{P}}{\perp\!\!\!\perp} \mathcal{D} | \mathcal{C}. \quad [(10.2), \text{Th. 16.22}]
\end{aligned}$$

### ***Proof of Theorem 16.32***

Case 1:  $X$  and  $Y$  nonnegative. If  $X: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  is a nonnegative random variable, then it is also  $(\sigma(X), \mathcal{B})$ -measurable, and therefore, according to Theorem 3.19 (ii), there is a sequence  $A_1, A_2, \dots \in \sigma(X)$  and a sequence of nonnegative real numbers  $\alpha_1, \alpha_2, \dots$  such that  $X = \sum_{i=1}^{\infty} \alpha_i 1_{A_i}$ . Analogously, a nonnegative random variable  $Y$  can be represented as  $Y = \sum_{j=1}^{\infty} \beta_j 1_{B_j}$ ,  $\beta_j \geq 0$ ,  $B_j \in \sigma(Y)$ ,  $j \in \mathbb{N}$ . Hence

$$\begin{aligned}
E(X \cdot Y | \mathcal{C}) &\stackrel{\bar{P}}{=} E\left(\sum_{i=1}^{\infty} \alpha_i 1_{A_i} \cdot \sum_{j=1}^{\infty} \beta_j 1_{B_j} \middle| \mathcal{C}\right) \\
&\stackrel{\bar{P}}{=} E\left(\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \alpha_i \beta_j 1_{A_i} \cdot 1_{B_j} \middle| \mathcal{C}\right) \\
&\stackrel{\bar{P}}{=} \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \alpha_i \beta_j E(1_{A_i} \cdot 1_{B_j} | \mathcal{C}) \quad [\text{Th. 10.21 (ii), Box 10.1 (iii)}] \\
&\stackrel{\bar{P}}{=} \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \alpha_i \beta_j E(1_{A_i} | \mathcal{C}) \cdot E(1_{B_j} | \mathcal{C}) \quad [\sigma(X) \perp\!\!\!\perp \sigma(Y) | \mathcal{C}] \\
&\stackrel{\bar{P}}{=} E\left(\sum_{i=1}^{\infty} \alpha_i 1_{A_i} \middle| \mathcal{C}\right) \cdot E\left(\sum_{j=1}^{\infty} \beta_j 1_{B_j} \middle| \mathcal{C}\right) \quad [\text{Th. 10.21 (ii), Box 10.1 (iii)}] \\
&\stackrel{\bar{P}}{=} E(X | \mathcal{C}) \cdot E(Y | \mathcal{C}).
\end{aligned}$$

Case 2: Let  $X^+, Y^+$  and  $X^-, Y^-$  be the *positive* and *negative parts* of  $X$  and  $Y$  (see Def. 2.62). These functions are nonnegative and measurable. Note that  $\sigma(X^+), \sigma(X^-) \subset \sigma(X)$  and  $\sigma(Y^+), \sigma(Y^-) \subset \sigma(Y)$  (see Th. 2.66). Hence, (16.33) holds for the four pairs of random variables  $(X^+, Y^+)$ ,  $(X^+, Y^-)$ ,  $(X^-, Y^+)$ , and  $(X^-, Y^-)$ . If  $E(X)$  and  $E(Y)$ , then  $E(X^+)$ ,  $E(X^-)$ ,  $E(Y^+)$ , and  $E(Y^-)$  are finite as well, and according to Box 10.1 (x), there are real-valued versions of the conditional expectations that are used in the following equations. Because  $X = X^+ - X^-$  and  $Y = Y^+ - Y^-$ , it can be shown that (16.33) also holds if we assume finite expectations:

$$\begin{aligned}
& E(X|\mathcal{C}) \cdot E(Y|\mathcal{C}) \\
& \stackrel{=}{=} E(X^+ - X^-|\mathcal{C}) \cdot E(Y^+ - Y^-|\mathcal{C}) \\
& \stackrel{=}{=} [E(X^+|\mathcal{C}) - E(X^-|\mathcal{C})] \cdot [E(Y^+|\mathcal{C}) - E(Y^-|\mathcal{C})] \quad [\text{Box 10.1 (xvi)}] \\
& \stackrel{=}{=} E(X^+|\mathcal{C}) \cdot E(Y^+|\mathcal{C}) - E(X^+|\mathcal{C}) \cdot E(Y^-|\mathcal{C}) \\
& \quad - E(X^-|\mathcal{C}) \cdot E(Y^+|\mathcal{C}) + E(X^-|\mathcal{C}) \cdot E(Y^-|\mathcal{C}) \\
& \stackrel{=}{=} E(X^+ \cdot Y^+|\mathcal{C}) - E(X^+ \cdot Y^-|\mathcal{C}) \\
& \quad - E(X^- \cdot Y^+|\mathcal{C}) + E(X^- \cdot Y^-|\mathcal{C}) \quad [(16.33)] \\
& \stackrel{=}{=} E(X^+ \cdot Y^+ + X^+ \cdot Y^- - X^- \cdot Y^+ + X^- \cdot Y^-|\mathcal{C}) \quad [\text{Box 10.1 (xvi)}] \\
& \stackrel{=}{=} E[(X^+ - X^-) \cdot (Y^+ - Y^-)|\mathcal{C}] \\
& \stackrel{=}{=} E(X \cdot Y|\mathcal{C}).
\end{aligned}$$

### Proof of Theorem 16.34

Part 1. If  $Y: (\Omega, \mathcal{A}, P) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$  is nonnegative, then it is also  $(\sigma(Y), \overline{\mathcal{B}})$ -measurable, and therefore, according to Theorem 3.19 (ii), there is a sequence  $A_1, A_2, \dots \in \sigma(Y)$  and a sequence of nonnegative real numbers  $\alpha_1, \alpha_2, \dots$  such that

$$Y = \sum_{i=1}^{\infty} \alpha_i 1_{A_i}, \quad \alpha_i \geq 0, \quad A_i \in \sigma(Y).$$

Hence, if  $\sigma(Y) \subset \mathcal{E}$ , then

$$\begin{aligned}
\mathcal{E} \perp\!\!\!\perp \mathcal{D} | \mathcal{C} & \Rightarrow P(A|\mathcal{C}, \mathcal{D}) \stackrel{=}{=} P(A|\mathcal{C}), \quad \forall A \in \mathcal{E} \quad [(16.23)] \\
& \Rightarrow \sum_{i=1}^{\infty} \alpha_i E(1_{A_i}|\mathcal{C}, \mathcal{D}) \stackrel{=}{=} \sum_{i=1}^{\infty} \alpha_i E(1_{A_i}|\mathcal{C}) \quad [(10.2), (2.39)] \\
& \Rightarrow E\left(\sum_{i=1}^{\infty} \alpha_i 1_{A_i} \middle| \mathcal{C}, \mathcal{D}\right) \stackrel{=}{=} E\left(\sum_{i=1}^{\infty} \alpha_i E(1_{A_i}|\mathcal{C}) \middle| \mathcal{C}\right) \quad [\text{Th. 10.21 (ii)}] \\
& \Rightarrow E(Y|\mathcal{C}, \mathcal{D}) \stackrel{=}{=} E(Y|\mathcal{C}). \quad [\text{Th. 3.19 (ii)}]
\end{aligned}$$

Part 2. If  $Y$  has a finite expectation, we use the positive part  $Y^+$  and the negative part  $Y^-$  of

$$Y = Y^+ - Y^-$$

(see Rem. 2.62). Because both  $Y^+$  and  $Y^-$  are nonnegative and  $\sigma(Y^+) \subset \sigma(Y) \subset \mathcal{E}$ ,  $\sigma(Y^-) \subset \sigma(Y) \subset \mathcal{E}$ , we can use Part 1 of this proof. Hence,

$$\begin{aligned}
E(Y|\mathcal{C}, \mathcal{D}) & \stackrel{=}{=} E(Y^+ - Y^-|\mathcal{C}, \mathcal{D}) \quad [\text{Rem. 2.62}] \\
& \stackrel{=}{=} E(Y^+|\mathcal{C}, \mathcal{D}) - E(Y^-|\mathcal{C}, \mathcal{D}) \quad [\text{Box 10.1 (xvi)}] \\
& \stackrel{=}{=} E(Y^+|\mathcal{C}) - E(Y^-|\mathcal{C}) \quad [\text{Part 1}] \\
& \stackrel{=}{=} E(Y^+ - Y^-|\mathcal{C}) \quad [\text{Box 10.1 (xvi)}] \\
& \stackrel{=}{=} E(Y|\mathcal{C}). \quad [\text{Rem. 2.62}]
\end{aligned}$$

**Proof of Theorem 16.37**

In this proof we use the notation  $Y \in \langle \mathcal{E} \rangle^+$  for  $Y$  is nonnegative and  $\sigma(Y) \subset \mathcal{E}$ .

(i)  $\Rightarrow$  (ii) This implication immediately follows from Theorem 16.34.

(ii)  $\Rightarrow$  (i)

$$(ii) \Rightarrow \forall B \in \mathcal{E}: E(1_B | \mathcal{C}, \mathcal{D}) \stackrel{P}{=} E(1_B | \mathcal{C}) \quad [1_B \in \langle \mathcal{E} \rangle^+]$$

$$\Rightarrow \forall B \in \mathcal{E}: P(B | \mathcal{C}, \mathcal{D}) \stackrel{P}{=} P(B | \mathcal{C}) \quad [(10.2)]$$

$$\Rightarrow \mathcal{E} \perp\!\!\!\perp_P \mathcal{D} | \mathcal{C}. \quad [(16.23)]$$

(ii)  $\Leftrightarrow$  (iii) This immediately follows from Theorem 10.51.

(iv)  $\Rightarrow$  (iii) Let  $W, X, Y: (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$  be random variables. Then, for all  $X \in \langle \mathcal{E} \rangle^+$ ,  $W \in \langle \mathcal{D} \rangle^+$ ,  $Y \in \langle \mathcal{E} \rangle^+$ ,

$$\begin{aligned} E[X \cdot E(W \cdot Y | \mathcal{C})] &= E[E(X \cdot W \cdot Y | \mathcal{C})] && [\text{Box 10.1 (xiv)}] \\ &= E(X \cdot W \cdot Y) && [\text{Box 10.1 (iv)}] \\ &= E[E(X \cdot W \cdot Y | \mathcal{D})] && [\text{Box 10.1 (iv)}] \\ &= E[W \cdot E(X \cdot Y | \mathcal{D})] && [\text{Box 10.1 (xiv)}] \\ &= E[W \cdot E[E(X \cdot Y | \mathcal{C}) | \mathcal{D}]] && [(iv), \sigma(X \cdot Y) \subset \sigma(\mathcal{C}, \mathcal{E})] \\ &= E[E[W \cdot E(X \cdot Y | \mathcal{C}) | \mathcal{D}]] && [\sigma(W) \subset \sigma(\mathcal{D}), \text{Box 10.1 (xiv)}] \\ &= E[W \cdot E(X \cdot Y | \mathcal{C})] && [\text{Box 10.1 (iv)}] \\ &= E[E[W \cdot E(X \cdot Y | \mathcal{C}) | \mathcal{C}]] && [\text{Box 10.1 (iv)}] \\ &= E[X \cdot E(W | \mathcal{C}) \cdot E(Y | \mathcal{C})]. && [\sigma(X), \sigma(E(Y | \mathcal{C})) \subset \mathcal{C}, \text{Box 10.1 (xiv)}] \end{aligned}$$

Because this equation holds for all  $X \in \langle \mathcal{E} \rangle^+$ , it also holds for all  $1_C$ ,  $C \in \mathcal{C}$ , and using Definition 10.2 (b) we conclude:  $E(W \cdot Y | \mathcal{C}) \stackrel{P}{=} E(W | \mathcal{C}) \cdot E(Y | \mathcal{C})$ .

(i)  $\Rightarrow$  (iv) Note that we have already shown that (i) and (ii) are equivalent to each other. Therefore, we can use (ii) in this part of the proof. For all nonnegative random variables  $Y$  with finite expectation and  $\sigma(Y) \subset \sigma(\mathcal{C}, \mathcal{E})$ ,

$$\begin{aligned} &\mathcal{E} \perp\!\!\!\perp_P \mathcal{D} | \mathcal{C} \\ \Rightarrow &\sigma(\mathcal{C}, \mathcal{E}) \perp\!\!\!\perp_P \mathcal{D} | \mathcal{C} && [\text{Box 16.2 (vii)}] \\ \Rightarrow &E(Y | \mathcal{C}) \stackrel{P}{=} E(Y | \mathcal{C}, \mathcal{D}) && [(ii)] \\ \Rightarrow &E[E(Y | \mathcal{C}) | \mathcal{D}] \stackrel{P}{=} E[E(Y | \mathcal{C}, \mathcal{D}) | \mathcal{D}] && [\text{Box 10.1 (ix)}] \\ \Rightarrow &E[E(Y | \mathcal{C}) | \mathcal{D}] \stackrel{P}{=} E(Y | \mathcal{D}). && [\text{Box 10.1 (v)}] \end{aligned}$$

**Proof of Theorem 16.38**

$E^{Z=z}(Y | \mathcal{C})$  and  $E(Y | \mathcal{C})$  are both  $\mathcal{C}$ -measurable. Hence, it suffices to show that condition (b) of Definition 14.7 holds for  $E(Y | \mathcal{C})$ , i. e., it suffices to show:

$$E^{Z=z}[1_C E(Y | \mathcal{C})] = E^{Z=z}(1_C Y), \quad \forall C \in \mathcal{C}. \quad (16.57)$$

Now

$$E[1_A E(Y|\mathcal{C}, Z)] = E(1_A Y), \quad \forall A \in \sigma(\mathcal{C}, Z). \quad [\text{Def. 10.2 (b)}]$$

Because, for  $C \in \sigma(\mathcal{C}, Z)$ ,  $\{Z=z\} \cap C \in \sigma(\mathcal{C}, Z)$ , and  $1_{\{Z=z\} \cap C} = 1_{Z=z} 1_C$ , Definition 10.2 (b) implies

$$E[1_{Z=z} 1_C E(Y|\mathcal{C}, Z)] = E(1_{Z=z} 1_C Y), \quad \forall C \in \mathcal{C}.$$

Dividing both sides by  $P(Z=z)$  yields

$$\frac{1}{P(Z=z)} E[1_{Z=z} 1_C E(Y|\mathcal{C}, Z)] = \frac{1}{P(Z=z)} E(1_{Z=z} 1_C Y), \quad \forall C \in \mathcal{C},$$

which, according to Equation (9.11), is equivalent to

$$E^{Z=z}[1_C E(Y|\mathcal{C}, Z)] = E^{Z=z}(1_C Y), \quad \forall C \in \mathcal{C}.$$

According to Theorem 16.34, assuming  $Z \perp\!\!\!\perp Y | \mathcal{C}$  implies  $E(Y|\mathcal{C}, Z) \stackrel{p}{=} E(Y|\mathcal{C})$  and Corollary 5.22 yields  $E(Y|\mathcal{C}, Z) \stackrel{p^{Z=z}}{=} E(Y|\mathcal{C})$ . Now Equation (16.57) follows from Box 6.1 (viii).

### **Proof of Theorem 16.44**

$\Leftarrow$  This proposition immediately follows from Definition 16.41 (i).

$\Rightarrow$  If  $(\mathcal{E}_i, i \in I)$  is a family of  $\cap$ -stable set systems on  $\mathcal{A}$  and  $J \subset I$  is finite, then

$$\begin{aligned} \perp\!\!\!\perp_P (\mathcal{E}_i, i \in I) | \mathcal{C} &\Rightarrow \perp\!\!\!\perp_P (\mathcal{E}_i, i \in J) | \mathcal{C} && [(16.46)] \\ &\Rightarrow \perp\!\!\!\perp_P (\sigma(\mathcal{E}_i), i \in J) | \mathcal{C}. \end{aligned}$$

This last implication follows from repeatedly applying (ii) (b) of the proof of Lemma 16.12. Because the implications above hold for all finite  $J \subset I$ , this yields  $\perp\!\!\!\perp_P (\sigma(\mathcal{E}_i), i \in I) | \mathcal{C}$ .

### **Proof of Corollary 16.47**

$\perp\!\!\!\perp_P (X_i, i = 1, \dots, n) | \mathcal{C} \Rightarrow (16.51)$  immediately follows from Definitions 16.45 and 16.41, because  $\{X_i = x_i\} \in X_i^{-1}(\mathcal{A}'_i)$  for all  $x_i \in \Omega'_{i0}$ ,  $i = 1, \dots, n$ .

(16.51)  $\Rightarrow \perp\!\!\!\perp_P (X_i, i = 1, \dots, n) | \mathcal{C}$ . For  $i = 1, \dots, n$ , let  $A'_i \in \mathcal{A}'_i$  and  $A'_{i0} := A'_i \cap \Omega'_{i0}$ , which implies that the  $A'_{i0}$  are finite or countable. Then

$$1_{X_i \in A'_i} \stackrel{p}{=} 1_{X_i \in A'_{i0}} \quad \forall i = 1, \dots, n. \quad (16.58)$$

Furthermore,

$$\begin{aligned}
P(X_1 \in A'_1, \dots, X_n \in A'_n | \mathcal{C}) &\stackrel{p}{=} E(1_{X_1 \in A'_1} \cdots 1_{X_n \in A'_n} | \mathcal{C}) && [(16.58), \text{Box 10.1 (ix)}] \\
&\stackrel{p}{=} E\left(\sum_{x_1 \in A'_1} \cdots \sum_{x_n \in A'_n} 1_{X_1=x_1} \cdots 1_{X_n=x_n} | \mathcal{C}\right) && [(1.36) (1.37)] \\
&\stackrel{p}{=} \sum_{x_1 \in A'_1} \cdots \sum_{x_n \in A'_n} E(1_{X_1=x_1} \cdots 1_{X_n=x_n} | \mathcal{C}) && [\text{Th. 10.21 (ii)}] \\
&\stackrel{p}{=} \sum_{x_1 \in A'_1} \cdots \sum_{x_n \in A'_n} E(1_{X_1=x_1} | \mathcal{C}) \cdots E(1_{X_n=x_n} | \mathcal{C}) && [(16.51)] \\
&\stackrel{p}{=} \left(\sum_{x_1 \in A'_1} E(1_{X_1=x_1} | \mathcal{C})\right) \cdots \left(\sum_{x_n \in A'_n} E(1_{X_n=x_n} | \mathcal{C})\right) \\
&\stackrel{p}{=} E\left(\sum_{x_1 \in A'_1} 1_{X_1=x_1} | \mathcal{C}\right) \cdots E\left(\sum_{x_n \in A'_n} 1_{X_n=x_n} | \mathcal{C}\right) && [\text{Th. 10.21 (ii)}] \\
&\stackrel{p}{=} E(1_{X_1 \in A'_1} | \mathcal{C}) \cdots E(1_{X_n \in A'_n} | \mathcal{C}) && [(1.36) (1.37)] \\
&\stackrel{p}{=} P(X_1 \in A'_1 | \mathcal{C}) \cdots P(X_n \in A'_n | \mathcal{C}). && [(16.58), \text{Box 10.1 (ix)}]
\end{aligned}$$

### Proof of Lemma 16.49

This proof is analogous to the proof of Theorem 6.5 of Bauer (1996, p. 42). For all  $j \in J$ , consider

$$\mathcal{E}_j := \left\{ \bigcap_{k=1}^n A_{i_k} : A_{i_k} \in \sigma(X_{i_k}), \{i_1, \dots, i_n\} \subset I_j, n \in \mathbf{N} \right\}.$$

The set system  $\mathcal{E}_j$  is  $\cap$ -stable because the intersection of two finite intersections of sets is again a finite intersection, and the  $\sigma$ -algebras  $\sigma(X_{i_k})$  are  $\cap$ -stable as well. Furthermore, if  $\prod_p (\sigma(X_i), i \in I) | \mathcal{C}$ , then according to Definitions 16.45 (i) and 16.41,

$$\prod_p (\mathcal{E}_j, j \in J) | \mathcal{C},$$

because the intersection of finitely many sets that are finite intersections is again an intersection of finitely many sets. Because  $\mathcal{E}_j \subset \sigma(X_i, i \in I_j)$  for all  $j \in J$  and  $\sigma(X_i) \subset \mathcal{E}_j$  for all  $i \in I_j, j \in J$ , we can conclude  $\sigma(\mathcal{E}_j) = \sigma(X_i, i \in I_j)$  (see Rem. 2.39 and Def. 1.13). Therefore, applying Theorem 16.44 completes the proof.

### Proof of Corollary 16.50

(i)  $\Rightarrow$  (ii) For all  $i = 1, \dots, m$  and all  $y_i \in \Omega'_{i0}$ ,

$$\begin{aligned}
&\prod_p Y_1, \dots, Y_m | X \\
&\Rightarrow Y_i \prod_p \sigma(Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_m) | X && [\text{Def. 16.45, Lemma 16.49}] \\
&\Rightarrow 1_{Y_i=y_i} \prod_p \sigma(Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_m) | X && [\sigma(1_{Y_i=y_i}) \subset \sigma(Y_i)] \\
&\Rightarrow P(Y_i=y_i | X, \sigma(Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_m)) \stackrel{p}{=} P(Y_i=y_i | X) && [(10.4), (16.37)]
\end{aligned}$$

$$\Rightarrow P(Y_i = y_i | X, Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_m) \stackrel{p}{=} P(Y_i = y_i | X).$$

The last implication follows from

$$\sigma(\sigma(Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_m) \cup \sigma(X)) = \sigma(Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_m, X).$$

(ii)  $\Rightarrow$  (iv) For all  $i = 1, \dots, m$  and all  $y_i \in \Omega'_{i0}$ ,

$$\begin{aligned} P(Y_i = y_i | X, Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_m) &\stackrel{p}{=} P(Y_i = y_i | X) \\ \Rightarrow E(E(\mathbf{1}_{Y_i = y_i} | X, Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_m) | X, Y_{i+1}, \dots, Y_m) \\ &\stackrel{p}{=} E(E(\mathbf{1}_{Y_i = y_i} | X) | X, Y_{i+1}, \dots, Y_m) && [(10.4), \text{Box 10.2 (ix)}] \\ \Rightarrow E(\mathbf{1}_{Y_i = y_i} | X, Y_{i+1}, \dots, Y_m) &\stackrel{p}{=} E(\mathbf{1}_{Y_i = y_i} | X) && [\text{Box 10.2 (v), (xiv), (i)}] \\ \Rightarrow P(Y_i = y_i | X, Y_{i+1}, \dots, Y_m) &\stackrel{p}{=} P(Y_i = y_i | X). && [(10.4)] \end{aligned}$$

(iv)  $\Rightarrow$  (i) If

$$P(Y_i = y_i | X, Y_{i+1}, \dots, Y_m) \stackrel{p}{=} P(Y_i = y_i | X), \quad \forall y_i \in \Omega'_{i0}, \forall i = 1, \dots, m-1,$$

then (10.4) and Th. 16.37 (ii), (iii) [with  $\mathcal{E} = \sigma(\mathbf{1}_{Y_i = y_i})$ ,  $\mathcal{C} = \sigma(X)$ ,  $\mathcal{D} = \sigma(Y_{i+1}, \dots, Y_m)$ ] imply

$$\begin{aligned} E(\mathbf{1}_{Y_i = y_i} \cdot \mathbf{1}_{Y_{i+1} = y_{i+1}} \cdot \dots \cdot \mathbf{1}_{Y_m = y_m} | X) &\stackrel{p}{=} E(\mathbf{1}_{Y_i = y_i} | X) \cdot E(\mathbf{1}_{Y_{i+1} = y_{i+1}} \cdot \dots \cdot \mathbf{1}_{Y_m = y_m} | X), \\ \forall y_j \in \Omega'_{j0}, j = i+1, \dots, m, \forall i = 1, \dots, m-1. \end{aligned}$$

Sequential application of this implication for  $i = 1, \dots, m-1$  yields

$$E(\mathbf{1}_{Y_1 = y_1} \cdot \dots \cdot \mathbf{1}_{Y_m = y_m} | X) \stackrel{p}{=} E(\mathbf{1}_{Y_1 = y_1} | X) \cdot \dots \cdot E(\mathbf{1}_{Y_m = y_m} | X), \quad \forall y_j \in \Omega'_{j0}, j = 1, \dots, m.$$

According to (10.4) this in turn implies

$$P(Y_1 = y_1, \dots, Y_m = y_m | X) \stackrel{p}{=} \prod_{i=1}^m P(Y_i = y_i | X), \quad \forall y_j \in \Omega'_{j0}, j = 1, \dots, m.$$

Now Corollary 16.47 yields

$$\perp\!\!\!\perp_p Y_1, \dots, Y_m | X.$$

(ii)  $\Rightarrow$  (iii) This proof is analogous to the proof of (ii)  $\Rightarrow$  (iv), where the role of  $Y_{i+1}, \dots, Y_m$  is taken by  $Y_1, \dots, Y_{i-1}$ , and the role of  $i = 1, \dots, m-1$  by  $i = 2, \dots, m$ .

(iii)  $\Rightarrow$  (i) This proof is analogous to the proof of (iv)  $\Rightarrow$  (i), where again the role of  $Y_{i+1}, \dots, Y_m$  is taken by  $Y_1, \dots, Y_{i-1}$ , and the role of  $i = 1, \dots, m-1$  by  $i = 2, \dots, m$ .

## 16.8 Exercises

▷ **Exercise 16-1** Prove proposition (16.9).

▷ **Exercise 16-2** Prove proposition (16.12).

▷ **Exercise 16-3** Prove the propositions of Box 16.2.

▷ **Exercise 16-4** Comparing the columns for  $P(Y=1|X, U)$  and  $P(Y=1|U)$  in Table 16.1 reveals that  $P(Y=1|X, U) \stackrel{P}{=} P(Y=1|U)$ . Show that this implies  $P(Y=0|X, U) \stackrel{P}{=} P(Y=0|U)$ .

▷ **Exercise 16-5** Show that Assumptions (13.36), (13.37), and the definition of  $\xi$  imply Equation (16.55).

▷ **Exercise 16-6** Show that Equations (16.55) and (16.56) are equivalent to each other.

## Solutions

▷ **Solution 16-1** If  $Z \stackrel{P}{=} \alpha$ ,  $\alpha \in \Omega'_Z$ , then

$$\begin{aligned} & A \stackrel{P}{\perp\!\!\!\perp} B | Z \\ \Leftrightarrow & P(A \cap B | Z = \alpha) = P(A | Z = \alpha) \cdot P(B | Z = \alpha) && [\text{Cor. 16.6, } P(Z = \alpha) = 1] \\ \Leftrightarrow & P(A \cap B) = P(A) \cdot P(B) && [\text{Rem. 10.35, Def. 4.12, } P(Z = \alpha) = 1] \\ \Leftrightarrow & A \stackrel{P}{\perp\!\!\!\perp} B. && [\text{Def. 4.37 (i)}] \end{aligned}$$

▷ **Solution 16-2**

$$\begin{aligned} \mathcal{D} \stackrel{P}{\perp\!\!\!\perp} \mathcal{E} | Z & \Leftrightarrow \forall (A, B) \in \mathcal{D} \times \mathcal{E}: A, B \stackrel{P}{\perp\!\!\!\perp} | Z && [\text{Def. 16.10 (ii)}] \\ & \Leftrightarrow \forall (A, B) \in \mathcal{D} \times \mathcal{E}: A \stackrel{P}{\perp\!\!\!\perp} B && [Z \stackrel{P}{=} \alpha, (16.9)] \\ & \Leftrightarrow \mathcal{D} \stackrel{P}{\perp\!\!\!\perp} \mathcal{E}. && [\text{Def. 4.40}] \end{aligned}$$

▷ **Solution 16-3** Proposition (i) of Box 16.2 directly follows from Definitions 16.2 (i) and 16.10 (i) of conditional independence and commutativity of  $\cap$  and multiplication.

Proposition (ii) of Box 16.2 follows from monotonicity of generated  $\sigma$ -algebras (see Def. 1.13 and Rem. 1.23).

(iii) If  $\stackrel{P}{\perp\!\!\!\perp} \mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$ , then, according to Equation (4.23),

$$\mathcal{D}_i \stackrel{P}{\perp\!\!\!\perp} \mathcal{D}_j, \quad i \neq j, \quad i, j = 1, 2, 3. \quad (16.59)$$

Furthermore, Equations (4.23) and Equation (4.24) yield

$$\{A \cap B: (A, B) \in \mathcal{D}_1 \times \mathcal{D}_2\} \stackrel{P}{\perp\!\!\!\perp} \mathcal{D}_3. \quad (16.60)$$

Hence,

$$\begin{aligned} \forall (A, B) \in \mathcal{D}_1 \times \mathcal{D}_2: P(A \cap B | \mathcal{D}_3) & \stackrel{P}{=} E(\mathbf{1}_{A \cap B} | \mathcal{D}_3) && [(10.2)] \\ & \stackrel{P}{=} E(\mathbf{1}_{A \cap B}) && [(16.60), \text{Box 10.1 (vi)}] \\ & = E(\mathbf{1}_A \cdot \mathbf{1}_B) && [(1.33)] \\ & = E(\mathbf{1}_A) \cdot E(\mathbf{1}_B) && [(16.59), \text{Box 6.1 (x)}] \\ & \stackrel{P}{=} E(\mathbf{1}_A | \mathcal{D}_3) \cdot E(\mathbf{1}_B | \mathcal{D}_3) && [(16.59), \text{Box 10.1 (vi)}] \\ & \stackrel{P}{=} P(A | \mathcal{D}_3) \cdot P(B | \mathcal{D}_3), && [(10.2)] \end{aligned}$$

i. e.,  $\mathcal{D}_1 \stackrel{P}{\perp\!\!\!\perp} \mathcal{D}_2 | \mathcal{D}_3$ .

(iv) This is an implication of Corollary 16.4.

(v) This is a special case of (iv).

(vi)

$$\begin{aligned} \mathcal{E} \perp\!\!\!\perp \mathcal{F} | \mathcal{C} &\Rightarrow A \perp\!\!\!\perp B | \mathcal{C}, \quad \forall (A, B) \in \mathcal{E} \times \mathcal{F} && [\text{Def. 16.10 (i)}] \\ &\Rightarrow A \perp\!\!\!\perp B | \mathcal{G}, \quad \forall (A, B) \in \mathcal{E} \times \mathcal{G} && [\mathcal{G} \subset \mathcal{F}] \\ &\Rightarrow \mathcal{E} \perp\!\!\!\perp \mathcal{G} | \mathcal{C}. && [\text{Def. 16.10 (i)}] \end{aligned}$$

(vii) Using  $\sigma[\mathcal{C}, \sigma(\mathcal{C}, \mathcal{D}_2)] = \sigma(\mathcal{C}, \mathcal{D}_2)$  [see Eq. (1.13)] yields:

$$\begin{aligned} \mathcal{D}_1 \perp\!\!\!\perp \mathcal{D}_2 | \mathcal{C} &\Leftrightarrow P(A | \mathcal{C}, \mathcal{D}_2) \stackrel{=}{=} P(A | \mathcal{C}), \quad \forall A \in \mathcal{D}_1 && [\text{Th. 16.22}] \\ &\Leftrightarrow P[A | \mathcal{C}, \sigma(\mathcal{C}, \mathcal{D}_2)] \stackrel{=}{=} P(A | \mathcal{C}), \quad \forall A \in \mathcal{D}_1 \\ &\Leftrightarrow \mathcal{D}_1 \perp\!\!\!\perp_P \sigma(\mathcal{C}, \mathcal{D}_2) | \mathcal{C}. && [\text{Th. 16.22}] \end{aligned}$$

(viii)  $\Rightarrow$

$$\mathcal{E} \perp\!\!\!\perp_P \sigma(\mathcal{D}_1, \mathcal{D}_2) | \mathcal{C} \Rightarrow \mathcal{E} \perp\!\!\!\perp_P \mathcal{D}_1 | \mathcal{C}, \quad [\text{Box 16.2 (vi)}]$$

and using  $\sigma[\mathcal{C}, \sigma(\mathcal{D}_1, \mathcal{D}_2)] = \sigma[\mathcal{D}_2, \sigma(\mathcal{D}_1, \mathcal{C})]$  (see Rem. 1.16) yields:

$$\begin{aligned} &\mathcal{E} \perp\!\!\!\perp_P \sigma(\mathcal{D}_1, \mathcal{D}_2) | \mathcal{C} \wedge \mathcal{E} \perp\!\!\!\perp_P \mathcal{D}_1 | \mathcal{C} \\ \Rightarrow &P[A | \mathcal{C}, \sigma(\mathcal{D}_1, \mathcal{D}_2)] \stackrel{=}{=} P(A | \mathcal{C}) \wedge P[A | \mathcal{C}, \mathcal{D}_1] \stackrel{=}{=} P(A | \mathcal{C}), \quad \forall A \in \mathcal{E} && [\text{Th. 16.22}] \\ \Rightarrow &P[A | \mathcal{D}_2, \sigma(\mathcal{D}_1, \mathcal{C})] \stackrel{=}{=} P(A | \mathcal{C}, \mathcal{D}_1), \quad \forall A \in \mathcal{E} && [\text{Rem. 2.73 (iii)}] \\ \Rightarrow &\mathcal{E} \perp\!\!\!\perp \mathcal{D}_2 | \sigma(\mathcal{D}_1, \mathcal{C}). && [16.2 (ii)] \\ &\Leftarrow \end{aligned}$$

$$\begin{aligned} &\mathcal{E} \perp\!\!\!\perp_P \mathcal{D}_1 | \mathcal{C} \wedge \mathcal{E} \perp\!\!\!\perp_P \mathcal{D}_2 | \sigma(\mathcal{C}, \mathcal{D}_1) \\ \Rightarrow &P(A | \mathcal{C}, \mathcal{D}_1) \stackrel{=}{=} P(A | \mathcal{C}) \wedge P[A | \mathcal{C}, \sigma(\mathcal{D}_1, \mathcal{D}_2)] \stackrel{=}{=} P(A | \mathcal{C}, \mathcal{D}_1), \quad \forall A \in \mathcal{E} && [(1.13), \text{Th. 16.22}] \\ \Rightarrow &P[A | \mathcal{C}, \sigma(\mathcal{D}_1, \mathcal{D}_2)] \stackrel{=}{=} P(A | \mathcal{C}), \quad \forall A \in \mathcal{E} && [\text{Rem. 2.73 (iii)}] \\ \Rightarrow &\mathcal{E} \perp\!\!\!\perp_P \sigma(\mathcal{D}_1, \mathcal{D}_2) | \mathcal{C}. && [\text{Th. 16.22}] \end{aligned}$$

(ix) This proposition immediately follows from (viii) for  $\mathcal{C} = \{\Omega, \emptyset\}$ .

(x)

$$\begin{aligned} \mathcal{F} \perp\!\!\!\perp_P \mathcal{G} | \mathcal{C} &\Rightarrow \sigma(\mathcal{F}, \mathcal{F}_0) \perp\!\!\!\perp_P \mathcal{G} | \mathcal{C} && [\mathcal{F}_0 \subset \mathcal{F}, \sigma(\mathcal{F}_0, \mathcal{F}) = \mathcal{F}] \\ &\Rightarrow \mathcal{F} \perp\!\!\!\perp_P \mathcal{G} | \sigma(\mathcal{C}, \mathcal{F}_0) && [(viii)] \\ &\Rightarrow \mathcal{F} \perp\!\!\!\perp_P \sigma(\mathcal{G}, \mathcal{G}_0) | \sigma(\mathcal{C}, \mathcal{F}_0) && [\mathcal{G}_0 \subset \mathcal{G}, \sigma(\mathcal{G}_0, \mathcal{G}) = \mathcal{G}] \\ &\Rightarrow \mathcal{F} \perp\!\!\!\perp_P \mathcal{G} | \sigma(\mathcal{C}, \mathcal{F}_0, \mathcal{G}_0). && [(viii), \sigma[\sigma(\mathcal{C}, \mathcal{F}_0), \mathcal{G}_0] = \sigma(\mathcal{C}, \mathcal{F}_0, \mathcal{G}_0), \text{Rem. 1.16}] \end{aligned}$$

(xi) First of all, note that for  $i = 1, 2$ ,

$$\forall A' \in \mathcal{A}'_X, \forall \omega \in \Omega: \mathbf{1}_{A'}[X_i(\omega)] = \mathbf{1}_{X_i^{-1}(A')}(\omega),$$

and

$$\{\omega \in \Omega: \mathbf{1}_{A'}[X_1(\omega)] \neq \mathbf{1}_{A'}[X_2(\omega)]\} \subset \{\omega \in \Omega: X_1(\omega) \neq X_2(\omega)\}. \quad (16.61)$$

Now, for all  $A \in \sigma(X_2)$ , let  $A'$  denote an element of  $\mathcal{A}'_X$  for which  $A = X_2^{-1}(A')$ . Then,

$$\forall A \in \sigma(X_2): \mathbf{1}_A = \mathbf{1}_{X_2^{-1}(A')}, \quad (16.62)$$

and, for all  $A \in \sigma(X_2)$  and all  $B \in \mathcal{E}$ ,

$$\begin{aligned}
 P(A \cap B | \mathcal{C}) &\stackrel{p}{=} E(\mathbf{1}_A \cdot \mathbf{1}_B | \mathcal{C}) && [(10.2), (1.33)] \\
 &\stackrel{p}{=} E(\mathbf{1}_{X_2^{-1}(A')} \cdot \mathbf{1}_B | \mathcal{C}) && [(16.62)] \\
 &\stackrel{p}{=} E(\mathbf{1}_{X_1^{-1}(A')} \cdot \mathbf{1}_B | \mathcal{C}) && [X_1 \stackrel{p}{=} X_2, (16.61), \text{Box 10.1 (ix)}] \\
 &\stackrel{p}{=} E(\mathbf{1}_{X_1^{-1}(A')} | \mathcal{C}) \cdot E(\mathbf{1}_B | \mathcal{C}) && [X_1 \perp\!\!\!\perp \mathcal{E} | \mathcal{C}] \\
 &\stackrel{p}{=} E(\mathbf{1}_{X_2^{-1}(A')} | \mathcal{C}) \cdot E(\mathbf{1}_B | \mathcal{C}) && [X_1 \stackrel{p}{=} X_2, (16.61), \text{Box 10.1 (ix)}] \\
 &\stackrel{p}{=} E(\mathbf{1}_A | \mathcal{C}) \cdot E(\mathbf{1}_B | \mathcal{C}) && [(16.62)] \\
 &\stackrel{p}{=} P(A | \mathcal{C}) \cdot P(B | \mathcal{C}). && [(10.2)]
 \end{aligned}$$

(xii) If  $X = \alpha$ , then  $\sigma(X) = \{\Omega, \emptyset\}$  (see Example 2.10) and Box 16.2 (v) implies  $X \perp\!\!\!\perp \mathcal{E} | \mathcal{C}$ . If  $X \stackrel{p}{=} \alpha$ , then applying Rule (xi) of Box 16.2 completes the proof.

▷ **Solution 16-4** Because  $P(Y=1 | X, U) \stackrel{p}{=} 1 - P(Y=0 | X, U)$  and  $P(Y=1 | U) \stackrel{p}{=} 1 - P(Y=0 | U)$  [see (10.2), Box 10.2 (i), (xv)],

$$\begin{aligned}
 P(Y=1 | X, U) \stackrel{p}{=} P(Y=1 | U) &\Leftrightarrow 1 - P(Y=0 | X, U) \stackrel{p}{=} 1 - P(Y=0 | U) && [\text{Rem. 2.73 (iii)}] \\
 &\Leftrightarrow P(Y=0 | X, U) \stackrel{p}{=} P(Y=0 | U). && [(2.36)]
 \end{aligned}$$

▷ **Solution 16-5** Using  $\sigma(\xi) \subset \sigma(U)$ , the notation introduced in Equation (10.4), and Remark (10.7): For all  $i = 1, \dots, m$ ,

$$\begin{aligned}
 &P(Y_i=1 | \xi, Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_m) \\
 &\stackrel{p}{=} E(P(Y_i=1 | U, Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_m) | \xi, Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_m) && [(13.37), \text{Box 10.2 (v)}] \\
 &\stackrel{p}{=} E(P(Y_i=1 | U) | \xi, Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_m) && [(13.37), \text{Box 10.2 (ix)}] \\
 &\stackrel{p}{=} E\left(\frac{\exp(\xi - \beta_i)}{1 + \exp(\xi - \beta_i)} \mid \xi, Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_m\right) && [(13.40)] \\
 &\stackrel{p}{=} \frac{\exp(\xi - \beta_i)}{1 + \exp(\xi - \beta_i)} && [(13.41), \text{Box 10.2 (vii)}] \\
 &\stackrel{p}{=} P(Y_i=1 | U) && [(13.40)] \\
 &\stackrel{p}{=} P(Y_i=1 | \xi). && [(13.41)]
 \end{aligned}$$

▷ **Solution 16-6** According to Corollary 16.50 (i) and (ii), Equation (16.55) is equivalent to  $\perp\!\!\!\perp_p Y_1, \dots, Y_m | \xi$ . Applying Corollary 16.47 yields that  $\perp\!\!\!\perp_p Y_1, \dots, Y_m | \xi$  is equivalent to (16.56).



## References

- Agresti, A. (2015). *Foundations of linear and generalized linear models*. John Wiley & Sons.
- Bauer, H. (1996). *Probability theory*. Berlin, Germany and New York, NY: de Gruyter.
- Bauer, H. (2001). *Measure and integration theory*. Berlin, Germany and New York, NY: de Gruyter.
- Billingsley, P. (1995). *Probability and measure* (3rd ed.). New York, NY: Wiley-Interscience.
- Bronshstein, I. N., Semendyayev, K. A., Musiol, G., & Mühlig, H. (2015). *Handbook of mathematics* (6th ed.). New York, NY: Springer.
- Ellis, R., & Gulick, D. (2006). *Calculus* (6<sup>th</sup> ed.). Mason, OH: Thomson.
- Elstrodt, J. (2007). *Maß- und Integrationstheorie [Measure and integration theory]* (4th ed.). Berlin, Germany: Springer.
- Fisz, M. (1963). *Probability theory and mathematical statistics*. New York, NY: Wiley.
- Fristedt, B., & Gray, L. (1997). *A modern approach to probability theory*. Boston, MA: Birkhäuser.
- Georgii, H.-O. (2008). *Stochastics – Introduction to probability and statistics*. Berlin, Germany: de Gruyter.
- Harris, J. W., & Stocker, H. (1998). *Handbook of mathematics and computational science*. New York, NY: Springer.
- Hoffmann-Jørgensen, J. (1994). *Probability with a view toward statistics* (Vol. 1). New York, NY: Chapman & Hall.
- Horn, R. A., & Johnson, C. R. (1991). *Matrix analysis*. Cambridge, UK: Cambridge University Press.
- Johnson, N. L., Kemp, A. W., & Kotz, S. (2005). *Univariate discrete distributions* (3rd ed.). New York, NY: Wiley.
- Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions* (2nd ed., Vol. 2). New York, NY: Wiley.
- Kheyfits, A. (2010). *A primer in combinatorics*. Berlin/New York: De Gruyter.
- Klenke, A. (2013). *Probability theory – A comprehensive course* (2nd ed.). London, England: Springer. doi: 10.1007/978-1-4471-5361-0
- Kolmogoroff, A. N. (1933/1977). *Grundbegriffe der Wahrscheinlichkeitsrechnung* (Reprinted ed.). Berlin, Germany: Springer.
- Kolmogorov, A. N. (1933/1977). *Grundbegriffe der Wahrscheinlichkeitsrechnung*

- [*Foundations of the Theory of Probability*] (Reprinted ed.). Berlin, Germany: Springer.
- Kolmogorov, A. N. (1956). *Foundations of the theory of probability* (2nd ed.; N. Morrison, Trans.). New York, NY: Chelsea.
- Mayer, A., Thoemmes, F., Rose, N., Steyer, R., & West, S. G. (2014). Theory and analysis of total, direct, and indirect causal effects. *Multivariate Behavioral Research*, 49(5), 425–442.
- McCullagh, P., & Nelder, J. A. (1989). *Monographs on statistics and applied probability: Vol. 37. Generalized linear models* (2nd ed.; D. R. Cox, D. V. Hinkley, N. Reid, D. B. Rubin, & D. V. Silverman, Eds.). Chapman & Hall.
- Michel, H. (1978). *Maß- und Integrationstheorie I. [Measure and integration theory I]*. Berlin: VEB Deutscher Verlag der Wissenschaften.
- Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York, NY: Wiley.
- Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests* (Expanded ed.). Chicago, IL: University of Chicago. (Originally published by Nissen & Lydicke, 1960. Kopenhagen.)
- Rosen, K. (2012). *Discrete mathematics and its applications* (7<sup>th</sup> ed.). New York, NY: McGraw-Hill.
- Steyer, R. (2003). *Wahrscheinlichkeit und Regression [Probability and regression]*. Berlin, Germany: Springer.
- Steyer, R., Mayer, A., & Fiege, C. (2014). Causal inference on total, direct, and indirect effects. In Michalos, A. C. (Ed.), *Encyclopedia of Quality of Life and Well-Being Research* (pp. 606–631). Dordrecht, Netherlands: Springer.
- Süli, E., & Mayers, D. F. (2003). *An introduction to numerical analysis*. Cambridge: Cambridge University Press.
- Tong, Y. L. (1990). *The multivariate normal distribution*. New York, NY: Springer.
- Wise, G. L., & Hall, E. B. (1993). *Counterexamples in probability and real analysis*. New York: Oxford University Press.

## Author Index

- Agresti, A., 382, 388, 541
- Balakrishnan, N., 264, 541
- Bauer, H., VI, 16, 24, 91–93, 98, 101, 102, 104, 106, 107, 109, 133, 174, 206, 251, 298, 319, 326, 426, 485, 498, 523, 541
- Billingsley, P., VI, 541
- Bronshtein, I. N., VIII, 541
- Ellis, R., VIII, 104, 188, 239, 394, 514, 541
- Elstrodt, J., 106, 541
- Fiege, C., 133, 542
- Fisz, M., 260, 263, 265, 267, 514, 541
- Fristedt, B., VI, 541
- Georgii, H.-O., 4, 7, 12, 145, 256, 258–260, 262, 264, 266, 541
- Gray, L., VI, 541
- Gulick, D., VIII, 104, 188, 239, 394, 514, 541
- Hall, E. B., 7, 23, 542
- Harris, J. W., VIII, 273, 541
- Hoffmann-Jørgensen, J., VI, 541
- Horn, R. A., 373, 541
- Johnson, C. R., 373, 541
- Johnson, N. L., 253, 264, 265, 267, 541
- Kemp, A. W., 253, 541
- Kheyfits, A., 8, 541
- Klenke, A., VI, 12, 13, 15, 22, 25, 27, 50, 51, 59, 62, 64, 80, 86, 90, 104, 105, 109, 110, 133, 160, 217, 252, 257, 306, 307, 325, 378, 498, 522, 541
- Kolmogoroff, A. N., V, VI, 541
- Kolmogorov, A. N., 127, 295, 541, 542
- Kotz, S., 253, 264, 541
- Mayer, A., 133, 542
- Mayers, D. F., 374, 542
- McCullagh, P., 382, 388, 542
- Michel, H., 11, 542
- Mühlig, H., VIII, 541
- Musiol, G., VIII, 541
- Nelder, J. A., 382, 388, 542
- Rao, C. R., 370, 542
- Rasch, G., 392, 542
- Rose, N., 133, 542
- Rosen, K., VII, VIII, 542
- Semendyayev, K. A., VIII, 541
- Steyer, R., VI, 133, 152, 542
- Stocker, H., VIII, 273, 541
- Süli, E., 374, 542
- Thoemmes, F., 133, 542
- Tong, Y. L., 261, 542
- West, S. G., 133, 542
- Wise, G. L., 7, 23, 542

# Subject Index

- absolute continuity, 109
  - and  $P$ -equivalence, 164
  - and independence, 172
  - of a conditional-probability measure, 141
  - of marginal distributions, 181
- absolute value function, 63
- additivity of a measure
  - $\sigma$ -additivity, 17
  - finite additivity, 17
- adjusted conditional effect function, 444
- adjusted effect, 444
- adjusted logit effect, 448
- almost all, 161
- almost everywhere, 100
- average effect, 444
  
- Bayes' theorem
  - for densities, 516
  - for events, 139
- Bernoulli distribution, 246
- Bernoulli variable, 246
- binomial coefficient, 247
- binomial distribution, 247
  - approximation by Poisson distribution, 251
- bivariate normal distribution, 260
- Borel  $\sigma$ -algebra on  $\mathbb{R}$ , 12
- Borel  $\sigma$ -algebra on  $\mathbb{R}^n$ , 13, 15
- Borel set, 12
  
- cardinality, 8, 21
- Cartesian product, 3
- Cauchy distribution, 264
- Cauchy-Schwarz inequality, 223
- central  $F$ -distribution, 266
- central  $t$ -distribution, 264
  
- central limit theorem, 257
- central moment, 208
- $\chi^2$ -distribution, 262
- closed interval, 6
- CNS-uniqueness, 499
- codomain of a mapping, 41
- coefficient of determination
  - of a conditional expectation, 334
  - of a linear quasi-regression, 218, 232
- coefficient of variation, 211
- common null set uniqueness, 499
- complement of a set, 5
- composition of two mappings, 58
- conditional correlation
  - given a value of a random variable, 346
- conditional covariance
  - given a  $\sigma$ -algebra, 338
  - given a random variable, 339
  - given a value of a random variable, 340
  - rules of computation, 343
- conditional density, 515, 517
- conditional distribution, 492
  - and  $P$ -equivalence, 502
  - and conditional independence, 508
  - and conditional independence given a random variable, 507
  - and independence, 497
  - and mean independence, 509
  - existence, 498
  - of a random variable given a  $\sigma$ -algebra, 492
  - of a random variable given a random variable, 492

- of a random variable given a value
    - of a random variable, 494
- conditional distribution function, 515
- conditional effect function, 439
- conditional expectation
  - w.r.t. a conditional-probability measure, 406–408
  - and conditional densities, 516, 517
  - and joint distribution, 312
  - and mean-squared error, 307
  - coefficient of determination, 334
  - convergence, 305, 306
  - discrete, 283
  - given a  $\sigma$ -algebra, 296
  - given a random variable, 296
  - linear parametrization, 360
  - marginalization, 512
  - monotonicity, 305
  - rules of computation, 303, 304
  - uniqueness w.r.t. a probability measure, 298
- conditional expectation value
  - and conditional distribution, 511
  - given a value of a random variable, 278, 309
  - given an event, 278
  - of a composition, 280
  - rules of computation, 283, 284
- conditional independence
  - of random variables given an event, 174
- conditional independence given a  $\sigma$ -algebra
  - and conditional mean independence, 470
  - characterizations, 472
  - family of  $\cap$ -stable set systems, 474
  - family of events, 473
  - family of random variables, 475
  - family of set systems, 474
  - notation, 465
  - of two events, 460
  - of two random variables, 462
  - of two set systems, 461
  - properties, 469
- conditional independence given a random variable
  - and conditional distributions, 506
  - family of events, 473
  - family of random variables, 475
  - family of set systems, 474
  - of two events, 460
  - of two random variables, 462
  - of two set systems, 462
  - properties, 470
- conditional independence given an event
  - family of  $\cap$ -stable set systems, 148
  - family of set systems, 146
  - of two events, 145
- conditional intercept function, 439
- conditional logit effect function, 448
- conditional logit intercept function, 448
- conditional mean independence, 313
  - and conditional independence, 470
  - and independence, 314
  - from a  $\sigma$ -algebra, 313
  - from random variable, 313
- conditional probability
  - w.r.t. a conditional-probability measure, 407, 408
  - given a  $\sigma$ -algebra, 296
  - given a random variable, 296
  - given a value of a random variable, 309
  - given an event, 132
  - of an event given a value of a random variable, 279
  - of an event given an event, 279
- conditional probability density, *see* conditional density
- conditional standard deviation
  - given a  $\sigma$ -algebra, 339
  - given a random variable, 339
  - given a value of a random variable, 340
- conditional variance
  - given a  $\sigma$ -algebra, 339
  - given a random variable, 339
  - given a value of a random variable, 340
  - rules of computation, 344
- conditional-probability measure, 140, 503
- constant mapping, 48
- continuity of a measure from above, 25
- continuity of a measure from below, 25
- continuous random variable, 186

- without expectation, 202
- continuous uniform distribution, 130, 254
- convergence of conditional expectations, 305, 306
- convergence of integrals, 103
- correlation, 225
  - and slope of a linear quasi-regression, 226
  - invariance under linear transformations, 225
- countable intersection, 6
- countable union, 5
- countably generated  $\sigma$ -algebra, 11
- counting measure, 21
- covariance, 221
  - rules of computation, 222
- covariance matrix, 230
  - rules of computation, 231
- density, *see also* probability density 186
  - and probability function, 180
  - of the  $\chi^2$ -distribution, 263
  - of the  $F$ -distribution, 266
  - of the  $t$ -distribution, 264
  - of the bivariate normal distribution, 260
  - of the Cauchy distribution, 264
  - of the multivariate normal distribution, 259
  - of the standard normal distribution, 108
  - of the univariate normal distribution, 256
- density of a measure w.r. t. another measure, 106
- dichotomous function, 48
- dichotomous random variable, 158
- Dirac measure, 21
- discrete conditional expectation given a random variable, 283
- discrete conditional probability given a random variable, 284
- discrete distribution, 175
- discrete random variable, 175
  - without expectation, 201
- discrete regression, 285
- discrete uniform distribution, 245
- disjoint sets, 4
- distribution, 156
  - and distribution function, 182, 185
  - Bernoulli, 246
  - binomial, 247
  - $\chi^2$ , 262
  - Cauchy, 264
  - discrete uniform, 245
  - $F$ , 266
  - geometric, 252
  - multivariate normal, 259
  - of a composition, 157
  - of an indicator, 157
  - Poisson, 250
  - $t$ , 264
  - univariate normal, 256
  - univariate standard normal, 256
- distribution function, 182
  - and distribution, 182, 185
  - and independence, 185
  - joint and marginal distribution function, 185
  - of a binomial distribution, 247
  - of the geometric distribution, 253
  - of the Poisson distribution, 250
  - of the standard normal distribution, 256
  - of the univariate normal distribution, 256
- domain of a mapping, 41
- dominated convergence of integrals, 104
- Dynkin system, 16
- effect function, 439
- elementary event, 128
- elementary function, 83
- equivalence w.r. t. a measure, 65
  - necessary and sufficient conditions, 107
- equivalence w.r. t. a probability measure of random variables, 161
  - of two factorizations of a conditional expectation, 308
- equivalence class w.r. t. a measure, 66
- equivalence of probability densities w.r. t. a measure, 179
- equivalence relation, 65
- event, 127
- existence
  - of a conditional distribution, 498

- of a conditional expectation given a  $\sigma$ -algebra, 297
  - of an expectation, 199
- expectation
  - w.r.t.  $P$ , 199
  - w.r.t. a conditional-probability measure, 200, 279
  - of a distribution, 204
  - of a random matrix, 228
  - of a random variable, 199
  - of a random variable with a countable number of real values, 201
  - of a random variable with a finite number of real values, 200
  - of a random variable with density, 202
  - of a random vector, 228
  - of a sample mean, 206
  - of an indicator, 200
  - of the product of random variables under independence, 206
  - of the product of two random variables, 221
  - rules of computation, 207
- expectation of a random matrix
  - rules of computation, 229
- factorial of an integer, 247
- factorization
  - equivalence of two versions w.r.t. a probability measure, 308
  - of a conditional expectation given a random variable, 307
  - uniqueness, 308
- factorization of a composition, 59
- $F$ -distribution, 266
- filtration, 134
- final  $\sigma$ -algebra, 54
- finite additivity of a measure, 17
- finite intersection, 6
- finite measure, 23
- finite union, 5
- Fubini's theorem, 110
- $\Gamma$ -function, 262
- generating system
  - of a  $\sigma$ -algebra, 9
  - of a product  $\sigma$ -algebra, 16
- geometric distribution, 252
- distribution function, 253
- identically distributed random variables, 157
- identification
  - linear logit regression, 388
  - linear regression, 368
- identity
  - of mappings, 46
  - of random variables, 161
- identity mapping, 48
- i. i. d., 171
- image measure, 68
  - under a step function, 69
- image of a set under a mapping, 41, 43
- increasing sequence of nonnegative step functions, 91
- independence
  - and  $P$ -equivalence, 172
  - and absolute continuity, 172
  - and conditional mean independence, 314
  - and distribution function, 185
  - and probability densities, 187
  - and product measure, 172
  - conditional, *see* conditional independence
  - family of  $\sigma$ -algebras, 144
  - family of events, 143
  - family of random variables, 171
  - family of set systems, 144
  - measurable mappings, 174
  - of  $\cap$ -stable set systems, 145
  - of  $n$  random variables, 171
  - of a constant and a set of events, 173
  - of a random variable and a set system, 171
  - of an event and a set system, 144
  - of three events, 144
  - of two events, 143
  - of two random variables, 169
- independent and identically distributed, 171
- indicator, 20
- infinite measure, 23
- integrable, 94
- integral
  - w.r.t. a Dirac measure, 92

- w.r.t. a finite weighted sum of measures, 101
- w.r.t. a measure with density, 107
- w.r.t. a weighted sum of Dirac measures, 101
- w.r.t. a weighted sum of measures, 101
- w.r.t. an image measure, 102
- w.r.t. the Lebesgue measure, 94
- of  $\mu$ -equivalent functions, 99
- of a constant, 86
- of a function with a finite number of values, 103
- of a measurable function, 94
- of a nonnegative measurable function, 92
- of a nonnegative step function, 86
- of a positive measurable function, 98
- over a null set, 98
- over a subset, 86, 96
- over the union of two sets, 97
- intercept
  - of a simple linear quasi-regression, 220
  - of a simple linear regression, 362
- intercept function, 439
- intersection
  - of countably many sets  $A_1, A_2, \dots$ , 6
  - of finitely many sets  $A_1, \dots, A_n$ , 6
- invariance of regression coefficients, 375
- inverse image of a set under a mapping, 41
- joint distribution, 168
- joint distribution function, 184
- Lebesgue integral and Riemann integral, 104
- Lebesgue measure, 22
- linear combination of two functions, 62
- linear logistic regression, 388
- linear logit parametrization, 384
  - identification, 388
- linear logit regression, 388
- linear parametrization
  - differences between means as coefficients, 371
  - identification, 368
  - means as coefficients, 370
- linear parametrization of a conditional expectation, 360
- linear quasi-regression, 217
  - and linear regression, 366
  - and regression, 308
  - equivalent characterizations, 223
- linear regression, 369
  - and linear quasi-regression, 366
  - and normal distribution, 370
  - identification, 368
- linearity
  - of the integral, 96
- link function, 382
- log odds, 386
- log odds ratio, 386
- log-odds, 382
- logit effect function, 448
- logit intercept function, 448
- logit of a conditional probability, 382
- logit transformation, 382
- marginal density, 181
- marginal distribution, 168
- marginal distribution function, 515
- marginal probability density, 515
- marginalization of a conditional expectation, 512
- mean centered random variable, 209
- mean independence, 313
  - and conditional distributions, 509
  - and uncorrelatedness, 318
- mean squared error, 217, 232
- measurability w.r.t. a mapping, 56
- measurable mapping, 46
- measurable set, 5
- measurable space, 6
- measure, 17
- measure space, 17
- measure with density, 106
- mixture of probability measures, 132
- moment, 208
- moment of a numerical random variable, 208
- monotonicity
  - of a conditional expectation, 305
  - of a measure, 19
  - of an integral, 93, 100
- MSE*, mean squared error function, 217

- $\mu$ -almost everywhere, 100
- multiple correlation, 336
- multiple linear logistic regression, 389
- multiple linear quasi-regression, 232
- multiple linear regression, 369
- multiplication rule for probabilities, 135
- multivariate Bernoulli distribution, 247
- multivariate Bernoulli variable, 246
- multivariate mapping, 54
- multivariate normal distribution, 259
- multivariate random variable, 167
  
- negative part of a function, 62
- nonnegative step function, 83
  - and counting measure, 88
  - and Dirac measure, 87
  - normal representation, 84
- normal distribution, 256
  - and linear regression, 370
  - bivariate, 260
  - density, 256
  - multivariate, 259
  - standard, 256
  - univariate, 256
- normal representation of a nonnegative step function, 84
- null set, 28
  - integral over a null set, 98
- null-set equivalence of two measures, 109
  
- odds ratio, 387
- outcome of a random experiment, 128
  
- pairwise independence, 143
- $P$ -almost all, 161
- partial  $(X, Z=z)$ -conditional expectation, 412
- partial correlation, 345
- partition of a set, 10
- $P$ -equivalence, 161
  - and absolute continuity, 164
  - and conditional distributions, 502
  - and distributions, 163
  - and independence, 172
  - of compositions, 163
  - of random variables, 161
- $P^B$ -expectation, 200
- pointwise convergence, 91
- Poisson distribution, 250
  - approximation of the binomial distribution, 251
  - distribution function, 250
- positive part of a function, 62
- power set, 6
- probability density, *see also* density 179
  - of a continuous real-valued random variable, 186
  - of a probability measure, 179
  - of a random variable, 179
- probability function, 175
  - and density, 180
  - of a marginal distribution, 177
- probability measure, 127
  - with density, 178
- probability of an event, 127
- probability space, 127
- product  $\sigma$ -algebra, 14, 56
- product measure, 24
  - and independence, 172
- product set, 3
- projection mapping, 56
- $P$ -uniqueness, 298
  - of a conditional expectation, 298
  - of a conditional expectation w.r.t. conditional-probability measure, 421
- $P^B$ -uniqueness, 411
  
- quantile, 183
- quantile function, 183
  - and inverse distribution function, 183
- quasi-integrable, 94
- quasi-regression
  - linear, 217
  
- Radon-Nikodym derivative, 109
- Radon-Nikodym theorem, 109
  - and probability density, 179
- random sample, 171
- random variable, 155
  - continuous, 186
  - discrete, 175
  - discrete real-valued, 176
  - numerical, 156
  - real-valued, 156
- regressand, 285, 307
- regression, 307

- and linear quasi-regression, 308
  - discrete, 285
- regression coefficients, 369
  - invariance, 375
- regressor, 285, 307
- residual
  - w.r.t. a conditional expectation, 330
  - w.r.t. a linear quasi-regression, 223, 232
- residual w.r.t. a conditional expectation
  - rules of computation, 332
- restriction of a measure, 23
- Riemann integral and Lebesgue integral, 104
- risk ratio, 387
- rules of computation
  - for a residual w.r.t. a conditional expectation, 332
  - for conditional covariances, 343
  - for conditional expectation values given a value of a random variable, 284
  - for conditional expectation values given an event, 283
  - for conditional expectations given a  $\sigma$ -algebra, 303
  - for conditional expectations given a random variable, 304
  - for conditional variances, 344
  - for covariance matrices, 231
  - for covariances, 222
  - for expectations of random matrices, 229
  - for measures, 19
  - for probabilities, 129
  - for the expectation of a random variable, 207
  - for variances, 209
- set system, 4
- $\sigma$ -additivity of a measure, 17
- $\sigma$ -algebra, 5
  - Borel, 12
  - countably generated, 11
  - final, 54
  - generated by a composition, 59
  - generated by a family of mappings, 55
  - generated by a mapping, 51
  - generated by a multivariate mapping, 54
  - generated by a set system, 9
  - generated by an indicator, 52
  - trivial, 28
- $\sigma$ -field, *see*  $\sigma$ -algebra
- $\sigma$ -finite measure, 23
- $\sigma$ -subadditivity, 19
- sign function, 64
- simple function, 83
- simple linear logistic regression, 389
- simple linear regression, 362, 369
  - identification of the intercept, 370
  - identification of the slope, 370
  - intercept, 362
  - slope, 362
- singleton, 6
- skewness, 212
- slope
  - of a linear quasi-regression and correlation, 226
  - of a simple linear quasi-regression, 220
  - of a simple linear regression, 362
- stability of a set system w.r.t. intersections,  $\cap$ -stability, 15
- standard deviation, 208
  - of the sample mean, 210
- standard error
  - of the sample mean, 210
- standard normal distribution, 256
- step function, 49
  - normal representation, 84
- strictly diagonally dominant matrix, 372
- $t$ -distribution, 264
- theorem
  - B. Levi, 103
  - Bayes' theorem for densities, 516
  - Bayes' theorem for events, 139
  - central limit, 257
  - Fubini, 110
  - of total probability, 137
  - Radon-Nikodym, 109
  - transformation theorem for a conditional expectation value, 280
  - transformation theorem for an integral, 102

- transformation theorem of an expectation, 203
- time order
  - between events, 134
  - between random variables, 160
- trace  $\sigma$ -algebra, 8
- trace of a set system, 8
- transformation theorem
  - for a conditional expectation value, 280
  - for an expectation, 203
  - for an integral, 102
  - for conditional distributions, 511
- triple-wise independence, 144
- trivial  $\sigma$ -algebra, 28
- trivial  $\sigma$ -algebra w.r.t. a measure, 28
  
- uncorrelated random variables, 221
- uncorrelatedness
  - mean independence, 318
- uncountable union, 6
- uniform distribution, 254
- union
  - of countably many sets, 5
  - of finitely many sets, 5
- uniqueness
  - of a conditional expectation w.r.t. a probability measure, 298
  - of a conditional expectation given a  $\sigma$ -algebra, 297
  - of a factorization, 308
- univariate normal distribution, 256
  - density, 256
  - distribution function, 256
  
- variance, 208
  - of an indicator, 210
  - of the sample mean, 210
  - rules of computation, 209
  
- Z-transformation, 211



## List of Symbols

$\neg$	Aa not, IX
$\vee$	Ab or, IX
$\wedge$	Ab and, IX
$\Rightarrow$	Aca implies, IX
$\Leftrightarrow$	Acb equivalent to, IX
$\exists$	Ada there is (synonymously, there exists), IX
$\forall$	Adb for all, IX
$\in$	Sacaaa element of, IX
$1_A$	SadaaaVmqaqa indicator (function) of the set $A$ , 20
$\cup$	Saeaaa union of sets, IX
$\cap$	Safaaa intersection of sets, IX
$\setminus$	Sahaaa set difference, IX
$A^c := \Omega \setminus A$	Sahcaa complement of a set $A$ with respect to a set $\Omega$ , IX
$\subset$	Sajaaa subset or equal, IX
$\times$	Sajbaa Cartesian product or product set, IX
$\otimes$	Sajbba product $\sigma$ -algebra, 14
$\otimes$	Sajbba product measure, 24
$\circ$	Sajbbb composition of two mappings, 58
$\odot$	SajbbbUfoaaa measure with density, 106
$\stackrel{\mu}{=}$	SajbbbVmaaaa $\mu$ -equivalence of mappings, 65
$\stackrel{\mu}{<}$	SajbbbVmhaaa smaller than except for a $\mu$ -null set, 67
$\stackrel{\mu}{>}$	SajbbbVmkaaa greater than except for a $\mu$ -null set, 67
$\stackrel{\mu}{\leq}$	SajbbbVmmeta smaller than or equal except for a $\mu$ -null set, 67
$\stackrel{\mu}{\geq}$	SajbbbVmoaaa greater than or equal except for a $\mu$ -null set, 67
$\stackrel{P}{=}$	SajbbbVmogaa $P$ -equivalence of random variables, 161
$\stackrel{\mu}{=} \mu\text{-a.a.}$	SajbbbVmsaaa equal for $\mu$ -almost all $\omega \in \Omega$ , 65
$\ll_{\mathcal{A}}$	SajbbbVpuaaa absolute continuity of a measure with respect to another measure, 109
$\approx_{\mathcal{A}}$	SajbbbVpvaaa null-set equivalence of two measures, 109

$\perp\!\!\!\perp_P$	SajbbbVtvaaa independence with respect to the probability measure $P$ , 143
$[a, b]$	Sajbca closed interval between real numbers $a$ and $b$ , 6
$]a, b]$	Sajbcc half-open interval including $b$ but not $a$ , 12
$\{x\}$	Sajbcg singleton, i. e., the set that contains $x$ as the only element, 6
$\emptyset$	Sajbcj empty set, IX
$(A_i, i \in I)$	Sajhcg family of sets $A_i, i \in I$ , where the index set $I$ may be finite, countable, or uncountable, IX
$\bigcup_{i \in I} A_i$	Sakaaa union of the sets $A_i, i \in I$ , IX
$\bigcap_{i \in I} A_i$	Samaaa intersection of the sets $A_i, i \in I$ , IX
$\bigcup_{i=1}^n A_i$	Samcaa union of finitely many sets $A_1, \dots, A_n$ , 5
$\bigcup_{i=1}^{\infty} A_i$	Sameaa union of countably many sets $A_1, A_2, \dots$ , 5
$\bigcap_{i=1}^n A_i$	Samfba intersection of finitely many sets $A_1, \dots, A_n$ , 6
$\bigcap_{i=1}^{\infty} A_i$	Samfca intersection of countably many sets $A_1, A_2, \dots$ , 6
$\prod_{i=1}^n A_i$	Samfga Cartesian product or product set of $n$ sets $A_i$ , IX
$\lim_{n \rightarrow \infty} a_n$	Sbaaaa limit of a sequence $a_1, a_2, \dots$ of real numbers, IX
$\sum_{i=1}^n a_i$	Scaaaa sum of the numbers $a_1, \dots, a_n$ , IX
$\sum_{i=1}^{\infty} a_i$	Sccaaaa $\lim_{n \rightarrow \infty} \sum_{i=1}^n a_i$ , IX
$\prod_{i=1}^n a_i$	Sdaaaa product of the real numbers $a_1, \dots, a_n$ , IX
$\mathcal{A}_1 \otimes \dots \otimes \mathcal{A}_n$	Sdchaa product $\sigma$ -algebra of the $\mathcal{A}_1, \dots, \mathcal{A}_n$ , 14
$\bigotimes_{i=1}^n \mathcal{A}_i$	Seaaaa product $\sigma$ -algebra of the $\mathcal{A}_1, \dots, \mathcal{A}_n$ , 14
$\int f d\mu$	Siaaaa integral of a measurable function $f: (\Omega, \mathcal{A}, \mu) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ with respect to the measure $\mu$ , 94
$\int_A f d\mu$	Sigaaa integral of a measurable function $f: (\Omega, \mathcal{A}, \mu) \rightarrow (\mathbb{R}, \mathcal{B})$ with respect to the measure $\mu$ over a subset $A$ of $\Omega$ , 96
$\int_a^b f(x) dx$	Simaaa Riemann integral of the function $f$ from $a$ to $b$ , 105
$\mathcal{A}  _{\Omega_0}$	Svcaaa trace of the set system $\mathcal{A}$ in the set $\Omega_0$ , 8
$\mathcal{C}'_f$	Svgaaa final $\sigma$ -algebra of $\mathcal{C}$ under $f$ , 54
$(\Omega, \mathcal{A})$	Tcaaaa measurable space, 6
$(\Omega, \mathcal{A}, \mu)$	Tcfaaa measure space, 17
$(\Omega, \mathcal{A}, P)$	Tcjaaa probability space, 127
$f: A \rightarrow B$	Uaaaaa mapping $f$ assigning to each $a \in A$ one and only one element $b \in B$ , IX
$f(A)$	Udaaaa image of the set $A$ under $f$ , 41
$f^{-1}(A')$	Ugaaaa inverse image of the set $A'$ under $f$ , 41
$\{f \in A'\}$	Uhaaaa $:= f^{-1}(A')$ , 41
$\{f = \omega'\}$	Ujaaaa $:= f^{-1}(\{\omega'\})$ , 42
$(f_i, i \in I)$	Umaaaa family of mappings, 55
$f^{-1}(\mathcal{E}')$	Umaaaa set of the inverse images of all sets $A' \in \mathcal{E}'$ . If $\mathcal{E}'$ is a $\sigma$ -algebra, then $f^{-1}(\mathcal{E}')$ is the $\sigma$ -algebra generated by $f$ , 49
$g \circ f, g(f)$	Umgaaa composition of $f$ and $g$ , 58

$ f $	Umiaaa absolute value function of $f$ , 63
$f^+$	Umkaaa positive part of the function $f$ , 62
$f^-$	Umlaaa negative part of the function $f$ , 62
$f_n \uparrow f$	Umooaaa the sequence $f_1, f_2, \dots$ of functions converges point-wise and monotonically from below to $f$ , 91
$f: (\Omega, \mathcal{A}) \rightarrow (\Omega', \mathcal{A}')$	Vaaaaa $(\Omega, \mathcal{A}), (\Omega', \mathcal{A}')$ are measurable spaces and the mapping $f: \Omega \rightarrow \Omega'$ is $(\mathcal{A}, \mathcal{A}')$ -measurable, 47
$f: (\Omega, \mathcal{A}, \mu) \rightarrow \Omega'$	Veaaaa $(\Omega, \mathcal{A}, \mu)$ is a measure space and $f: \Omega \rightarrow \Omega'$ is a mapping, 64
$f: (\Omega, \mathcal{A}, \mu) \rightarrow (\Omega', \mathcal{A}')$	Vhaaaa $f: \Omega \rightarrow \Omega'$ is an $(\mathcal{A}, \mathcal{A}')$ -measurable mapping and $\mu$ is a measure on $(\Omega, \mathcal{A})$ , 64
$f \underset{\mu}{=} g$	Vmaaaa the mappings $f$ and $g$ are $\mu$ -equivalent, 65
$f \underset{\mu}{<} g$	Vmhaaa $f$ is smaller than $g$ except for a set $A$ of arguments with $\mu(A) = 0$ , 67
$f \underset{\mu}{>} g$	Vmkaaa $f$ is greater than $g$ except for a set $A$ of arguments with $\mu(A) = 0$ , 67
$f \underset{\mu}{\leq} g$	Vmmaaa $f$ is smaller than or equal to $g$ except for a set $A$ of arguments with $\mu(A) = 0$ , 67
$f \underset{\mu}{\geq} g$	Vmoaaa $f$ is greater than or equal to $g$ except for a set $A$ of arguments with $\mu(A) = 0$ , 67
$f(\omega) \underset{\mu\text{-a.a.}}{=} g(\omega)$	Vmsaaa $f(\omega) = g(\omega)$ , for $\mu$ -almost all $\omega \in \Omega$ , 65
$\mu_f$	Vptaaa image measure of $\mu$ under $f$ , 68
$\nu \ll_{\mathcal{A}} \mu$	Vpuaaa the measure $\nu$ is absolutely continuous with respect to the measure $\mu$ , 109
$\nu \underset{\mathcal{A}}{\approx} \mu$	Vpvaaa the measures $\mu$ and $\nu$ are null-set equivalent, i. e., they are absolutely continuous with respect to each other, 109
$A \underset{P}{\perp} B$	Vtvaaa the events $A$ and $B$ are independent with respect to the probability measure $P$ , 143
$\underset{P}{\perp} (A_i, i \in I)$	Vtvea a family of events $A_i$ that are independent with respect to the probability measure $P$ , 143
$A \underset{P}{\perp} C   B$	Vtvdaa the events $A$ and $C$ are $B$ -conditionally independent with respect to the probability measure $P$ , 145
$\mathcal{E}_1 \underset{P}{\perp} \mathcal{E}_2$	Vtvgaa the set systems $\mathcal{E}_1$ and $\mathcal{E}_2$ are independent with respect to the probability measure $P$ , 144
$\underset{P}{\perp} (\mathcal{E}_i, i \in I)$	Vtvhaa a family of set systems $\mathcal{E}_i$ that are independent with respect to the probability measure $P$ , 144
$A \underset{P}{\perp} \mathcal{E}$	Vtvoaa the event $A$ and the set system $\mathcal{E}$ are independent with respect to the probability measure $P$ , 144
$\underset{P}{\perp} (X_i, i \in I)$	Vtvtha a family of random variables $X_i$ that are independent with respect to the probability measure $P$ , 171

$\perp_P X_1, \dots, X_n$	Vtvta the random variables $X_1, \dots, X_n$ are independent with respect to the probability measure $P$ , 171
$\mathcal{E} \perp_P X$	Vtvtoa the set system $\mathcal{E}$ and the random variable $X$ are independent with respect to the probability measure $P$ , 171
$X \perp_P (Y_i, i \in I)$	Vtvtua the random variable $X$ and the ( $\sigma$ -algebra generated by the) family of random variables are independent with respect to the probability measure $P$ , 171
$\perp_P (\mathcal{E}_i, i \in I)   B$	Vtwaa a family of set systems $\mathcal{E}_i$ that are $B$ -conditionally independent with respect to the probability measure $P$ , 146
$\bar{Y}$	Xtyaaa arithmetic mean (sample mean) of the real-valued random variables $Y_1, \dots, Y_n$ , 206
$\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}, \mathcal{E}$	Zababe set systems, sometimes $\sigma$ -algebras, 4
$\mathcal{A}_1 \otimes \dots \otimes \mathcal{A}_n$	Zaeabe product $\sigma$ -algebra of the $\mathcal{A}_1, \dots, \mathcal{A}_n$ , 14
$\mathcal{A}  _{\Omega_0}$	Zahabe trace of the set system $\mathcal{A}$ in the set $\Omega_0$ , 8
$\mathcal{B}$	Zbcaaa Borel $\sigma$ -algebra on $\mathbb{R}$ , 12
$\mathcal{B}_n$	Zbccaa Borel $\sigma$ -algebra on $\mathbb{R}^n$ , 13
$\bar{\mathcal{B}}$	Zbcfaa Borel $\sigma$ -algebra on $\bar{\mathbb{R}}$ , 13
$\bar{\mathcal{B}}_n$	Zbgaaa Borel $\sigma$ -algebra on $\bar{\mathbb{R}}^n$ , 13
$\mathcal{B}_{n,p}$	Zbvaaa binomial distribution with parameters $n$ and $p$ , 247
$b_{n,p}$	Zbvaaab probability function of the binomial distribution with parameters $n$ and $p$ , 247
$Cov(X, Y)$	Zcmeta covariance of the random variables $X$ and $Y$ , 221
$Corr(X, Y)$	Zcgaaa correlation of the random variables $X$ and $Y$ , 225
$\mathcal{C}'_f$	Zchaaa final $\sigma$ -algebra of $\mathcal{C}$ under $f$ , 54
$\chi_n^2$	Zcvaaa central $\chi^2$ -distribution with $n$ degrees of freedom, 263
$\frac{d\nu}{d\mu}$	Zdgaaa Radon-Nikodym density (also Radon-Nikodym derivative) of $\nu$ with respect to $\mu$ , 109
$\delta_\omega$	Zdmeta Dirac measure at (point) $\omega$ , 21
$E(Y)$	Zemeta expectation of the random variable $Y$ , 199
$E(\mathbf{x})$	Zefaaa column vector of expectations, 228
$E^B(Y)$	Zehaaa expectation of the random variable $Y$ with respect to the probability measure $P^B$ , 200
$E(Y B)$	Zeiaaa conditional expectation value of $Y$ given the event $B$ , 278
$E(Y X=x)$	Zeifaa conditional expectation value of $Y$ given the event $\{X=x\}$ , also denoted by $E(Y \{X=x\})$ , 278
$E(\mathbf{X})$	Zejaaa matrix of expectations, 228
$E_Y(g)$	Zejaaa expectation of the random variable $g$ with respect to the distribution of the random variable $Y$ , 203
$E_Y^{X=x}(g)$	Zelaaa expectation of $g$ with respect to the distribution $P_Y^{X=x}$ , 280
$E^{X=x}(Y)$	Zemaaa expectation of $Y$ with respect to the conditional-probability measure $P^{X=x}$ , 279

$E^{\{X=x\}}(Y)$	Zemaaa expectation of $Y$ with respect to the conditional-probability measure $P^{X=x}$ , 279
$\epsilon$	Zepbaa residual with respect to a (multiple) linear quasi-regression, 232
$F_X$	Zfcaaa distribution function of a real-valued random variable $X$ , 182
$F_{X_1, \dots, X_n}$	Zfcgaa joint distribution function of $X_1, \dots, X_n$ , 184
$F(x) \Big _a^b$	Zfgaaa $F(b) - F(a)$ , 105
$f_X$	Zfgbaa probability density of a continuous real-valued random variable $X$ , 186
$ f $	Zfgjaa absolute value function of $f$ , 63
$f^+$	Zfhaaa positive part of the function $f$ , 62
$f^-$	Zfiaaa negative part of the function $f$ , 62
$(f_i, i \in I)$	Zflaaa family of mappings, 55
$f \circ \mu$	Zfoaaa measure with density $f$ with respect to $\mu$ , 106
$F_{m,n}$	Zfvaaa $F$ -distribution with $m$ and $n$ degrees of freedom, 266
$g \circ f, g(f)$	Zgfaaa composition of $f$ and $g$ , 58
$\Gamma$	Zggaaa gamma function, 262
$\mathcal{G}_p$	Zgvaaa geometric distribution with parameter $p$ , 252
$1_A$	Ziaaaa indicator (function) of the set $A$ , 20
$1_{X \in A'}$	Ziahaa indicator of the event $\{X \in A'\}$ , i. e., $1_{X \in A'} = 1_{X^{-1}(A')}$ , 158
$1_{X=x}, 1_{\{X=x\}}$	Zialaa indicator variable of the event $\{X=x\}$ , 279
$\mathcal{I}_1$	Zicaaa set system of all half-open intervals in $\mathbb{R}$ , 12
$id$	Zidaaa identity mapping, 48
$\mathcal{I}_n$	Zigaaa set system of all half-open cuboids in $\mathbb{R}^n$ , 13
$\lambda, \lambda_n$	Zlraaa Lebesgue measure on $(\mathbb{R}^n, \mathcal{B}_n)$ , where $\lambda := \lambda_1$ , 22
$\mu, \nu$	Zmaaaa general symbols for measures, 17
$\mu_{\#}$	Zmaaaa counting measure, 21
$\mu_f$	Zmfaaa image measure of $\mu$ under $f$ , 68
$\mu_1 \otimes \dots \otimes \mu_n$	Zmgaaa product measure of $\mu_1, \dots, \mu_n$ , 24
$\nu \underset{\mathcal{A}}{\ll} \mu$	Zncabe the measure $\nu$ is absolutely continuous with respect to the measure $\mu$ , 109
$\nu \underset{\mathcal{A}}{\approx} \mu$	Zngabe the measures $\mu$ and $\nu$ are null-set equivalent, i. e., they are absolutely continuous with respect to each other, 109
$\mathbb{N}$	Znhabe set of all positive integers without zero, i. e., $\mathbb{N} = \{1, 2, \dots\}$ , 5
$\mathbb{N}_0$	Znhcbe set of all nonnegative integers including zero, i. e., $\mathbb{N}_0 = \{0, 1, 2, \dots\}$ , 4
$\mathcal{N}_{\mu, \sigma^2}$	Znvaaa univariate normal distribution with parameters $\mu$ and $\sigma^2$ , 256
$\mathcal{N}_{\mu, \Sigma}$	Znvcaa multivariate normal distribution with parameters $\mu$ and $\Sigma$ , 259

$(\Omega, \mathcal{A}, P)$	Zohaaa probability space, 127
$\mathcal{P}(\Omega)$	Zpbhaa power set of $\Omega$ , 6
$P$	Zpcaaa probability measure, 127
$P(A)$	Zpceaa probability of the event $A$ , 127
$P(X=x)$	Zpcgaa probability of the event $\{X=x\} = X^{-1}(\{x\})$ , 156
$P(X \in A')$	Zpcjaa probability of the event $\{X \in A'\} = X^{-1}(A')$ , 156
$P(A B)$	Zpcjaa conditional probability of $A$ given $B$ (with respect to the probability measure $P$ ), 132
$P(X_1 \in A', X_2 \in B')$	Zpcjca probability of the event $\{X_1 \in A'\} \cap \{X_2 \in B'\}$ , 169
$P(A X=x)$	Zpcnca conditional probability of the event $A$ given the event $\{X=x\}$ with $P(\{X=x\}) > 0$ , also denoted by $P(A \{X=x\})$ , 279
$P(Y=y \{X=x\})$	Zpcncfa conditional probability of the event $\{Y=y\}$ given the event $\{X=x\}$ with $P(\{X=x\}) > 0$ , also denoted by $P(Y=y X=x)$ , 280
$P(A X)$	Zpcnia conditional probability of the event $A$ (with respect to $P$ ) given the discrete random variable $X$ , 284
$P^{X=x}$	Zpcpaa the $(X=x)$ -conditional probability measure on $(\Omega, \mathcal{A})$ , 164
$P_X$	Zpcpca distribution of the random variable $X$ (with respect to $P$ ), 156
$P_{X_1, \dots, X_n}$	Zpcpha joint distribution of the random variables $X_1, \dots, X_n$ , the distribution of the multivariate random variable $X = (X_1, \dots, X_n)$ , 168
$(P_X)_g$	Zpcqaa image measure of $P_X$ under $g$ , 157
$\Phi$	Zpdcaa distribution function of the standard normal distribution, 256
$P^B$	Zpcnaa $B$ -conditional-probability measure, 140
$P_X^B$	Zpfgaa distribution of $X$ with respect to the conditional-probability measure $P^B$ , 164
$p_X$	Zpiaaa probability function of a discrete random variable $X$ , 175
$p_{X_1, X_2}$	Zpifaa probability function of the bivariate random variable $X = (X_1, X_2)$ , 177
$\pi_j$	Zpkaaa $j$ th projection mapping, 56
$\mathcal{P}_\lambda$	Zpvaaa Poisson distribution with parameter $\lambda$ , 250
$Q_X$	Zqfaaa quantile function of a real-valued random variable $X$ , 183
$\mathbb{Q}$	Zqhaaa set of all rational numbers, 13
$Q_{lin}(Y X)$	Zqlaaa the composition of $X$ and the linear quasi-regression (or linear least-squares regression of $Y$ on $X$ ), 218
$Q_{lin}(Y X_1, \dots, X_n)$	Zqmaaa linear quasi-regression of $Y$ on $X_1, \dots, X_n$ , 232
$\mathbb{R}$	Zrjcbe set of all real numbers, 3
$\mathbb{R}^2$	Zrjfaa Cartesian product $\mathbb{R} \times \mathbb{R}$ , 3
$\mathbb{R}^n$	Zrjfaaa $n$ -fold Cartesian product $\mathbb{R} \times \dots \times \mathbb{R}$ , 3
$\bar{\mathbb{R}}$	Zrjjaa extended set of all real numbers, i. e., $\bar{\mathbb{R}} = \mathbb{R} \cup \{\infty, -\infty\}$ , 13
$SD(Y)$	Zscaaa standard deviation of the random variable $Y$ , 208

$SE(\bar{Y})$	Zscdaa standard error of the sample mean of the random variables $Y_1, \dots, Y_n$ , 210
$\text{sgn}(f)$	Zscdha sign function of $f$ , 64
$\sigma(\mathcal{E})$	Zsdhaa $\sigma$ -algebra generated by the set system $\mathcal{E}$ , 9
$\sigma(f)$	Zsdhda $\sigma$ -algebra generated by the mapping $f$ , 51
$\sigma(f_1, \dots, f_n)$	Zsdhfa $\sigma$ -algebra generated by the mappings $f_1, \dots, f_n$ , 54
$\sigma(X)$	Zsdhja $\sigma$ -algebra generated by the random variable $X$ , 169
$\sigma(f, \mathcal{A})$	Zsdmca the $\sigma$ -algebra generated by the union of $\sigma(f)$ and $\mathcal{A}$ , 52
$\Sigma_{xy}$	Zsiaaa covariance matrix of $\mathbf{x}$ and $\mathbf{y}$ , 230
$\Sigma_{xx}$	Zsmaaa variance-covariance matrix of $\mathbf{x}$ , 230
$t_n$	Ztvaaa $t$ -distribution with $n$ degrees of freedom, 264
$\mathcal{U}_B$	Zuvaaa continuous uniform distribution on the set $B$ , 254
$\text{Var}(Y)$	Zvaaaa variance of the random variable $Y$ , 208
$\mathbf{x}$	Zxcaaa column vector of numerical random variables, 227
$\mathbf{X}$	Zxdaaa matrix of numerical random variables, 228
$X^{-1}(\mathcal{A}'_X)$	Zxdhjb $\sigma$ -algebra generated by the random variable $X$ , 169
$X \sim \mathcal{L}$	Zxgaaa the random variable $X$ has the distribution $\mathcal{L}$ . Examples for $\mathcal{L}$ are $\mathcal{B}_{n,p}$ , $\mathcal{P}_\lambda$ , $\mathcal{N}_{0,1}$ , or $F_{m,n}$ , 247
$X \stackrel{P}{=} Y$	Zxhaaa $X$ and $Y$ are $P$ -equivalent, 161
$X(\omega) \stackrel{P\text{-a.a.}}{=} Y(\omega)$	Zxhjaa $X$ and $Y$ are identical for $P$ -almost all $\omega \in \Omega$ , 161
$\bar{X}, \bar{Y}$	Zyjaaa arithmetic mean (sample mean) of the random variables $X_1, \dots, X_n$ or $Y_1, \dots, Y_n$ , respectively, 206
$Z_Y$	Zzaaaa $Z$ -transformation of the random variable $Y$ , 211