Probability and Inferential Statistics

Lecture SS 16

Test of Significance

Prof. Dr. Rolf Steyer

Examples of Hypotheses in a Significance Test

A *significance test* is a statistical procedure for testing *a null hypothesis* (H_0) about the size of a parameter.

Examples of such a null hypotheses:

$$\begin{aligned} P(A) - 0.5 &= 0 \\ E(Y) - 100 &= 0 \\ E(Y_1) - E(Y_2) &= 0 \\ Var(Y) &= 0 \\ Cov(X,Y) &= 0 \\ Corr(X,Y) - 1 &= 0. \end{aligned}$$

The role of test statistics in a significance test

Testing a *null hypothesis* (H_0) requires an appropriate *test statistic*. A test statistic is a random variable that takes on an extreme value only with a small probability if H_0 holds, and with a larger probability if H_0 does not hold. If H_0 does not hold we say that the alternative hypothesis (H_1) holds. Usually, we choose a test statistic whose distribution under the null hypothesis is known.

Questions:

1. Which test statistic is appropriate for testing the null hypothesis $H_0 : P(A) = .50$ in our parapsychological experiment, where $P(A)$ denotes the probability that our medium predicts the result of a coin flip?

2. Which test statistic is appropriate for testing the null hypothesis $H_0 : E(D) = 0$, where $D := Y_1 - Y_2$ is the difference between a pre-test and a post-test of the *Coloured Progressive Matrices Test* (CPM) in the treatment condition?

Basic idea of a significance test

Using the distribution of an appropriate test statistic under the null hypothesis, we compute the probability that a test statistic occurs that is as big as or bigger than the value of the test statistic observed in our data sample. This probability is called the *p-value*.

If this p-value is smaller than the *significance level* or α-*level*, a fixed probability — usually $\alpha = .05$, $\alpha = .01$ or $\alpha = .001$ — then the assumption that the null hypothesis holds is not very plausible and we decide to reject the null hypothesis. As mentioned before, the alternativ hypothesis H_1 is that H_0 *does not hold*.

If the test statistic in our data sample is not extreme, then we have no reason to reject the null hypothesis, because the occurrence of such a non-extreme value of the test statistic is to be expected under the assumption that H_0 holds.

Example of a One-Sided Hypothesis

In the *parapsychological experiment* we want to test, if a person is a medium that can predict the outcome of a coin flip with a probability of hits that is greater than .5. This can be tested using a significance test.

One-sided null hypothesis. The first step is to formulate the *null hypothesis* H_0 specifying more precisely, what we mean that a person is *not a medium*. An example for such a precise specification is:

$$H_0 : P(A) \leq 0.5.$$

This is a *one-sided hypothesis*, where $P(A)$ denotes the probability of the event that our putative medium has a hit in a single trial. Only a probability $P(A) > .5$ means that the putative medium is actually a medium.

Alternative hypothesis. Hence, in this case the alternative hypothesis is:

$$H_1 : P(A) > 0.5.$$

Example: Parapsychological Experiment 2

Two-sided null hypothesis.

$$H_0 : P(A) = 0.5.$$

This is a *two-sided null hypothesis*. A probability $P(A) \neq .5$ means that the putative medium is a medium.

Alternative hypothesis. Hence, in this case the alternative hypothesis is:

$$H_1 : P(A) \neq 0.5.$$

Example: parapsychological experiment — continued 3

Test statistic. If we repeat the guessing experiment n times, then an appropriate test statistic is

$$X := \textit{number of hits}.$$

An unexpectedly high number of hits would question the null hypothesis H_0.

Significance level. If we choose the significance level $\alpha = .05$ and consider $n = 10$ trials, then 9 hits or 10 hits are values of the test statistic X that are *statistically significant at the* .05*-level* and would lead to reject the one-sided null hypothesis. *Statistically significant* means $P(X \in \{9, 10\}) \leq \alpha$. All numbers of hits other than 9 or 10 are to be expected under the one-sided null hypothesis given the chosen significance level. In this example, the *rejection area* or *rejection region* is the set $\{9, 10\}$.

Parapsychological Experiment — continued 4

The probability of 10 hits is

$$P(X = 10) = \binom{10}{10} .50^{10} (1 - .50)^{10-10} = 0.50^{10} = 0.0009765625.$$

The probability of 9 hits is

$$P(X = 9) = \binom{10}{9} .50^{9} (1 - .50)^{10-9} = 10 \cdot 0.50^{9} \cdot 0.50^{1} = 0.009765625$$

and the probability of 8 hits is

$$P(X = 8) = \binom{10}{8} .50^{8} (1 - .50)^{10-8} = 45 \cdot 0.50^{8} \cdot 0.50^{2} = 0.04394531.$$

These values of X are disjoint. Therefore,

$$P(X \in \{9, 10\}) = P_X(\{9, 10\}) = 0.50^{10} + 10 \cdot 0.50^{10} = 0.01074219,$$

$$\begin{aligned} P(X \in \{8, 9, 10\}) = P_X(\{8, 9, 10\}) &= 0.50^{10} + 10 \cdot 0.50^{10} + 45 \cdot 0.50^{8} \cdot 0.50^{2} \\ &= 0.0546875. \end{aligned}$$

Parapsychological Experiment — continued 5

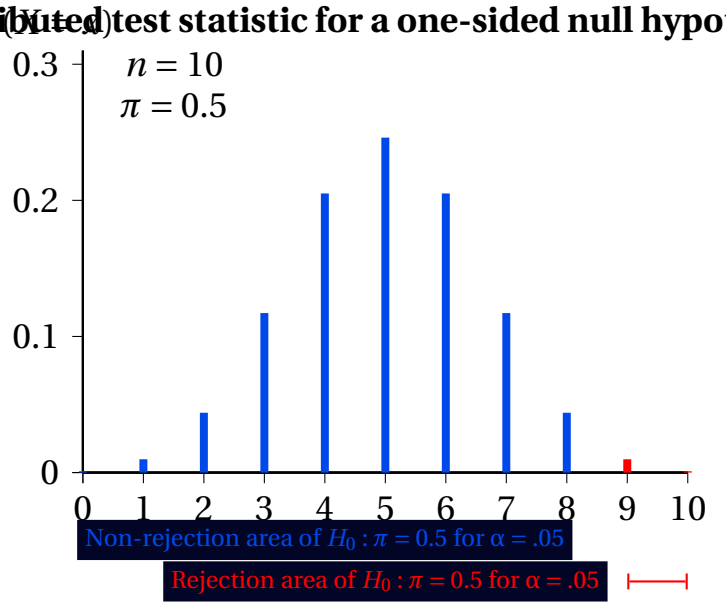
Computing these probabilities we presume $P(A) := \pi = .5$. Under this assumption, extreme values such of the test statistic such as $X = 9$ or $X = 10$ have a larger probability, if we compare $\pi = .5$ to other parameters such as $\pi = .40$, which also belong to the one-sided hypothesis. In other words, for all probabilities $\pi < .50$ — which belong to the one-sided null hypothesis — the probabilities for $X = 9$ or $X = 10$ are smaller than for $\pi = .50$. Therefore, these other probabilities belonging to the one-sided null hypothesis can be neglected.

Parapsychological Experiment — continued 6

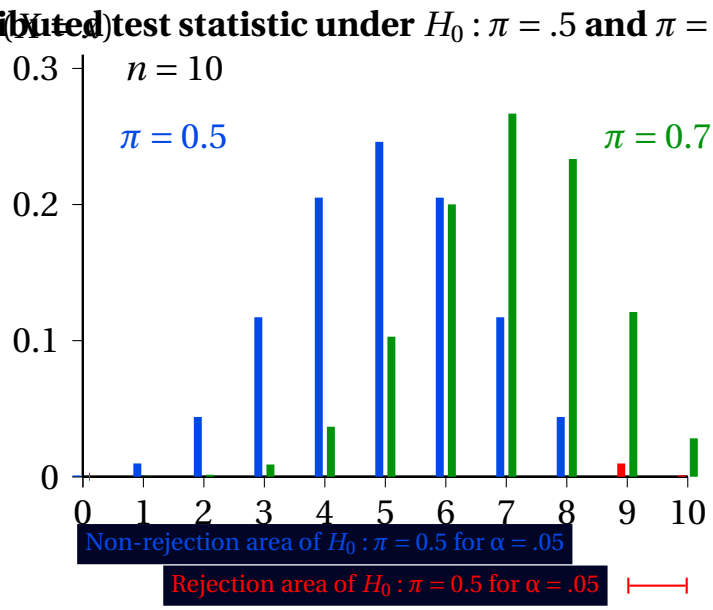
Hence, for $n = 10$ trials all value of *number of hits* X — this is our test statistic — that are greater than or equal to 9 are *statistically significant* at the .05-level and lead to a rejection of our one-sided null hypothesis. Eight hits would not be statistically significant any more at the .05-level, because $P_X(\{8,9,10\}) = 0.0546875$ is greater than the chosen significance level $\alpha = .05$. For 8 or less than 8 hits, there is no reason to reject the one-sided null hypothesis.

The set $\{9,10\}$ is called the *rejection area* or *rejection region* of the test of the one-sided null hypothesis at .05-level and the set $\{0,1,\ldots,8\}$ the *non-rejection area* of this null hypothesis (see the figure). Note that non-rejection does not mean acceptance. Instead it means to abstain from a decision.

Binomially distributed test statistic for a one-sided null hypothesis



Binomially distributed test statistic under $H_0 : \pi = .5$ and $\pi = .7$



Two-Sided Test

Null hypothesis. The hypothesis that the putative medium is *not a medium* can also be specified more precisely by

$$H_0 : P(A) := \pi = 0.5.$$

This is a *two-sided null hypothesis*, where again $P(A)$ denotes the probability of the event that the putative medium has a hit. Given this *two-sided null hypothesis*, a probability $P(A) \neq .50$ means that the putative medium is a medium.

Which of the two null hypotheses — the one-sided or the two-sided one — should be chosen is a matter of psychological theory.

Example: Parapsychological experiment — continued 7

Test statistic. Also for the test of the two-sided hypothesis

$$X := \textit{number of hits}$$

is an appropriate test statistic. An unexpectedly high or low number of hits given the null hypothesis $H_0 : \pi = .5$ would question the H_0.

Significance level. Choosing again $\alpha = .05$, then each number of hits smaller than or equal to 1 or greater than or equal to 9 is significant *at the* .05*-level* and leads to a rejection of H_0. All other numbers of hits would occur with a relatively high probability and not give reason to reject the H_0.

Example: Parapsychological experiment — continued 8

The probabilities for 0 hits and for 10 hits are

$$P(X = 0) \;=\; P(X = 10) \;=\; 0.50^{10} \;=\; 0.0009765625.$$

The probabilities for 1 hit and for 9 hits are

$$P(X = 1) \;=\; P(X = 9) \;=\; 10 \cdot 0.50^{9} \cdot 0.50^{1} \;=\; 0.009765625$$

and the probabilities for 2 hits and for 8 hits are

$$\begin{aligned} P(X = 2) \;=\; P(X = 8) \;&=\; \binom{10}{8} .50^{8}\,(1 - .50)^{10-8} \;=\; \binom{10}{2} .50^{2}\,(1 - .50)^{10-2} \\ &=\; 45 \cdot 0.50^{2} \cdot 0.50^{8} \;=\; 0.04394531. \end{aligned}$$

These values of X are disjoint. Therefore,

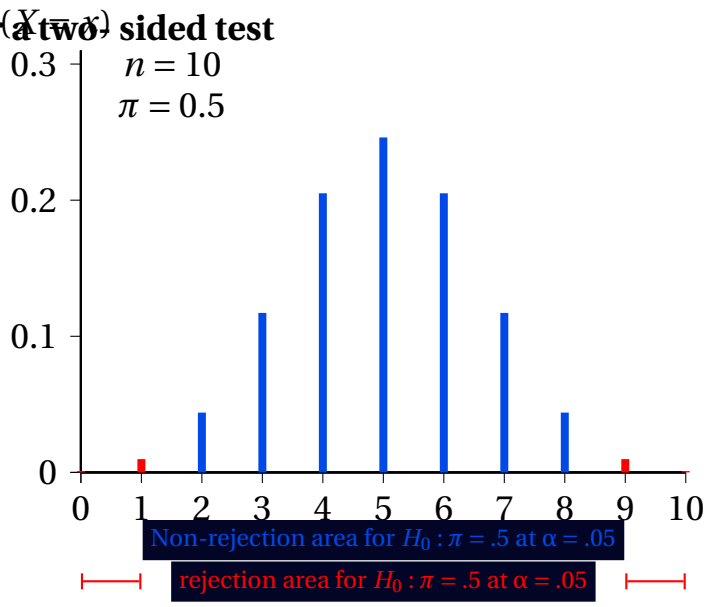
$$P_X(\{0, 1, 9, 10\}) \;=\; 2 \cdot 0.50^{10} \;+\; 2 \cdot 10 \cdot 0.50^{10} \;=\; 0.02148438.$$

Example: parapsychological experiment — continued 9

Hence, according to these probabilities (computed for $n = 10$ trials with $\pi = .5$), all values of the test statistic X (number of hits) less than or equal to 1 or greater that or equal to 9 are significant at the .05-level and lead to a rejection of the two-sided null hypothesis. Eight hits would not be significant any more at the .05-level, because $P_X(\{0,1,2,8,9,10\}) = 0.109375$ ist greater than the chosen significance level $\alpha = .05$. For at least 2 or at most 8 hits, there is no reason to reject the null hypothesis.

Therefore, the set $\{0,1,9,10\}$ is called the *rejection area* of the two-sided test at .05-level and the set $\{2,3\ldots,7,8\}$ is called the *non-rejection area* (see the figure).

Binomial test for a two-sided test



Two types of error in significance testing

Two kinds of error:

- rejection of H_0 if it is in fact true: α-error (error of type I)
- non-rejection of H_0 if it is in fact wrong: β-error (error of type II)

	true is	
decision about H_0	H_0	H_1
non-rejection	✓	β-error
rejection	α-error	✓

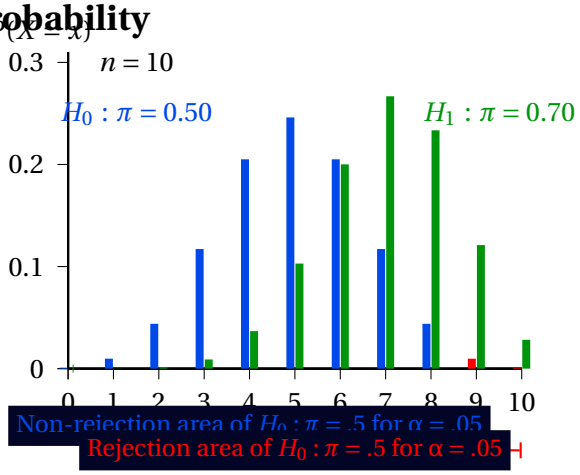
Error probabilities in significance testing

Under the null hypothesis, there are probabilities for the two cases

- α: the probability to reject the H_0 although it is true
- $1-\alpha$: the probability not to reject the H_0 if it is true
- β: If H_1 is a point hypothesis (such as $H_1 : \pi = .7$ or $H_1 = \mu_1 - \mu_2 = 10$), then this is the probability that H_0 is not rejected
- $1-\beta$: If H_1 is a point hypothesis, then this is the probability to reject the H_0. The probability $1-\beta$ is also called the *power* of the test.

Note that β and $1-\beta$ can only be computed for point hypotheses. For a range null hypothesis, we cannot compute β.

Example of β-error probability



Which are the probabilities α, $1-\alpha$, β and $1-\beta$ in this example?

Example of β-error probability etc.

Which are the four probabilities in this example?

- α is the probability to reject H_0 although it is true. We choose $\alpha = .05$, which implies that we reject H_0 if $X = 9$ or $X = 10$.
- $1-\alpha$ is the probability not to reject H_0 although it is true. If $\alpha = .05$, then $1-\alpha = .95$.
- Under $H_1 : \pi = .70$, β is the probability not to reject the H_0. In our example with $\alpha = .05$, we do not reject the H_0, if $X \leq 8$. In this example, computing β yields: $\beta = 0.8506917$ (see below).
- Under H_1, $1-\beta$ is the probability to reject the H_0. In our example: $1-\beta = 1-0.8506917 = 0.1493083$.

The probability $1-\beta$ to reject the one-sided H_0, if $H_1 : \pi = .70$ holds is only .149! The probability β is computed as follows:

$$\beta = \sum_{x=0}^{8} b_{10,.70}(x) = 0.8506917.$$

Using the program *R* we obtain: $pbinom(8, 10, .70) = 0.8506917$.
If $H_1 : \pi = .90$ holds and we consider $n = 10$ trials, then we obtain $\beta = pbinom(8, 10, .90) = 0.2639011$ and the power $1-\beta = 1-0.2639011 = 0.7360989$.

Exercise: Determine the rejection area of the one-sided test for $n = 20$ trials and $\alpha = .01$. Compute the power of this test using the program R. Repeat these computation for the two-sided hypothesis.

Measures to reduce the β-error probability

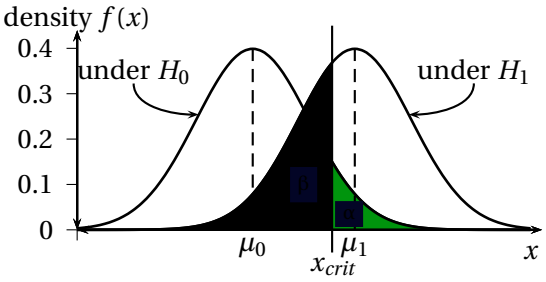
The probability to reject the one-sided H_0, if $H_1 : \pi = .70$ is true, is only .149! This means: Only in 15 out of 100 cases we would reject the H_0, although it is wrong. Hence, the power is pretty low.

What can be done to increase power?

- To increase the sample size n.
- To increase the α-error probability.

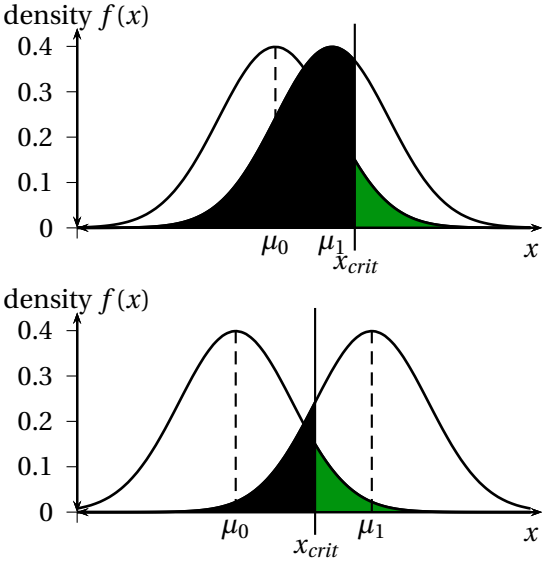
Furthermore, if the difference between $\pi = .50$ under the null hypothesis and π under the alternative hypothesis would be greater, then the power of the test would be greater. Usually, the probability π under alternative hypothesis is unknown and cannot be changed. (In our example we would have to choose a medium that has a hit with $\pi > .70$).

α-error and β-error for a normally distributed test statistic

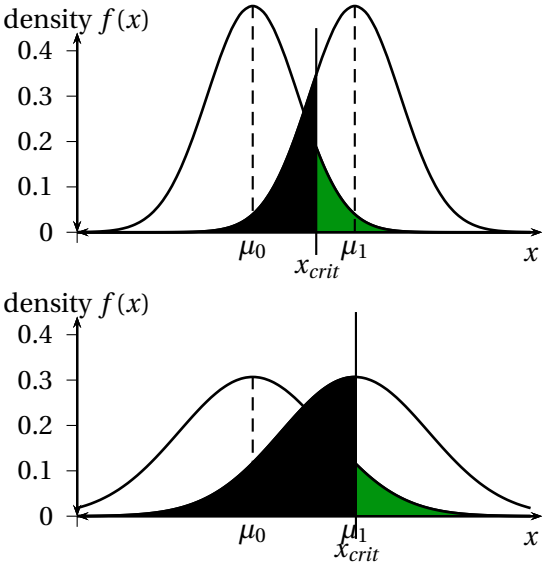


It is presumed that H_1 is a point hypothesis

α- and β-error for a normally distributed test statistic: role of the expectations



α-error and β-error for a normally distribtuted test statistic: role of the standard deviations



Note

- Neither H_0 nor H_1 are events. Therefore, they do not have probabilities. They are either true or wrong. In particular α is *not* the probability that H_0 is true and $1-\alpha$ is *not* the probability that H_1 holds.
- The β-error probability can be computed only if we presume that H_1 is a point hypothesis about the parameter considered, for example, $H_1 : \pi = .70$ or $H_1 : \mu_1 = 2$.
- There is a probability to commit an $\alpha-$error, but no probability to commit an error in significance testing (i.e., an α-error *or* a β-error). Such a probability could only exist, if H_0 and H_1 would be events. However, this does not mean that we can exclude that we commit an error in significance testing or it is unlikely to commit an error.