7 Hypothesis Testing

Hypothesis tests belong to the most commonly used methods in empirical research. Put bluntly, such a test provides answers to the questions: Does the data confirm my assumption or could what I observe also be a consequence of random fluctuations? Would I find the same thing with new data? If, for example, in an opinion poll, 55% are in favour of a specific proposal, is it justified to assume that the proposal would also find a majority of the people in the actual vote?

If the data originates from a random sample, then the issue is whether one can take a chance to draw a conclusion about the data in the whole population from where the sample was taken. The branch of statistics which deals with such inferences is called *inferential statistics* or *inductive statistics*.

Questions which can be examined using hypothesis tests include:

- Which is more effective in breast cancer treatment: chemotherapy or hormone therapy?
- Can elephants count?
- Does smoking marihuana cigarettes have a direct impact on short-term memory?
- Are there more fish deformities after the nuclear reactor incident in Fukushima, Japan?
- Is social competence more developed in girls compared to boys?

In statistics there are many different test procedures with which diverse hypotheses can be investigated. In this chapter, the basic ideas and terminology of hypothesis testing as relates to the binomial distribution will be presented. Understanding the basic ideas in this special situation will be a good preparation to learn the test methodology for other, more complex cases.

Opening problem: The lottery ticket seller

A lottery ticket seller has several thousand tickets. He maintains that 20% of his tickets are winning tickets.

a) 100 tickets are bought of which only 15 are winning ones.
   - Is making a complaint to the ticket seller appropriate?
   - How high is the risk that the complaint is not justified?

   Use suitable calculations to find answers to these questions.

b) Do the calculations change if 200 tickets are bought of which 30 are winning ones? Would a complaint be adequate now?

c) Assume that i) 100 tickets, ii) 200 tickets, iii) 500 tickets are bought.
   How low must the respective number of winning tickets be in order to be able to assume that the actual proportion of winning tickets lies below 20%? Justify the answers also by calculation.

7.1 Checking a Conjecture by Way of a Prediction Interval

Example: The game 'Ludo' (engl. version of 'Eile mit Weile'), part I

Sarah likes to play the board game *Ludo* with her brothers and sisters. The last time she played, however, she became very impatient. She had to throw the die 16 times until she finally got the six which allowed her to place one of her tokens on the starting square. 'Next time I am going to check before the game starts whether the six really appears with a probability of $\frac{1}{6}$', she declared. She decides to test roll the die 100 times and asks herself the question how many sixes there have to be for her to reject the die as unfair.

Recently in her mathematics class, she learnt how to calculate the prediction interval for random variables with a binomial distribution. This allows one to predict which deviations around the expected value can 'normally' be expected to occur: the number of 'sixes' obtained in multiple throws with a fair die lie in this interval with a probability of 90%.

She decides to conduct 100 throws and calculates the 90% prediction interval for $n = 100$ and $p = \frac{1}{6}$.

$$k_L \approx 100 \cdot \frac{1}{6} - 1.64 \cdot \sqrt{100 \cdot \frac{1}{6} \cdot \frac{5}{6}} \approx 16.67 - 1.64 \cdot 3.72 \approx 10.56 \approx 11$$

$$k_R \approx 100 \cdot \frac{1}{6} + 1.64 \cdot \sqrt{100 \cdot \frac{1}{6} \cdot \frac{5}{6}} \approx 16.67 + 1.64 \cdot 3.72 \approx 22.78 \approx 23$$

Since $\sigma \approx 3.72 > 3$, the approximate calculation is sufficiently precise. Rounding the approximated prediction limits to integers, Sarah arrives at the 90% prediction interval $\{11, 12, \ldots, 23\}$. So, in 100 throws with a fair die, the six usually appears between 11 and 23 times. Only in a mere 10% of the cases does the number of thrown sixes lie outside this interval.

Consequently, Sarah decides not to use any die which shows the six less than 11 or more than 23 times, since this is a strong indicator that the die is not fair. ⌟

This example clearly shows how the prediction interval for the binomial distribution can be used for a hypothesis test. First, a hypothesis states an assumption for the probability p. When throwing a regular die, the probability for a 'six' can be assumed to be $p = \frac{1}{6}$. Next, the number of experimental runs for the test should be determined. In the example it is $n = 100$ throws. Then the prediction interval for these values of n and p is calculated. In order to be able to use the approximation for the prediction interval, n should be large enough so that $\sigma = \sqrt{npq} > 3$.

The decision whether the hypothesis is rejected or retained is then based on the number of successes in the n experimental runs.

- If the observed number of successes lies in the calculated prediction interval, then nothing contradicts the assumed value p for the probability of success. Hence, the assumed hypothesis about the value of the probability of success is retained.
- However, if the observed number of successes lies outside the prediction interval, then the hypothesis that the actual probability of success is equal to p is rejected.

Binomial Test

A *binomial test* is a decision rule whether a conjecture about the probability of success is to be retained or rejected. Here the following definitions are used:

- The conjecture about the probability of success is called the *null hypothesis* H_0.
- The number n of repetitions in the test is called the *sample size*.
- For the decision whether the conjecture about the probability of success should be retained or rejected, a prediction interval is calculated. It consists of the outcomes which do not contradict the null hypothesis. If the number of successes lies outside of this prediction interval, then the null hypothesis is rejected; otherwise the null hypothesis is retained.
- The set of all possible values outside the prediction interval comprises the outcomes which are poorly compatible with the null hypothesis. They form the *rejection region* V_{H_0} of the binomial test.

Example: The game 'Ludo', part II

For the test where Sarah examined the die the following is true:

- The *null hypothesis* H_0: 'the probability for a six is $p = \frac{1}{6}$',
- The *sample size* $n = 100$,
- The *90% prediction interval* associated with $p = \frac{1}{6}$ equals $\{11, 12, \ldots, 22, 23\}$,
- The *rejection region* of H_0 is the complement of the prediction interval, that is $V_{H_0} = \{0, \ldots, 10\} \cup \{24, \ldots, 100\}$.

7.2 The Significance Level

Example: The game 'Ludo', part III

Sarah now throws the die from the *Ludo* game 100 times. A six appears nine times. Consequently, the null hypothesis is rejected that the probability for a six with this die is $\frac{1}{6}$.

Is this proof enough that the die is not fair? The answer to this question is clear: even with a probability of success $p = \frac{1}{6}$, it is possible that the number of successes does not lie in the calculated prediction interval; it is simply rather unlikely. The sum of the probabilities of all outcomes from the 90% prediction interval amounts to 90% or a bit more. The probability that a different outcome in other words an outcome from the rejection region is obtained, is consequently slightly less than 10%. This means that when testing ten fair dice, Sarah should expect to obtain an outcome from the rejection region in one of these ten tests. If this happens, she wrongly decides that for this die the probability of a six is not $\frac{1}{6}$.

A hypothesis test cannot decide with absolute certainty whether the null hypothesis is correct or false. With a test on the basis of the 90% prediction interval, the probability of rejecting the null hypothesis even though it actually is true, is at most 10%. When testing dice, the erroneous rejection of the null hypothesis does not have any serious consequences – a fair die is simply binned as unfair. In other applications, the consequences which can arise from rejection of the null hypothesis are

more far-reaching: when testing a new type of medication, the null hypothesis is, for example, 'the new medication is not more effective than the old one'. If this null hypothesis is erroneously rejected, then in the future the new and mostly more expensive medication is prescribed.

For this reason, the tests are usually designed in such a way that the probability of erroneously rejecting the null hypothesis is at most 5% or 1%. This value is called the significance level α of the test. For this, instead of the 90% prediction interval, a 95% or 99% prediction interval is calculated. This will be discussed in detail in the next section.

Significance Level

A hypothesis test has the *significance level* α if the probability that the null hypothesis is rejected, although it is actually true, is at most α.

The significance level is the basic idea of inferential statistics. The smaller the significance level chosen, the more the data has to speak against the null hypothesis until it can be rejected.

In empirical sciences like medicine, psychology, sociology, or economics, the testing of hypotheses using samples is a very commonly used procedure. When studying these subjects, it is important to understand the meaning of the significance level of a test. The concept of significance is not only used in academic journals but also frequently in the science and technology section of daily newspapers, so much so that it can almost be considered as general knowledge.

7.3 The $(1-\alpha)$ Prediction Interval

The rejection region of a binomial test associated with the significance level α includes all possible outcomes which do not lie in the prediction interval associated with the so-called coverage probability $1-\alpha$. Below the calculation of such prediction intervals is explained.

$(1-\alpha)$ Prediction Interval

An interval $\{k_L, k_L+1, \ldots, k_R-1, k_R\}$ is called a $(1-\alpha)$ *prediction interval* for a random variable X with binomial(n,p)-distribution, if, with a probability greater than or equal to $1-\alpha$, X is at least k_L and at most k_R:

$$P(X \in \{k_L, k_L+1, \ldots, k_R\}) = P(k_L \leq X \leq k_R) = \sum_{k=k_L}^{k_R} B_{n,p}(k) \geq 1-\alpha$$

The probability that a random variable X with binomial(n,p)-distribution assumes a value outside the $(1-\alpha)$ prediction interval is at most equal to α:

$$P(X \notin \{k_L, k_L+1, \ldots, k_R\}) = P(X < k_L) + P(X > k_R) \leq \alpha$$

In a binomial test of the null hypothesis H_0: 'the actual probability of success is equal to p' the values outside of the $(1-\alpha)$ prediction interval form the rejection region. Consequently, the probability to erroneously reject a correct null hypothesis with a binomial test is at most equal to α. This means that the corresponding test has the significance level α.

Example: 95% Prediction interval for $n = 12$ and $p = \frac{1}{2}$

Given is

$$\sum_{k=4}^{8} B_{12,0.5}(k) = 0.854, \quad \sum_{k=3}^{9} B_{12,0.5}(k) = 0.961$$

This means that $\{3, 4, \ldots, 9\}$ is a 95% prediction interval for $n = 12$ and $p = \frac{1}{2}$. However, $\{4, 5, \ldots, , 8\}$ is not a 95% prediction interval for $n = 12$ and $p = \frac{1}{2}$.

$(1-\alpha)$ prediction intervals for random variables with binomial distribution are calculated in exactly the same way as in the case $1 - \alpha = 90\%$ (p. 114 ◀ The 90% Prediction Interval). For the approximate calculation of the interval limits, the standard deviation $\sigma = \sqrt{npq}$ is multiplied with a factor $u_{1-\alpha}$ that varies with α. The most important values can be read off the table by calculating the coverage probability $s = 1-\alpha$, which we also call the *degree of certainty* or simply the *certainty*.

Certainty s	80%	90%	95%	98%	99%
Factor u_s	$u_{80\%} = 1.28$	$u_{90\%} = 1.64$	$u_{95\%} = 1.96$	$u_{98\%} = 2.33$	$u_{99\%} = 2.58$

For the interval limits it follows that

$$k_L \approx np - u_s\sqrt{npq}, \quad k_R \approx np + u_s\sqrt{npq}$$

As a rule of thumb, the approximation is suitable for $\sigma > 3$ as before. For the exact calculation, a calculator or a table with the binomial probabilities is needed. The procedure is analogous to the case $1 - \alpha = 90\%$ in chapter 6.5 (p. 117 ◀ Calculating the Exact 90% Prediction Interval).

Example: Is a coin fair? part I

On the basis of 100 test tosses of a coin, the assumption that the coin is fair will be examined - this means that the probability for the coin to show 'heads' is 50%. The aim is to identify the rejection region for the test with significance level $\alpha = 5\%$.

The null hypothesis H_0 is as follows: the probability of 'heads' is $p = \frac{1}{2}$.

The complementary probability to $\alpha = 5\%$ is $1-\alpha = 95\%$. Therefore the 95% prediction interval for the binomial distribution associated with sample size $n = 100$ and probability $p = \frac{1}{2}$ is calculated. The required u-value is $u_{95\%} = 1.96$.

$$100 \cdot \frac{1}{2} - 1.96 \cdot \sqrt{100 \cdot \frac{1}{2} \cdot \frac{1}{2}} = 50 - 1.96 \cdot 5 = 40.2 \approx 40 = k_L$$

$$100 \cdot \frac{1}{2} + 1.96 \cdot \sqrt{100 \cdot \frac{1}{2} \cdot \frac{1}{2}} = 50 + 1.96 \cdot 5 = 59.8 \approx 60 = k_R$$

The required 95% prediction interval is $\{40, 41, \ldots, 60\}$.

The rejection region therefore includes all values from 0 to 39 and from 61 to 100.

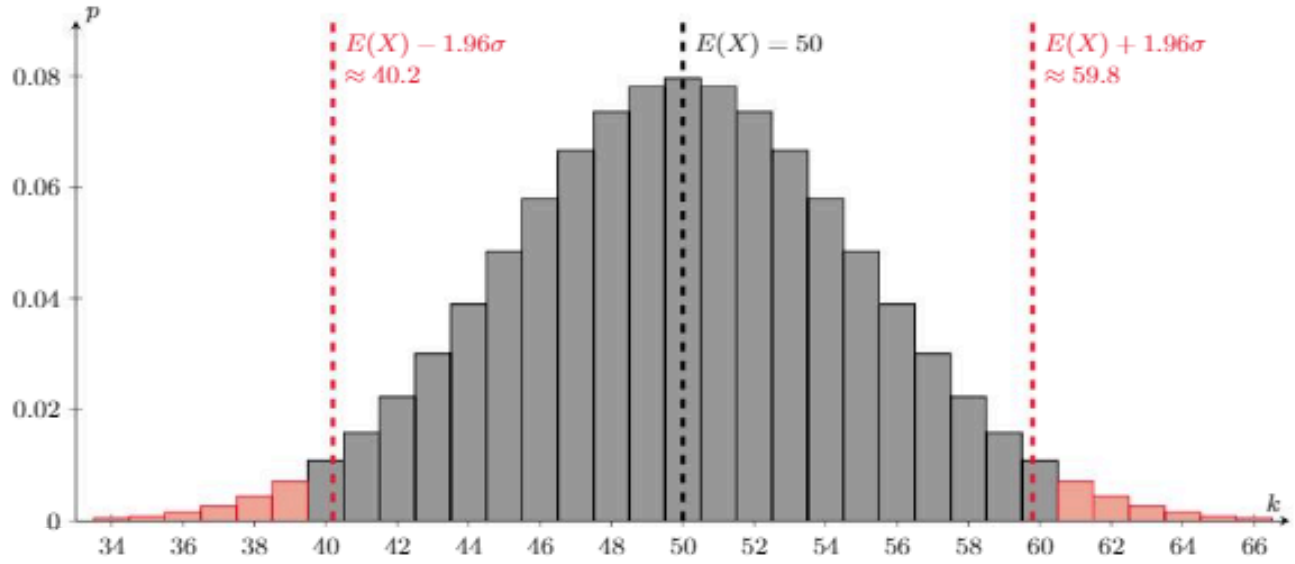
In the following, the rounding of the interval limits is justified. To this aim, the binomial distribution is represented as a histogram rather than as a bar chart.

Histogram for Random Variables

A *histogram* for random variables with integer values displays probabilities as *areas of rectangles* ('columns') instead of heights of bars used by a bar chart. The bases of the rectangles go from $k-\frac{1}{2}$ to $k+\frac{1}{2}$ and therefore have the width 1. The heights of the rectangles are equal to $P(X=k)$.

Such histograms differ from bar charts only in as much as that for the histogram the bars are extended to rectangles with width 1 that touch each other.

The histogram suggests how to proceed when the approximated limits of the prediction intervals are not integers. The two rectangles in which the unrounded prediction limits $np \pm u_{1-\alpha}\sqrt{npq}$ lie should belong to the prediction interval. In the example (p. 128 ◄ Is a coin fair? part l), the approximated lower limit 40.2 lies in the rectangle over $[39.5, 40.5]$, which means $k_L = 40$. In general, this rule corresponds to standard rounding.



In the illustration, the area marked in grey corresponds to the probability that a result within the 95% prediction interval occurs. The area is therefore greater than 0.95 area units. The two areas marked in red correspond to the complementary probability. These areas together are smaller than 0.05 units. The approximation is designed such that the area of the histogram between the two unrounded limits $E(X) \pm 1.96\sigma$ is approximately 95% area units. If cut rectangles are included in the prediction interval, its probability is at least 0.95. This is achieved by determining the limits k_L and k_R of the prediction interval by conventional rounding. This consideration is explained in more detail in chapter 8 (► The Normal Distribution as an Approximation to the Binomial Distribution, p. 145).

7.4 Indirect Procedure and Alternative Hypothesis

Example: The game 'Ludo', part IV

As Sarah had the impression that the die used was not fair, she decided to test the die. With the test she wants to demonstrate that the die under investigation does not show a six in one sixth of the throws in the long run. Her actual assumption is therefore $p \neq \frac{1}{6}$, the opposite of the null hypothesis $p = \frac{1}{6}$. As in a court of law, she has to start out from the position of 'presumed innocence', in other words from the null hypothesis $p = \frac{1}{6}$. Next, she determines which outcomes are to be expected if the null hypothesis were true. To that end, the prediction interval is calculated. If the test throws lead to a result that does not lie in the prediction interval, then this result is hardly consistent with the null hypothesis. The result of the 100 test throws corresponds to the evidence in a court case. If the evidence is strong enough, then the null hypothesis can be rejected and consequently the opposite of the null hypothesis is indirectly confirmed, in other words the original assumption $p \neq \frac{1}{6}$.

When testing a new medication, the null hypothesis H_0 would, for example, state: 'the new medication is not more effective than the old one'. The manufacturer of the new medication, which is carrying out the study, hopes to be able to reject this null hypothesis with a hypothesis test. This seems strange and a roundabout approach: why is the *alternative hypothesis* H_1 that the new medication is more effective not proven directly?

One reason for this is that hasty decisions should be avoided. A competitor or the insurer who would have to pay for the new medication could express justified doubts about the effectiveness of the new medication and support the null hypothesis. Using a hypothesis test, such doubts should be refuted on the basis of objective facts.

Another reason for the indirect approach when testing is a matter of principle: hypotheses and theories can be refuted more easily with experiments and data than confirmed. If the predictions of a theory are not consistent with the experiment, then the theory cannot be correct. If the predictions of the theory are consistent with the experiment, then the theory is still not proven. There can be several theories which all make the same predictions for this experiment.

A hypothesis test uses an *indirect procedure*. The person who carries out the test is really striving to confirm a conjecture. The *opposite* of this conjecture is then formulated as the null hypothesis. If the null hypothesis is rejected by the hypothesis test, then this counts as a confirmation of the original conjecture.

In this book the diction 'the null hypothesis is retained' is used when the observed outcome lies in the prediction interval and not 'the null hypothesis is accepted'. This is a way of demonstrating that the null hypothesis cannot actually be proven. What is retained for the moment can still be rejected later when new facts come to light.

7.5 The Procedure of a Binomial Test

When carrying out a binomial test, one adheres to the following fixed procedure.

The Procedure of a Binomial Test

1. *Formulating the null hypothesis H_0 and determining the significance level α*

   Given is a value p_0 for the probability of success which, as a rule, describes the situation that no interesting phenomenon exists. Hence, the null hypothesis is $H_0 : p = p_0$ and the alternative hypothesis $H_1 : p \neq p_0$. It is common to take the significance level to be $\alpha = 5\%$. This is chosen when no other information exists.

2. *Determining the rejection region*

   If the null hypothesis is correct then the test variable X = 'number of successes' has a binomial(n, p_0)-distribution. Hence, the (1-α) prediction interval $\{k_L, k_L+1, \ldots, k_R-1, k_R\}$ for a random variable with binomial(n, p_0)-distribution is calculated. The values outside of this interval form the rejection region of the test:

   $$V_{H_0} = \{0, 1, \ldots, k_L - 1\} \cup \{k_R + 1, \ldots, n - 1, n\}$$

   In the case that $\sigma > 3$, the limits k_L and k_R are obtained from the following approximation:

   $$k_{R,L} = \mu \pm u_s \cdot \sigma = np_0 \pm u_s \cdot \sqrt{np_0(1 - p_0)}$$

   Here, $s = 1 - \alpha$ and u_s can be extracted from the table on page 128. Otherwise, the exact $(1 - \alpha)$ prediction interval should be calculated. The procedure for this is analogous to the case of $\alpha = 10\%$ in chapter 6.5 (p. 117 ◀ Calculating the Exact 90% Prediction Interval).

3. *Making a test decision*

   Once the rejection region has been determined, the observations are examined and the test decision is made.

   - If the observed number of successes lies in the rejection region V_{H_0}, then the null hypothesis is rejected; the result serves as evidence for the alternative hypothesis H_1.
   - If the observed number of successes does not lie in the rejection region V_{H_0}, then the null hypothesis is retained; the result does not contradict the null hypothesis.

Example: Irregular die

A corner of a die is chipped. On the basis of 200 throws, an assessment is to be made as to whether or not the six still has a probability of $\frac{1}{6}$: in other words, if, in the long run, the six still appears as before with the frequency of a sixth of all throws. Design a corresponding hypothesis test with the significance level $\alpha = 0.01$. What conclusion can be drawn if the die shows a six 22 times in the 200 throws?

1. *Formulating the null hypothesis H_0 and determining the significance level α*

   H_0: $p = \frac{1}{6}$, H_1: $p \neq \frac{1}{6}$, $\alpha = 0.01$.

2. *Determining the rejection region*

   Test variable X = 'number of throws where the die shows a six'.

   In the case that the null hypothesis applies, X has a binomial$(200, \frac{1}{6})$-distribution. The expected value μ is equal to $200 \cdot \frac{1}{6} \approx 33.33$ and the standard deviation $\sigma = \sqrt{200 \cdot \frac{1}{6} \cdot \frac{5}{6}} \approx 5.27 > 3$. The limits of the 99% prediction interval can be calculated with the approximation:

   $$np \pm 2.58 \cdot \sigma \approx 33.33 \pm 2.58 \cdot 5.27 = 19.74 \text{ or } 46.93, \text{respectively.}$$

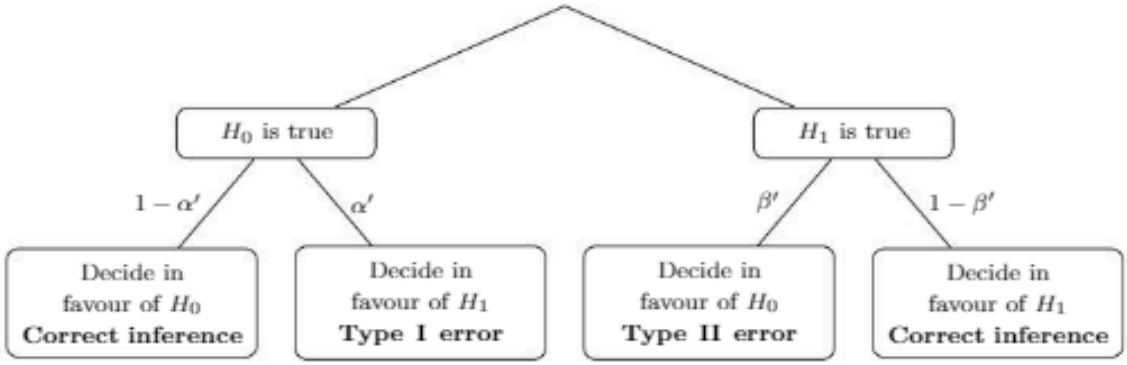
   Then $k_L = 20$ and $k_R = 47$, and the rejection region is $V = \{0, 1, \ldots, 19\} \cup \{48, 49, \ldots, 200\}$.

3. *Making a test decision*

   Because $22 \notin V$, the null hypothesis is retained. There is no clear indication that the chipped corner has changed the probability of a six appearing. ⌟

7.6 Type I and Type II Errors

A test is only carried out when it is not clear whether in reality the null hypothesis H_0 or its opposite, the alternative hypothesis H_1, is true. Even after the test has been carried out, there is no absolute certainty that the test result corresponds to reality. A wrong decision is possible both when the null hypothesis has been rejected and when it has been retained.



Type I and II Errors

In a statistical test, two error types are possible:

– *Type I error (α-error):* the null hypothesis is rejected even though it is true in reality.

– *Type II error (β-error):* the null hypothesis is retained even though it is false in reality.

The probability for a type I error is less than or equal to the significance level α and is therefore denoted by α'. The probability for a type II error is denoted by β'.

These two types of error do not only occur in statistics but also in general when deciding between two alternatives. The following example shows this quite well:

Example: Mushroom picker

A mushroom picker finds a mushroom. Now he has to decide whether this mushroom is edible or poisonous. The null hypothesis here is: 'the mushroom is poisonous'. The mushroom picker can then assess the mushroom correctly or make one of the following errors:

– *Type I error (or α-error):* the mushroom is indeed poisonous, but he considers it edible.

– *Type II error (or β-error):* the mushroom is indeed edible, but he considers it poisonous.

From this example, it is clear that, because of the consequences, both errors must be weighted differently. In this case, the consequences of a type I error are serious: a poisonous mushroom ends up among the mushrooms collected for consumption and will be eaten unless it is controlled a second time. The consequences of a type II error are much less serious: there is simply one mushroom less in the basket. The type I error must, in this case, be kept as small as possible, the size of the type II error plays a minor part. ⌟

The calculation for the probability of a type II error is now shown in an example:

Example: Is a coin fair? – part II

Consider again the test of the null hypothesis 'The coin is fair', in other words H_0: $p = \frac{1}{2}$, at the significance level α =5%. With $n = 100$ throws, the 95% prediction interval is $\{40, 41, ..., 60\}$ and the rejection region $V = \{0, 1, ..., 39\} \cup \{61, 62, ..., 100\}$.

If the null hypothesis is false, then $p \neq \frac{1}{2}$ and a type II error is made when the null hypothesis is nevertheless retained. This happens when an outcome from the 95% prediction interval occurs:

$$\beta' = P(40 \leq X \leq 60) = \sum_{k=40}^{60} B_{100,p}(k)$$

The probability β' depends on how large the true probability of success p is. However, p is unknown, otherwise a hypothesis test would be superfluous. It is nevertheless meaningful to assume that p takes a value $p_1 \neq p_0$ and use it to calculate the probability of a type II error. The deviation should be large enough to be considered relevant. If, for example, the test should reliably assess a coin which shows heads with a probability $p = \frac{2}{3}$ as unfair, then β' is calculated in the following way:

$$\beta' = \sum_{k=40}^{60} B_{100,2/3}(k) \approx 0.097$$

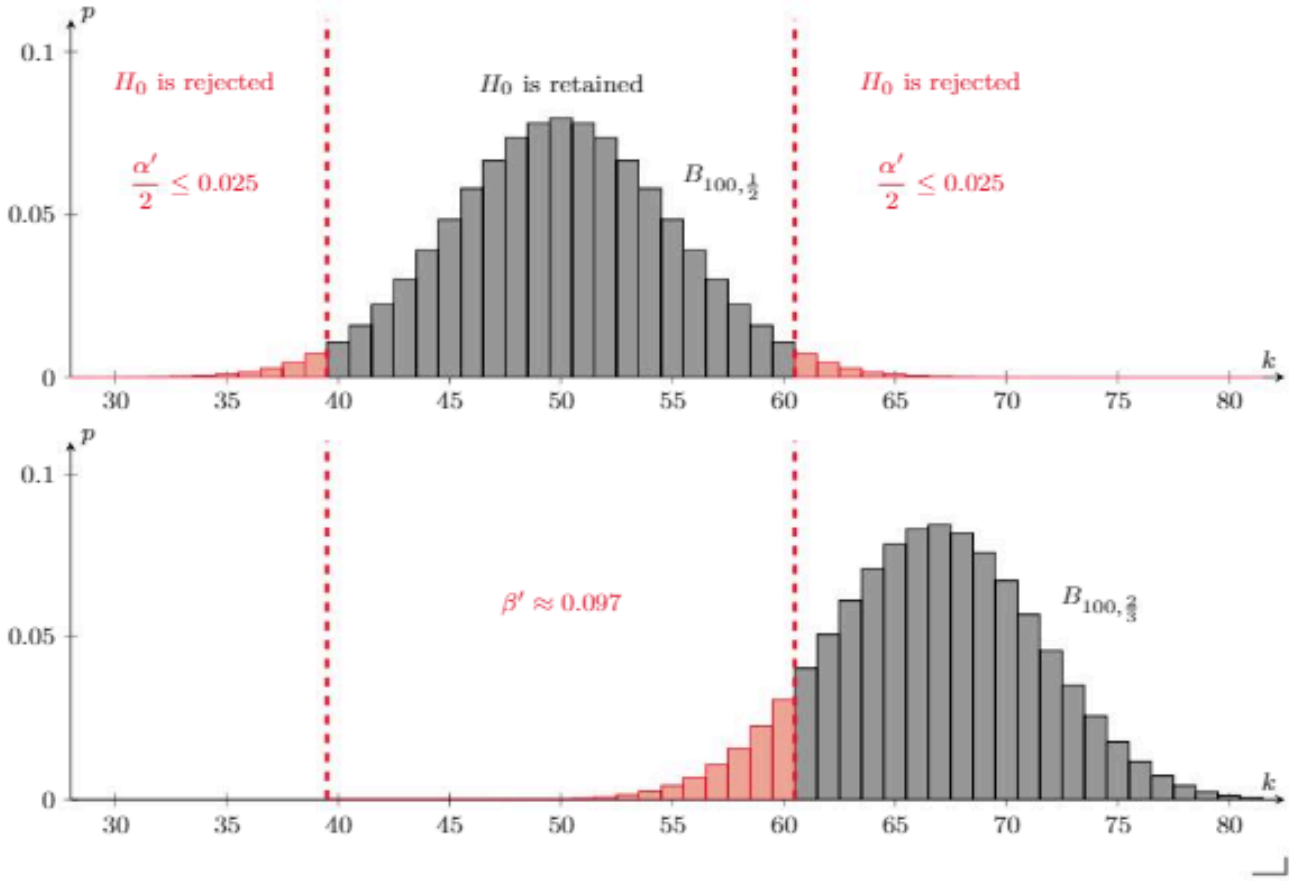
This means the probability that such a deviation of the probability for 'heads' is not detected using this test is about 10%. This suffices in most cases.

If, however, a coin with a probability of success $p = 60\%$ is considered as not acceptable, then β' has to be calculated as follows:

$$\beta' = \sum_{k=40}^{60} B_{100,0.6}(k) \approx 0.538$$

The test thus fails to detect this smaller deviation in more than half of all cases, which is certainly not very satisfactory. The only way to reduce the probability of a type II error while keeping the significance level is to increase the number of trials n.

The illustration shows the binomial$(100, \frac{1}{2})$-distribution with the probability of a type I error as well as the binomial$(100, \frac{2}{3})$-distribution with the probability of a type II error.



Type I and Type II Errors

With a *type I error*, the given probability of success p_0 of the null hypothesis is valid, however the test result does not lie in the prediction interval $\{k_L, k_L + 1, \ldots, k_R\}$, but in the rejection region V_{H_0}. Consequently:

$$\alpha' = P_{p_0}(X \in V) = 1 - \sum_{k=k_L}^{k_R} B_{n,p_0}(k)$$

The index p_0 of P_{p_0} should make it clear that the null hypothesis is assumed to be correct. By definition, the significance level α is an upper bound of the probability of the type I error: $\alpha' \leq \alpha$.

To calculate the *probability of a type II error*, an actual probability of success p_1 for the considered event has to be assumed. It differs from the value p_0, which was given with the null hypothesis. In most cases, p_1 represents a deviation from the null hypothesis which is considered as relevant. Then, using the probability of success p_1, one calculates with which probability an outcome in the prediction interval under p_0 occurs and thereby the null hypothesis is wrongly retained:

$$\beta'_{p_1} = P_{p_1}(X \notin V) = \sum_{k=k_L}^{k_R} B_{n,p_1}(k)$$

By changing the rejection region, a test can be improved only with regard to one of the two errors:

- If the rejection region is made smaller, then the probability for the type I error is reduced but the probability for the type II error is increased.
- If the rejection region is made larger, then the probability of the type II error is reduced but the probability for the type I error is increased.

In a hypothesis test, the probability of a type I error is bounded by the significance level. The probability of a type II error also implicitly depends on the significance level but one does not have any control over it. Avoiding a type I error therefore has a greater weight. To reduce the probability of type I and type II errors at the same time, the number of repetitions n has to be increased.

7.7 One-Tailed Tests

Example: The game 'Ludo', part V

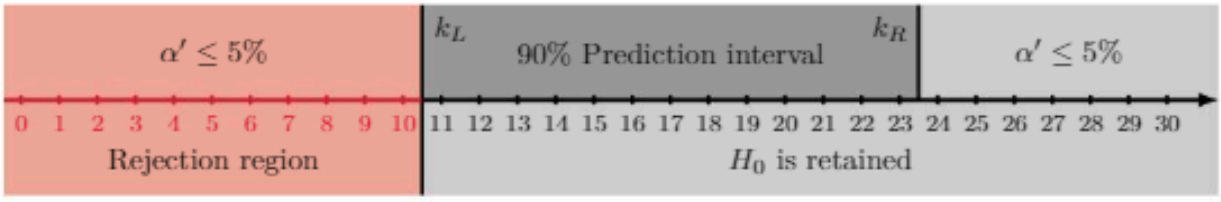
Sarah is actually only interested in getting her token on the starting position as soon as possible. If the probability for a six with the die to be used is greater than $\frac{1}{6}$, then this does not bother her. If, after 100 test throws, the probability of a six is greater than the upper limit 23 of the 90% prediction interval, then she is prepared to use the die nevertheless, even though this is a sign that the die is not fair. This means, on the basis of the indirect procedure of the test, that Sarah actually considers the null hypothesis H_0 '$p \geq \frac{1}{6}$' and she rejects this null hypothesis if the frequency of six in the 100 throws is less than $k_L = 11$.

What is then the probability of a type I error?

A type I error occurs when the null hypothesis is rejected even though it is correct. Since the rejection region only includes the values $0, 1, \ldots, 10$, this means:

$$\alpha' \leq \sum_{k=0}^{10} B_{n,1/6}(k) = 0.043 < 5\%$$

This means that if the rejection region only includes those values which lie to the left of the 90% prediction interval, then the test has the significance level 5%. This is illustrated in the figure below.



Lower-Tailed Test

If, from the outset, one is only interested in whether the probability of success is smaller than a given value p_0, then the null hypothesis is $H_0 : p \geq p_0$ and the alternative hypothesis is $H_1 : p < p_0$. In this case a *lower-tailed test* is used, which only rejects the null hypothesis if the number of successes lies too far to the left of the expected value np_0. The rejection region of the lower-tailed test with significance level α is $V_{H_0} = \{0, 1, \ldots, k_L - 1\}$, where k_L is the lower limit of the $(1 - 2\alpha)$ prediction interval of the binomial(n, p_0)-distribution.

Upper-Tailed Test

Accordingly, an *upper-tailed test* is carried out when one is only interested in whether the probability of success is greater than a given value p_0. The rejection region of the upper-tailed test with significance level α is $V_{H_0} = \{k_R + 1, k_R + 2, \ldots, n\}$, where k_R is the upper limit of the $(1 - 2\alpha)$ prediction interval of the binomial(n, p_0)-distribution.

Approximated Limits with One-Tailed Tests

A lower-tailed test has the approximated lower limit $k_L \approx np_0 - u_s\sqrt{np_0(1-p_0)}$ where $s = 1 - 2\alpha$ and u_s is taken from the table on page 128.

Accordingly, for the upper-tailed test, $k_R \approx np_0 + u_s\sqrt{np_0(1-p_0)}$ with $s = 1 - 2\alpha$.

Remark: With one-tailed tests, the null hypothesis is not described by a single value p_0 but by an entire interval of probabilities. Accordingly, the probability for a type I error also depends on which value of p from the null hypothesis is correct. Intuitively it is, however, clear that the probability for X to be in a one-tailed rejection region V_{H_0} is greatest for $p = p_0$. Therefore, the rejection region is determined by the prediction interval of the binomial(n, p_0)-distribution.

The probability of a type I error is, in the case of $p = p_0$, for the one-tailed as well as for the two-tailed binomial test at most equal to α. The probability for a type II error is smaller for the one-tailed test than for the two-tailed test. For this reason the former is preferred. It is, however, important that the one-tailed test is only used if the direction is defined in advance, before the data have been looked at, and not afterwards when it is already known whether the number of successes is larger or smaller than the expected value np_0.

Example: The game 'Ludo', part VI

Sarah wants to know the probability that a test with significance level 10% fails to detect a die which shows a six with a probability of only 10%. In a two-tailed test, the probability of a type II error is

$$\beta' = \sum_{k=11}^{23} B_{100,0.1}(k) \approx 0.417$$

In a one-tailed test, first the rejection region at the significance level 10% has to be determined. By approximation and rounding, $k_L = 16.67 - 1.28 \cdot 3.72 \approx 11.90 \approx 12$. Therefore, the probability of a type II error in the case of $p = 10\%$ equals

$$\beta' = \sum_{k=12}^{100} B_{100,0.1}(k) \approx 0.297$$

It is thus lower than for the two-tailed test. A one-tailed test detects a biased die with a greater probability than a two-tailed test. It will, however, still miss it in approximately 30% of the cases.

⌟

7.8 The Procedure of a One-Tailed Binomial Test

The procedure outlined for a binomial test in chapter 7.5 (p. 131 ◄ The Procedure of a Binomial Test) is adapted as follows for one-tailed tests.

Procedure of a One-Tailed Binomial Test

1. *Formulating the null hypothesis H_0 and determining the significance level α*

   Based on the question, it is decided whether a two-tailed, lower-tailed or an upper-tailed test is appropriate. The following null and alternative hypotheses belong to the two one-tailed tests:

   - Upper-tailed: $H_0 : p \leq p_0$, $H_1 : p > p_0$
   - Lower-tailed: $H_0 : p \geq p_0$, $H_1 : p < p_0$

2. *Determining the rejection region for one-tailed cases*

   The starting point is once again the test variable $X =$ 'number of successes'.

   - Upper-tailed: calculate the upper limit k_R of the $(1-2\alpha)$ prediction interval for a random variable with binomial(n, p_0)-distribution. The rejection region of the test is $V_{H_0} = \{k_R + 1, k_R + 2, \ldots, n\}$.
   - Lower-tailed: calculate the lower limit k_L of the $(1-2\alpha)$ prediction interval for a random variable with binomial(n, p_0)-distribution. The rejection region of the test is $V_{H_0} = \{0, 1, \ldots, k_L - 1\}$.

3. *Making a test decision*

   Same procedure as in the two-tailed case.

Example: Quality control

A company buys SD memory cards of a certain type in large quantities. The manufacturer declares that, at most, 4% of the memory cards are faulty. In a statistical test, a random sample of 400 memory cards is examined to check if the manufacturers specifications are correct.

What is the test result if 20 (or 24) of the examined memory cards are faulty?

1. *Formulating the null hypothesis H_0 and determining the significance level α*

   The upper limit, provided by the manufacturer, for the rate of faulty cards is the given value p_0 of the null hypothesis. Only deviations greater than this value are relevant. So, an upper-tailed test is carried out. H_0: $p \leq 0.04$, H_1: $p > 0.04$. No information about the significance level is given, so $\alpha = 5\%$.

2. *Determining the rejection region*

   Test variable $X =$ 'number of faulty memory cards in the sample'.

   In the case that H_0 applies, X has a binomial$(400, 0.04)$-distribution with expected value $\mu = 400 \cdot 0.04 = 16$ and standard deviation $\sigma = \sqrt{400 \cdot 0.04 \cdot 0.96} \approx 3.92$. Because the test is upper-tailed, the upper limit of the 90% prediction interval is needed, and the approximation can be used:

   $$k_R \approx \mu + 1.64\sigma = 16 + 1.64 \cdot 3.92 \approx 22.4 \approx 22$$

   The rejection region is therefore $V_{H_0} = \{23, 24, ..., 400\}$.

3. *Making a test decision*

   Because $20 \notin V_{H_0} = \{23, 24, ..., 400\}$, the null hypothesis is retained with 20 faulty memory cards in the sample. It would be different if there were 24 faulty memory cards. Because $24 \in V_{H_0} = \{23, 24, ..., 400\}$ the null hypothesis would then be rejected. ⌟

Remark: In the last example, the random variable X (number of faulty memory cards) does not have a binomial distribution, strictly speaking, because in practice, a sample is always drawn without replacement. Because of the large number of items, this can, however, be disregarded. In such examples, as well as in opinion polls, the binomial distribution can still be used. Only if the sample is larger than a tenth of the entire population from which the sample is taken does another distribution have to be used.

7.9 Exercises

1. Intuitively, it is plausible to assume: 'at birth, the probability of a boy is the same as that of a girl'. Suppose you only have one of the following pieces of information available. Can these data be seen as a confirmation or contradiction to the assumption?

   a) A family has three children. Two of them are boys.

   b) The ratio of boys to girls in your class at the present school or at primary school.

   c) Of 1000 new births in a hospital i) 490, ii) 501, iii) 520, iv) 550 are boys.

   d) Of 80,290 births in Switzerland in 2010, 41,111 were boys (p. 34 ◂).

   e) Of 8029 births in a city, 4111 are boys. Does the change in data, by a factor of 10 compared to part d), affect the conclusion about the assumption or the uncertainty of this conclusion?

2. Sarah wants to check, by throwing a die n times, whether the die shows a 'six' with a probability of $p = \frac{1}{6}$. Determine, on the basis of the 90% prediction interval, the rejection region of the null hypothesis H_0: $p = \frac{1}{6}$.

   a) $n = 120$ b) $n = 200$ c) $n = 500$

   d) How will Sarah decide in parts a), b), and c) if she throws ten 'sixes' more than the respective expected value?

3. A wheel of fortune has 16 sectors of equal size, of which one is painted red. If the dial stops in the red sector after spinning, then the contestant wins a prize. Using a test, the conjecture that the probability of winning a prize on this wheel of fortune is not $\frac{1}{16}$ will be examined.

   Design such a test: state the null hypothesis and determine its rejection region for the given sample size n on the basis of the 90% prediction interval.

   a) $n = 200$ b) $n = 1000$

4. Water pollution caused by chemical substances can lead to changes in the sex distribution of fish in the long term. Using a sample catch of 50 fish, we shall examine whether the sex distribution of a certain species of fish in a chemically polluted river has changed. Usually, males and females of this species of fish are found in equal numbers.

   a) Construct a corresponding test on the basis of the 90% prediction interval.

   b) If the test decides that the distribution has changed, due to the pollution caused by the factory, then the factory has to pay a fine. Which arguments could the company management present to dispute this fine?

5. Determine the prediction interval associated with the given certainty $1 - \alpha$.

   a) $n = 50$, $p = \frac{1}{4}$, $1 - \alpha = 90\%$

   b) $n = 50$, $p = \frac{1}{4}$, $1 - \alpha = 95\%$

   c) $n = 50$, $p = \frac{1}{4}$, $1 - \alpha = 99\%$

   d) $n = 100$, $p = 20\%$, $1 - \alpha = 90\%$

   e) $n = 100$, $p = 20\%$, $1 - \alpha = 95\%$

   f) $n = 100$, $p = 20\%$, $1 - \alpha = 99\%$

   g) $n = 500$, $p = 5\%$, $1 - \alpha = 90\%$

   h) $n = 500$, $p = 5\%$, $1 - \alpha = 95\%$

   i) $n = 500$, $p = 5\%$, $1 - \alpha = 99\%$

   j) $n = 200$, $p = 73\%$, $1 - \alpha = 90\%$

   k) $n = 200$, $p = 73\%$, $1 - \alpha = 95\%$

   l) $n = 200$, $p = 73\%$, $1 - \alpha = 99\%$

6. Using a statistical test, we shall examine whether or not a coin is fair. Determine for the given sample size n and significance level α the rejection region for the null hypothesis H_0: $p = 50\%$.

   a) $n = 50$, $\alpha = 5\%$ b) $n = 250$, $\alpha = 5\%$ c) $n = 500$, $\alpha = 1\%$

7. The faces of a regular polyhedron are numbered. It is thrown like a die. The following null hypothesis is being tested: a 'one' is thrown with the probability $p = (\text{number of faces})^{-1}$. Determine the rejection region for the given quantities.

   a) Throwing a tetrahedron: $n = 100$, $\alpha = 5\%$

   b) Throwing an octahedron: $n = 200$, $\alpha = 5\%$

   c) Throwing a dodecahedron (12 faces): $n = 250$, $\alpha = 10\%$

   d) Throwing an icosahedron (20 faces): $n = 1000$, $\alpha = 1\%$

8. Gregor Mendel (1822–1884) the 'father of genetics', crossed two pure strains of yellow and green peas in one of his experiments. Since yellow is dominant, all peas are yellow in the first generation, and in the second generation, according to his theory, the probability of a green pea is $\frac{1}{4}$. In the second generation, he accrued 8023 peas, of which 2001 were green.

   a) State the null hypothesis and the rejection region of a test which can check Mendel's theory with the significance level $\alpha = 5\%$.

   b) What conclusions can you draw from the observed frequency of 2001 green peas?

9. A machine mixes two coarse-grained substances. The mix should comprise $\frac{1}{4}$ of substance A and $\frac{3}{4}$ of substance B. Using samples, the blending is examined. Large deviations from the desired mixing ratio show that the blending is unsatisfactory.

   a) Check, at the significance level $\alpha = 5\%$, whether the following samples point to an unsatisfactory blending.

      i) In test sample 1, one finds 34 grains of substance A and 105 of substance B.

      ii) In test sample 2, one finds 27 grains of substance A and 110 of substance B.

   b) Find the number of grains of substance A for which, at the significance level $\alpha = 5\%$

      i) in a test sample of 150 grains it can only just be concluded that the desired mixing ratio of $\frac{1}{4}$ is not maintained.

      ii) in a test sample of 173 grains it can only just be concluded that the desired mixing ratio of $\frac{1}{4}$ is maintained.

10. Clairvoyance? In the year 1973, CHARLES TART carried out an experiment at the University of California in Davis to demonstrate extrasensory perception. He used a random generator on a computer which selected one of four possible symbols. The test person did not see the result and had to find out the correct symbol. 15 people who claimed to have clairvoyant ability were tested, each trying 500 times to find the correct symbol. In total, the right symbol was detected in 2006 out of $15 \cdot 500 = 7500$ trials (Freedman et al., 1978).

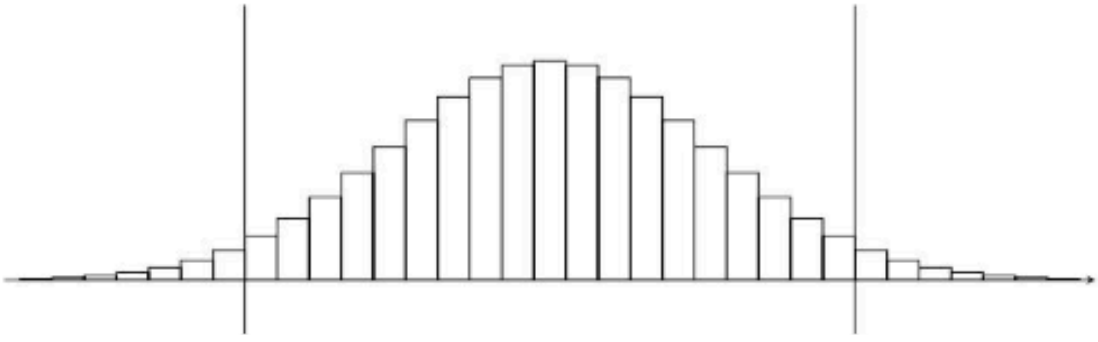
    a) Determine the rejection region of the null hypothesis 'the tested person guessed the symbols by chance' for the significance level $\alpha = 1\%$.

    b) What conclusion did the constructed test come to?

    c) What is the probability that in 7500 repetitions people without clairvoyant ability guess the correct symbol purely by chance in 2006 or more cases?

11. At a secondary school, 63 of the 109 failing mathematics marks are from girls: in other words, the percentage of girls among the students with failing mathematics marks is 57.8%. The percentage of girls at the whole school is 62.8%. Is chance a plausible explanation for this difference? Calculate the 90% prediction interval for the proportion of girls in a randomly selected sample of 109 students at this school. What conclusion can you draw?

12. Explain, using the diagram, the principle of a statistical test. Mark important values and regions in the diagram.



13. In the last local elections in the city of Berne, of the total female part of the population, 24.2% voted, 36.3% were not eligible to vote, and 39.5% were eligible to vote but did not cast their vote. An opinion research institute carried out a survey on voting behaviour. 210 females were selected at random from the population of Berne. They were asked if they voted in the last local elections.

    a) Design a statistical test ($\alpha = 10\%$) where you can examine whether the number of women in the random sample, who claimed to have voted, is consistent with the actual female voter participation figure of 24.2%.

    b) 64 of the people surveyed stated that they had voted in the last local elections. What do you make of this finding?

    c) What could have caused this outcome to turn out somewhat different than expected?

    d) Design a new test ($\alpha = 5\%$) with the assumption that, for the random sample, 210 women *who are actually eligible to vote* are chosen. The aim is to examine if the proportion of women eligible to vote and who actually voted in the last local elections is consistent with the information provided in the random sample.

14. Among a recently cultured type of flowers, there are some with single-coloured and others with two-coloured petals. One of the two variants is the dominant hereditary characteristic and occurs with probability $p = \frac{3}{4}$. The other, recessive characteristic occurs with probability $q = \frac{1}{4}$.

    a) A crossing experiment produces nine progeny. The decision rule is defined such that the colour variant which is more frequent is considered dominant. What is the probability that the wrong characteristic is taken for dominant?

    b) A crossing experiment produces ten progeny. If at least six have two-coloured petals, one will assume that the two-coloured variant is dominant. Describe the two possible errors and determine their probability.

    c) A crossing experiment produces 25 progeny. If more than 10 have two-coloured petals, then one will assume that the two-coloured variant is dominant. Determine the probabilities of the two possible errors.

15. A die is to be examined in 50 throws at a significance level of 5%, to see if the probability of a 'six' is really $p = \frac{1}{6}$. The rejection region is chosen as $V_{H_0} = \{0, \ldots, 3\} \cup \{15, \ldots, 50\}$.

    a) Determine the probability for a type I error.

    b) Actually, the probability of a 'six' equals $p = 0.25$ for this die. With which probability is the tested die still classified as fair?

16. With 100 tosses a coin is to be tested at a significance level of 10% regarding the probability of heads and tails being the same.

    a) Indicate the null hypothesis and the alternative hypothesis and determine the rejection region.

    b) Determine the probability for a type I error.

    c) The coin shows tails with a probability of $p = 0.4$. What is the probability that the test leads to a wrong decision?

17. The following tests are carried out for the null hypothesis H_0: $p = 0.5$ (see exercise 6):

    a) $n = 50$, $\alpha = 5\%$, rejection region $V_{H_0} = \{0, \ldots, 17\} \cup \{33, \ldots, 50\}$

    b) $n = 250$, $\alpha = 5\%$, rejection region $V_{H_0} = \{0, \ldots, 109\} \cup \{141, \ldots, 250\}$

    c) $n = 500$, $\alpha = 1\%$, rejection region $V_{H_0} = \{0, \ldots, 220\} \cup \{280, \ldots, 500\}$

    Determine the respective probabilities of a type II error under the assumption of an actual probability of success of

    i) $p = 0.7$. ii) $p = 0.6$. iii) $p = 0.45$.

18. Determine for the tests in exercise 7, the probability of a type II error assuming $p = \frac{1}{6}$.

    a) Throwing a tetrahedron: $n = 100$, $\alpha = 5\%$, $V_{H_0} = \{0, \ldots, 16\} \cup \{34, \ldots, 100\}$

    b) Throwing an octahedron: $n = 200$, $\alpha = 5\%$, $V_{H_0} = \{0, \ldots, 15\} \cup \{35, \ldots, 200\}$

    c) Throwing a dodecahedron: $n = 250$, $\alpha = 10\%$, $V_{H_0} = \{0, \ldots, 13\} \cup \{29, \ldots, 250\}$

    d) Throwing an icosahedron: $n = 1000$, $\alpha = 1\%$, $V_{H_0} = \{0, \ldots, 31\} \cup \{69, \ldots, 1000\}$

19. According to information provided by the manufacturer, side effects occur in at most 10% of patients who take a certain type of medication. A head physician at a hospital suspects that, in the case of her patients, this ratio is somewhat higher. To test this conjecture, 120 patients who have been prescribed the medication are examined.

    a) Set up a null hypothesis and an alternative hypothesis for a corresponding test.

    b) Describe what consequences type I and type II errors have in this test.

    c) Calculate the rejection regions for the following significance levels: $\alpha = 10\%$, $\alpha = 5\%$, and $\alpha = 1\%$. Determine the respective probability β of a type II error under the assumption that the percentage of patients with side-effects is actually 20%.

    d) The physician discovers side effects in 19 patients. What conclusions can she draw from this? Distinguish the three cases $\alpha = 10\%$, $\alpha = 5\%$, and $\alpha = 1\%$.

    e) What do you have to consider when determining α in this medical test? Which α is the most appropriate here and why?

20. A casino maintains that, in the case of its gambling machines, the probability of making a profit is at least 25%. This statement is to be tested using a series of n repetitions.

    – devise a statistical test with significance level $\alpha = 5\%$ for the alternative hypothesis and the sample size n specified below.

    – determine, for the constructed tests, the probability of a type II error in case that the probability of the casino making a profit actually equals $p_1 = 15\%$ (or $p_2 = 20\%$).

    a) $H_1 : p \neq 25\%$ (two-tailed), $n = 100$

    b) $H_1 : p \neq 25\%$ (two-tailed), $n = 300$

    c) $H_1 : p < 25\%$ (one-tailed), $n = 100$

    d) $H_1 : p < 25\%$ (one-tailed), $n = 300$

21. Can elephants count? Researchers addressed this question in the following experiment (Irie-Sugimoto et al., 2009). A keeper dropped a number of fruits in each of two metal buckets which were placed two metres away from the elephant. The elephant could not see how many fruits were dropped into each of the buckets, but he could hear the sound. Then the elephant was allowed to fetch one of the buckets and eat the fruit in it. The experiment was repeated 54 times with each of four elephants and it was taken note of whether or not the elephant chose the container with more fruit in it. The difference in the number of fruits in the two buckets varied between 1 and 3 and the larger number between 3 and 6 (with a total of 9 combinations).

    a) Discuss whether it is justified to use the binomial distribution for the number of successes in the $4 \cdot 54 = 216$ repetitions.

    b) A total of 163 successes were registered. How high do you estimate the probability that an elephant chooses the container with more fruit in it? Is it plausible that this probability is equal to 0.5?

    c) Construct a test where you can examine whether elephants choose the container with more fruit in it more often. Choose a suitable significance level α. How does the test decision turn out if there are 163 successes?

    d) The number of successes of the four elephants were 39, 37, 44 and 43. Would you conclude from this that the four elephants vary in their ability to choose the fuller container? How would you go about formulating an answer to this question?

22. A sweets manufacturer produces jelly babies in the colours red, green, yellow, white, and orange and packages them in 100 g packets. There are usually 41 or 42 jelly babies in such a packet. Most people like eating the red jelly babies most. For this reason, the proportion of red jelly babies in the packets is higher compared to the other coloured jelly babies. The manufacturer declares that at least one third of his jelly babies are red. David suspects that this information from the manufacturer is wrong and that the proportion of red jelly babies is less than a third. He buys three 100 g packets and counts the number of jelly babies.

	red	green	yellow	white	orange	total
Packet 1	12	8	7	5	9	41
Packet 2	10	7	4	9	11	41
Packet 3	14	4	10	6	8	42
Total	36	19	21	20	28	124

    a) Construct a test for $n = 124$ at the significance level $\alpha = 5\%$, so that David can check whether the information from the manufacturer is correct. Can Davids suspicion be confirmed on the basis of the existing colour distribution?

    b) Construct a two-tailed test with sample size $n = 124$ and significance level $\alpha = 10\%$, so that you can check whether the proportion of green jelly babies is one sixth.

    c) If you now evaluate all colours except red with the test constructed in b) then you detect that the number of orange jelly babies lies in the rejection region. Can you therefore reject the hypothesis that all colours, with the exception of red, occur with a probability of $p = \frac{1}{6}$ at the 10% significance level?

    d) How many green (and yellow or orange, respectively) jelly babies in a 100g packet with $n = 41$ jelly babies are compatible with the hypothesis $p = \frac{1}{6}$ ($\alpha = 0.05$)?

23. The risk for many illnesses also depends on the blood type (Bruhns 2015). Some studies indicate that people with the blood type B are more susceptible to chronic inflammations of the pancreas. To check this conjecture, the blood types of 1083 people were evaluated who suffer from this illness. 8% of the population have blood type B.

    a) Formulate the null hypothesis and construct a statistical test with $\alpha = 5\%$.

    b) Interpret the result if 112 people from the sample have blood type B.

24. Pierre Simon Laplace studied the question whether the probability for a female birth and a male birth is exactly the same (Barth and Haller, 1983).

    a) Devise a test for $n = 10{,}000$ ($n = 20{,}000$ and $n = 30{,}000$) with $\alpha = 0.01$.

    b) On the basis of the analysis of birth data from several cities and years, Laplace assumed that the probability for a male birth is close to $p_1 = \frac{22}{43} \approx 0.5116$. Determine the probability for the β-error in the devised test in a) for this value p_1.

    c) In Paris in the years 1745 – 1770, 241,945 female births and 251,527 male births were registered. Is this data, at the significance level of $\alpha = 0.01$, consistent with the assumption that the probability for a male birth is $p = \frac{22}{43}$?

    d) In London in the years 1664 – 1757, 737,629 male births and 698,958 female births were registered. Is this data, at the significance level of $\alpha = 0.01$, consistent with the assumption that the probability for a male birth is $p = \frac{22}{43}$?

    Remark: The detected deviation from the normal percentage troubled Laplace. He later discovered that it was linked to the fact that in Paris the percentage of girls amongst the foundlings was disproportionally high. Presumably, the population in the surrounding area of Paris abandoned more girls than boys in the city.