# Hypothesis Testing

# 10

## 10.1 Introduction

We introduced point and interval estimation of parameters in the previous chapter. Sometimes, the research question is less ambitious in the sense that we are not interested in precise estimates of a parameter, but we only want to examine whether a statement about a parameter of interest or the research hypothesis is true or not (although we will see later in this chapter that there is a connection between confidence intervals and statistical tests, called *duality*). Another related issue is that once an analyst estimates the parameters on the basis of a random sample, (s)he would like to infer something about the value of the parameter in the population. Statistical hypothesis tests facilitate the comparison of estimated values with hypothetical values.

*Example 10.1.1* As a simple example, consider the case where we want to find out whether the proportion of votes for a party $P$ in an election will exceed 30 % or not. Typically, before the election, we will try to get representative data about the election proportions for different parties (e.g. by telephone interviews) and then make a statement like "yes", we expect that $P$ will get more than 30 % of the votes or "no", we do not have enough evidence that $P$ will get more than 30 % of the votes. In such a case, we will only know after the election whether our statement was right or wrong. Note that the term representative data only means that the sample is similar to the population with respect to the distributions of some key variables, e.g. age, gender, and education. Since we use one sample to compare it with a fixed value (30 %), we call it a **one-sample problem**.

*Example 10.1.2* Consider another example in which a clinical study is conducted to compare the effectiveness of a new drug ($B$) to an established standard drug ($A$) for a specific disease, for example too high blood pressure. Assume that, as a first step, we want to find out whether the new drug causes a higher reduction in blood



C. Heumann et al., *Introduction to Statistics and Data Analysis*,
DOI 10.1007/978-3-319-46162-5_10

pressure than the already established older drug. A frequently used study design for this question is a randomized (i.e. patients are randomly allocated to one of the two treatments) controlled clinical trial (double blinded, i.e. neither the patient nor the doctor know which of the drugs a patient is receiving during the trial), conducted in a fixed time interval, say 3 months. A possible hypothesis is that the average change in the blood pressure in group $B$ is higher than in group $A$, i.e. $\delta_B > \delta_A$ where $\delta_j = \mu_{j0} - \mu_{j3}$, $j = A, B$ and $\mu_{j0}$ is the average blood pressure at baseline before measuring the blood pressure again after 3 months ($\mu_{j3}$). Note that we expect both the differences $\delta_A$ and $\delta_B$ to be positive, since otherwise we would have some doubt that either drug is effective at all. As a second step (after statistically proving our hypothesis), we are interested in whether the improvement of $B$ compared to $A$ is relevant in a medical or biological sense and is valid for the entire population or not. This will lead us again to the estimation problems of the previous chapter, i.e. quantifying an effect using point and interval estimation. Since we are comparing two drugs, we need to have two samples from each of the drugs; hence, we have a **two-sample problem**. Since the patients receiving $A$ are different from those receiving $B$ in this example, we refer to it as a "two-*independent*-samples problem".

*Example 10.1.3* In another example, we consider an experiment in which a group of students receives extra mathematical tuition. Their ability to solve mathematical problems is evaluated before and after the extra tuition. We are interested in knowing whether the ability to solve mathematical problems increases after the tuition, or not. Since the same group of students is used in a pre–post experiment, this is called a "two-*dependent*-samples problem" or a "paired data problem".

## 10.2 Basic Definitions

### 10.2.1 One- and Two-Sample Problems

In one-sample problems, the data is usually assumed to arise as *one* sample from a defined population. In two-sample problems, the data originates in the form of *two samples* possibly from two different populations. The heterogeneity is often modelled by assuming that the two populations only differ in some parameters or key quantities such as expectation (i.e. mean), median, or variance. As in our introductory example, the samples can either be independent (as in the drug Example 10.1.2) or dependent (as in the evaluation Example 10.1.3).

### 10.2.2 Hypotheses

A researcher may have a research question for which the truth about the population of interest is unknown. Suppose data can be obtained using a survey, observation, or

an experiment: if, given a prespecified uncertainty level, a statistical test based on the data supports the hypothesis about the population, we say that this hypothesis is statistically proven. Note that the research question has to be operationalized before it can be tested by a statistical test. Consider the drug Example 10.1.2: we want to examine whether the new drug $B$ has a greater blood pressure lowering effect than the standard drug $A$. We have several options to operationalize this research question into a statistical set-up. One is to test whether the *average* reduction (from baseline to 3 months) of the blood pressure is higher (and positive) for drug $B$ than drug $A$. We then state our hypotheses in terms of expected values (i.e. $\mu$). Why do we have to use the expected values $\mu$ and not simply compare the arithmetic means $\bar{x}$? The reason is that the superiority of $B$ shown in the sample will only be valid for this sample and not necessarily for another sample. We need to show the superiority of $B$ in the entire population, and hence, our hypothesis needs to reflect this. Another option would be, for example, to use median changes in blood pressure values instead of mean changes in blood pressure values. An important point is that the research hypothesis which we want to prove has to be formulated as the statistical alternative hypothesis, often denoted by $H_1$. The reason for this will become clearer later in this chapter. The opposite of the research hypothesis has to be formulated as the statistical null hypothesis, denoted by $H_0$. In the drug example, the alternative and null hypotheses are, respectively,

$$H_1 : \delta_B > \delta_A$$

and

$$H_0 : \delta_B \leq \delta_A.$$

We note that the two hypotheses are disjoint and the union of them covers all possible differences of $\delta_B$ and $\delta_A$. There is a boundary value ($\delta_B = \delta_A$) which separates the two hypotheses. Since we want to show the superiority of $B$, the hypothesis was formulated as a one-sided hypothesis. Note that there are different ways to formulate two-sample hypotheses; for example, $H_1 : \delta_B > \delta_A$ is equivalent to $H_1 : \delta_B - \delta_A > 0$. In fact, it is very common to formulate two-sample hypotheses as differences, which we will see later in this chapter.

### 10.2.3 One- and Two-Sided Tests

We distinguish between one-sided and two-sided hypotheses and tests. In the previous section, we gave an example of a one-sided test.

For an unknown population parameter $\theta$ (e.g. $\mu$) and a fixed value $\theta_0$ (e.g. 5), the following three cases have to be distinguished:

| Case | Null hypothesis | Alternative hypothesis | |
|---|---|---|---|
| (a) | $\theta = \theta_0$ | $\theta \neq \theta_0$ | Two-sided test problem |
| (b) | $\theta \geq \theta_0$ | $\theta < \theta_0$ | One-sided test problem |
| (c) | $\theta \leq \theta_0$ | $\theta > \theta_0$ | One-sided test problem |

*Example 10.2.1* One-sample problems often test whether a target value is achieved or not. For example, consider the null hypothesis as

- $H_0$ : average filling weight of packages of flour = 1 kg
- $H_0$ : average body height (men) = 178 cm.

The alternative hypothesis $H_1$ is formulated as deviation from the target value. If deviations in both directions are interesting, then $H_1$ is formulated as a two-sided hypothesis,

- $H_1$ : average body height (men) $\neq$ 178 cm.

If deviations in a specific direction are the subject of interest, then $H_1$ is formulated as a one-sided hypothesis, for example,

- $H_1$ : average filling weight of flour packages is lower than 1 kg.
- $H_1$ : average filling weight of flour packages is greater than 1 kg.

Two-sample problems often examine differences of two samples. Suppose the null hypothesis $H_0$ is related to the average weight of flour packages filled by two machines, say 1 and 2. Then, the null hypothesis is

- $H_0$ : average weight of flour packages filled by machine 1 = average weight of flour packages filled by machine 2.

Then, $H_1$ can be formulated as a one-sided or two-sided hypothesis. If we want to prove that machine 1 and machine 2 have different filling weights, then $H_1$ would be formulated as a two-sided hypothesis

- $H_1$ : average filling weight of machine 1 $\neq$ average filling weight of machine 2.

If we want to prove that machine 1 has lower average filling weight than machine 2, $H_1$ would be formulated as a one-sided hypothesis

- $H_1$ : average filling weight of machine 1 $<$ average filling weight of machine 2.

If we want to prove that machine 2 has lower filling weight than machine 1, $H_1$ would be formulated as a one-sided hypothesis

- $H_1$ : average filling weight of machine 1 $>$ average filling weight of machine 2.

*Remark 10.2.1* Note that we have *not* considered the following situation: $H_0 : \theta \neq \theta_0$, $H_1 : \theta = \theta_0$. In general, with the tests described in this chapter, we cannot prove the equality of a parameter to a predefined value and neither can we prove the equality of two parameters, as in $H_0 : \theta_1 \neq \theta_2$, $H_1 : \theta_1 = \theta_2$. We can, for example,

not prove (statistically) that machines 1 and 2 in the previous example provide equal filling weight. This would lead to the more complex class of equivalence tests, which is a topic beyond the scope of this book.

### 10.2.4 Type I and Type II Error

If we undertake a statistical test, two types of error can occur.

- The hypothesis $H_0$ is true but is rejected; this error is called **type I error**.
- The hypothesis $H_0$ is not rejected although it is wrong; this is called **type II error**.

When a hypothesis is tested, then the following four situations are possible:

| | $H_0$ is true | $H_0$ is not true |
|---|---|---|
| $H_0$ is not rejected | Correct decision | Type II error |
| $H_0$ is rejected | Type I error | Correct decision |

The significance level is the probability of type I error, $P(H_1|H_0) = \alpha$, which is the probability of rejecting $H_0$ (accepting $H_1$) if $H_0$ is true. If we construct a test, the significance level $\alpha$ is prespecified, e.g. $\alpha = 0.05$. A significance test is constructed such that the probability of a type I error does not exceed $\alpha$ while the probability of a type II error depends on the true but unknown parameter values in the population(s) and the sample size. Therefore, the two errors are not symmetrically treated in a significance test. In fact, the type II error $\beta$, $P(H_0|H_1) = \beta$ is not controlled by the construction of the test and can become very high, sometimes up to $1 - \alpha$. This is the reason why a test not rejecting $H_0$ is not a (statistical) proof of $H_0$. In mathematical statistics, one searches for the best test which maintains $\alpha$ and minimizes $\beta$. Minimization of both $\alpha$ and $\beta$ simultaneously is not possible. The reason is that when $\alpha$ increases then $\beta$ decreases and vice versa. So one of the errors needs to be fixed and the other error is minimized. Consequently, the error which is considered more serious is fixed and then the other error is minimized. The tests discussed in the below sections are obtained based on the assumption that the type I error is more serious than the type II error. So the test statistics are obtained by fixing $\alpha$ and then minimizing $\beta$. In fact, the null hypothesis is framed in such a way that it implies that the type I error is more serious than the type II error. The probability $1 - \beta = P(H_1|H_1)$ is called the **power** of the test. It is the probability of making a decision in favour of the research hypothesis $H_1$, if it is true, i.e. the probability of detecting a correct research hypothesis.

### 10.2.5 How to Conduct a Statistical Test

In general, we can follow the steps described below to test a hypothesis about a population parameter based on a sample of data.

(1) Define the distributional assumptions for the random variables of interest, and specify them in terms of population parameters (e.g. $\theta$ or $\mu$ and $\sigma$). This is necessary for parametric tests. There are other types of tests, so-called nonparametric tests, where the assumptions can be relaxed in the sense that we do not have to specify a particular distribution, see Sect. 10.6ff. Moreover, for some tests the distributional assumptions can be relaxed if the sample size is large.
(2) Formulate the null hypothesis and the alternative hypothesis as described in Sects. 10.2.2 and 10.2.3.
(3) Fix a significance value (often called type I error) $\alpha$, for example $\alpha = 0.05$, see also Sect. 10.2.4.
(4) Construct a test statistic $T(\mathbf{X}) = T(X_1, X_2, \ldots, X_n)$. The distribution of $T$ has to be known under the null hypothesis $H_0$. We note again that $(X_1, X_2, \ldots, X_n)$ refers to the random variables before drawing the actual sample and $x_1, x_2, \ldots, x_n$ are the realized values (observations) in the sample.
(5) Construct a critical region $K$ for the statistic $T$, i.e. a region where—if $T$ falls in this region—$H_0$ is rejected, such that

$$P_{H_0}(T(\mathbf{X}) \in K) \leq \alpha \, .$$

The notation $P_{H_0}(\cdot)$ means that this inequality must hold for all parameter values $\theta$ that belong to the null hypothesis $H_0$. Since we assume that we know the distribution of $T(\mathbf{X})$ under $H_0$, the critical region is defined by those values of $T(\mathbf{X})$ which are unlikely (i.e. with probability of less than $\alpha$) to be observed under the null hypothesis. Note that although $T(X)$ is a random variable, $K$ is a well-defined region, see Fig. 10.1 for an example.
(6) Calculate $t(x) = T(x_1, x_2, \ldots, x_n)$ based on the realized sample values $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$.
(7) Decision rule: if $t(x)$ falls into the critical region $K$, the null hypothesis $H_0$ is rejected. The alternative hypothesis is then statistically proven. If $t(x)$ falls outside the critical region, $H_0$ is not rejected.

$$t(x) \in K : H_0 \text{ rejected} \Rightarrow H_1 \text{ is statistically significant,}$$
$$t(x) \notin K : H_0 \text{ not rejected and therefore accepted.}$$

The next two paragraphs show how to arrive at the test decisions from step 7 in a different way. Readers interested in an example of a statistical test may jump to Sect. 10.3.1 and possibly also Example 10.3.1.

### 10.2.6 Test Decisions Using the $p$-Value

Statistical software usually does not show us all the steps of hypothesis testing as outlined in Sect. 10.2.5. It is common that instead of calculating and reporting the critical values, the test statistic is printed together with the so-called $p$-value. It is possible to use the $p$-value instead of critical regions for making test decisions. The $p$-value of the test statistic $T(\mathbf{X})$ is defined as follows:

$$\begin{aligned} \text{two-sided case:} \quad & P_{H_0}(|T| \geq t(x)) = p\text{-value} \\ \text{one-sided case:} \quad & P_{H_0}(T \geq t(x)) = p\text{-value} \\ & P_{H_0}(T \leq t(x)) = p\text{-value} \end{aligned}$$

It can be interpreted as the probability of observing results equal to, or more extreme than those actually observed if the null hypothesis was true. Then, the decision rule is

$H_0$ is rejected if the $p$-value is smaller than the prespecified significance level $\alpha$.
Otherwise, $H_0$ cannot be rejected.

*Example 10.2.2* Assume that we are dealing with a two-sided test and assume further that the test statistic $T(x)$ is $N(0, 1)$-distributed under $H_0$. The significance level is $\alpha = 0.05$. If we observe, for example, $t = 3$, then the $p$-value is $P_{H_0}(|T| \geq 3)$. This can be calculated in $R$ as

```
2*(1-pnorm(3))
```

because `pnorm()` is used to calculate $P(X \leq x)$, and therefore, `1-pnorm()` can be used to calculate $P(X > x)$. We have to multiply with two because we are dealing with a two-sided hypothesis. The result is $p = 0.002699796$. Therefore, $H_0$ is rejected. The one-sided $p$-value is half of the two-sided $p$-value, i.e. $P(T \geq 3) = P(T \leq 3) = 0.001349898$, and is not necessarily reported by $R$. It is therefore important to look carefully at the $R$ output when dealing with one-sided hypotheses.

The $p$-value is sometimes also called the *significance*, although we prefer the term $p$-value. We use the term *significance* only in the context of a test result: a test is (statistically) significant if (and only if) $H_0$ can be rejected.

Unfortunately, the $p$-value is often over-interpreted: both a test and the $p$-value can only provide a yes/no decision: either $H_0$ is rejected or not. Interpreting the $p$-value as the probability that the null hypothesis is true is wrong! It is also incorrect to say that the $p$-value is the probability of making an error during the test decision. In our (frequentist) context, hypotheses are true or false and no probability is assigned to them. It can also be misleading to speak of "highly significant" results if the $p$-value is very small. A last remark: the $p$-value itself is a random variable: under the null hypothesis, it follows a uniform distribution, i.e. $p \sim U(0, 1)$.

### 10.2.7 Test Decisions Using Confidence Intervals

There is an interesting and useful relationship between confidence intervals and hypothesis tests. If the null hypothesis $H_0$ is rejected at the significance level $\alpha$, then there exists a $100(1-\alpha)\,\%$ confidence interval which yields the same conclusion as the test: if the appropriate confidence interval does not contain the value $\theta_0$ targeted in the hypothesis, then $H_0$ is rejected. We call this **duality**. For example, recall Example 10.1.2 where we were interested in whether the average change in blood pressure for drug $B$ is higher than for drug $A$, i.e. $H_1 : \delta_B > \delta_A$. This hypothesis is equivalent to $H_1 : \delta_B - \delta_A > \delta_0 = 0$. In the following section, we develop tests to decide whether $H_1$ is statistically significant or not. Alternatively, we could construct a $100(1-\alpha)\%$ confidence interval for the difference $\delta_B - \delta_A$ and evaluate whether the interval contains $\delta_0 = 0$ or not; if yes, we accept $H_0$; otherwise, we reject it. For some of the tests introduced in following section, we refer to the confidence intervals which lead to the same results as the corresponding test.

## 10.3 Parametric Tests for Location Parameters

### 10.3.1 Test for the Mean When the Variance is Known (One-Sample Gauss Test)

We develop a hypothesis test to test whether the unknown mean (expectation) $\mu$ of a $N(\mu, \sigma^2)$-distributed random variable $X$ either differs from a specific value $\mu = \mu_0$ or is smaller (or greater) than $\mu_0$. We assume that the variance $\sigma^2 = \sigma_0^2$ is known. We apply the scheme of Sect. 10.2.5 step by step to develop the test procedure and then give an illustrative example.

*1. Distributional assumption*: The random variable $X$ follows a $N(\mu, \sigma_0^2)$- distribution with known variance $\sigma_0^2$. We assume that an i.i.d. random sample is drawn from $X_1, X_2, \ldots, X_n$ where the $X_i$s follow the same distribution as $X$, $i = 1, 2, \ldots, n$.

*2. Define any of the following set of hypotheses $H_0$ and $H_1$*:

$$\begin{aligned} H_0 &: \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0, \qquad \text{(two-sided test)} \\ H_0 &: \mu \leq \mu_0 \quad \text{versus} \quad H_1 : \mu > \mu_0, \qquad \text{(one-sided test)} \\ H_0 &: \mu \geq \mu_0 \quad \text{versus} \quad H_1 : \mu < \mu_0, \qquad \text{(one-sided test).} \end{aligned}$$

*3. Specify the probability of a type I error* $\alpha$: Often $\alpha = 0.05 = 5\,\%$ is chosen.

*4. Construct a test statistic*: The unknown mean, i.e. the expectation $\mu$, is usually estimated by the sample mean $\bar{x}$. We already know that if the $X_i$s are i.i.d., then the sample mean is normally distributed. Under the assumption that $H_0$ is true,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \overset{H_0}{\sim} N(\mu_0, \sigma_0^2/n),$$

where $\overset{H_0}{\sim}$ means the "distribution under $H_0$". If we standardize the mean under $H_0$, we get a $N(0, 1)$-distributed test statistic

$$T(\mathbf{X}) = \frac{\bar{X} - \mu_0}{\sigma_0}\sqrt{n} \overset{H_0}{\sim} N(0, 1),$$

see also Theorem 7.3.2. Note that $T(\mathbf{X})$ follows a normal distribution even if the $X_i$s are *not* normally distributed and if $n$ is large enough which follows from the Central Limit Theorem (Appendix C.3). One can conclude that the distributional assumption from step 1 is thus particularly important for small samples, but not necessarily important for large samples. As a rule of thumb, $n \geq 30$ is considered to be a large sample. This rule is based on the knowledge that a $t$-distribution with more than 30 degrees of freedom gets very close to a $N(0, 1)$-distribution.

*5. Critical region*: Since the test statistic $T(\mathbf{X})$ is $N(0, 1)$-distributed, we get the following critical regions, depending on the hypothesis:

| Case | $H_0$ | $H_1$ | Critical region $K$ |
|---|---|---|---|
| (a) | $\mu = \mu_0$ | $\mu \neq \mu_0$ | $K = (-\infty, -z_{1-\alpha/2}) \cup (z_{1-\alpha/2}, \infty)$ |
| (b) | $\mu \leq \mu_0$ | $\mu > \mu_0$ | $K = (z_{1-\alpha}, \infty)$ |
| (c) | $\mu \geq \mu_0$ | $\mu < \mu_0$ | $K = (-\infty, z_\alpha = -z_{1-\alpha})$ |

For case (a) with $H_0$: $\mu = \mu_0$ and $H_1$: $\mu \neq \mu_0$, we are interested in extreme values of the test statistic on both tails: very small values and very large values of the test statistic give us evidence that $H_0$ is wrong (because the statistic is mainly driven by the difference of the sample mean and the test value $\mu_0$ for a fixed variance), see Fig. 10.1. In such a two-sided test, when the distribution of the test statistic is symmetric, we divide the critical region into two equal parts and assign each region of size $\alpha/2$ to the left and right tails of the distribution. For $\alpha = 0.05$, 2.5 % of the most extreme values towards the right end of the distribution and 2.5 % of the most extreme values towards the left end of the distribution give us enough evidence that $H_0$ is wrong and can be rejected and that $H_1$ is accepted. It is also clear why $\alpha$ is

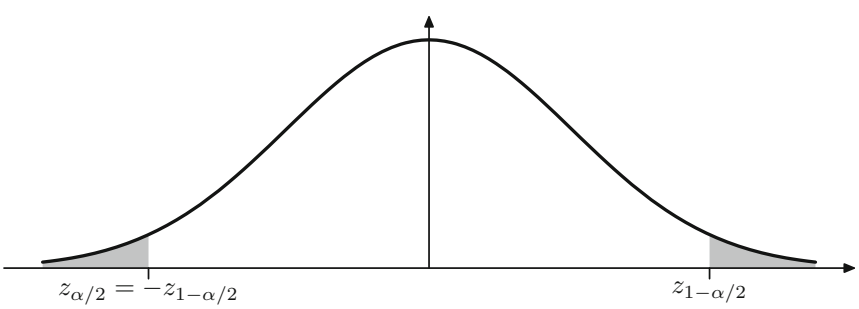


**Fig. 10.1** Critical region of a two-sided one-sample Gauss-test $H_0$: $\mu = \mu_0$ versus $H_1$: $\mu \neq \mu_0$. The critical region $K = (-\infty, -z_{1-\alpha/2}) \cup (z_{1-\alpha/2}, \infty)$ has probability mass $\alpha$ if $H_0$ is true*

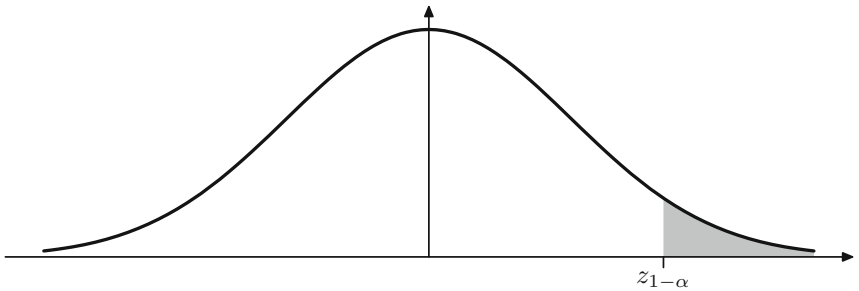


**Fig. 10.2** Critical region of a one-sided one-sample Gauss test $H_0$: $\mu \leq \mu_0$ versus $H_1$: $\mu > \mu_0$. The critical region $K = (z_{1-\alpha}, \infty)$ has probability mass $\alpha$ if $H_0$ is true*

the probability of a type I error: the most extreme values in the two tails together have 5 % probability and are just the probability that the test statistic falls into the critical region although $H_0$ is true. Also, these areas are those which have the least probability of occurring if $H_0$ is true. For $\alpha = 0.05$, we get $z_{1-\frac{\alpha}{2}} = 1.96$.

For case (b), only one direction is of interest. The critical region lies on the right tail of the distribution of the test statistic. A very large value of the test statistic has a low probability of occurrence if $H_0$ is true. An illustration is given in Fig. 10.2: for $\alpha = 0.05$, we get $z_{1-\alpha} = 1.64$ and any values greater than 1.64 are unlikely to be observed under $H_0$. Analogously, the critical region for case (c) is constructed. Here, the shaded area (critical region) is on the left-hand side. In this case, for $\alpha = 0.05$, we get $z_\alpha = -z_{1-\alpha} = -1.64$.

*6. Realization of the test statistic*: For an observed sample $x_1, x_2, \ldots, x_n$, the arithmetic mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

is used to calculate the realized (observed) test statistic $t(x) = T(x_1, x_2, \ldots, x_n)$ as

$$t(x) = \frac{\bar{x} - \mu_0}{\sigma_0} \sqrt{n}.$$

*7. Test decision*: If the realized test statistic from step 6 falls into the critical region, $H_0$ is rejected (and therefore, $H_1$ is statistically proven). Table 10.1 summarizes the test decisions depending on $t(x)$ and the quantiles defining the appropriate critical regions.

*Example 10.3.1* A bakery supplies loaves of bread to supermarkets. The stated selling weight (and therefore the required minimum expected weight) is $\mu = 2$ kg. However, not every package weighs exactly 2 kg because there is variability in the weights. It is therefore important to find out if the average weight of the loaves

**Table 10.1** Rules to make test decisions for the one-sample Gauss test (and the two-sample Gauss test, the one-sample approximate binomial test, and the two-sample approximate binomial test—which are all discussed later in this chapter)

| Case | $H_0$ | $H_1$ | Reject $H_0$ if |
|---|---|---|---|
| (a) | $\mu = \mu_0$ | $\mu \neq \mu_0$ | $\lvert t(x)\rvert > z_{1-\alpha/2}$ |
| (b) | $\mu \geq \mu_0$ | $\mu < \mu_0$ | $t(x) < z_\alpha$ |
| (c) | $\mu \leq \mu_0$ | $\mu > \mu_0$ | $t(x) > z_{1-\alpha}$ |

is significantly smaller than 2 kg. The weight $X$ (measured in kg) of the loaves is assumed to be normally distributed. We assume that the variance $\sigma_0^2 = 0.1^2$ is known from experience. A supermarket draws a sample of $n = 20$ loaves and weighs them. The average weight is calculated as $\bar{x} = 1.97$ kg. Since the supermarket wants to be sure that the weights are, on average, not lower than 2 kg, a one-sided hypothesis is appropriate and is formulated as $H_0$: $\mu \geq \mu_0 = 2$ kg versus $H_1$: $\mu < \mu_0 = 2$ kg. The significance level is specified as $\alpha = 0.05$, and therefore, $z_{1-\alpha} = 1.64$. The test statistic is calculated as

$$t(x) = \frac{\bar{x} - \mu_0}{\sigma_0}\sqrt{n} = \frac{1.97 - 2}{0.1}\sqrt{20} = -1.34.$$

The null hypothesis is not rejected, since $t(x) = -1.34 > -1.64 = -z_{1-0.05} = z_{0.05}$.

*Interpretation*: The sample average $\bar{x} = 1.97$ kg is below the target value of $\mu = 2$ kg. But there is not enough evidence to reject the hypothesis that the sample comes from a $N(2, 0.1^2)$-distributed population. The probability to observe a sample of size $n = 20$ with an average of at most 1.97 in a $N(2, 0.1^2)$-distributed population is greater than $\alpha = 0.05 = 5\,\%$. The difference between $\bar{x} = 1.97$ kg and the target value $\mu = 2$ kg is not statistically significant.

*Remark 10.3.1* The Gauss test assumes the variance to be known, which is often not the case in practice. The $t$-test (Sect. 10.3.2) assumes that the variance needs to be estimated. The $t$-test is therefore commonly employed when testing hypotheses about the mean. Its usage is outlined below. In *R*, the command `Gauss.test` from the library `compositions` offers an implementation of the Gauss test.

### 10.3.2 Test for the Mean When the Variance is Unknown (One-Sample *t*-Test)

If the variance $\sigma^2$ is unknown, hypotheses about the mean $\mu$ of a normal random variable $X \sim N(\mu, \sigma^2)$ can be tested in a similar way to the one-sample Gauss test. The difference is that the unknown variance is estimated from the sample. An

unbiased estimator of $\sigma^2$ is the sample variance

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

The test statistic is therefore

$$T(\mathbf{X}) = \frac{\bar{X} - \mu_0}{S_X}\sqrt{n},$$

which follows a $t$-distribution with $n-1$ degrees of freedom if $H_0$ is true, as we know from Theorem 8.3.2.

**Critical regions and test decisions**
Since $T(\mathbf{X})$ follows a $t$-distribution under $H_0$, the critical regions refer to the regions of the $t$-distribution which are unlikely to be observed under $H_0$:

| Case | $H_0$ | $H_1$ | Critical region $K$ |
|---|---|---|---|
| (a) | $\mu = \mu_0$ | $\mu \neq \mu_0$ | $K = (-\infty, -t_{n-1;1-\alpha/2}) \cup (t_{n-1;1-\alpha/2}, \infty)$ |
| (b) | $\mu \geq \mu_0$ | $\mu < \mu_0$ | $K = (-\infty, -t_{n-1;1-\alpha})$ |
| (c) | $\mu \leq \mu_0$ | $\mu > \mu_0$ | $K = (t_{n-1;1-\alpha}, \infty)$ |

The hypothesis $H_0$ is rejected if the realized test statistic, i.e.

$$t(x) = \frac{\bar{x} - \mu_0}{s_X}\sqrt{n},$$

falls into the critical region. The critical regions are based on the appropriate quantiles of the $t$-distribution with $(n-1)$ degrees of freedom, as outlined in Table 10.2.

*Example 10.3.2* We again consider Example 10.3.1. Now we assume that the variance of the loaves is unknown. Suppose a random sample of size $n = 20$ has an arithmetic mean of $\bar{x} = 1.9668$ and a sample variance of $s^2 = 0.0927^2$. We want to test whether this result contradicts the two-sided hypothesis $H_0$: $\mu = 2$, that is case (a). The significance level is fixed at $\alpha = 0.05$. For the realized test statistic $t(x)$, we calculate

$$t(x) = \frac{\bar{x} - \mu_0}{s_X}\sqrt{n} = \frac{1.9668 - 2}{0.0927}\sqrt{20} = -1.60.$$

**Table 10.2** Rules to make test decisions for the one-sample $t$-test (and the two-sample $t$-test, and the paired $t$-test, both explained below)

| Case | $H_0$ | $H_1$ | Reject $H_0$, if |
|---|---|---|---|
| (a) | $\mu = \mu_0$ | $\mu \neq \mu_0$ | $\lvert t(x)\rvert > t_{n-1;1-\alpha/2}$ |
| (b) | $\mu \geq \mu_0$ | $\mu < \mu_0$ | $t(x) < -t_{n-1;1-\alpha}$ |
| (c) | $\mu \leq \mu_0$ | $\mu > \mu_0$ | $t(x) > t_{n-1;1-\alpha}$ |

$H_0$ is not rejected since $|t| = 1.60 < 2.09 = t_{19;0.975}$, where the quantiles $\pm 2.09$ are defining the critical region (see Table C.2 or use *R*: `qt(0.975,19)`). The same results can be obtained in *R* using the `t.test()` function, see Example 10.3.3 for more details. Or, we can directly calculate the (two-sided) $p$-value as

```
2*(1-pt(abs(1.6), df=19))
```

This yields a $p$-value of 0.1260951 which is not smaller than $\alpha$, and therefore, $H_0$ is not rejected.

### 10.3.3 Comparing the Means of Two Independent Samples

In a two-sample problem, we may be interested in comparing the means of two *independent* samples. Assume that we have two samples of two normally distributed variables $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ of size $n_1$ and $n_2$, i.e. $X_1, X_2, \ldots, X_{n_1}$ are i.i.d. with the same distribution as $X$ and $Y_1, Y_2, \ldots, Y_{n_2}$ are i.i.d. with the same distribution as $Y$. We can specify the following hypotheses:

| Case | Null hypothesis | Alternative hypothesis | |
|---|---|---|---|
| (a) | $\mu_X = \mu_Y$ | $\mu_X \neq \mu_Y$ | Two-sided test problem |
| (b) | $\mu_X \geq \mu_Y$ | $\mu_X < \mu_Y$ | One-sided test problem |
| (c) | $\mu_X \leq \mu_Y$ | $\mu_X > \mu_Y$ | One-sided test problem |

We distinguish another three cases:

1. $\sigma_X^2$ and $\sigma_Y^2$ are known.
2. $\sigma_X^2$ and $\sigma_Y^2$ are unknown, but they are assumed to be equal, i.e. $\sigma_X^2 = \sigma_Y^2$.
3. Both $\sigma_X^2$ and $\sigma_Y^2$ are unknown and unequal ($\sigma_X^2 \neq \sigma_Y^2$).

**Case 1: The variances are known (two-sample Gauss test).**
If the null hypothesis $H_0$: $\mu_X = \mu_Y$ is true, then, using the usual rules for the normal distribution and the independence of the samples,

$$\bar{X} \sim N\left(\mu_X, \frac{\sigma_X^2}{n_1}\right),$$

$$\bar{Y} \sim N\left(\mu_Y, \frac{\sigma_Y^2}{n_2}\right),$$

and

$$(\bar{X} - \bar{Y}) \sim N\left(\mu_Y - \mu_Y, \frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}\right).$$

It follows that the test statistic

$$T(\mathbf{X}, \mathbf{Y}) = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}} \tag{10.1}$$

follows a standard normal distribution, $T(\mathbf{X}, \mathbf{Y}) \sim N(0, 1)$. The realized test statistic is

$$t(x, y) = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}}. \tag{10.2}$$

The test procedure is identical to the procedure of the one-sample Gauss test introduced in Sect. 10.3.1; that is, the test decision is based on Table 10.1.

**Case 2: The variances are unknown, but equal (two-sample $t$-test).**
We denote the unknown variance of both distributions as $\sigma^2$ (i.e. both the populations are assumed to have variance $\sigma^2$). We estimate $\sigma^2$ by using the pooled sample variance where each sample is assigned weights relative to the sample size:

$$S^2 = \frac{(n_1 - 1)S_X^2 + (n_2 - 1)S_Y^2}{n_1 + n_2 - 2}. \tag{10.3}$$

The test statistic

$$T(\mathbf{X}, \mathbf{Y}) = \frac{\bar{X} - \bar{Y}}{S} \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}} \tag{10.4}$$

with $S$ as in (10.3) follows a $t$-distribution with $n_1 + n_2 - 2$ degrees of freedom if $H_0$ is true. The realized test statistic is

$$t(x, y) = \frac{\bar{x} - \bar{y}}{s} \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}}. \tag{10.5}$$

The test procedure is identical to the procedure of the one-sample $t$-test; that is, the test decision is based on Table 10.2.

**Case 3: The variances are unknown and unequal (Welch test).**
We test $H_0$: $\mu_X = \mu_Y$ versus $H_1$: $\mu_X \neq \mu_Y$ given $\sigma_X^2 \neq \sigma_Y^2$ and both $\sigma_X^2$ and $\sigma_Y^2$ are unknown. This problem is also known as the Behrens–Fisher problem and is the most frequently used test when comparing two means in practice. The test statistic can be written as

$$T(\mathbf{X}, \mathbf{Y}) = \frac{\left|\bar{X} - \bar{Y}\right|}{\sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}}, \tag{10.6}$$

which is approximately $t$-distributed with $v$ degrees of freedom:

$$v = \left(\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}\right)^2 \Bigg/ \left(\frac{\left(s_x^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_y^2/n_2\right)^2}{n_2 - 1}\right) \tag{10.7}$$

where $s_x^2$ and $s_y^2$ are the estimated values of $S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ and $S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$, respectively. The test procedure, using the observed test statistic

$$t(x, y) = \frac{|\bar{x} - \bar{y}|}{\sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}}}, \tag{10.8}$$

is identical to the procedure of the one-sample $t$-test; that is, the test decision is based on Table 10.2 except that the degrees of freedom are not $n - 1$ but $v$. If $v$ is not an integer, it can be rounded off to an integer value.

*Example 10.3.3* A small bakery sells cookies in packages of 500 g. The cookies are handmade and the packaging is either done by the baker himself or his wife. Some customers conjecture that the wife is more generous than the baker. One customer does an experiment: he buys packages of cookies packed by the baker and his wife on 16 different days and weighs the packages. He gets the following two samples (one for the baker, one for his wife).

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Weight (wife) ($X$) | 512 | 530 | 498 | 540 | 521 | 528 | 505 | 523 |
| Weight (baker) ($Y$) | 499 | 500 | 510 | 495 | 515 | 503 | 490 | 511 |

We want to test whether the complaint of the customers is justified. Let us start with the following simple hypotheses:

$$H_0 : \mu_x = \mu_y \quad \text{versus} \quad H_1 : \mu_x \neq \mu_y,$$

i.e. we only want to test whether the weights are different, not that the wife is making heavier cookie packages. Since the variances are unknown, we assume that case 3 is the right choice. We calculate and obtain $\bar{x} = 519.625$, $\bar{y} = 502.875$, $s_X^2 = 192.268$, and $s_Y^2 = 73.554$. The test statistic is:

$$t(x, y) = \frac{|\bar{x} - \bar{y}|}{\sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}}} = \frac{|519.625 - 502.875|}{\sqrt{\frac{192.268}{8} + \frac{73.554}{8}}} \approx 2.91.$$

The degrees of freedom are:

$$v = \left(\frac{192.268}{8} + \frac{73.554}{8}\right)^2 \Big/ \left(\frac{(192.268/8)^2}{7} + \frac{(73.554/8)^2}{7}\right) \approx 11.67 \approx 12.$$

Since $|t(x)| = 2.91 > 2.18 = t_{12;0.975}$, it follows that $H_0$ is rejected. Therefore, $H_1$ is statistically significant. This means that the mean weight of the wife's packages is different from the mean weight of the baker's packages. Let us refine the hypothesis and try to find out whether the wife's packages have a higher mean weight. The hypotheses are now:

$$H_0 : \mu_x \leq \mu_y \quad \text{versus} \quad H_1 : \mu_x > \mu_y.$$

The test statistic remains the same but the critical region and the degrees of freedom change. Thus, $H_0$ is rejected if $t(x, y) > t_{v;1-\alpha}$. Using $t_{v;1-\alpha} = t_{12;0.95} \approx 1.78$ and $t(x, y) = 2.91$, it follows that the null hypothesis can be rejected. The mean weight of the wife's packages is greater than the mean weight of the baker's packages.

In *R*, we would have obtained the same result using the `t.test` command:

```
x <- c(512,530,498,540,521,528,505,523)
y <- c(499,500,510,495,515,503,490,511)
t.test(x,y,alternative='greater')
```

```
        Welch Two-Sample t-test

data:  x and y
t = 2.9058, df = 11.672, p-value = 0.006762
alternative hypothesis: true difference in means is greater
than 0...
```

Note that we have to specify the *alternative* hypothesis under the option `alternative`. The output shows us the test statistic (2.9058), the degrees of freedom (11.672), the alternative hypothesis—but not the decision rule. We know that $H_0$ is rejected if $t(x, y) > t_{v;1-\alpha}$, so the decision is easy in this case: we simply have to calculate $t_{12;0.95}$ using `qt(0.95,12)` in *R*. A simpler way to arrive at the same decision is to use the $p$-value. We know that $H_0$ is rejected if $p < \alpha$ which is the case in this example. It is also worthwhile mentioning that *R* displays the hypotheses slightly differently from ours: our alternative hypothesis is $\mu_x > \mu_y$ which is identical to the statement $\mu_x - \mu_y > 0$, as shown by *R*, see also Sect. 10.2.2.

If we specify `two.sided` as an alternative (which is the default), a confidence interval for the mean *difference* is also part of the output:

```
t.test(x,y,alternative='two.sided')
```

```
...
95 % confidence interval:
  4.151321 29.348679
```

It can be seen that the confidence interval of the difference does not cover the "0". Therefore, the null hypothesis is rejected. This is the duality property referred to earlier in this chapter: the test decision is the same, no matter whether one evaluates (i) the confidence interval, (ii) the test statistic, or (iii) the $p$-value.

Any kind of $t$-test can be calculated with the `t.test` command: for example, the two-sample $t$-test requires to specify the option `var.equal=TRUE` while the Welch test is calculated when the (default) option `var.equal=FALSE` is set. We can also conduct a one-sample $t$-test. Suppose we are interested in whether the mean

weight of the wife's packages of cookies is greater than 500 g; then, we could test the hypotheses:

$$H_0 : \mu_x \leq 500 \quad \text{versus} \quad H_1 : \mu_x > 500.$$

In *R*, we simply have to specify $\mu_0$:

```
t.test(x,mu=500,alternative='greater')
```

which gives us

```
        One-Sample t-test

data:  x
t = 4.0031, df = 7, p-value = 0.002585
alternative hypothesis: true mean is greater than 500
...
```

### 10.3.4 Test for Comparing the Means of Two Dependent Samples (Paired *t*-Test)

Suppose there are two dependent continuous random variables $X$ and $Y$ with $\mathrm{E}(X) = \mu_X$ and $\mathrm{E}(Y) = \mu_Y$. They could be dependent because we measure the same variable twice on the same subjects at different times. Typically, this is the case in pre–post experiments, for example when we measure the weight of a person before starting a special diet and after finishing the diet; or when evaluating household expenditures on electronic appliances in two consecutive years. We then say that the samples are *paired*, or dependent. Since the same variable is measured twice on the same subject, it makes sense to calculate a difference between the two respective values. Let $D = X - Y$ denote the random variable "difference of $X$ and $Y$". If $H_0$: $\mu_X = \mu_Y$ is true, then the expected difference is zero, and we get $\mathrm{E}(D) = \mu_D = 0$. This means testing $H_0 : \mu_X = \mu_Y$ is identical to testing $\mu_X - \mu_Y = \mu_D = 0$. We further assume that $D$ is normally distributed if $H_0$: $\mu_X = \mu_Y$ is true (or equivalently if $H_0$: $\mu_D = 0$ is true), i.e. $D \sim N(0, \sigma_D^2)$. For a random sample $(D_1, D_2, \ldots, D_n)$ of the differences, the test statistic

$$T(\mathbf{X}, \mathbf{Y}) = T(\mathbf{D}) = \frac{\bar{D}}{S_D}\sqrt{n} \tag{10.9}$$

is $t$-distributed with $n - 1$ degrees of freedom. The sample mean is $\bar{D} = \sum_{i=1}^{n} / D_i n$ and the sample variance is

$$S_D^2 = \frac{\sum_{i=1}^{n}(D_i - \bar{D})^2}{n - 1}$$

which is an estimator of $\sigma_D^2$. The realized test statistic is thus

$$t(d) = \frac{\bar{d}}{s_d}\sqrt{n} \tag{10.10}$$

where $\bar{d} = \sum_{i=1}^{n} d_i/n$ and $s_d^2 = \sum_{i=1}^{n}(d_i - \bar{d})^2/n - 1$.

The two-sided test $H_0$: $\mu_D = 0$ versus $H_1$: $\mu_D \neq 0$ and the one-sided tests $H_0$: $\mu_D \leq 0$ versus $H_1$: $\mu_D > 0$ or $H_0$: $\mu_D \geq 0$ versus $H_1$: $\mu_D < 0$ can be derived as in Sect. 10.3.2; that is, the test decision is based on Table 10.2. In fact, the paired $t$-test is a one-sample $t$-test on the differences of $X$ and $Y$.

*Example 10.3.4* In an experiment, $n = 10$ students have to solve different tasks before and after drinking a cup of coffee. Let $Y$ and $X$ denote the random variables "number of points before/after drinking a cup of coffee". Assume that a higher number of points means that the student is performing better. Since the test is repeated on the same students, we have a paired sample. The data is given in the following table:

| $i$ | $y_i$ (before) | $x_i$ (after) | $d_i = x_i - y_i$ | $(d_i - \bar{d})^2$ |
|---|---|---|---|---|
| 1 | 4 | 5 | 1 | 0 |
| 2 | 3 | 4 | 1 | 0 |
| 3 | 5 | 6 | 1 | 0 |
| 4 | 6 | 7 | 1 | 0 |
| 5 | 7 | 8 | 1 | 0 |
| 6 | 6 | 7 | 1 | 0 |
| 7 | 4 | 5 | 1 | 0 |
| 8 | 7 | 8 | 1 | 0 |
| 9 | 6 | 5 | −1 | 4 |
| 10 | 2 | 5 | 3 | 4 |
| Total | | | 10 | 8 |

We calculate

$$\bar{d} = 1 \quad \text{and} \quad s_d^2 = \frac{8}{9} = 0.943^2,$$

respectively. For the realized test statistic $t(d)$, using $\alpha = 0.05$, we get

$$t(d) = \frac{1}{0.943}\sqrt{10} = 3.35 > t_{9;0.95} = 1.83,$$

such that $H_0$: $\mu_X \leq \mu_Y$ is rejected and $H_1$: $\mu_X > \mu_Y$ is accepted. We can conclude (for this example) that drinking coffee significantly increased the problem-solving capacity of the students.

In *R*, we would have obtained the same results using the `t.test` function and specifying the option `paired=TRUE`:

```
yp <- c(4,3,5,6,7,6,4,7,6,2)
xp <- c(5,4,6,7,8,7,5,8,5,5)
t.test(xp,yp,paired=TRUE)
```

```
        Paired t-test

data:  xp and yp
t = 3.3541, df = 9, p-value = 0.008468
alternative hypothesis: true difference in means != 0
95 % confidence interval:
 0.325555 1.674445
sample estimates:
mean of the differences
                      1
```

We can make the test decision using the *R* output in three different ways:

(i) We compare the test statistic ($t = -3.35$) with the critical value (1.83, obtained via `qt(0.95,9)`).
(ii) We evaluate whether the $p$-value (0.008468) is smaller than the significance level $\alpha = 0.05$.
(iii) We evaluate whether the confidence interval for the mean difference covers "0" or not.

## 10.4 Parametric Tests for Probabilities

### 10.4.1 One-Sample Binomial Test for the Probability $p$

**Test construction and hypotheses.**
Let $X$ be a Bernoulli $B(1; p)$ random variable with the two possible outcomes 1 and 0, which indicate occurrence and non-occurrence of an event of interest $A$. The probability for $A$ in the population is $p$. From the sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ of independent $B(1; p)$-distributed random variables, we calculate the mean (relative frequency) as $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ which is an unbiased estimate of $p$. The following hypotheses may thus be of interest:

| Case | Null hypothesis | Alternative hypothesis | |
|---|---|---|---|
| (a) | $p = p_0$ | $p \neq p_0$ | Two-sided problem |
| (b) | $p \geq p_0$ | $p < p_0$ | One-sided problem |
| (c) | $p \leq p_0$ | $p > p_0$ | One-sided problem |

In the following, we describe two possible solutions, one exact approach and an approximate solution. The approximate solution is based on the approximation of the binomial distribution by the normal distribution, which is appropriate if $n$ is sufficiently large and the condition $np(1-p) \geq 9$ holds (i.e. $p$ is neither too small nor too large). First, we present the approximate solution and then the exact one.

**Test statistic and test decisions.**

(a) **Approximate binomial test**. We define the standardized test statistic as

$$T(\mathbf{X}) = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}. \tag{10.11}$$

It holds approximately that $T(\mathbf{X}) \sim N(0,1)$, given that the conditions that (i) $n$ is sufficiently large and (ii) $np(1-p) \geq 9$ are satisfied. The test can then be conducted along the lines of the Gauss test in Sect. 10.3.1; that is, the test decision is based on Table 10.1.

*Example 10.4.1* We return to Example 10.1.1. Let us assume that a representative sample of size $n = 2000$ has been drawn from the population of eligible voters, from which 700 (35 %) have voted for the party of interest $P$. The research hypothesis (which has to be stated as $H_1$) is that more than 30 % (i.e. $p_0 = 0.3$) of the eligible voters cast their votes for party $P$. The sample is in favour of $H_1$ because $\hat{p} = 35\,\%$, but to draw conclusions for the proportion of voters of party $P$ in the population, we have to conduct a binomial test. Since $n$ is large and $np(1-p) = 2000 \cdot 0.35 \cdot 0.65 = 455 \geq 9$, the assumptions for the use of the test statistic (10.11) are satisfied. We can write down the realized test statistic as

$$t(x) = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n} = \frac{0.35 - 0.3}{\sqrt{0.3(1-0.3)}}\sqrt{2000} = 4.8795.$$

Using $\alpha = 0.05$, it follows that $T(X) = 4.8795 > z_{1-\alpha} = 1.64$, and thus, the null hypothesis $H_0 : p \leq 0.3$ can be rejected. Therefore, $H_1 : p > 0.3$ is statistically significant; that is, the proportion of votes for party $P$ is greater than 30 %.

(b) The **exact binomial test** can be constructed using the knowledge that under $H_0$, $Y = \sum_{i=1}^{n} X_i$ (i.e. the number of successes) follows a binomial distribution. In fact, we can use $Y$ directly as the test statistic:

$$T(\mathbf{X}) = Y \sim B(n, p_0)\ .$$

The observed test statistic is $t(x) = \sum_i x_i$. For the two-sided case (a), the two critical numbers $c_l$ and $c_r$ ($c_l < c_r$) which define the critical region, have to be found such that

$$P_{H_0}(Y \leq c_l) \leq \frac{\alpha}{2} \quad \text{and} \quad P_{H_0}(Y \geq c_r) \leq \frac{\alpha}{2}.$$

The null hypothesis is rejected if the test statistic, i.e. $Y$, is greater than or equal to $c_r$ or less than or equal to $c_l$. For the one-sided case, a critical number $c$ has to be found such that

$$P_{H_0}(Y \leq c) \leq \alpha$$

for hypotheses of type (b) and

$$P_{H_0}(Y \geq c) \leq \alpha$$

for hypotheses of type (c). If $Y$ is less than the critical value $c$ (for case (b)) or greater than the critical value (for case (c)), the null hypothesis is rejected.

*Example 10.4.2* We consider again Example 10.1.1 where we looked at the population of eligible voters, from which 700 (35 %) have voted for the party of interest $P$. The observed test statistic is $t(x) = \sum_i x_i = 700$ and the alternative hypothesis is $H_1 : p \geq 0.3$, as in case (c). There are at least two ways in which we can obtain the results:

(i) *Long way*: We can calculate the test statistic and compare it to the critical region. To get the critical region, we search $c$ such that

$$P_{p=0.3}(Y \geq c) \leq 0.05\ ,$$

which equates to

$$P_{p=0.3}(Y < c) \geq 0.95$$

and can be calculated in *R* as:

```
qbinom(p=0.95, prob=0.3, size=2000)
[1] 634
```

Since $Y = 700 > c = 634$ we reject the null hypothesis. As in Example 10.4.1, we conclude that there is enough evidence that the proportion of votes for party $P$ is greater than 30 %.

(ii) *Short way*: The above result can be easily obtained in *R* using the `binom.test()` command. We need to specify the number of "successes" (here: 700), the number of "failures" ($2000 - 700 = 1300$), and the alternative hypothesis:

```
binom.test(c(700,1300),p=0.3,alternative='greater')
```

```
data:  c(700, 1300)
number of successes = 700, number of trials = 2000,
p-value = 8.395e-07
alternative hypothesis: true probability of success
is greater than 0.3
95 % confidence interval:
 0.332378 1.000000
probability of success
                  0.35
```

Both the $p$-value (which is smaller than $\alpha = 0.05$) and the confidence interval (for which we do not show the calculation) confirm the rejection of the null hypothesis.
Note that

```
binom.test(x=700,n=2000,p=0.3, alternative='greater')
```

returns the same result.

### 10.4.2 Two-Sample Binomial Test

**Test construction and hypotheses.**
We consider now the case of two independent i.i.d. samples from Bernoulli distributions with parameters $p_1$ and $p_2$.

$$\mathbf{X} = (X_1, X_2, \ldots, X_{n_1}), \quad X_i \sim B(1; p_1)$$
$$\mathbf{Y} = (Y_1, Y_2, \ldots, Y_{n_2}), \quad Y_i \sim B(1; p_2).$$

The sums

$$X = \sum_{i=1}^{n_1} X_i \sim B(n_1; p_1), \quad Y = \sum_{i=1}^{n_2} Y_i \sim B(n_2; p_2)$$

follow binomial distributions. One of the following hypotheses may be of interest:

| Case | Null hypothesis | Alternative hypothesis | |
|---|---|---|---|
| (a) | $p_1 = p_2$ | $p_1 \neq p_2$ | Two-sided problem |
| (b) | $p_1 \geq p_2$ | $p_1 < p_2$ | One-sided problem |
| (c) | $p_1 \leq p_2$ | $p_1 > p_2$ | One-sided problem |

Similar to the one-sample case, both exact and approximate tests exist. Here, we only present the approximate test. The **exact test of Fisher** is presented in Appendix C.5, p. 428. Let $n_1$ and $n_2$ denote the sample sizes. Then, $X/n_1$ and $Y/n_2$ are approximately normally distributed:

$$\frac{X}{n_1} \overset{approx.}{\sim} N\left(p_1, \frac{p_1(1-p_1)}{n_1}\right),$$
$$\frac{Y}{n_2} \overset{approx.}{\sim} N\left(p_2, \frac{p_2(1-p_2)}{n_2}\right).$$

Their difference $D$

$$D \overset{approx.}{\sim} N\left(0, p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

is normally distributed too under $H_0$ (given $p = p_1 = p_2$ holds). Since the probabilities $p_1$ and $p_2$ are identical under $H_0$, we can pool the two samples and estimate $p$ by

$$\hat{p} = \frac{X + Y}{n_1 + n_2}\,. \tag{10.12}$$

**Test statistic and test decision.**
The test statistic

$$T(\mathbf{X}, \mathbf{Y}) = \frac{D}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \tag{10.13}$$

follows a $N(0, 1)$-distribution if $n_1$ and $n_2$ are sufficiently large and $p$ is not near the boundaries 0 and 1 (one could use, for example, again the condition $np(1-p) > 9$ with $n = n_1 + n_2$). The realized test statistic can be calculated using the observed difference $\hat{d} = \hat{p}_1 - \hat{p}_2$. The test can be conducted for the one-sided and the two-sided case as the Gauss test introduced in Sect. 10.3.1; that is, the decision rules from Table 10.1 can be applied.

*Example 10.4.3* Two competing lotteries claim that every fourth lottery ticket wins. Suppose we want to test whether the probabilities of winning are different for the two lotteries, i.e. $H_0 : p_1 = p_2$ and $H_1 : p_1 \neq p_2$. We have the following data

| | $n$ | Winning | Not winning |
|---|---|---|---|
| Lottery A | 63 | 14 | 49 |
| Lottery B | 45 | 13 | 32 |

We can estimate the probabilities of a winning ticket for each lottery, as well as the respective difference, as

$$\hat{p}_A = \frac{14}{63}, \quad \hat{p}_B = \frac{13}{45}, \quad \hat{d} = \hat{p}_A - \hat{p}_B = -\frac{1}{15}.$$

Under $H_0$, an estimate for $p$ following (10.12) is

$$\hat{p} = \frac{14 + 13}{63 + 45} = \frac{27}{108} = 0.25.$$

The test statistic can be calculated as

$$t(x, y) = \frac{-\frac{1}{15}}{\sqrt{0.25(1-0.25)\left(\frac{1}{63} + \frac{1}{45}\right)}} = -0.79.$$

$H_0$ is not rejected since $|t(x, y)| = 0.79 < 1.96 = z_{1-0.05/2}$. Thus, there is no statistical evidence for different winning probabilities for the two lotteries. These hypotheses can be tested in *R* using the Test of Fisher, see Appendix C.5, p. 428, for more details.

## 10.5 Tests for Scale Parameters

There are various tests available to test hypotheses about scale parameters. Such tests are useful when one is interested in the dispersion of a variable, for example in quality control where the variability of a process may be of interest. One-sample tests of hypotheses for the variance of a normal distribution, e.g. hypotheses such as $H_0 : \sigma^2 = \sigma_0^2$, can be tested by the $\chi^2$-test for the variance, see Appendix C.5, p. 430. Two-sample problems can be addressed by the $F$-test (which is explained in Appendix C.5, p. 431); or by other tests such as the Levene test or Bartlett's test, which are also available in *R* (`leveneTest` in the package `car`, `bartlett` in the base distribution of *R*).

## 10.6 Wilcoxon–Mann–Whitney (WMW) U-Test

**Test construction and hypotheses.**

The WMW $U$-test is often proposed as an alternative to the $t$-test because it also focuses on location but not on the expected value $\mu$. It is a *nonparametric* test and useful in situations where skewed distributions are compared with each other. We consider two independent random samples $\mathbf{X} = (X_1, X_2, \ldots, X_{n_1})$ and $\mathbf{Y} = (Y_1, Y_2, , \ldots, Y_{n_2})$ from two populations with observed values $(x_1, x_2, \ldots, x_{n_1})$ and $(y_1, y_2, \ldots, y_{n_2})$, respectively. In this case, the null hypothesis $H_0$ considering the location can be formulated as

$$H_0 : P(X > Y) = P(Y > X) = \frac{1}{2} .$$

The null hypothesis can be interpreted in the following way: the probability that a randomly drawn observation from the first population has a value $x$ that is greater (or lower) than the value $y$ of a randomly drawn subject from the second population is $\frac{1}{2}$. The alternative hypothesis $H_1$ is then

$$H_1 : P(X > Y) \neq P(Y > X) .$$

This means we are comparing the entire distribution of two variables. If there is a location shift in the sense that one distribution is shifted left (or right) compared with the other distribution, the null hypothesis will be rejected because this shift can be seen as part of the alternative hypothesis $P(X > Y) \neq P(Y > X)$. In fact, under some assumptions, the hypothesis can even be interpreted as comparing two medians, and this is what is often done in practice.

**Observed test statistic.**

To construct the test statistic, it is necessary to merge $(x_1, x_2, \ldots, x_{n_1})$ and $(y_1, y_2, \ldots, y_{n_2})$ into one sorted sample, usually in ascending order, while keeping the information which value belongs to which sample. For now, we assume that all values of the two samples are distinct; that is, no ties are present. Then, each observation has

a rank between 1 and $(n_1 + n_2)$. Let $R_{1+}$ be the sum of ranks of the $x$-sample and let $R_{2+}$ be the sum of ranks of the $y$-sample. The test statistic is defined as $U$, where $U$ is the minimum of the two values $U_1, U_2$, $U = \min(U_1, U_2)$ with

$$U_1 = n_1 \cdot n_2 + \frac{n_1(n_1+1)}{2} - R_{1+}, \tag{10.14}$$

$$U_2 = n_1 \cdot n_2 + \frac{n_2(n_2+1)}{2} - R_{2+}. \tag{10.15}$$

**Test decision.**
$H_0$ is rejected if $U < u_{n_1,n_2;\alpha}$. Here, $u_{n_1,n_2;\alpha}$ is the critical value derived from the distribution of $U$ under the null hypothesis. The exact (complex) distribution can, for example, be derived computationally (in $R$). We are presenting an approximate solution together with its implementation in $R$.

Since $U_1 + U_2 = n_1 \cdot n_2$, it is sufficient to compute only $R_{i+}$ and $U = \min\{U_i, n_1 n_2 - U_i\}$ ($i = 1$ or $i = 2$ are chosen such that $R_{i+}$ is calculated for the sample with the lower sample size). For $n_1, n_2 \geq 8$, one can use the approximation

$$T(\mathbf{X}, \mathbf{Y}) = \frac{U - \frac{n_1 \cdot n_2}{2}}{\sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12}}} \overset{approx.}{\sim} N(0,1) \tag{10.16}$$

as the test statistic. For two-sided hypotheses, $H_0$ is rejected if $|t(x, y)| > z_{1-\alpha/2}$; for one-sided hypotheses $H_0$ is rejected if $|t(x, y)| > z_{1-\alpha}$. In the case of ties, the denominator of the test statistic in (10.16) can be modified as

$$T(\mathbf{X}, \mathbf{Y}) = \frac{U - \frac{n_1 \cdot n_2}{2}}{\sqrt{\left[\frac{n_1 \cdot n_2}{n(n-1)}\right]\left[\frac{n^3 - n}{12} - \sum_{j=1}^{G} \frac{t_j^3 - t_j}{12}\right]}} \overset{approx.}{\sim} N(0,1),$$

where $G$ is the number of different (groups of) ties and $t_j$ denotes the number of tied ranks in tie group $j$.

*Example 10.6.1* In a study, the reaction times (in seconds) to a stimulus were measured for two groups. One group drank a strong coffee before the stimulus and the other group drank only the same amount of water. There were 9 study participants in the coffee group and 10 participants in the water group. The following reaction times were recorded:

| Reaction time | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Coffee group (C) | 3.7 | 4.9 | 5.2 | 6.3 | 7.4 | 4.4 | 5.3 | 1.7 | 2.9 | |
| Water group (W) | 4.5 | 5.1 | 6.2 | 7.3 | 8.7 | 4.2 | 3.3 | 8.9 | 2.6 | 4.8 |

We test with the $U$-test whether there is a location difference between the two groups. First, the ranks of the combined sample are calculated as:

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Value (C) | 3.7 | 4.9 | 5.2 | 6.3 | 7.4 | 4.4 | 5.3 | 1.7 | 2.9 | | |
| Rank (C) | 5 | 10 | 12 | 15 | 17 | 7 | 13 | 1 | 3 | | 83 |
| Value (W) | 4.5 | 5.1 | 6.2 | 7.3 | 8.7 | 4.2 | 3.3 | 8.9 | 2.6 | 4.8 | |
| Rank (W) | 8 | 11 | 14 | 16 | 18 | 6 | 4 | 19 | 2 | 9 | 107 |

With $R_{C+} = 83$ and $R_{W+} = 107$, we get

$$U_1 = n_1 \cdot n_2 + \frac{n_1(n_1+1)}{2} - R_{C+} = 9 \cdot 10 + \frac{9 \cdot 10}{2} - 83 = 52,$$
$$U_2 = n_1 \cdot n_2 + \frac{n_2(n_2+1)}{2} - R_{W+} = 9 \cdot 10 + \frac{10 \cdot 11}{2} - 107 = 38.$$

With $n_1, n_2 \geq 8$ and $U = U_2 = 38$,

$$t(x,y) = \frac{U - \frac{n_1 \cdot n_2}{2}}{\sqrt{\frac{n_1 \cdot n_2 \cdot (n_1+n_2+1)}{12}}} = \frac{38 - \frac{9 \cdot 10}{2}}{\sqrt{\frac{9 \cdot 10 \cdot (9+10+1)}{12}}} \approx -0.572.$$

Since $|t(x,y)| = 0.572 < z_{1-\alpha/2} = 1.96$, the null hypothesis cannot be rejected; that is, there is no statistical evidence that the two groups have different reaction times.

In *R*, one can use the `wilcox.test` command to obtain the results:

```
coffee <- c(3.7, 4.9, 5.2, 6.3, ..., 2.9)
water <- c(4.5, 5.1, 6.2, ..., 4.8)
wilcox.test(coffee, water)
```

The output is

```
    Wilcoxon rank sum test

data:  coffee.sample and water.sample
W = 38, p-value = 0.6038
alternative hypothesis: true location shift is not equal to 0
```

We can see that the null hypothesis is not rejected because $p = 0.6038 > \alpha = 0.05$. The displayed test statistic is $W$ which equates to our statistic $U_2$. The alternative hypothesis in *R* is framed as location shift, an interpretation which has already been given earlier in the chapter. Note that the test also suggests that the medians of the two samples are not statistically different.

## 10.7 $\chi^2$-Goodness-of-Fit Test

**Test construction.**
The $\chi^2$-goodness-of-fit test is one of the most popular tests for testing the goodness of fit of the observed data to a distribution. The construction principle is very general and can be used for variables of any scale. The test statistic is derived such that the *observed* absolute frequencies are compared with the *expected* absolute frequencies *under the null hypothesis* $H_0$.

*Example 10.7.1* Consider an experiment where a die is rolled $n = 60$ times. Under the null hypothesis $H_0$, we assume that the die is fair, i.e. $p_i = \frac{1}{6}, i = 1, 2, \ldots, 6$, where $p_i = P(X = i)$. We could have also said that $H_0$ is the hypothesis that the rolls are following a discrete uniform distribution. Thus, the expected absolute frequencies under $H_0$ are $np_i = 60 \cdot \frac{1}{6} = 10$, while the observed frequencies in the sample are $N_i, i = 1, 2, \ldots, 6$. The $N_i$ generally deviate from $np_i$. The $\chi^2$-statistic is based on the squared differences, $\sum_{i=1}^{6}(N_i - np_i)^2$, and becomes large as the differences between the observed and the expected frequencies become larger. The $\chi^2$-test statistic is a modification of this sum by scaling each squared difference by the expected frequencies, $np_i$, and is explained below.

With a nominal variable, we can proceed as in Example 10.7.1. If the scale of the variable is ordinal or continuous, the number of different values can be large. Note that in the most extreme case, we can have as many different values as observations ($n$), leading to $N_i = 1$ for all $i = 1, 2, \ldots, n$. Then, it is necessary to group the data into $k$ intervals before applying the $\chi^2$-test. The reason is that the general theory of the $\chi^2$-test assumes that the number $k$ (which was 6 in Example 10.7.1 above) is fixed and does not grow with the number of observations $n$; that is, the theory says that the $\chi^2$-test only works properly if $k$ is fixed and $n$ is large. For this reason, we group the sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ into $k$ classes as shown in Sect. 2.1.

| Class | 1 | 2 | $\cdots$ | $k$ | Total |
|---|---|---|---|---|---|
| Number of observations | $n_1$ | $n_2$ | $\cdots$ | $n_k$ | $n$ |

The choice of the class intervals is somewhat arbitrary. As a rule of thumb $np_i > 5$ should hold for most class intervals. The general hypotheses can be formulated in the form of distribution functions:

$$H_0 : F(x) = F_0(x) \text{ versus } H_1 : F(x) \neq F_0(x).$$

**Test statistic.**
The test statistic is defined as

$$T(\mathbf{X}) = t(x) = \chi^2 = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i}. \tag{10.17}$$

Here,

- $N_i$ $(i = 1, 2, \ldots, k)$ are the absolute frequencies of observations of the sample $\mathbf{X}$ in class $i$, $N_i$ is a random variable with realization $n_i$ in the observed sample;
- $p_i$ $(i = 1, 2, \ldots, k)$ are calculated from the distribution under $H_0$, $F_0(x)$, and are the (hypothetical) probabilities that an observation of $X$ falls in class $i$;
- $np_i$ are the expected absolute frequencies in class $i$ under $H_0$.

**Test decision.**
For a significance level $\alpha$, $H_0$ is rejected if $t(x)$ is greater than the $(1-\alpha)$-quantile of the $\chi^2$-distribution with $k-1-r$ degrees of freedom, i.e. if

$$t(x) = \chi^2 > c_{k-1-r,1-\alpha}.$$

Note that $r$ is the number of parameters of $F_0(x)$, if these parameters are estimated from the sample. The $\chi^2$-test statistic is only asymptotically $\chi^2$-distributed under $H_0$.

*Example 10.7.2* Let $F_0(x)$ be the distribution function of the test distribution. If one specifies a normal distribution such as $F_0(x) = N(3, 10)$, or a discrete uniform distribution with $p_i = 0.25$ $(i = 1, 2, 3, 4)$, then $r = 0$, since no parameters have to be estimated from the data. Otherwise, if we simply want to test whether the data is generated from a normal distribution $N(\mu, \sigma^2)$ or the data follows a normal distribution $N(\mu, \sigma^2)$, then $\mu$ and $\sigma^2$ may be estimated from the sample by $\bar{x}$ and $s^2$. Then, $r = 2$ and the number of degrees of freedom is reduced.

*Example 10.7.3* Gregor Mendel (1822–1884) conducted crossing experiments with pea plants of different shape and colour. Let us look at the outcome of a pea crossing experiment with the following results:

| Crossing result | Round Yellow | Round Green | Edged Yellow | Edged Green |
|---|---|---|---|---|
| Observations | 315 | 108 | 101 | 32 |

Mendel had the hypothesis that the four different types occur in proportions of 9:3:3:1, that is

$$p_1 = \frac{9}{16},\ p_2 = \frac{3}{16},\ p_3 = \frac{3}{16},\ p_4 = \frac{1}{16}.$$

The hypotheses are

$$H_0 : P(X = i) = p_i \quad \text{versus} \quad H_1 : P(X = i) \neq p_i, \qquad i = 1, 2, 3, 4.$$

With $n = 556$ observations, the test statistic can be calculated from the following observed and expected frequencies:

| $i$ | $N_i$ | $p_i$ | $np_i$ |
|---|---|---|---|
| 1 | 315 | $\frac{9}{16}$ | 312.75 |
| 2 | 108 | $\frac{3}{16}$ | 104.25 |
| 3 | 101 | $\frac{3}{16}$ | 104.25 |
| 4 | 32 | $\frac{1}{16}$ | 34.75 |

The $\chi^2$-test statistic is calculated as

$$t(x) = \chi^2 = \frac{(315 - 312.75)^2}{312.75} + \cdots + \frac{(32 - 34.75)^2}{34.75} = 0.47.$$

Since $\chi^2 = 0.47 < 7.815 = \chi^2_{0.95,3} = c_{0.95,3}$, the null hypothesis is not rejected. Statistically, there is no evidence that Mendel was wrong with his 9:3:3:1 assumption. In *R*, the test can be conducted by applying the `chisq.test` command:

```
chisq.test(c(315, 108, 101, 32),
p=c(9/16,3/16,3/16,1/16))
qchisq(df=3, p=0.95)
```

which leads to the following output

```
        Chi-squared test for given probabilities

data:  c(315, 108, 101, 32)
X-squared = 0.47, df = 3, p-value = 0.9254
```

and the critical value is

```
[1] 7.814728
```

*Remark 10.7.1* In this example, the data was already summarized in a frequency table. For raw data, the `table` command can be used to preprocess the data, i.e. we can use `chisq.test(table(var1,var2))`.

Another popular goodness-of-fit test is the test of Kolmogorov–Smirnov. There are two different versions of this test, one for the one-sample scenario and one for the two-sample scenario. The null hypothesis for the latter is that the two independent samples come from the same distribution. In *R*, the command `ks.test()` can be used to perform Kolmogorov–Smirnov tests.

## 10.8 $\chi^2$-Independence Test and Other $\chi^2$-Tests

In Chap. 4, we introduced different methods to describe the association between two variables. Several association measures are possibly suitable if the variables are categorical, for example Cramer's $V$, Goodman's and Kruskal's $\gamma$, Spearman's rank correlation coefficient, and the odds ratio. If we are not interested in the strength of association but rather in finding out whether there is an association at all, one can use the $\chi^2$-independence test.

**Test construction.**
In the following we assume that we observe a sample from a bivariate discrete distribution of two variables $X$ and $Y$ which can be summarized in a contingency table with absolute frequencies $n_{ij}$, $(i = 1, 2, \ldots, I;\ j = 1, 2 \ldots, J)$:

| | | $Y$ | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | $\cdots$ | $J$ | |
| $X$ | 1 | $n_{11}$ | $n_{12}$ | $\cdots$ | $n_{1J}$ | $n_{1+}$ |
| | 2 | $n_{21}$ | $n_{22}$ | $\cdots$ | $n_{2J}$ | $n_{2+}$ |
| | $\vdots$ | $\vdots$ | | $\vdots$ | $\vdots$ | $\vdots$ |
| | $I$ | $n_{I1}$ | $n_{I2}$ | $\cdots$ | $n_{IJ}$ | $n_{I+}$ |
| | | $n_{+1}$ | $n_{+2}$ | $\cdots$ | $n_{+J}$ | $n$ |

Remember that

$n_{i+}$ is the $i$th row sum,
$n_{+j}$ is the $j$th column sum, and
$n$ is the total number of observations.

The hypotheses are $H_0$: $X$ and $Y$ are independent versus $H_1$ : $X$ and $Y$ are not independent. If $X$ and $Y$ are independent, then the expected frequencies $m_{ij}$ are

$$\hat{m}_{ij} = n\hat{\pi}_{ij} = \frac{n_{i+}n_{+j}}{n}. \tag{10.18}$$

**Test statistic.**
Pearson's $\chi^2$-test statistic was introduced in Chap. 4, Eq. (4.6). It is

$$T(\mathbf{X}, \mathbf{Y}) = \chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - m_{ij})^2}{m_{ij}},$$

where $m_{ij} = n\pi_{ij} = n\pi_{i+}\pi_{+j}$ (expected absolute cell frequencies under $H_0$). Strictly speaking, $m_{ij}$ are the true, unknown expected frequencies under $H_0$ and are estimated by $\hat{m}_{ij} = n\pi_{i+}\pi_{+j}$, such that the realized test statistic equates to

$$t(x, y) = \chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}. \tag{10.19}$$

**Test decision.**
The number of degrees of freedom under $H_0$ is $(I-1)(J-1)$, where $I-1$ are the parameters which have to be estimated for the marginal distribution of $X$, and $J-1$ are the number of parameters for the marginal distribution of $Y$. The test decision is:

$$\text{Reject } H_0, \text{ if } t(x, y) = \chi^2 > c_{(I-1)(J-1);1-\alpha}.$$

Note that the alternative hypothesis $H_1$ is very general. If $H_0$ is rejected, nothing can be said about the structure of the dependence of $X$ and $Y$ from the $\chi^2$-value itself.

*Example 10.8.1* Consider the following contingency table. Here, $X$ describes the educational level (1: primary, 2: secondary, 3: tertiary) and $Y$ the preference for a specific political party (1: Party A, 2: Party B, 3: Party C). Our null hypothesis is that the two variables are independent, and we want to show the alternative hypothesis which says that there is a relationship between them.

| | | $Y$ | | | Total |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | |
| $X$ | 1 | 100 | 200 | 300 | 600 |
| | 2 | 100 | 100 | 100 | 300 |
| | 3 | 80 | 10 | 10 | 100 |
| | Total | 280 | 310 | 410 | 1000 |

For the (estimated) expected frequencies $\hat{m}_{ij} = \frac{n_{i+}n_{+j}}{n}$, we get

| | $Y$ | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| $X$ 1 | 168 | 186 | 246 |
| 2 | 84 | 93 | 123 |
| 3 | 28 | 31 | 41 |

For example: $\hat{m}_{11} = 600 \cdot 280/1000 = 168$. The test statistic is

$$\begin{aligned} t(x, y) &= \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}} \\ &= \frac{(100-168)^2}{168} + \cdots + \frac{(10-41)^2}{41} \approx 182.54. \end{aligned}$$

Since $\chi^2_{4;0.95} = 9.49 < t(x, y) = 182.54$, $H_0$ is rejected.

In *R*, either the summarized data (as shown below) can be used to calculate the test statistic or the raw data (summarized in a contingency table via `table(var1,var2)`):

```
ct <- matrix(nrow=3,ncol=3,byrow=T,
data=c(100,200,300,100,100,100,80,10,10))
chisq.test(ct)
qchisq(df=(3-1)*(3-1), p=0.95)
```

The output is

```
    Pearson's Chi-squared test

data:  contingency.table
X-squared = 182.5428, df = 4, p-value < 2.2e-16
```

with the critical value

```
[1] 9.487729
```

which confirms our earlier manual calculations. The $p$-value is smaller than $\alpha = 0.05$ which further confirms that the null hypothesis has to be rejected.

For a binary outcome, the $\chi^2$-test of independence can be formulated as a test for the null hypothesis that the proportions of the binary variable are equal in several ($\geq 2$) groups, i.e. for a $K \times 2$ (or $2 \times K$) table. This test is called the **$\chi^2$-test of homogeneity**.

*Example 10.8.2* Consider two variables $X$ and $Y$, where $X$ is describing the rating of a coffee brand with the categories "bad taste" and "good taste" and $Y$ denotes three age subgroups, e.g. "18–25", "25–35", and "35–45". The observed data is

| | | $Y$ | | | |
|---|---|---|---|---|---|
| | | 18–25 | 25–35 | 35–45 | Total |
| $X$ | Bad | 10 | 30 | 65 | 105 |
| | Good | 90 | 70 | 35 | 195 |
| | Total | 100 | 100 | 100 | 300 |

Assume $H_0$ is the hypothesis that the probabilities $P(X = \text{'good'}|Y = \text{'18–25'})$, $P(X = \text{'good'}|Y = \text{'25–35'})$, and $P(X = \text{'good'}|Y = \text{'35–45'})$ are all equal. Then, we can use the function either `prop.test` or `chisq.test` in *R* to test this hypothesis:

```
prop.test(x=rbind(c(10,30,65), c(90,70,35) ))
chisq.test(x=rbind(c(10,30,65), c(90,70,35) ))
```

This produces the following outputs:

```
    3-sample test for equality of proportions

data:  cbind(c(10, 30, 65), c(90, 70, 35))
X-squared = 68.1319, df = 2, p-value = 1.605e-15
alternative hypothesis: two.sided
sample estimates:
prop 1 prop 2 prop 3
  0.10   0.30   0.65
```

and

```
    Pearson's Chi-squared test

data:  cbind(c(10, 30, 65), c(90, 70, 35))
X-squared = 68.1319, df = 2, p-value = 1.605e-15
```

The results (test statistic, $p$-value) are identical and $H_0$ is rejected. Note that `prop.test` strictly expects a $K \times 2$ table (i.e. exactly 2 columns).

*Remark 10.8.1* For $2 \times 2$-tables with small sample sizes and therefore small cell frequencies, it is recommended to use the exact test of Fisher as described in Appendix C.5.

*Remark 10.8.2* The test described in Example 10.8.2 is a special case (since one variable is binary) of the general $\chi^2$-test of homogeneity. The $\chi^2$-test of homogeneity is valid for any $K \times C$ table, where $K$ is the number of subgroups of a variable $Y$ and $C$ is the number of values of the outcome $X$ of interest. The null hypothesis $H_0$ assumes that the conditional distributions of $X$ given $Y$ are identical in all subgroups, i.e.

$$P(X = x_c|Y = y_k) = P(X = x_c|Y = y_{k'})$$

forall $c = 1, 2, \ldots, C$; $k, k' = 1, 2, \ldots, K, k \neq k'$. Again, the usual $\chi^2$-test statistic can be used.

## 10.9 Key Points and Further Issues

**Note:**

- ✓ A graphical summary on when to use the tests introduced in this chapter is given in Appendices D.2 and D.3.
- ✓ To arrive at a test decision, i.e. accept $H_0$ or reject it, it does not matter whether one compares the test statistic to the critical region, one uses the $p$-value obtained from statistical software, or one evaluates the appropriate confidence interval. However, it is important not to misinterpret the $p$-value (see Sect. 10.2.6) and to choose the correct confidence interval.
- ✓ There is a difference between relevance and significance. A test might be significant, but the point estimate of the quantity of interest may not be relevant from a substantive point of view. Similarly, a test might not be significant, but the point and interval estimates may still yield relevant conclusions.
- ✓ The test statistic of the $t$-test (one-sample, two-sample, paired) is *asymptotically* normally distributed. This means that for relatively large $n$ (as a rule of thumb $>30$ per group) the sample does not need to come from a normal distribution. However, the application of the $t$-test makes sense only when the expectation $\mu$ can be interpreted meaningfully; this may not be the case for skewed distributions or distributions with outliers.

## 10.10 Exercises

*Exercise 10.1* Two people, A and B, are suspects for having committed a crime together. Both of them are interrogated in separate rooms. The jail sentence depends on who confesses to have committed the crime, and who does not:

| | B does not confess | B does confess |
|---|---|---|
| A does not confess | Each serves 1 year | A: 3 years; B: goes free |
| A does confess | A: goes free; B: 3 years | Each serves 2 years |

A has two hypotheses:

$$H_0: \text{ B does not confess} \quad \text{versus} \quad H_1: \text{ B does confess.}$$

Given the possible sentences he decides to not confess if $H_0$ is true and to confess otherwise. Explain the concepts of type I error and type II error for this situation. Comment on the consequences if these errors are made.

*Exercise 10.2* A producer of chocolate bars hypothesizes that his production does not adhere to the weight standard of 100 g. As a measure of quality control, he weighs 15 bars and obtains the following results in grams:

96.40, 97.64, 98.48, 97.67, 100.11, 95.29, 99.80, 98.80, 100.53, 99.41, 97.64, 101.11, 93.43, 96.99, 97.92

It is assumed that the production process is standardized in the sense that the variation is controlled to be $\sigma = 2$.

(a) What are the hypotheses regarding the expected weight $\mu$ for a two-sided test?
(b) Which test should be used to test these hypotheses?
(c) Conduct the test that was suggested to be used in (b). Use $\alpha = 0.05$.
(d) The producer wants to show that the expected weight is smaller than 100 g. What are the appropriate hypotheses to use?
(e) Conduct the test for the hypothesis in (d). Again use $\alpha = 0.05$.

*Exercise 10.3* Christian decides to purchase the new CD by Bruce Springsteen. His first thought is to buy it online, via an online auction. He discovers that he can also buy the CD immediately, without bidding at an auction, from the same online store. He also looks at the price at an internet book store which was recommended to him by a friend. He notes down the following prices (in €):

**Internet book store** 16.95

**Online store, no auction** 18.19, 16.98, 19.97, 16.98, 18.19, 15.99, 13.79, 15.90, 15.90, 15.90, 15.90, 15.90, 19.97, 17.72

**Online store, auction** 10.50, 12.00, 9.54, 10.55, 11.99, 9.30, 10.59, 10.50, 10.01, 11.89, 11.03, 9.52, 15.49, 11.02

(a) Calculate and interpret the arithmetic mean, variance, standard deviation, and coefficient of variation for the online store, both for the auction and non-auction offers.
(b) Test the hypothesis that the mean price at the online store (no auction) is unequal to €16.95 ($\alpha = 0.05$).
(c) Calculate a confidence interval for the mean price at the online store (no auction) and interpret your findings in the light of the hypothesis in (b).
(d) Test the hypothesis that the mean price at the online store (auction) is less than €16.95 ($\alpha = 0.05$).
(e) Test the hypothesis that the mean non-auction price is higher than the mean auction price. Assume that (i) the variances are equal in both samples and (ii) the variances are unequal ($\alpha = 0.05$).
(f) Test the hypothesis that the variance of the non-auction price is unequal to the variance of the auction price ($\alpha = 0.05$).

(g) Use the $U$-test to compare the location of the auction and non-auction prices. Compare the results with those of (e).
(h) Calculate the results of (a)–(g) with $R$.

*Exercise 10.4* Ten of Leonard's best friends try a new diet: the "Banting" diet. Each of them weighs him/herself before and after the diet. The data is as follows:

| Person ($i$) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Before diet ($x_i$) | 80 | 95 | 70 | 82 | 71 | 70 | 120 | 105 | 111 | 90 |
| After diet ($y_i$) | 78 | 94 | 69 | 83 | 65 | 69 | 118 | 103 | 112 | 88 |

Choose a test and a confidence interval to test whether there is a difference between the mean weight before and after the diet ($\alpha = 0.05$).

*Exercise 10.5* A company producing clothing often finds deficient T-shirts among its production.

(a) The company's controlling department decides that the production is no longer profitable when there are more than 10 % deficient shirts. A sample of 230 shirts yields 30 shirts which contain deficiencies. Use the approximate binomial test to decide whether the T-shirt production is profitable or not ($\alpha = 0.05$).
(b) Test the same hypothesis as in (a) using the exact binomial test. You can use $R$ to determine the quantiles needed for the calculation.
(c) The company is offered a new cutting machine. To test whether the change of machine helps to improve the production quality, 115 sample shirts are evaluated, 7 of which have deficiencies. Use the two-sample binomial test to decide whether the new machine yields improvement or not ($\alpha = 0.05$).
(d) Test the same hypothesis as in (c) using the test of Fisher in $R$.

*Exercise 10.6* Two friends play a computer game and each of them repeats the same level 10 times. The scores obtained are:

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Player 1 | 91 | 101 | 112 | 99 | 108 | 88 | 99 | 105 | 111 | 104 |
| Player 2 | 261 | 47 | 40 | 29 | 64 | 6 | 87 | 47 | 98 | 351 |

(a) Player 2 insists that he is the better player and suggests to compare their mean performance. Use an appropriate test ($\alpha = 0.05$) to test this hypothesis.
(b) Player 1 insists that he is the better player. He proposes to not focus on the mean and to use the $U$-test for comparison. What are the advantages and disadvantages of using this test compared with (a)? What are the results ($\alpha = 0.05$)?

*Exercise 10.7* Otto loves gummy bears and buys 10 packets at a factory store. He opens all packets and sorts them by their colour. He counts 222 white gummy bears, 279 red gummy bears, 251 orange gummy bears, 232 yellow gummy bears, and 266 green ones. He is disappointed since white (pineapple flavour) is his favourite flavour. He hypothesizes that the producer of the bears does not uniformly distribute the bears into the packets. Choose an appropriate test to find out whether Otto's speculation could be true.

*Exercise 10.8* We consider Exercise 4.4 where we evaluated which of the passengers from the *Titanic* were rescued. The data was summarized as follows:

| | 1. Class | 2. Class | 3. Class | Staff | Total |
|---|---|---|---|---|---|
| Rescued | 202 | 125 | 180 | 211 | 718 |
| Not rescued | 135 | 160 | 541 | 674 | 1510 |

(a) The hypothesis derived from the descriptive analysis was that travel class and rescue status are not independent. Test this hypothesis.

(b) Interpret the following *R* output:

```
4-sample test for equality of proportions
data:  titanic
X-squared = 182.06, df = 3, p-value < 2.2e-16
alternative hypothesis: two.sided
sample estimates:
   prop 1    prop 2    prop 3    prop 4
0.5994065 0.4385965 0.2496533 0.2384181
```

(c) Summarize the data in a $2\times2$ table: passengers from the first and second class should be grouped together, and third class passengers and staff should be grouped together as well. Is the probability of being rescued higher in the first and second class? Provide an answer using the following three tests: exact test of Fisher, $\chi^2$-independence test, and $\chi^2$-homogeneity test. You can use *R* to conduct the test of Fisher.

*Exercise 10.9* We are interested in understanding how well the $t$-test can detect differences with respect to the mean. We use *R* to draw 3 samples each of 20 observations from three different normal distributions: $X \sim N(5, 2^2)$, $Y_1 \sim N(4, 2^2)$, and $Y_2 \sim N(3.5, 2^2)$. The summary statistics of this experiment are as follows:

- $\bar{x} = 4.97$, $s_x^2 = 2.94$,
- $\bar{y}_1 = 4.55$, $s_{y_1}^2 = 2.46$,
- $\bar{y}_2 = 3.27$, $s_{y_2}^2 = 3.44$.

(a) Use the $t$-test to compare the means of $X$ and $Y_1$.
(b) Use the $t$-test to compare the means of $X$ and $Y_2$.
(c) Interpret the results from (a) and (b).

*Exercise 10.10* Access the theatre data described in Appendix A.4. The data summarizes a survey conducted on visitors of a local Swiss theatre in terms of age, sex, annual income, general expenditure on cultural activities, expenditure on theatre visits, and the estimated expenditure on theatre visits in the year before the survey was done.

(a) Compare the mean expenditure on cultural activities for men and women using the Welch test ($\alpha = 0.05$).
(b) Would the conclusions change if the two-sample $t$-test or the $U$-test were used for comparison?
(c) Test the hypothesis that women spend on average more money on theatre visits than men ($\alpha = 0.05$).
(d) Compare the mean expenditure on theatre visits in the year of the survey and the preceding year ($\alpha = 0.05$).

*Exercise 10.11* Use *R* to read in and analyse the pizza data described in Appendix A.4 (assume $\alpha = 0.05$).

(a) The manager's aim is to deliver pizzas in less than 30 min and with a temperature of greater than 65 °C. Use an appropriate test to evaluate whether these aims have been reached on average.
(b) If it takes longer than 40 min to deliver the pizza, then the customers are promised a free bottle of wine. This offer is only profitable if less than 15 % of deliveries are too late. Test the hypothesis $p < 0.15$.
(c) The manager wonders whether there is any relationship between the operator taking the phone call and the pizza temperature. Assume that a hot pizza is defined to be one with a temperature greater 65 °C. Use the test of Fisher, the $\chi^2$-independence test, and the $\chi^2$-test of homogeneity to test his hypothesis.
(d) Each branch employs the same number of staff. It would thus be desirable if each branch receives the same number of orders. Use an appropriate test to investigate this hypothesis.
(e) Is the proportion of calls taken by each operator the same in each branch?
(f) Test whether there is a relationship between drivers and branches.

*Exercise 10.12* The authors of this book went to visit historical sites in India. None of them has a particularly strong interest in photography, and they speculated that each of them would take about the same number of pictures on their trip. After returning home, they counted 110, 118, and 105 pictures, respectively. Use an appropriate test to find out whether their speculation was correct ($\alpha = 0.01$).

→ Solutions to all exercises in this chapter can be found on p. 393

*Source* Toutenburg, H., Heumann, C., *Induktive Statistik*, 4th edition, 2007, Springer, Heidelberg