Chapter 6
Statistical Decisions

Data are often collected and analysed to answer a particular question. For instance, when modifying an existing product, one wants to know whether said modification has improved the product's quality. In the social or medical sciences, one wants to know whether an intervention has a positive (or negative) impact. One might also want to know whether the collected data come from a particular distribution. All of these questions can be tackled with the principles of statistical testing. It follows that testing implies deciding whether a particular conjecture holds or not, which we will introduce in the next section.

6.1 The Idea of Testing

To introduce the principles behind hypothesis testing, let us consider a binary decision problem with two disjoint possibilities, which we label as H_0 and H_1. We call H_0 the null hypothesis and H_1 the alternative hypothesis, or sometimes just the alternative. Generally, we do not know whether the hypothesis H_0 or the alternative H_1 holds, but given our data we want to choose one or the other, which may or may not be erroneous. If we decide for H_1 when H_0 is valid, and similarly for H_0 when H_1 is valid, we have made a mistake.

To exemplify this, let us assume a company questions whether they should introduce a new marketing strategy for a given product. This decision requires a comparison between the new and old strategies. Let us label with H_0 the hypothesis that the new strategy does not increase sales, while letting the alternative H_1 represent that in fact it does increase sales. The consequences of this decision depend on the true effect of the new marketing strategy. If the new strategy is better and the company introduces it, or it is not better and the company does not introduce it, then they made the correct decision. However, there are two possible types of incorrect decision. When the company decides to introduce the new strategy, despite the old strategy being better, it means they spend money implementing the new system (and also lose the benefits of the previous system). On the other hand, when the company refrains from introducing a new, better strategy, they have missed an opportunity. Both bad decisions have a completely different meaning and need to be treated as such. It should also be clear that if we try to avoid making one type of mistake, we run the risk of seeing more of the other. For instance, if we wanted to minimise the risk of falsely changing the existing marketing strategy, we could simply refrain from any change at all, no matter the results. By doing so, we clearly increase the risk that we will not take advantage of a newer, better strategy.

Let us formalise the above framework. We already introduced the two possible states as the null H_0 and alternative H_1 hypothesis. The decisions that we actually make are denoted with quotes, i.e. "H_0" and "H_1". In the example, this means that H_0 stands for the new strategy not being better, while "H_0" expresses our decision to refrain from making a change. Similarly, H_1 represents that the new strategy is actually better, while "H_1" denotes the corresponding decision to adopt the new strategy. Both of the states and decisions are binary and we may write these in the following decision matrix:

	Decision	
States	"H_0"	"H_1"
H_0	Correct decision	Type I error
H_1	Type II error	Correct decision

We can commit two types of errors that we label as type I and type II. It is clearly impossible to simultaneously avoid both. Statistical tests are built upon the idea that the type I error cannot be avoided but should only occur with a small probability. This asymmetrical approach requires a careful choice of H_0 and H_1.

In many applications, researchers aim to prove a hypothesis like "this drug has an effect" or "there are differences between these two groups". To use a statistical test for such questions, the null hypothesis is framed as the opposite of the research hypothesis, e.g. "this drug has no effect" or "there are no differences between these groups". That is to say, the hypothesis H_1 is usually the research hypothesis one wants to prove. This setting allows us to directly control the type I error, which is generally considered more harmful in research applications.

Definition 6.1 A statistical **significance level α-test** assumes two states H_0 and H_1 and two possible decisions "H_0" and "H_1". The decision rule is thereby constructed such that

$$P\big(\text{“}H_1\text{”}\big|H_0\big) \leq \alpha$$

for a fixed small value of α.

Example 19 We illustrate the above idea with a classical example. Assume we have observations $Y_1, \ldots, Y_n$, which are sampled *i.i.d.* and originate from a $N(\mu, \sigma^2)$ distribution. For simplicity, we assume that σ^2 is known, but μ is not. For instance, let Y_i be a quality control measurement of a process, which on average should not exceed a particular value μ_0. We want to run a test on this value and question whether $H_0 : \mu \leq \mu_0$ or $H_1 : \mu > \mu_0$ applies for a given μ_0. Note that H_0 has the inequality, while the strict inequality is in H_1. This is necessary because we need to calculate probability statements under H_0 and hence need boundedness under H_0. We take the average values of $Y_1, \dots, Y_n$ and calculate $\overline{Y} = \sum\limits_{i=1}^{n} Y_i/n$. Clearly, if $\overline{Y}$ takes (very) large values, this is in favour of H_1. This suggests the decision rule

$$\text{“}H_1\text{”} \Leftrightarrow \overline{Y} > c,$$

where the threshold c is often called the **critical value**. To determine c, we need to relate it to our probability threshold α. Under $H_0 : \mu \leq \mu_0$, it holds that

$$P\big(\text{“}H_1\text{”}\big|H_0\big) \leq \alpha \Leftrightarrow P\big(\overline{Y} > c\big|\mu \leq \mu_0\big) \leq \alpha. \tag{6.1.1}$$

The second probability statement can be solved explicitly with

$$\begin{aligned}
&P\left(\overline{Y} > c\big|\mu \leq \mu_0\right) \leq P\left(\overline{Y} > c\big|\mu = \mu_0\right) \leq \alpha \\
\Leftrightarrow &P\left(\overline{Y} \leq c\big|\mu = \mu_0\right) \geq 1-\alpha \\
\Leftrightarrow &P\left(\frac{\overline{Y}-\mu_0}{\sigma/\sqrt{n}} \leq \frac{c-\mu_0}{\sigma/\sqrt{n}}\middle|\mu = \mu_0\right) \geq 1-\alpha \\
\Leftrightarrow &\Phi\left(\frac{c-\mu_0}{\sigma/\sqrt{n}}\right) \geq 1-\alpha,
\end{aligned}$$

where Φ is the distribution function of the $N(0, 1)$ distribution. This gives

$$z_{1-\alpha} = \frac{c-\mu_0}{\sigma/\sqrt{n}} \Leftrightarrow c = \mu_0 + z_{1-\alpha}\,\sigma/\sqrt{n},$$

where $z_{1-\alpha}$ is the $1-\alpha$ quantile of the $N(0, 1)$ distribution. This approach can easily be extended to the **two-sided testing problem**, where we are interested in

$$H_0 : \mu = \mu_0 \text{ and } H_1 : \mu \neq \mu_0.$$

In this case, we look at the difference between $\overline{Y}$ and μ_0. If this difference is large, it speaks in favour of H_1, which can be formulated as the following decision rule:

$$\text{“}H_1\text{”} \Leftrightarrow \left|\overline{Y} - \mu_0\right| > c.$$

Following the same calculation as above gives

$$P\left(\left|\overline{Y}-\mu_0\right|>c\middle|H_0\right)\leq\alpha\Leftrightarrow P\left(\left|\overline{Y}-\mu_0\right|\leq c\middle|H_0\right)\geq 1-\alpha,$$

and the critical value c is given by

$$\begin{aligned}
P\left(-c\leq\overline{Y}-\mu_0\leq c\middle|\mu=\mu_0\right) && =1-\alpha\\
\Leftrightarrow P\left(\frac{-c}{\sigma/\sqrt{n}}\leq\frac{\overline{Y}-\mu_0}{\sigma/\sqrt{n}}\leq\frac{c}{\sigma/\sqrt{n}}\right) && =1-\alpha\\
\Leftrightarrow\Phi\left(\frac{c}{\sigma/\sqrt{n}}\right)-\Phi\left(-\frac{c}{\sigma/\sqrt{n}}\right) && =1-\alpha,
\end{aligned}$$

such that

$$\frac{c}{\sigma/\sqrt{n}}=z_{1-\alpha/2}\Leftrightarrow c=z_{1-\alpha/2}\,\sigma/\sqrt{n}.$$

We see that the critical value c can be derived similarly for both one-sided $H_0:\mu\leq\mu_0$ and two-sided hypotheses $H_0:\mu=\mu_0$.

▷

6.2 Classical Tests

In this section, we will introduce a number of statistical tests that are so prevalent that they require individual attention. Statistical tests can focus not only on a single parameter but also on multiple components of a multidimensional parameter. However, in order to maintain notational simplicity, we will present our classical statistical tests for one-dimensional parameters only. Let

$$Y_i\sim f(y;\theta)\quad i.i.d.,$$

and, for simplicity, let us assume that the distribution is Fisher-regular (see Definition 3.12). We formulate our hypothesis as

$$H_0:\theta\in\Theta_0\text{ and }H_1:\theta\in\Theta_1,$$

where Θ_0 and Θ_1 are a disjoint decomposition of the parameter space Θ, i.e. $\Theta=\Theta_0\cup\Theta_1$ and $\Theta_0\cap\Theta_1=\varnothing$. We restrict the subsequent presentation to the cases where $\Theta_0=\{\theta_0\}$ and $\Theta_1=\Theta\setminus\{\theta_0\}$ or $\Theta_0=\{\theta:\theta\leq\theta_0\}$ and $\Theta_1=\Theta\setminus\Theta_0$.

6.2.1 t-Test

The t-test and t-distribution are used when we would like to do inference on the mean of normally distributed random variables when the variance is unknown. Let us consider normal random variables

$$Y_i \sim N(\mu, \sigma^2) \quad i.i.d., i = 1, \ldots, n,$$

and we want to test $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$. We obtain the decision rule

$$\text{“}H_1\text{”} \Leftrightarrow \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i > c,$$

where c can be calculated with the normal distribution, i.e. $c = \mu_0 + z_{1-\alpha}\sigma/\sqrt{n}$. However, to calculate c, we need to know the variance σ^2 of the underlying distribution, which in most practical applications is unknown. In this case we can rely on the t-distribution which we will now describe. Note that under H_0, i.e. if $\mu = \mu_0$,

$$\frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1). \tag{6.2.1}$$

If we replace σ with its estimate

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2},$$

then the standard normal distribution in (6.2.1) becomes a t-distribution, that is,

$$\frac{\bar{Y} - \mu_0}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1},$$

where t_{n-1} denotes the t-distribution with $n - 1$ degrees of freedom. The t-distribution resembles the normal distribution but has heavier tails, i.e., it has a higher probability mass for more extreme observations. As n increases, that is, with increasing degrees of freedom, the t-distribution converges to the normal distribution, which can be seen in Fig. 6.1. The curious reader can also refer to Dudewicz and Mishra (1988) for more technical detail.

Example 20 One typical application of the simple t-test, also called the one-sample t-test, is to assess whether the difference between two variables has mean zero. In a psychiatric study, conducted by Leuzinger-Bohleber et al. (2019), the values of a depression score (BDI) for 149 patients were recorded at the start (baseline) and

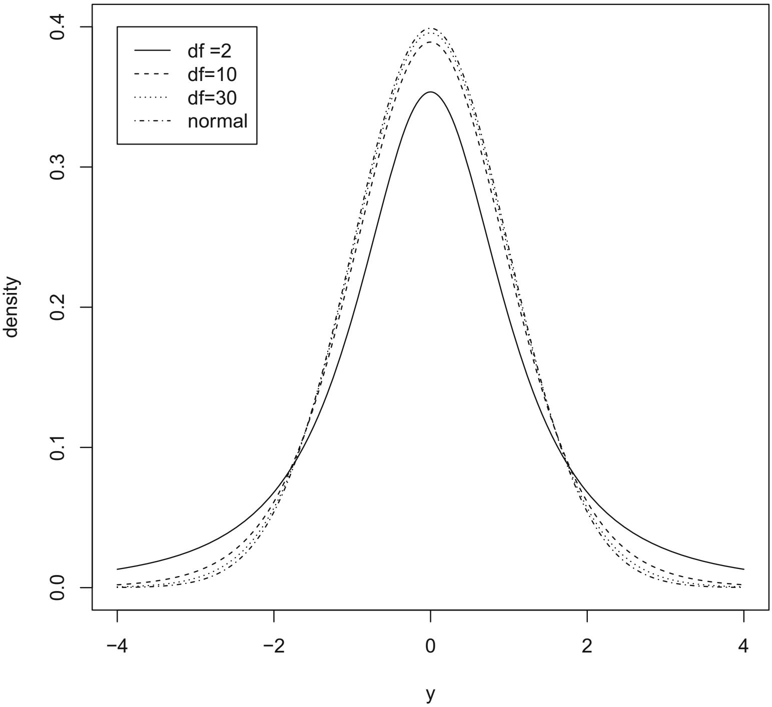

Fig. 6.1 t-Distribution with different degrees of freedom

after 3 years of a particular treatment. The boxplot of the difference between the two values for each subject can be seen in Fig. 6.2. The mean difference was -17.2, indicating a reduced average depression rating after treatment, i.e. an improvement. The standard deviation was 13. The t-value was $t = \frac{Y-0}{\sigma/\sqrt{149}} = -18.5$. Let μ_D be the mean of the difference of the depression score for each individual at baseline and after treatment. We pursue a two-sided test, i.e. $H_0 : \mu_D = 0$ against $H_1 : \mu_D \neq 0$. Then the t-value has to be compared with the critical value $t_{0.0975,149} = 1.97$. Therefore, we can reject the null hypothesis and say that we have a statistically significant change in the depression score.

▷

In many applications, we want to compare two means, e.g. the performance of male and female students or the blood pressure of two treatment groups. In such scenarios, we consider two $i.i.d.$ samples

$$Y_{1i} \sim N(\mu_1, \sigma_1^2), i = 1, \ldots, n_1 \text{ and } Y_{2j} \sim N(\mu_2, \sigma_2^2), j = 1, \ldots, n_2,$$

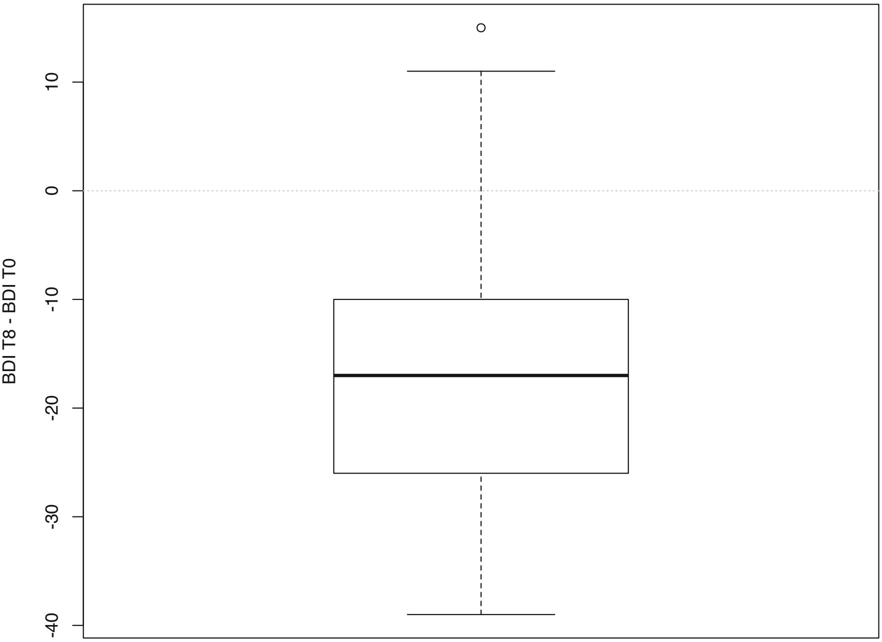

Fig. 6.2 The difference between pre-treatment and post-treatment depression score

where we allow the two groups to have different variances (variance heterogeneity). The hypothesis $H_0 : \mu_1 = \mu_2$ is tested against $H_1 : \mu_1 \neq \mu_2$. From two independent $i.i.d.$ samples, the decision rule is given by

$$\text{“}H_1\text{”} : |\bar{y}_1 - \bar{y}_2| > c.$$

As $\bar{Y}_1 - \bar{Y}_2$ has a normal distribution with mean $\mu_1 - \mu_2$ and variance $\sigma^2(1/n_1 + 1/n_2)$, the critical value c can be obtained, as in the above example, with

$$c = z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

If we replace σ_k^2 with its pooled estimate

$$\hat{\sigma}_k^2 = \frac{1}{n_k - 1} \sum_{k=1,2} (y_{ki} - \bar{y}_k)^2,$$

the $z_{1-\frac{\alpha}{2}}$-quantile of the normal distribution is replaced by the quantile of a t-distribution with df degrees of freedom, where

$$df = \Big(\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}\Big) / \Big(\frac{1}{n_1 - 1}(\frac{\hat{\sigma}_1^2}{n_1})^2 + \frac{1}{n_2 - 1}(\frac{\hat{\sigma}_2^2}{n_2})^2\Big).$$

Due to the central limit theorem and the convergence of the t-distribution to a standard normal distribution, the standard normal distribution can be applied for large sample sizes. In fact, the central limit theorem also justifies the use of a normal approximation, even if the original data are not normally distributed. This is demonstrated with the following two examples.

Example 21 Let us now see how to approach a binomially distributed random variable. Assume we want to check a promotion offered by a chocolate company, which claims that every 7th chocolate contains a prize. We buy a random sample of 350 chocolates and we expect to get 50 prizes. Let p be the proportion of chocolates containing a prize. We perform a significance test for the hypothesis $H_0 : p \geq \frac{1}{7}$ versus $H_1 : p < \frac{1}{7}$. Assuming that we have an independent sample $Y_1, \ldots, Y_{350}$ where $Y_i = 1$ if we win a prize with the i-th chocolate and $Y_i = 0$ if we do not, then under the null hypothesis

$$T = \sum_{i=1}^{350} Y_i \sim B(350, \frac{1}{7}).$$

For constructing a test with significance level α, we need to find a value for c such that $P(T \leq c|H_0) \leq \alpha$. We may now make use of the central limit theorem and get

$$\bar{Y} = \frac{T}{350} \overset{a}{\sim} N(p, \frac{p(1-p)}{350}).$$

The critical value is now directly given by

$$\begin{aligned} & P(\text{“}H_1\text{”}|H_0) \leq \alpha \\ \Leftrightarrow & P(\bar{Y} \leq c|p_0 = 1/7) \leq \alpha \Leftrightarrow \quad P\left(\frac{\bar{Y} - p_0}{\sqrt{p_0(1-p_0)/350}} \leq \underbrace{\frac{c - p_0}{\sqrt{p_0(1-p_0)/350}}}_{\approx -z_{1-\alpha} = z_\alpha}\right) \leq \alpha. \end{aligned}$$

▷

Example 22 Let us continue with Example 20, from Leuzinger-Bohleber et al. (2019). One aim of the study was to compare different types of treatments. The first group was given cognitive behavioural therapy (CBT) and the second psychoanalytic therapy (PAT). We are interested in comparing the improvements

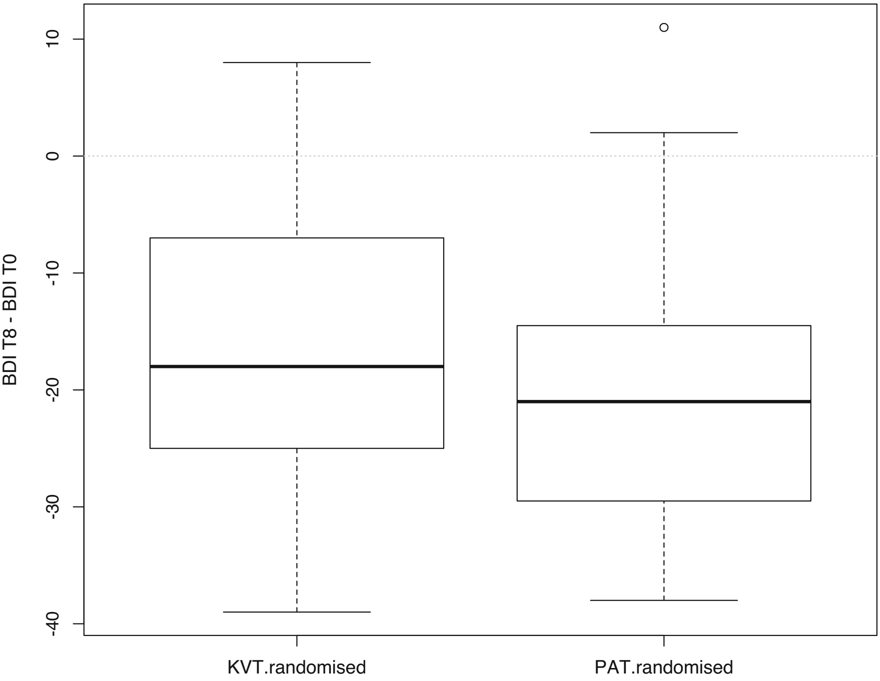

Fig. 6.3 The reduction in BDI scores after treatment with either cognitive behavioural therapy (CBT) or psychoanalytic therapy (PAT)

in depression scores of the two different groups. Denoting μ_{D1} as the difference in scores of Group 1 and μ_{D2} as the difference in scores of Group 2, we want to check the null hypothesis $H_0 : \mu_{D1}{=}\mu_{D2}$. The result for $n_1 = 30$ patients receiving cognitive behavioural therapy and $n_2 = 27$ patients receiving a long-term psycho analytic therapy is displayed in Fig. 6.3. The means of each group were $\bar{x}_1 = -17.5$ and $\bar{x}_2 = -20.1$. Using the sample variances gives the t-statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}} = 0.81.$$

This value is lower than the corresponding critical value of 1.96, and therefore the null hypothesis cannot be rejected and a difference between the two therapies cannot be claimed.

▷

6.2.2 Wald Test

We already discovered in Sect. 4.2 that the ML estimate $\hat{\theta}$ is asymptotically normal with the true parameter θ as mean, i.e.,

$$\hat{\theta} \overset{a}{\sim} N(\theta, I^{-1}(\theta)).$$

Hence, if n is large, any hypothesis on θ can be treated as a hypothesis on the mean of a normal distribution. With the hypotheses $H_0 : \theta = \theta_0$ and $H_1 : \theta \neq \theta_0$, we obtain the following decision rule:

$$\text{“}H_1\text{”} \Leftrightarrow \left|\hat{\theta} - \theta_0\right| > c,$$

where c is given by $c = z_{1-\alpha/2}\sqrt{I^{-1}(\theta_0)}$. Tests of this form are known as **Wald tests** and were proposed in Wald (1943). The Wald test is by far the most frequently used testing principle in statistics. However, it requires that we have an estimate of a parameter and its variance. Often, the variance is calculated with the Fisher information not at its hypothetical value θ_0 but at the estimated value $\hat{\theta}$. In this case, the critical value is derived from $c = z_{1-\alpha/2}\sqrt{I^{-1}(\hat{\theta})}$. Because under H_0 we know that $\hat{\theta}$ converges to θ_0, it is plausible to use $I(\hat{\theta})$ instead of $I(\theta_0)$. Given that standard software packages also give $I(\hat{\theta})$ as an output, it is also more convenient to work with the Fisher information calculated at $\hat{\theta}$ and not at θ_0. Generally speaking, however, both versions are equivalent in large samples. Note that the Wald test runs completely comparable to test in a normal distribution, as discussed in Example 19.

6.2.3 Score Test

Another very common test in statistics is the **score test**. From Eq. (4.1.3), we know that the score $s(\theta; Y_1, \ldots, Y_n)$ has mean value zero when evaluated at the true parameter θ. Assuming $\theta = \theta_0$ allows us to calculate $s(\theta_0; y_1, \ldots, y_n)$ from the data. Clearly, if the hypothesis holds, then the random score $s(\theta_0; Y_1, \ldots, Y_n)$ is asymptotically normal with mean value zero. If the value of $s(\theta_0; y_1, \ldots, y_n)$ lies far away from zero, it speaks against the hypothesis that θ_0 is the true value. Hence, we can again apply asymptotic normality to derive a test. To be specific, we test whether $s(\theta_0; y_1, \ldots, y_n)$ is a random variable drawn from a normal distribution with zero mean. Making use of Eq. (4.2.2), we derive the decision rule

$$\text{“}H_1\text{”} \Leftrightarrow \left|s(\theta_0; y_1, \ldots, y_n)\right| > c,$$

where $c = z_{1-\alpha/2}\sqrt{I(\theta_0)}$. The score test is comparable to the Wald test, but there is a subtle difference. For the Wald test, we first need to calculate the ML estimate $\hat{\theta}$ to test whether H_0 holds. This is not the case for the score test. Here we simply have to calculate the score function, but not the ML estimate. The score test is therefore particularly useful if the calculation of the ML estimate is cumbersome or numerically demanding.

6.2.4 Likelihood-Ratio Test

Also in common use is the **likelihood-ratio test**. We saw in the previous chapter that if θ is the true parameter, the likelihood-ratio $lr(\hat{\theta}; \theta) = 2\{l(\hat{\theta}) - l(\theta)\}$ is asymptotically χ^2 distributed with p degrees of freedom, with p being the dimension of the parameter. Hence, large values of the likelihood-ratio $lr(\hat{\theta}; \theta)$ calculated at θ_0 speak against the hypothesis $H_0 : \theta = \theta_0$. This allows us to derive the decision rule

$$\text{“}H_1\text{”} \Leftrightarrow 2\{l(\hat{\theta}) - l(\theta_0)\} > c.$$

The critical value c can thereby be calculated from the $1 - \alpha$ quantiles of the χ^2 distribution with p degrees of freedom. For the normal distribution, these three tests, i.e. the Wald test, the score test and the likelihood-ratio test, are equivalent, which we will demonstrate in the following example. Generally, the three tests give different results, although for a large sample size n, all three tests are asymptotically equivalent.

Example 23 In this example, we show that the Wald, score and likelihood-ratio tests are all equivalent for normally distributed random variables. Let $Y_i \sim N(\mu, \sigma^2)$ *i.i.d.* for $i = 1, \dots n$ and σ^2 be known. We want to test the null hypothesis $H_0 : \mu = \mu_0$. Note that the ML estimate is $\hat{\mu} = \sum_{i=1}^n y_i/n = \bar{y}$ with a corresponding Fisher matrix $I(\mu) = n/\sigma^2$. The Wald test gives the decision rule

$$\text{“}H_1\text{”} \Leftrightarrow |\bar{y} - \mu_0| > z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}}. \tag{6.2.2}$$

The score function is given by

$$s(\mu; y_1, \dots, y_n) = \sum_{i=1}^n \frac{y_i - \mu}{\sigma^2} = \frac{\bar{y} - \mu}{\sigma^2/n},$$

which gives the score test the decision rule

$$\begin{aligned}
\text{“}H_1\text{”} &\Leftrightarrow \left|\frac{\bar{y}-\mu_0}{\sigma^2/n}\right| > z_{1-\frac{\alpha}{2}}\sqrt{\frac{n}{\sigma^2}} \\
&\Leftrightarrow |\bar{y}-\mu_0| > z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}}.
\end{aligned}$$

Clearly, this is the same rule as that of the Wald test. Finally, the likelihood-ratio for $\mu = \mu_0$ is defined by

$$\begin{aligned}
lr(\bar{y};\mu_0) &= 2\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\bar{y})^2}{\sigma^2}+\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu_0)^2}{\sigma^2}\right\} \\
&= \frac{1}{\sigma^2}\left\{-\sum_{i=1}^{n}y_i^2+2n\bar{y}^2-n\bar{y}^2+\sum_{i=1}^{n}y_i^2-2n\bar{y}\mu_0+n\mu_0^2\right\} \\
&= \frac{1}{\sigma^2}\{n\bar{y}^2-2n\bar{y}\mu_0+n\mu_0^2\} \\
&= \frac{1}{\sigma^2/n}(\bar{y}-\mu_0)^2.
\end{aligned}$$

The decision rule for the likelihood-ratio test is now

$$\text{“}H_1\text{”} \Leftrightarrow \frac{(\bar{y}-\mu_0)^2}{\sigma^2/n} > \mathcal{X}^2_{1,1-\alpha} \tag{6.2.3}$$

$$\Leftrightarrow |(\bar{y}-\mu_0)| > \sqrt{\mathcal{X}^2_{1,1-\alpha}}\sqrt{\sigma^2/n}. \tag{6.2.4}$$

Note that for a standard normal random variable $Z \sim N(0, 1)$,

$$\begin{aligned}
&P(|Z| \le z_{1-\frac{\alpha}{2}}) = 1-\alpha \\
\Leftrightarrow &P(\underbrace{Z^2}_{\sim\mathcal{X}^2_1} \le z^2_{1-\frac{\alpha}{2}}) = 1-\alpha.
\end{aligned}$$

Hence, the $(1-\alpha)$ quantile of the chi-squared distribution is given by $\mathcal{X}^2_{1,1-\alpha} = z^2_{1-\frac{\alpha}{2}}$, and therefore (6.2.4) is equal to the decision rule of the Wald test, and the three tests are equivalent for normally distributed random variables.

▷

6.3 Power of a Test and Neyman–Pearson Test

By this point, it should be clear that statistical tests are built with the intent of limiting our type I error to α. However, this condition can be trivially met by always choosing H_0, giving $P(\text{“}H_1\text{”}|H_0) = 0$ and $P(\text{“}H_0\text{”}|H_1) = 1$. By forcing the type I error to occur with zero probability, we have made the type II error occur with probability one. That is to say, although we have effectively bounded the type I error with α, we have not defined our chance of correctly rejecting the null hypothesis in the presence of true effects. In fact, it seems that these two objectives oppose each other in some fashion.

For fixed α, the type II error changes with the magnitude of the effect. This can be demonstrated using the normal distribution with $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$. If H_1 is true, but the true mean is very close to μ_0, i.e. $\mu_0 + \delta$ for some infinitely small $\delta > 0$, then the probability $P(\bar{Y} > c|H_0)$ is almost exactly 0.05 and the probability of the type II error is close to 0.95. On the other hand, the probability of the type II error is low if the true value is much larger than μ_0. To quantify how our significance level α and the true value of μ influence our probability of a type II error, we define here the power of a test.

Definition 6.2 The power of a statistical significance test is defined as $P(\text{“}H_1\text{”}|H_1)$.

That is, the power tells us how likely we are to correctly reject the null hypothesis when it is void. We will now investigate the power and derive a test that has maximal power at a given significance level. Let us demonstrate the power function with a normally distributed example, and assume that $Y_i \sim N(\mu, \sigma^2)$, where $\sigma^2 = 1$ is known. As a hypothesis, we consider $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$. From that, we get the test decision

$$\text{“}H_1\text{”} \Leftrightarrow \frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} \geq z_{1-\alpha}.$$

We can now calculate the power in relation to μ, the true mean with

$$\begin{aligned}
P(\text{“}H_1\text{”}|\mu) &= P\left(\frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} \geq z_{1-\alpha}|\mu\right) \\
&= P\left(\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \geq z_{1-\alpha} + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}|\mu\right) \\
&= 1 - \Phi\left(z_{1-\alpha} + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}\right).
\end{aligned}$$

We can plot $P(\text{“}H_1\text{”}|\mu)$ against μ, which can be seen for $\mu \in \mathbb{R}$ in the left-hand plot of Fig. 6.4. We plot the function for two values of the sample size and see that the power increases with sample size. Despite the sample size, however, the minimum

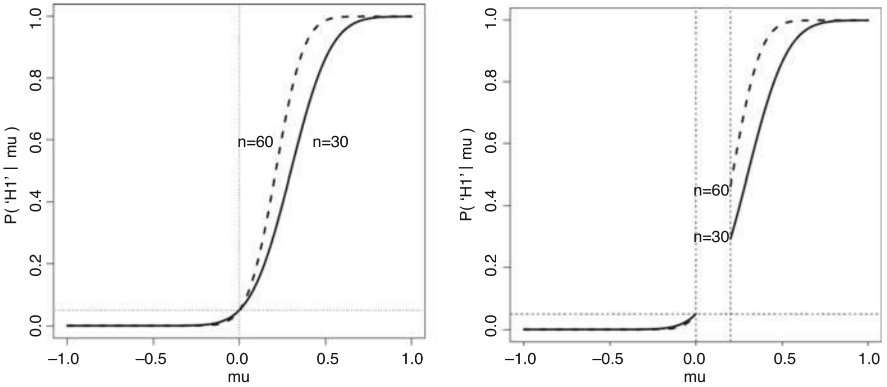

Fig. 6.4 Power function for $H_0 : \mu \leq 0$ for different sample sizes and alternatives $H_1 : \mu > 0$ (left-hand side) and $H_1 : \mu > \delta$ (right-hand side)

power is about 0.05, or to be specific

$$\sup P(\text{“}H_0\text{”}|H_1) = 1 - \inf P(\text{“}H_1\text{”}|\mu > \mu_0) = 1 - \alpha.$$

That is to say that the type II error still can occur with a high probability. Note that H_1 holds if $\mu = \mu_0 + \delta$ as long as $\delta > 0$. Hence even for $\delta = 10^{-1000}$, the hypothesis is void and H_1 holds. This brings us to the question of whether a value $\mu = \mu_0 + \delta$ is even of interest. It should be clear that the distance δ is dependent upon the problem at hand. Nanometre differences are certainly relevant in quantum physics, while in astronomy differences of thousands of kilometres might be of little interest. Thus, depending upon the problem at hand, we can set an appropriate δ and reformulate our testing problem as

$$H_0 : \mu \leq \mu_0 \text{ versus } H_1 : \mu > \mu_0 + \delta.$$

The resulting power function is shown on the right in Fig. 6.4. This time, we see that increasing the sample size has an effect on the minimum power. We also see that we can focus on two particular points in the parameter space, namely μ_0 and $\mu_1 = \mu_0 + \delta$. This simplification of the power function to just two points will be used in the following construction of an optimal test.

Property 6.1 (Neyman–Pearson Lemma) Let $H_0 : \theta = \theta_0$ be tested against $H_1 : \theta = \theta_1$ with a statistical significance test using level α. The most powerful test has the decision rule

$$\text{“}H_1\text{”} \Leftrightarrow l(\theta_0) - l(\theta_1) \leq c,$$

where c is determined such that $P(\text{“}H_1\text{”}\big|H_0) \leq \alpha$. This test is called the **Neyman–Pearson test**.

The lemma was published by Neyman and Pearson (1933). By defining a relationship between likelihood and optimal testing, the Neyman–Pearson lemma underlines once again the importance of the likelihood function in statistical reasoning. However, it does not specify the critical value c, which we will soon calculate. Using the results from the previous section, we know that twice the likelihood-ratio is asymptotically chi-squared. Hence, we can derive the critical value c based on quantiles of the chi-squared distribution and obtain the optimal test.

Proof Let y be the data, which gives $l(\theta) = \log f(y;\theta)$ as the log-likelihood. The above decision rule can then be rewritten as

$$l(\theta_0) - l(\theta_1) \leq c \Leftrightarrow \frac{f(y;\theta_0)}{f(y;\theta_1)} \leq k = \exp(c).$$

We define with $\varphi(y)$ the outcome of the Neyman–Pearson test with

$$\varphi(y) = \begin{cases} 1 & \text{if } \dfrac{f(y;\theta_0)}{f(y;\theta_1)} \leq k \\ 0 & \text{otherwise.} \end{cases}$$

Hence, if $\varphi(y) = 1$, we decide "H_1", and if $\varphi(y) = 0$, we conclude with decision "H_0". Now we take an arbitrary statistical significance test for which we similarly write the test outcome as a function $\psi(y) \in \{0, 1\}$, where $\psi(y) = 1$ means we decide for "H_1" and $\psi(y) = 0$ for "H_0". Let us now assume that $\theta = \theta_1$, i.e. H_1 holds. We need to prove that

$$P\Big(\psi(Y) = 1; \theta_1\Big) \leq P\Big(\varphi(Y) = 1; \theta_1\Big);$$

that is, if H_1 holds, then the Neyman–Pearson test decides for H_1 with a higher probability than any other arbitrarily chosen test with the same significance level. Note that $P(\psi(Y) = 1; \theta) = E_\theta(\psi(Y))$ and the same holds for the Neyman–Pearson test $\varphi(Y)$, such that

$$P\Big(\varphi(Y) = 1; \theta_1\Big) - P\Big(\psi(Y) = 1; \theta_1\Big) = \int \{\varphi(y) - \psi(y)\} f(y;\theta_1) dy.$$

We have to prove that this integral is greater than or equal to zero. The above integral can be labelled over three regions as

$$R_1 = \left\{ y : \frac{f(y;\theta_0)}{f(y;\theta_1)} < k \right\}$$

$$R_2 = \left\{ y : \frac{f(y;\theta_0)}{f(y;\theta_1)} > k \right\}$$

$$R_3 = \left\{ y : \frac{f(y;\theta_0)}{f(y;\theta_1)} = k \right\}.$$

For region R_1, we have $\varphi(y) \equiv 1$ and $f(y;\theta_1) > f(y;\theta_0)/k$, such that

$$\int\limits_{y\in R_1} \Big[\varphi(y) - \psi(y)\Big] f(y;\theta_1)dy \geq \frac{1}{k} \int\limits_{y\in R_1} \Big[\varphi(y) - \psi(y)\Big] f(y;\theta_0)dy.$$

For region R_2, we have $\varphi(y) \equiv 0$ and $-f(y;\theta_1) > -f(y;\theta_0)/k$, such that

$$\int\limits_{y\in R_2} \Big[\varphi(y) - \psi(y)\Big] f(y;\theta_1)dy = -\int\limits_{y\in R_2} \psi(y) f(y;\theta_1)dy$$

$$\geq -\frac{1}{k} \int\limits_{y\in R_2} \psi(y) f(y;\theta_0)dy = \frac{1}{k} \int\limits_{y\in R_2} \Big[\varphi(y) - \psi(y)\Big] f(y;\theta_0)dy.$$

And finally for $y \in R_3$, we have $f(y;\theta_1) = f(y;\theta_0)/k$, such that

$$\int\limits_{y\in R_3} \Big[\varphi(y) - \psi(y)\Big] f(y;\theta_1)dy = \frac{1}{k} \int\limits_{y\in R_3} \Big[\varphi(y) - \psi(y)\Big] f(y;\theta_0)dy.$$

Collecting the right-hand sides of the three regions, we can conclude

$$\int \Big[\varphi(y) - \psi(y)\Big] f(y;\theta_1)dy \geq \frac{1}{k} \int \Big[\varphi(y) - \psi(y)\Big] f(y;\theta_0)dy. \tag{6.3.1}$$

As both tests have a significance level of α, one obtains

$$\alpha = P(\text{“}H_1\text{”}\big| H_0) = \int \varphi(y) f(y;\theta_0)dy = \int \psi(y) f(y;\theta_0)dy,$$

such that the right-hand side in (6.3.1) is equal to zero and the proof is completed.

□

6.4 Goodness-of-Fit Tests

Goodness-of-fit tests are fundamentally different from the tests that we have discussed so far. Here we do not look at particular parameter values but instead test the entire distributional model itself. The hypothesis can therefore be formulated as

$$H_0 : Y \sim f(y;\theta),$$

where $f(\cdot;\cdot)$ is a known distribution, which might depend on the unknown, possibly multivariate, parameter θ. Assume now that we have *i.i.d.* data $Y_1, \dots, Y_n$ and question whether Y_i is drawn from $f(y;\theta)$. There are two classical statistical tests that are used for this purpose: the chi-squared goodness-of-fit test and the Kolmogorov–Smirnov test.

6.4.1 Chi-Squared Goodness-of-Fit Test

The chi-squared test for goodness of fit goes all the way back to Fisher (1925). To motivate the idea, assume that Y takes discrete values $1, \dots, K$. From a sample $Y_1, \dots, Y_n$, we get the following contingency table:

	1	2		K
Observed	n_1	n_2	...	n_K
Expected	e_1	e_2	...	e_K

where $n_k = \sum\limits_{i=1}^{n} 1_{\{Y_i=k\}}$ is the number of samples with the given value. We have $n = n_1 + n_2 + \dots + n_K$ and define with

$$e_k = E(n_k) = nE\big(1_{\{y=k\}}\big) = n\big\{P(Y=k;\theta)\big\},$$

the expected number of elements in each cell. We assume that the data are generated by a candidate distribution $P(k;\theta)$, whose parameter θ is known or can be estimated from the data. The difference between n_k and e_k speaks to the appropriateness of the model. If the expected numbers e_k and the observed numbers n_k deviate substantially, this gives evidence that the distributional model might not hold. On the other hand, if the expected numbers e_k and the observed numbers n_k are close to each other, it speaks in favour of the model. These differences can be assessed with the chi-squared measure

$$X^2 = \sum_{k=1}^{K} \frac{(n_k - e_k)^2}{e_k}.$$

The distribution of X^2 can be approximated by a chi-squared distribution, where the degrees of freedom are the number of cells, corrected by the number of parameters in θ that have been estimated. This correction was, in fact, the basis for a controversial discussion between Pearson and Fisher, see Baird (1983). Pearson proposed the chi-squared test for measuring goodness of fit but gave a different calculation for the degrees of freedom of the chi-squared distribution. Fisher then correctly set the degrees of freedom equal to the number of cells minus the side constraints in the cells minus the number of estimated parameters in θ. In the example above, this means the degrees of freedom are

$$K - 1 - p.$$

Here, K is the number of cells, the fact that the probability of all cells sums to 1 is a side constraint and p is the number of parameters in θ. The decision rule is now given by

$$\text{“}H_1\text{”} \Leftrightarrow X^2 \geq \mathcal{X}^2_{K-1-p,1-\alpha}.$$

The test can also be applied to continuous random variables Y. In this case, we discretise Y into “bins”. To this end, let $-\infty = c_0 < c_1 < c_2 < \ldots < c_{K-1} < c_K = \infty$ be threshold values fixed on the real axis. We define the discrete-valued random variable Z with

$$Z = k \Leftrightarrow c_{k-1} < Y \leq c_k.$$

With this trick, we have discretised the continuous variable. By defining $e_k = n \int_{c_{k-1}}^{c_k} f(\tilde{y}; \hat{\theta}) d\tilde{y}$ as the (estimated) expected cell frequencies, we can simply use the previous decision rule.

It is not difficult to see that increasing K, i.e. working with a finer grid of values, makes the test more accurate as the discrete values are closer to the continuous distribution. However, this comes at the price of a reduced number of observations per bin, such that the asymptotic distribution of X^2 may no longer be valid. In other words, one needs to find a compromise when choosing K between the discrete approximation and the asymptotic approximation. A generally accepted rule of thumb is that e_k should be larger than 5 and smaller than $(n - 5)$.

6.4.2 Kolmogorov–Smirnov Test

The Kolmogorov–Smirnov test directly compares the difference between the empirical distribution function and the hypothetical distribution. Let $F(y; \theta) = \int_{-\infty}^{y} f(\tilde{y}; \theta) d\tilde{y}$ be the distribution function, and let $F_n(y) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{Y_i \leq y\}}$ be its empirical counterpart. If $F(y; \theta)$ is the true distribution function, then the distance between $F(y; \theta)$ and $F_n(y)$ should be small, at least when n is large. We therefore construct a test statistic by looking at the difference between $F(y; \theta)$ and $F_n(y)$. To be specific, we consider the supremum of the absolute difference

$$D_n := \sup_y |F_n(y) - F(y; \theta)|.$$

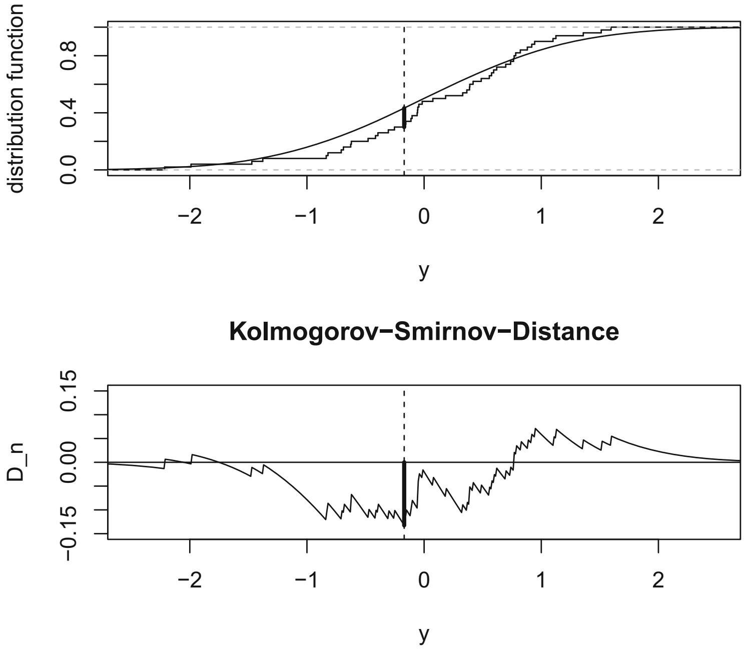

Fig. 6.5 Empirical distribution function and hypothetical distribution (top plot). Difference between these two distributions (bottom plot) with maximum (supreme) value indicated as vertical line

This quantity is also called the Kolmogorov–Smirnov distance. Figure 6.5 illustrates the construction of D_n with an example, with the empirical distribution function of a sample as $F_n(\cdot)$ and the standard normal distribution as hypothetical distribution $F(\cdot)$. This is shown in the top plot. The bottom plot gives the difference $F_n(y) - F(y)$ for all values of y, which takes its maximum (in absolute terms) at the locations indicated by the vertical line. This defines D_n. The next step is to find the distribution for D_n under H_0, i.e. if $F(y;\theta)$ is the true distribution. Let $F^{-1}(y;\theta)$ be the inverse of the distribution function, or more formally for $x \in [0, 1]$,

$$F^{-1}(x;\theta) = \min\{y : F(y,\theta) \geq x\}.$$

Then for $y = F^{-1}(x;\theta)$, we get $F(F^{-1}(x;\theta);\theta) = x$ such that

$$P(\sup_y |F_n(y) - F(y;\theta)| \leq t) = P(\sup_x |F_n(F^{-1}(x;\theta)) - x| \leq t), \tag{6.4.1}$$

where

$$F_n(F^{-1}(x;\theta)) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{Y_i \le F^{-1}(x;\theta)\}} = \frac{1}{n}\sum_{i=1}^{n} 1_{\{F(Y_i;\theta)\le x\}}.$$

Note that if $F(y;\theta)$ is the true distribution, then $F(Y_i;\theta)$ has a uniform distribution on [0,1]. This is easily seen as

$$P(F(Y_i;\theta) \le x) = P(Y_i \le F^{-1}(x;\theta)) = F(F^{-1}(x;\theta);\theta) = x.$$

Hence, we define $U_i = F(Y_i;\theta)$, which is uniform on [0,1], which in turn allows us to rewrite (6.4.1) as

$$P(\sup_x |\frac{1}{n}\sum_{i=1}^{n} 1_{\{U_i \le x\}} - x| \le t).$$

Note that this probability statement does not depend on the hypothetical distribution, and thus we have formulated a pivotal distribution. It can be shown that the probability above has a limit distribution, known as the Kolmogorov–Smirnov distribution. With this result, the test decision is now given by

$$\text{“}H_1\text{”} \Leftrightarrow D_n \ge KS_{1-\alpha},$$

where $KS_{1-\alpha}$ denotes the $1-\alpha$ quantile of the Kolmogorov–Smirnov distribution. However, the above test assumes that the parameter θ of the distribution is known, which in most practical settings is not the case. Hence, the parameter needs to be estimated, and the derivation of the exact distribution of the test statistic with estimated parameter

$$\hat{D}_n = \sup_y |F_n(y) - F(y;\hat{\theta})|$$

is thereby complicated and has only been determined for specific distributions, such as the normal or exponential distributions. For an unspecified distribution, the Kolmogorov–Smirnov distribution only holds asymptotically for D_n.

6.5 Tests on Independence

Assume multivariate data of the form

$$Y_i = (Y_{i1},\ldots,Y_{iq}) \sim F(y_1,\ldots,y_q) \quad i.i.d. \text{ for } i = 1,\ldots,n,$$

where $F(.) : \mathbb{R}^q \to [0, 1]$ is a multivariate distribution function of any form. A central question in this setting is to assess the dependence and independence structure among the components of the random vector Y_i. We denote with $F_j(.)$ the univariate marginal distribution of the j-th component of Y_i, that is,

$$P(Y_{ij} \leq y_j) = F_j(y_j) := F(\underbrace{\infty, \ldots, \infty}_{j-1 \text{ components}}, y_j, \underbrace{\infty, \ldots, \infty}_{q-j-1 \text{ components}}).$$

We can now formulate the independence of this component as a test with null hypothesis

$$H_0 : F(y_1, \ldots, y_q) = \prod_{j=1}^{q} F_j(y_j)$$

against

$$H_1 : F(y_1, \ldots, y_q) \neq \prod_{j=1}^{q} F_j(y_j).$$

This is clearly quite a general setting, but here we will limit our scope to a few classical tests. The topic is still an active research field, and we refer to Pfister et al. (2018) for an overview of modern approaches. Chapter 10 of this book also offers strategies for modelling the dependencies in multivariate data.

6.5.1 Chi-Squared Test of Independence

In Sect. 6.4.1, when testing for the goodness of fit of our distributional model, we described how to make use of a chi-squared test for discrete-valued data. This test was then extended towards discretised continuous random variables. This strategy can easily be extended towards testing for independence. For example, let us consider the bivariate case. To proceed, we discretise each random variable separately, that is, instead of $Y_i = (Y_{i1}, Y_{i2})$, we consider $Z_i = (Z_{i1}, Z_{i2})$, where

$$Z_{ij} = k \Leftrightarrow c_{j,k-1} < Y_{ij} \leq c_{j,k}$$

for $j = 1, 2$ and $k = 1, \ldots, K_j$. The threshold values fulfil $-\infty = c_{j,0} < c_{j,1} < \ldots < c_{j,K_j} = \infty$, and they are chosen such that the range of the j-th variable is well covered. If Y_{i1} and Y_{i2} are independent, that is,

$$\begin{aligned} P(Y_{i1} \leq y_1, Y_{i2} \leq y_2) &= F(y_1, y_2) = F_1(y_1)F_2(y_2) \\ &= P(Y_{i1} \leq y_1) \cdot P(Y_{i2} \leq y_2), \end{aligned}$$

then clearly Z_{i1} and Z_{i2} are also independent. We can now apply the chi-squared test to this discretised version in exactly the same fashion as in Sect. 6.4.1. This allows us to generate a table of counts:

		Z_2				
		1	2	$\dots$	K_2	
	1	n_{11}	n_{12}		n_{1K_2}	$n_{1\cdot}$
	2	n_{21}	n_{22}		n_{2K_2}	$n_{2\cdot}$
Z_1	$\vdots$					$\vdots$
	K_1	n_{K_11}	n_{K_12}		$n_{K_1K_2}$	$n_{K_1\cdot}$
		$n_{\cdot 1}$	$n_{\cdot 2}$	$\dots$	$n_{\cdot K_2}$	$n_{\cdot\cdot}$

With n_{kl}, we define the number of observations with $Z_1 = k$ and $Z_2 = l$, i.e.

$$\begin{aligned} n_{kl} &= \sum_{i=1}^{n} 1_{\{Z_{i1}=k, Z_{i2}=l\}} \\ &= \sum_{i=1}^{n} 1_{\{c_{1;k-1}<Y_{i1}\leq c_{1,k}, c_{2;l-1}<Y_{i2}\leq c_{2,l}\}}. \end{aligned}$$

Moreover, we define $n_{.l} = \sum_k n_{kl}$ as the column sum, $n_{k.}$ as the row sum and $n_{..} = n$ as the sample size. We now compare the observed counts to their expected versions under independence. Let $\pi_{kl} = P(Z_{i1} = k, Z_{i2} = l)$, which in case of independence decomposes to $\pi_{k.}\pi_{.l} = P(Z_{i1} = k) \cdot P(Z_{i2} = l)$. It is not difficult to see that the Maximum Likelihood estimates equal

$$\hat{\pi}_{k.} = n_{k.}/n_{..} \text{ and } \hat{\pi}_{.l} = n_{.l}/n_{..}$$

such that

$$e_{kl} = n_{..}\hat{\pi}_{k.}\hat{\pi}_{.l} = \frac{n_{k.}n_{.l}}{n_{..}}$$

gives the expected number of observations in the (k,l)-th cell of the table. This leads us to the chi-squared statistic

$$X^2 = \sum_{k=1}^{K_1}\sum_{l=1}^{K_2} \frac{(n_{kl} - e_{kl})^2}{e_{kl}}.$$

The distribution of X^2 can again be approximated with a chi-squared distribution, but the degrees of freedom need to be calculated. Following the results of Sect. 6.4.1, we get K_1K_2 as the number of cells and $(K_1 - 1) + (K_2 - 1)$ as the number of fitted parameters. Together with the side constraint that all cell probabilities sum up to 1, this allows us to calculate the degrees of freedom

$$K_1K_2 - 1 - (K_1 - 1) - (K_2 - 1) = (K_1 - 1)(K_2 - 1).$$

Note that the application and validity of the test require the same conditions as already discussed in Sect. 6.4.1. A rule of thumb to guarantee that the chi-squared approximation is valid is that e_{kl} should be larger than 5 and smaller than $(n_{..} - 5)$. If this does not hold, a coarser discretisation is advisable.

6.5.2 Fisher's Exact Test

In the case where we have two populations with a binary response, e.g. two samples under two different conditions, Fisher's exact test can be used as an alternative to the chi-squared test, if the sample sizes are small. Let us assume that the binary outcomes are coded as "success" and "failure". We consider two samples from the two populations:

$$\begin{aligned} X_1, X_2, \ldots, X_{n_1} &\sim B(1; p_1) \quad i.i.d. \\ Y_1, Y_2, \ldots, Y_{n_2} &\sim B(1; p_2) \quad i.i.d. \end{aligned}$$

Fisher's exact test is a test of the null hypothesis

$$H_0 : p_1 = p_2 = p$$

against the alternative hypothesis

$$H_1 : p_1 \neq p_2.$$

For the sums of these random variables, we get

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

For the following, we define $Z = X + Y$. Fisher's exact test uses the fact that the row sums n_1 and n_2 in the following 2×2 contingency table are treated as fixed.

	Success	Failure	Total
Sample 1	X	$n_1 - X$	n_1
Sample 2	$Z - X = Y$	$n_2 - (Z - X)$	n_2
Total	Z	$(n_1 + n_2 - Z)$	$n = n_1 + n_2$

Conditional on the total number of successes, $Z = z$, and thus also conditional on the column sums, the only remaining random variable is X. With $n = n_1 + n_2$ and under H_0, X is distributed hypergeometrically,

$$X \sim H(z, n_1, n) \ ,$$

i.e.

$$P(X = x|Z = z) = \frac{\binom{n_1}{x}\binom{n-n_1}{z-x}}{\binom{n}{z}} \ .$$

This allows us to derive critical values. Note that also one-sided hypotheses can be tested.

Proof The proof uses $Z = z$ and the assumption of independence between the two samples:

$$\begin{aligned} P(X = x|Z = z) &= \frac{P(X = x, Z = z)}{P(Z = z)} = \frac{P(X = x, Y = z - x)}{P(Z = z)} \\ &= \frac{P(X = x)P(Y = z - x)}{P(Z = z)} = \frac{\binom{n_1}{x}p^x(1-p)^{n_1-x}\binom{n_2}{z-x}p^{z-x}(1-p)^{n_2-(z-x)}}{\binom{n}{z}p^z(1-p)^{n-z}} \\ &= \frac{\binom{n_1}{x}\binom{n_2}{z-x}}{\binom{n}{z}} = \frac{\binom{n_1}{x}\binom{n-n_1}{z-x}}{\binom{n}{z}}. \end{aligned}$$

Here we also made use of the fact that under H_0, $Z = X + Y$ is $B(n, p)$, which comes simply from the additivity of the binomial distribution. □

Example 24 In a controlled experiment, the taste of mineral water was assessed, see Clausnitzer et al. (2004). The research question was whether the addition of oxygen changes the subjective taste of the water. There were two groups: the control group, who tasted the same water twice (type W1), and the treatment group, who tasted water of type W1 and then water with extra oxygen (type W2). The participants were asked whether there was a difference between the two samples. The experiment was double blind, i.e. both the participants and the observers did not know whether the second sample was of type W1 or type W2. The result was the following:

	"differ"	"equal"
Control group	76	24
Treatment group	89	11
	100	100

We test the hypothesis $H_0 : p_C = p_T$ vs. $H_1 : p_C \neq p_T$, where p_C and p_T denote the probability of a member of the control group and of the treatment group saying that the samples are different. The estimated probabilities are $\hat{p}_C = 0.76$ and

$\hat{p}_T = 0.89$. Fisher's exact test shows a p-value of 0.025. This indicates a significant effect on the 5% level.

▷

Let us demonstrate the idea of one-sided tests by considering 2 × 2 tables, which are more "extreme" than the one calculated from the observed data. In the case of a two-sided test, one calculates the corresponding probabilities for all possible entries of x, $x \in \{0, \ldots, \min(n_1, z)\}$. Now, all probabilities that are equal to or lower than the probability of the observed table are summed up to calculate the p-value.

Example 25 We want to assess whether a treatment works by testing if the response and treatment are independent (in which case the treatment has no effect). We look at 2 treatments (success or failure) and 10 patients per treatment:

	Success	Failure	Total
Treatment 1	0	10	10
Treatment 2	4	6	10
Total	4	16	20

Under H_0, the observed table ($x = 0$) can be evaluated using the hypergeometric distribution $H(4, 10, 20)$, which gives the probability 0.04334. Another possible table with $x = 4$, e.g.

	Success	Failure	Total
Treatment 1	4	6	10
Treatment 2	0	10	10
Total	4	16	20

has exactly the same probability (and thus belongs to the more "extreme" tables). All other fictitious and possible tables (with $x = 1, 2, 3$) fixing the row and column sums have higher probabilities. The two-sided p-value is therefore approximately 0.08669. As a consequence, H_0 would not be rejected in this example at $\alpha = 0.05$.

▷

Fisher's exact test can also be seen as a test of independence of two binary outcomes in a 2 × 2 table, i.e. the null hypothesis $H_0 : p_1 = p_2 = p$ is equivalent to the hypothesis that success and failure are independent from the conditions under which the samples have been drawn.

6.5.3 Correlation-Based Tests

A simple test on independence is carried out by directly looking at the correlation between two variables. A word of warning is required here, as independence implies zero correlation but the reverse most certainly is not true. Hence, having two uncorrelated variables does not imply that they are independent, which only holds if we additionally assume normality. In general, we can calculate the empirical correlation between two variables with

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(Y_{i1} - \bar{Y}_1)(Y_{i2} - \bar{Y}_2)}{\sqrt{\left(\sum_{i=1}^{n}(Y_{i1} - \bar{Y}_1)^2\right)\left(\sum_{i=1}^{n} Y_{i2} - \bar{Y}_2\right)^2}},$$

where $\bar{Y}_j = \sum_{i=1}^{n} Y_{ij}/n$ for $j = 1, 2$. As $\hat{\rho}$ is a statistic of the data, we can calculate its variance to derive confidence intervals. In the case of uncorrelated normally distributed random variables Y_{i1} and Y_{i2}, it can be shown (see Dudewicz and Mishra 1988) or Kendall and Stuart 1973) that

$$t = \hat{\rho}\sqrt{\frac{n-2}{1-\hat{\rho}}}$$

follows a t-distribution with $n - 2$ degrees of freedom. Fisher (1915) more than 100 years ago proposed the transformation of t with

$$z = \frac{1}{2}\log\left(\frac{1+\hat{\rho}}{1-\hat{\rho}}\right),$$

to obtain a more stable inference; see also Zimmerman et al. (2003) for a deeper discussion. A common alternative is to make use of resampling techniques to test the hypothesis $H_0 : \rho = 0$ directly. This has the advantage of being less dependent on the assumption of normality. We propose such techniques later in Chap. 8.

6.6 *p*-Value, Confidence Intervals and Test

6.6.1 The p-*Value*

When presenting the results of a statistical test, one usually does not report the outcome of the decision, that is, "H_0" or "H_1", but instead reports the ***p*-value**. This is also the case with results reported by statistical software. The p-value is closely connected to statistical tests. In fact, the idea is even older than statistical testing and was first proposed by Fisher (1925). Its calculation begins with the null hypothesis H_0. With this assumption in place, one then calculates the probability of seeing data that contradict the null hypothesis even more than the observed data. For this purpose, one also needs a discrepancy measure that quantifies how far the observed data lie from the hypothesis.

Let us make the concept more clear with a simple example. Assume we have $Y_i \sim N(\mu, \sigma^2)$ *i.i.d.* and consider the hypothesis $H_0 : \mu \leq \mu_0$. Taking $\bar{Y} = \frac{1}{n}\sum Y_i$, we can see that large values of $\bar{Y}$ speak against the hypothesis. Assume now that we have observed $\bar{y}_{obs} = \frac{1}{n}\sum y_i$. Then, we can calculate the probability that $\bar{Y}$ is larger than $\bar{y}_{obs}$, which is exactly the desired property, that is, the probability of observing data that contradict the hypothesis even more than the current data. We can now define the p-value as

$$\begin{aligned} p\text{-value } &= P(\bar{Y} \geq \bar{y}_{obs} | \mu \leq \mu_0) \\ &\leq P(\bar{Y} \geq \bar{y}_{obs} | \mu = \mu_0) \\ &= P\left(\frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} \geq \frac{\bar{y}_{obs} - \mu_0}{\sigma/\sqrt{n}} | \mu = \mu_0\right) \\ &= 1 - \Phi\left(\frac{\bar{y}_{obs} - \mu_0}{\sigma/\sqrt{n}}\right). \end{aligned}$$

Fisher argued that the smaller the p-value, the stronger the evidence against the null hypothesis H_0 and even proposed the following thresholds:

$$\begin{aligned} p\text{-value } &\leq 0.1 \Leftrightarrow \text{ weak evidence against } H_0 \\ p\text{-value } &\leq 0.05 \Leftrightarrow \text{ increased evidence against } H_0 \\ p\text{-value } &\leq 0.01 \Leftrightarrow \text{ strong evidence against } H_0. \end{aligned}$$

Note that the p-value exclusively expresses the evidence *against* the null hypothesis H_0, which is the core of statistical significance. In other words, a large p-value does not give evidence in favour of H_0; it simply means that for a large p-value, there is little evidence that H_0 might not hold. In contrast, the smaller the p-value, the more evidence there is against the validity of the hypothesis. Therefore, the p-value contains quantifiable information. In fact, p-values and statistical significance tests are closely connected. Neyman and Pearson (1933) developed the idea of testing further, leading to the significance test given in Definition 5.1. This is related to the p-value as follows.

Property 6.2 Assume a statistical significance test with significance level α, and then it holds

$$p\text{-value } \leq \alpha \Leftrightarrow \text{“}H_1\text{”}.$$

Proof The proof of this statement is rather simple. Assume we have a test statistic $t()$ such that we decide in favour of H_1 if $t(y_{obs}) > c$, where y_{obs} are the observed data, and the critical value c is determined such that

$$P(t(Y) > c | H_0) \leq \alpha.$$

Because

$$p\text{-value} = P(t(Y) > t(y_{obs})|H_0),$$

it is directly clear that if $t(y_{obs}) > c$, then the p-value $\leq \alpha$ and vice versa. □

Hence, p-values and significance tests are deeply related. However, the p-value, in explicitly quantifying how much evidence there is against the hypothesis, gives even more information. In fact, Efron (2010) writes "Fisher's famous $\alpha = 0.05$ direction for 'significance' has been overused, but has served a crucial purpose nevertheless in bringing order to scientific reporting". In Fig. 6.6, we visualise the connection and consequences between p-values and significance tests using the example of a traffic light. While the p-value is a continuous measure from green (large) to red (small), a significance test discretises this range with a fixed threshold (α), such that one either rejects or does not reject a hypothesis.

Arguments against $\alpha = 0.05$ and their misuse in scientific reporting have been collected by Goodman (2008). The p-value is unfortunately also often misinterpreted. For a recent discussion of p-values, see the American Statistical Association's statement on the issue (Wasserstein and Lazar 2016). One major misconception is that the p-value is sometimes interpreted as the size of an effect. As the p-value is also a function of the sample size, a small effect can be shown with

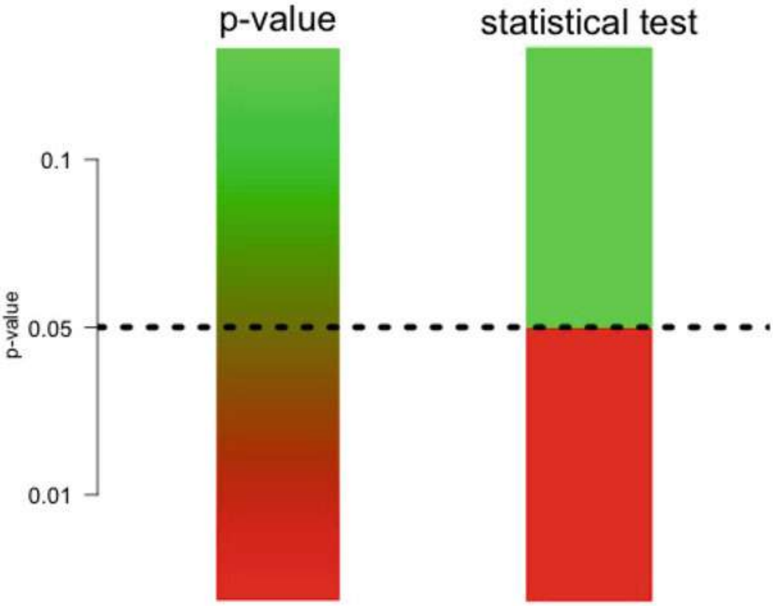

Fig. 6.6 p-value (left) and test decision (right) with threshold $\alpha = 0.05$

strong evidence ($p < 0.01$) when the sample size n is large. On the other hand, a p-value $p < 0.05$ can indicate a strong effect when the sample size is small.

Note that we can also comprehend the p-value as random variable, if we consider the observed data as random. Let us make this clear with the normal distribution example from above. In this case we denote with $\bar{Y}_{obs}$ the resulting random variable for the (later) observed data. We can calculate the random p-value for the hypothesis $H_0 : \mu \le \mu_0$ with

$$p\text{ -value } = P(\bar{Y} \ge \bar{Y}_{obs}|\mu = \mu_0) = 1 - \Phi\left(\frac{\bar{Y}_{obs} - \mu_0}{\sigma/\sqrt{n}}\right)$$

with Φ as distribution function of the $N(0, 1)$ distribution. The distribution of the p-value is then given by

$$\begin{aligned} P(p\text{-value } \le p) &= P\left(1 - \Phi\left(\frac{\bar{Y}_{obs} - \mu_0}{\sigma/\sqrt{n}}\right) \le p\right) \\ &= P\left(\Phi\left(\frac{\bar{Y}_{obs} - \mu_0}{\sigma/\sqrt{n}}\right) \ge 1 - p\right) \\ &= P\left(\frac{\bar{Y}_{obs} - \mu_0}{\sigma/\sqrt{n}} \ge \underbrace{\Phi^{-1}(1 - p)}_{z_{1-p}}\right) = p, \end{aligned}$$

where z_{1-p} is the $1 - \alpha$ quantile of the standard normal distribution. In other words, if the hypothesis holds, then the p-value has a uniform distribution on [0,1].

6.6.2 Confidence Intervals and Tests

Confidence intervals and statistical tests are deeply related. In fact, given a confidence interval, one can directly construct a corresponding test and vice versa. To demonstrate this, we require Definition 3.14, where a confidence interval was defined as $[t_l(y), t_r(y)]$ such that

$$P_\theta(t_l(Y) \le \theta \le t_r(Y)) \ge 1 - \alpha,$$

where $Y = (Y_1, \dots, Y_n)$. We define the corresponding test on the hypothesis $H_0 : \theta = \theta_0$ through the decision rule

$$\text{“}H_1\text{”} \Leftrightarrow \theta \notin [t_l(y), t_r(y)].$$

Hence, for all parameter values in the confidence interval, we accept the hypothesis and vice versa. This gives us a statistical test with significance level α. To see this, we can also define the function

$$\varphi_\theta(Y) = \begin{cases} 0 & \text{if } \theta \in [t_l(y), t_r(Y)] \\ 1 & \text{otherwise.} \end{cases}$$

Then,

$$\begin{aligned} P(\text{“}H_1\text{”}|H_0) &= 1 - P(\text{“}H_0\text{”}|H_0) \\ &= 1 - P(\varphi(Y) = 0|\theta = \theta_0) \\ &= 1 - \underbrace{P(t_l(Y) \le \theta_0 \le t_r(Y))}_{\ge 1-\alpha} \le \alpha. \end{aligned}$$

The opposite works in the same fashion. Assume we have a statistical test, which we can define as

$$\varphi_\theta(Y) = \begin{cases} 0 & \text{if “}H_0\text{”} \\ 1 & \text{if “}H_1\text{”.} \end{cases}$$

The corresponding confidence interval can then be calculated with

$$CI(y) = \{\theta : \varphi_\theta(y) = 0\}.$$

As $P(\text{“}H_1\text{”}|H_0) = P(\varphi_\theta(Y) = 1|\theta) \le \alpha$, we have

$$\begin{aligned} P(\theta \in CI(Y)|\theta) &= 1 - P(\theta \notin CI(Y)|\theta) \\ &= 1 - P(\varphi_\theta(Y) = 1|\theta) \\ &= 1 - P(\text{“}H_1\text{”}|H_0) \ge 1 - \alpha, \end{aligned}$$

which demonstrates the connection between tests and confidence intervals.

6.7 Bayes Factor

Although significance testing and the Bayesian paradigm have little in common, there is a plausible link based on what is called the Bayes factor. In some way, the Bayes factor can be seen as the Bayesian version of a p-value. Without putting too much emphasis on this link, we want to demonstrate how the Bayes factor can be used for decision-making. The Bayesian idea allows for a number of extensions by taking the Bayesian paradigm quite rigorously and making use of it in decision analytic questions. In fact, using the Bayesian view, we can also calculate the posterior probabilities of whole models to assess their validity. To demonstrate, let

M_0 and M_1 represent two different models. These models may differ because of different parameterisations or may be completely different. The former refers to the Bayesian parameter selection, while the latter relates to goodness-of-fit questions.

Generally, model selection is discussed in more detail in Chap. 9. Therefore, we will focus here on exploring a Bayesian view of parameter selection. To begin, assume that

$$Y \sim f(y|\theta),$$

where we have two different models for parameter θ. We either assume that $\theta \in \Theta_0$, which we call model M_0, or alternatively $\theta \in \Theta_1$, which we denote as model M_1. With

$$P(M_0|y) \quad \text{and} \quad P(M_1|y),$$

we can express the posterior probability that model M_0 or model M_1 holds. If we restrict ourselves to two models, we are able to calculate posterior model probabilities with

$$P(M_1|y) = 1 - P(M_0|y).$$

Using the Bayes theorem, we have

$$P(M_0|y) = \frac{f(y|M_0)P(M_0)}{f(y)} \tag{6.7.1}$$

and accordingly for $P(M_1|y)$, where $P(M_0)$ denotes the prior belief in model M_0. Clearly, each model is specified with parameters, which affect the above calculation through

$$f(y|M_0) = \int f(y|\vartheta) f_\theta(\vartheta|M_0) d\vartheta,$$

where $f(y|\theta)$ is the likelihood, as before, and $f_\theta(\theta|M_0)$ is the prior distribution of the parameter if model M_0 holds. The same holds for model M_1, i.e.

$$f(y|M_1) = \int f(y|\vartheta) f_\theta(\vartheta|M_1) d\vartheta, \tag{6.7.2}$$

such that different parameter spaces enter the comparison through different prior distributions $f_\theta(\theta|M_0)$ and $f_\theta(\theta|M_1)$. In the calculation of (6.7.1), one also has the prior distribution of the models, which expresses our prior belief in model M_0 relative to model M_1 before seeing any data. Dependent on the situation, it may be useful to set $P(M_0) = P(M_1) = \frac{1}{2}$, but this is not a requirement. Finally, we have the marginal distribution $f(y)$ in (6.7.1), which, as we have seen, is difficult to evaluate. Fortunately, we do not need to calculate $f(y)$, because instead of looking at (6) we look at the ratio

$$\frac{P(M_1|y)}{P(M_0|y)} = \underbrace{\frac{f(y|M_1)}{f(y|M_0)}}_{\text{Bayes-Factor}} \frac{P(M_1)}{P(M_0)}. \tag{6.7.3}$$

The first component in (6.7.3) is the **Bayes factor**, which gives the evidence for model M_1 relative to M_0. The larger the Bayes factor, the more evidence there is for model M_1. A rough guideline on how to interpret the values of the Bayes factor is given by Kass and Raftery (1995), who declare the following evidence in favour of M_1 based on the Bayes factor:

Bayes factor	Interpretation
1 to 3	Not worth mentioning
3 to 20	Evidence
20 to 150	Strong evidence
> 150	Very strong evidence

The calculation of the Bayes factor does require the numerical tools introduced in Chap. 6, because we need to integrate out the parameters. Looking at (6.7.2), we can solve the integral by sampling from the prior. The problem with this approach is that $f(y|\theta)$ might be (very) small for a range of values of θ, such that only a few values of θ determine the integral. It is therefore more useful to apply a sampling scheme from the posterior to calculate the integral. There is a close connection between the Bayes factor and the p-value discussed in Sect. 6.6.1. We refer to Held and Ott (2015) for a deeper discussion.

6.8 Multiple Testing

In many data analytic situations, we are not testing a single hypothesis but multiple hypotheses concurrently. A common case is when the parameter θ is multivariate of the form $\theta = (\theta_1, \ldots, \theta_p)$, in which case we test p individual hypotheses

$$H_{0j} : \theta_j = \theta_{0j}$$

for $\theta_0 = (\theta_{01}, \ldots, \theta_{0p})$ and $j = 1, \ldots, m$. For example, in a genetic experiment, we may be testing the influence of thousands of genes separately, with a single test for each individual gene. Equally, in a multicolumn dataset, we may be testing the mean of each individual column. This confronts us with the problem of multiple testing. Let us assume that we have a single test and its resulting p-value. If the hypothesis H_0 holds, then we know that the p-value is uniform on [0,1] and

$$P(p\text{ -value } \leq \alpha | H_0) = \alpha.$$

Hence, if we use the decision rule "H_1" $\Leftrightarrow$ p -value $\leq \alpha$, we make a type I error with the small probability α. Now, let us move on to two tests on the hypotheses H_{01} and H_{02}, with p-values p_1 and p_2. Assume that both H_{01} and H_{02} hold. If we define the decision rules

$$\text{"}H_{1j}\text{"} \Leftrightarrow p_j \leq \alpha \qquad j = 1, 2,$$

we can derive the probability of a type I error for either of our two hypotheses, that is, rejecting one or both of them when they both are, in fact, true. This probability is given by

$$P\Big((p_1 \leq \alpha) \vee (p_2 \leq \alpha)|H_{01} \wedge H_{02}\Big) \geq P\Big((p_1 \leq \alpha) \wedge (p_2 \leq \alpha)|H_{01} \wedge H_{02}\Big),$$

where "$\vee''$ denotes the logical "or" and "$\wedge''$ is the logical "and". If the two tests are independent, we can calculate the final probability with

$$\begin{aligned} P((p_1 > \alpha) \wedge (p_2 > \alpha)|H_{01} \text{ and } H_{02}) &= P(p_1 \geq \alpha|H_{01})P(p_2 \geq \alpha|H_{02}) \\ &= (1-\alpha)(1-\alpha) \\ &= 1 - 2\alpha + \alpha^2. \end{aligned} \tag{6.8.1}$$

Consequently, we get that

$$P(\text{"}H_{11}\text{" and/or "}H_{12}\text{"}|H_{01} \wedge H_{02}) = 2\alpha - \alpha^2 = \alpha(2-\alpha).$$

As $0 < \alpha < 1$, we get $\alpha(2-\alpha) > \alpha$. That is to say, applying two (independent) tests at significance level α leads to a significance level larger than α. In fact, one can show that the result of (6.8.1) is a lower bound, such that

$$\alpha \leq P\Big((p_1 \leq \alpha) \text{ and/or } (p_2 \leq \alpha)|H_{01} \wedge H_{02}\Big) \leq 2\alpha - \alpha^2 \leq 2\alpha. \tag{6.8.2}$$

By definition, α is small and therefore α^2 is negligible, and we can safely work with the limit 2α. In contrast, the leftmost limit occurs if the p-values are exactly the same, e.g. when we apply the same test twice to the same data. In reality, the true probability lies between these two limits.

Assume now that we have $m \geq 2$ tests for hypotheses H_{0j}, $j = 1, \ldots, m$. Then for p tests with resulting p-values p_j, we have

$$\alpha \leq P\left(\bigcup_{j=1}^{m}(p_j \leq \alpha)|H_{0j}, j = 1, \ldots, m\right) \leq m\alpha.$$

This stacking of individual tests can have powerful consequences. It means that if we make many individual tests (m large), we will falsely reject at least one hypothesis with high probability. To correct for this drawback, Bonferroni (1936) suggested the adjustment of the α value in case of multiple testing. To motivate his idea, we must first define the **family-wise error rate** as the probability of rejecting at least one of the hypotheses

$$FWER := P(\text{“}H_{11}\text{” and/or “}H_{12}\text{” and/or } \ldots \text{ and/or “}H_{1m}\text{”}|H_{01} \wedge H_{02} \wedge \ldots \wedge H_{0m}).$$

The intention is now to control the FWER instead of the significance level of each individual test.

Property 6.3 (Bonferroni Adjustment) The FWER is limited to α, if we set the significance level of each individual test to α/m, where m is the number of tests. The adjusted significance level is defined by $\alpha_{adjust} = \alpha/m$.

This correction is easily motivated, as we simply need to use the right-hand boundary in (6.8.2). Clearly, this correction is an approximation, and by again considering (6.8.1), we can find a less restrictive adjusted $\tilde{\alpha}_{\text{adjust}}$, by realising that under independence

$$P\left(\bigcap_{j=1}^{m}(p_j \geq \tilde{\alpha}_{adjust})|H_{01} \wedge H_{02} \wedge \ldots \wedge H_{1m}\right) = (1 - \tilde{\alpha}_{adjust})^m.$$

Hence, the FWER is given by

$$FWER = 1 - (1 - \tilde{\alpha}_{adjust})^m.$$

If we set $FWER = \alpha$, this gives

$$\alpha = 1 - (1 - \tilde{\alpha}_{adjust})^m \Leftrightarrow \tilde{\alpha}_{adjust} = 1 - (1 - \alpha)^{1/m}.$$

This approach is also known as the **Šidák procedure**, which improves the Bonferroni correction in the case of independent tests. Note that Bonferroni's adjustment works even for dependent tests, where the Šidák approach can fail.

However, both corrections do not take the explicit p-values of the various tests into account. On the other hand, **Holm's procedure**, originally presented in Holm (1979), makes use of the information contained in the p-values. For m tests, we derive m corresponding p-values, which we order such that

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)},$$

where $p_{(1)}$ is the smallest p-value and $p_{(m)}$ is the largest p-value obtained from the m tests. The ordering corresponds to the decision rule that if we reject the test on hypothesis $H_{0(i)}$ with p-value $p_{(i)}$, we also reject the hypotheses of all tests with p-values $p_{(j)}$ for $j = 1, \dots, i - 1$. Holm's testing procedure now proceeds as follows:

1. If $p_{(1)} > \alpha/m$, accept all hypotheses $H_{0(1)}, H_{0(2)}, \dots, H_{0(m)}$ and stop. If $p_{(1)} \leq \alpha/m$, reject hypothesis $H_{0(1)}$ and proceed to step 2.
2. If $p_{(2)} > \alpha/(m-1)$, accept all hypotheses $H_{0(2)}, \dots, H_{0(m)}$ and stop. If $p_{(2)} \leq \alpha/(m-1)$, reject hypothesis $H_{0(2)}$ and proceed to step 3.
3. If $p_{(3)} > \alpha/(m-2)$, accept all hypotheses $H_{0(3)}, \dots, H_{0(m)}$ and stop. If $p_{(3)} \leq \alpha/(m-2)$, reject hypotheses $H_{0(3)}$ and proceed to step 4.

$\vdots$

m. stop

With this procedure, we tend to first reject the tests with the smallest p-value while still keeping the FWER limited to significance level α. As $P(p_{(1)} \leq \alpha/m) = P(\min p_j \leq \alpha/m)$, the Bonferroni bound tells us that the FWER is bounded by α. In this respect, Holm's procedure does not appear to be an improvement. To fully understand the advantages of the procedure, we also need to look at the power of the test. So far we have assumed that all hypotheses are valid. In practice, however, some of the m hypotheses may not hold and, in fact, we *want* the tests on these hypotheses to lead to rejections. With that in mind, let us look at our matrix of errors again, but this time with counts of the number of times we made each decision. The following contingency table summarises the relevant values:

	Non rejected hypotheses i.e. "H_{0j}"	Rejected hypotheses i.e. "H_{1j}"	
True hypotheses H_{0j}	U	V	m_0
False hypotheses, i.e. H_{1j}	T	S	m_1
	$m - R$	R	m

Note that such tables are also used in classification problems. We stress, however, that the common arrangement of the cells in classification is different and the reader should not get confused. The table above is arranged to mimic the statistical decision matrix from above, but now for m tests. On the other hand, in classification, one labels hypothesis H_0 as "negative", while H_1 is labelled as "positive". This leads us to the quantities:

- true positive (TP) = S,
- true negative (TN) = U,
- false positive (FP) = V and
- false negative (FN) = T.

Although these are commonly used terms in classification, we stick to the notation used in the table above to exhibit the relation to statistical testing. We will return to this point in Sect. 6.9.2. Hence we assume that for m_0 of the m tests, the hypothesis is correct, while for m_1 tests, the hypothesis should be rejected. We reject R hypotheses, out of which V hypotheses are rejected incorrectly. Note that the capital letters in the inner part of the matrix are random variables, which we cannot observe, i.e. we can observe R, but neither V nor S is observable. We have defined the FWER with

$$FWER = P(V \geq 1 | \text{ true hypotheses } H_{0j}).$$

Assume now that $\Lambda_0 \subseteq \{1, \ldots, m\}$ is the index set of the true hypotheses, such that $|\Lambda_0| = m_0$. If $m_0 < m$, we have some hypotheses that are not true, and clearly we want tests on these hypotheses to reject them with high probability, that is, S should be positive and preferably as large as possible.

Coming back to Holm's procedure, we can see that the ordering of the p-values has the welcome effect that hypotheses with small p-value are more likely to be rejected, i.e. S may increase, while V is still controlled. This can be demonstrated as follows. Let j^* be the index of the smallest p-value for the valid hypotheses. That is, $p_{(j^*)} = \min_{j \in \Lambda_0} \{p_j\}$. An incorrect rejection of the true hypotheses occurs if the p-value leads to a rejection in the j^*-th step of Holm's procedure, i.e. if

$$p_{(j^*)} \leq \frac{\alpha}{m - j^* + 1}.$$

As $j^* \leq m_1 + 1 = m - m_0 + 1$, we have $m_0 \leq m - j^* + 1$, such that

$$\frac{\alpha}{m - j^* + 1} \leq \frac{\alpha}{m_0}.$$

As a consequence, the Bonferroni bound applies, but with the ordering of the p-values, it is more likely to reject false hypotheses compared to the simple Bonferroni adjustment and is therefore always recommended. The procedure has been extended in various ways, and we refer to Efron (2010) for further details.

Statistical testing overall is a conservative strategy, which means that the priority is to limit the probability of a type I error. Consequently, the probability of a type II error is uncontrolled in multiple testing. In particular, in a setting where one has a limited number of observations, but hundreds or thousands of hypotheses to test, multiple testing procedures do not suffice. A typical example is genetic data, where the sample size is usually limited, but the number of genes to be analysed is large. Statisticians sometimes call this the $n << p$ problem, meaning that the sample size n is much smaller than the number of variables p.

Benjamini and Hochberg (1995) introduced the **False Discovery Rate (FDR)** as a way to balance the type I and type II errors. The idea behind the FDR is that it is reasonable to accept a small number of false detections among the rejected hypotheses. This is expressed with the false discovery rate, which is defined as the proportion of falsely rejected hypotheses from the total. With the notation from the contingency table above, this means we are looking at the ratio

$$Q = \frac{V}{\max(R, 1)}.$$

This is an empirical number and hence random. Taking the expectation defines the FDR as

$$FDR = E(Q) = E\left(\frac{V}{R}\right).$$

The idea is now, instead of controlling V, to control Q. This is especially useful if R is large, that is, we are rejecting a substantial number of the hypotheses we are testing. It is also useful to allow V to be greater than 0, as long as we are also detecting a reasonable number of true positives.

The procedure was suggested by Benjamini and Hochberg (1995) and proceeds as follows. We again order the p-values of the m tests and get

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}.$$

We then fix $\alpha \in (0, 1)$ as the targeted FDR, where α should be small. The largest index j is then chosen such that $p_{(j)} \leq \alpha j/m$ and therefore $p_{(j+1)} > \alpha(j+1)/m$. Then all hypotheses $H_{0,(i)}$ are rejected, for which $i \leq j$. Clearly, this procedure is quite simple, which certainly explains why it is often used in practice. We omit the proof that the FDR resulting from the above strategy is in fact keeping the α level, i.e. $E(V/R) \leq \alpha$, but Benjamini and Hochberg (1995) show, in fact, that $E(V/R) \leq m_0\alpha/m$ with m_0 as the number of true hypotheses.

This still leaves the choice of α for the FDR. Note that the purpose of the FDR is *not* to control the type I error. Instead, it is to control the proportion of falsely rejected hypotheses (type I error) with respect to all rejected hypotheses. In other words, α percentage of the rejected hypotheses which are in fact false discoveries. It follows that α as FDR has a completely different interpretation from α as a significance level. It is important that one keeps this fact in mind when working with the FDR.

6.9 Significance and Relevance

6.9.1 Significance in Large Samples

It is becoming clear that in the age of "Big Data", the application of tests and confidence intervals on very large datasets might need to be reexamined. To demonstrate how the sample size affects our results, let us once again look at the power function. Assume $Y_i \sim N(\mu_0 + \delta_n, \sigma^2)$ *i.i.d.*, $i = 1, \ldots, n$, where our hypothesis is $H_0 : \delta_n = 0$ versus $H_1 : \delta_n > 0$. If we allow our effect size δ_n to be related to our sample size, we get some interesting results. To demonstrate, we set $\delta_n = \delta_1/\sqrt{n}$, where δ_1 is some constant. It is not difficult to show that this setting results in a test with constant power. Let the critical value c be defined by

$$P((\bar{Y} - \mu_0) \geq c|\delta_1 = 0) = \alpha \Leftrightarrow c = z_{1-\frac{\alpha}{2}}\sigma/\sqrt{n}$$

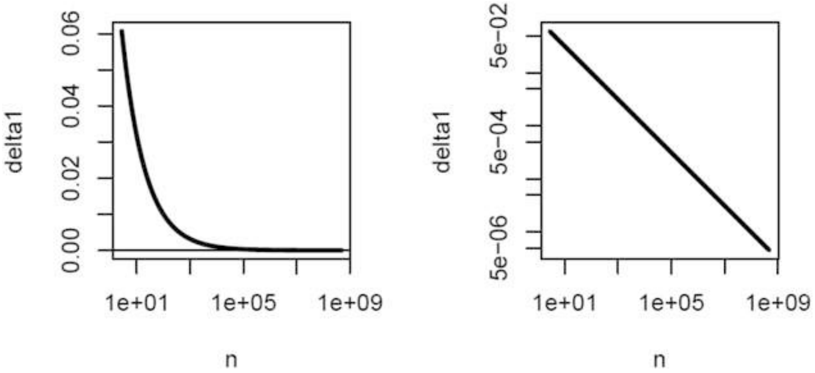

Fig. 6.7 Quantity δ_1 which leads to a power of order 0.25. Left-hand plot shows $\delta_1/\sqrt{n}$ with n on the log scale. The right-hand side shows both axes on the log scale

such that $P(\text{“}H_1\text{”}|H_0) = \alpha$. Then,

$$\begin{aligned} P(\text{“}H_1\text{”}|H_1) &= P\left((\bar{Y} - \mu_0) \geq z_{1-\alpha}\frac{\sigma}{\sqrt{n}}|\delta_1\right) \\ &= P\left(\frac{\bar{Y} - \mu_0 - \delta_1/\sqrt{n}}{\sigma/\sqrt{n}} \geq z_{1-\alpha} - \frac{\delta_1/\sqrt{n}}{\sigma/\sqrt{n}}\right) \\ &= 1 - \Phi(z_{1-\alpha} - \frac{\delta_1}{\sigma}), \end{aligned}$$

which does not depend on n.

The behaviour of $\delta_1/\sqrt{n}$ is visualised in Fig. 6.7, where on the left the sample size is given on a log scale and on the right both the sample size and $\delta_1/\sqrt{n}$ are shown on a log scale. We see that for increasing sample sizes, $\delta_1/\sqrt{n}$ goes to zero quite fast. This implies that, with a large enough sample, one could in principle detect any discrepancy from the hypothesis. If the database is large, nearly everything becomes significant. In this case, one must pose the question whether significant parameters are also relevant. We will tackle this problem in more depth in Chap. 9, when we talk about model validation. We will also look at this question from the perspective of data quality and quantity in Sect. 11.3. For now, we simply emphasise that principles of statistical testing have their limits when the sample size is very large.

6.9.2 Receiver Operating Characteristics

Let us return to the multiple testing problem and consider the decision matrix again. Note that the decision depends upon a decision rule, which is determined by the test-specific significance level $\tilde{\alpha}$. We test the j-th hypothesis H_{0j} with the j-th decision rule

$$\text{“}H_1\text{”} \Leftrightarrow p_j \leq \tilde{\alpha} \text{ for } j = 1, \ldots, m.$$

So far, we have chosen the test-specific level $\tilde{\alpha}$ either with the intention of controlling the FWER, i.e. postulating for V in the above contingency table

$$P(V \geq 1 | \text{ true hypotheses } H_{0j}) \leq \alpha,$$

or by limiting the false discovery rate, i.e.

$$FDR \leq \alpha.$$

But what happens if we increase α? Clearly, two things will happen in opposite directions. If we allow for a higher rate of the false positive cases V, we will also increase the number of true positives S. In the same fashion, decreasing the number of true negatives U will decrease the number of false negatives T. In other words, increasing the significance level α, which is the probability of a type I error, will decrease the probability of a type II error, that is, $P(\text{“}H_0\text{”}|H_1)$. To tackle this problem, let us now look at the two quantities at the same time and define the following terms.

Definition 6.3 The **specificity** of a test, or **true negative rate**, is given by

$$TNR = E\left(\frac{U}{m_0}\right) = 1 - E\left(\frac{V}{m_0}\right) = 1 - FPR,$$

which is the proportion of correct null hypotheses that are not rejected. The term $E(V/m_0)$ is also called the false positive rate (FPR). The **sensitivity** of a test, or **true positive rate**, is given by

$$TPR = E\left(\frac{S}{m_1}\right).$$

This is the proportion of incorrect null hypotheses that are correctly rejected.

One can now relate the false positive rate $FPR = 1 - TNR = E(V/m_0)$ to the true positive rate by changing the value of α. This leads us to a **Receiver Operating Curve (ROC)**, which is commonly used in classification. A typical ROC is shown in Fig. 6.8. Different values of α lead to different decisions, visualised in the behaviour of the curve. **The Area Under the Curve (AUC)** describes the quality of the test situation. Clearly, if the area is exactly $\frac{1}{2}$, the test is not recommendable, as it does not have any power. The larger the AUC, the better the test situation, which also means the higher the overall power for different values of the significance level α.

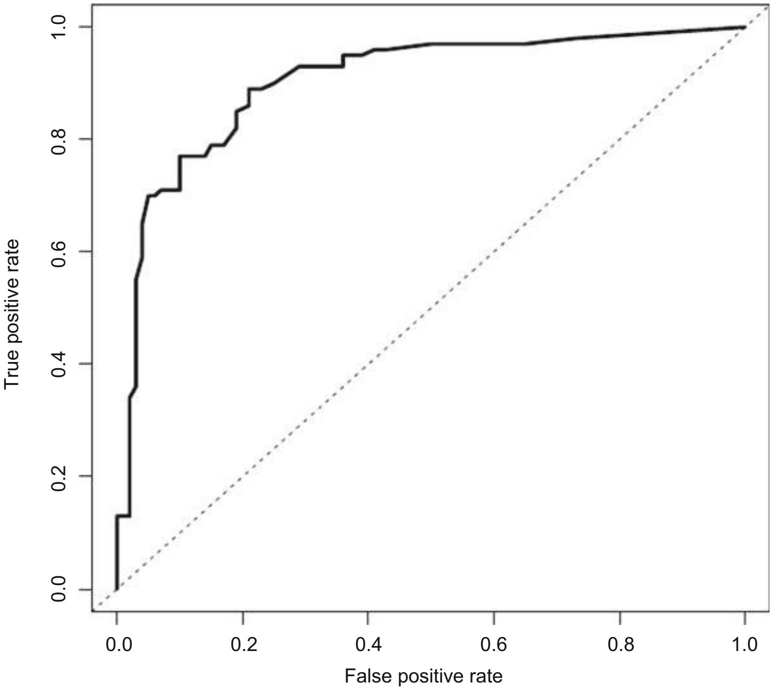

Fig. 6.8 Typical behaviour of ROC

One should note that all quantities involved are random numbers, and, of course, in reality, one cannot derive the ROC in an analytic form. We will see in Chap. 8, however, that resampling methods allow us to estimate the ROC.

6.10 Exercises

Exercise 1
We consider an $i.i.d.$ sample $Y_1, \ldots, Y_n$ from an exponential distribution $Exp(\lambda), \lambda > 0$, with density function $f(y|\lambda) = \lambda \exp(-\lambda y), y \geq 0$, and want to construct a statistical test for the hypotheses

$$H_0 : \lambda = 1 \quad \text{versus} \quad H_1 : \lambda \neq 1.$$

Construct the Wald, score and likelihood-ratio tests with the appropriate critical values and decision rules, i.e. when one decides for H_1.

Exercise 2 (Use R Statistical Software)
Simulate N vectors ("the samples") of length $n = 50$ of independent $N(0, 1)$ random numbers. Then modify the vectors by adding constants $c_1, \ldots, c_N$ to each of the vectors.

1. Use appropriate tests ($\alpha = 0.05$) for each of the N test problems

$$H_{0j} : \mu_j = 0 \quad \text{versus} \quad H_{1j} : \mu_j = c_j \ , j = 1, \ldots, N \ .$$

   Simulate the distribution of the p-values for the case that $c_j = 0, \forall j = 1, \ldots, N$, and the case that some $c_j \neq 0, \forall j = 1, \ldots, N$, by increasing the number N ($N \to \infty$).
2. Now set $N = 10$ and $\alpha = 0.05$. Repeat the generation of samples 100 times, i.e. 1000 vectors of length 50 are generated. Estimate the false discovery rate (FDR). What is the FDR after correcting the p-values with the Bonferroni procedure? Repeat the process for different values of the constants c_j.