8. Testing Statistical Hypotheses

8.1 Introduction

Often the problem of a statistical analysis does not involve determining an originally unknown parameter, but rather, one already has a preconceived opinion about the value of the parameter, i.e., a *hypothesis*. In a sample taken for quality control, for example, one might initially assume that certain critical values are normally distributed within tolerance levels around their nominal values. One would now like to test this hypothesis. To elucidate such test procedures, called *statistical tests*, we will consider such an example and for simplicity make the hypothesis that a sample of size 10 originates from a standard normal distribution.

Suppose the analysis of the sample resulted in the arithmetic mean $\bar{x} = 0.154$. Under the assumption that our hypothesis is correct, the random variable $\bar{x}$ is normally distributed with mean 0 and standard deviation $\frac{1}{\sqrt{10}}$. We now ask for the probability to observe a value $|\bar{x}| \geq 0.154$ from such a distribution. From (5.8.5) and Table I.3 this is

$$P(|\bar{x}| \geq 0.154) = 2\{1 - \psi_0(0.154\sqrt{10})\} = 0.62 \quad .$$

Thus we see that even if our hypothesis is correct, there is a probability of 62% that a sample of size 10 will lead to a sample mean that differs from the population mean by 0.154 or more.

We now find ourselves in the difficult situation of having to answer the simple question: "Is our hypothesis true or false?" A solution to this problem is provided by the concept of the *significance level*: One specifies before the test a (small) test probability α. Staying with our previous example, if $P(|\bar{x}| \geq 0.154) < \alpha$, then one would regard the occurrence of $\bar{x} = 0.154$ as improbable. That is, one would say that $\bar{x}$ differs significantly from the hypothesized value

S. Brandt, *Data Analysis: Statistical and Computational Methods for Scientists and Engineers*,
DOI 10.1007/978-3-319-03762-2_8, © Springer International Publishing Switzerland 2014

and the hypothesis would be rejected. The converse is, however, not true. If P does not fall below α, we cannot say that "the hypothesis is true", but rather "it is not contradicted by the result of the sample". The choice of the significance level depends on the problem being considered. For quality control of pencils one might be satisfied with 1%. If, however, one wishes to determine insurance premiums such that the probability for the company to go bankrupt is less than α, then one would probably still regard 0.01% as too high. In the analysis of scientific data α values of 5, 1, or 0.1% are typically used. From Table I.3 we can obtain limiting values for $|\bar{x}|$ such that a deviation in excess of these values corresponds to the given probabilities. These are

$$\begin{aligned} 0.05 &= 2\{1-\psi_0(1.96)\} = 2\{1-\psi_0(0.62\sqrt{10})\} \quad , \\ 0.01 &= 2\{1-\psi_0(2.58)\} = 2\{1-\psi_0(0.82\sqrt{10})\} \quad , \\ 0.001 &= 2\{1-\psi_0(3.29)\} = 2\{1-\psi_0(1.04\sqrt{10})\} \quad . \end{aligned}$$

At these significance levels the value $|\bar{x}|$ would have to exceed the values 0.62, 0.82, 1.04 before we could reject the hypothesis.

In some cases the sign of $\bar{x}$ is important. In many production processes, deviations in a positive and negative direction are of different importance. (If a baker's rolls are too heavy, this reduces profits; if they are too light they cost the baker his license.) If one sets, e.g.,

$$P(\bar{x} \geq x'_\alpha) < \alpha \quad ,$$

then this is referred to as a *one-sided test* in contrast to the *two-sided test*, which we have already considered (Fig. 8.1).

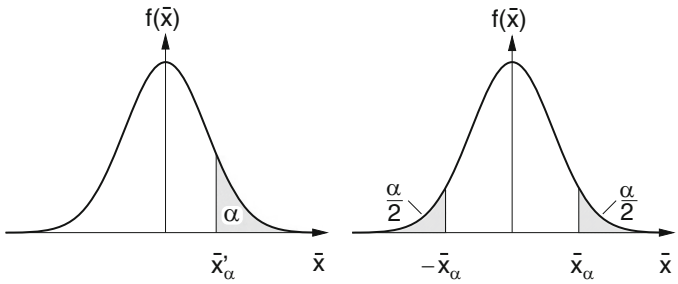

Fig. 8.1: One-sided and two-sided tests

For many tests one does not construct the sample mean but rather a certain function of the sample called a *test statistic*, which is particularly suited for tests of certain hypotheses. As above one specifies a certain significance level α and determines a region U in the space of possible values of the test statistic T in which

$$P_H(T \in U) = \alpha \quad .$$

The index H means that the probability was computed under the assumption that the hypothesis H is valid. One then obtains a sample, which results in a

particular value T' for the test statistic. If T' is in the region U, the *critical region* of the test, then the hypothesis is rejected.

In the next sections we will discuss some important tests in detail and then turn to a more rigorous treatment of test theory.

8.2 F-Test on Equality of Variances

The problem of comparing two variances occurs frequently in the development of measurement techniques or production procedures. Consider two populations with the same expectation value; e.g., one measures the same quantity with two different devices without systematic error. One may then ask if they also have the same variance.

To test this hypothesis we take samples of size N_1 and N_2 from both populations, which we assume to be normally distributed. We construct the sample variance (6.2.6) and consider the ratio

$$F = \mathsf{s}_1^2/\mathsf{s}_2^2 \quad . \tag{8.2.1}$$

If the hypothesis is true, then F will be near unity. It is known from Sect. 6.6 that for every sample we can construct a quantity that follows a χ^2-distribution:

$$\begin{aligned} X_1^2 &= \frac{(N_1-1)\mathsf{s}_1^2}{\sigma_1^2} = \frac{f_1\mathsf{s}_1^2}{\sigma_1^2} \quad , \\ X_2^2 &= \frac{(N_2-1)\mathsf{s}_2^2}{\sigma_2^2} = \frac{f_2\mathsf{s}_2^2}{\sigma_2^2} \quad . \end{aligned}$$

The two distributions have $f_1 = (N_1 - 1)$ and $f_2 = (N_2 - 1)$ degrees of freedom.

Under the assumption that the hypothesis ($\sigma_1^2 = \sigma_2^2$) is true, one has

$$F = \frac{f_2}{f_1}\,\frac{X_1^2}{X_2^2} \quad .$$

The probability density of a χ^2-distribution with f degrees of freedom is [see (6.6.10)]

$$f(\chi^2) = \frac{1}{\Gamma(\frac{1}{2}f)2^{\frac{1}{2}f}}(\chi^2)^{\frac{1}{2}(f-2)}\mathrm{e}^{-\frac{1}{2}\chi^2} \quad .$$

We now compute the probability*

*We use here the symbol W for a distribution function in order to avoid confusion with the ratio F.

$$W(Q) = P\left(\frac{X_1^2}{X_2^2} < Q\right)$$

that the ratio X_1^2/X_2^2 is smaller than Q:

$$W(Q) = \frac{1}{\Gamma(\frac{1}{2}f_1)\Gamma(\frac{1}{2}f_2)2^{\frac{1}{2}(f_1+f_2)}} \int\int_{\substack{x>0\\ y>0\\ x/y>Q}} x^{\frac{1}{2}f_1-1}\mathrm{e}^{-\frac{1}{2}x} y^{\frac{1}{2}f_2-1}\mathrm{e}^{-\frac{1}{2}y}\,\mathrm{d}x\,\mathrm{d}y \quad .$$

Calculating the integral gives

$$W(Q) = \frac{\Gamma(\frac{1}{2}f)}{\Gamma(\frac{1}{2}f_1)\Gamma(\frac{1}{2}f_2)} \int_0^Q t^{\frac{1}{2}f_1-1}(t+1)^{-\frac{1}{2}f}\,\mathrm{d}t \quad , \tag{8.2.2}$$

where

$$f = f_1 + f_2 \quad .$$

Finally if one sets

$$F = Q\, f_2/f_1 \quad ,$$

then the distribution function of the ratio F can be obtained from (8.2.2),

$$W(F) = P\left(\frac{\mathsf{s}_1^2}{\mathsf{s}_2^2} < F\right) \quad .$$

This is called the Fisher F-distribution.[†] It depends on the parameters f_1 and f_2. The probability density for the F-distribution is

$$f(F) = \left(\frac{f_1}{f_2}\right)^{\frac{1}{2}f_1} \frac{\Gamma(\frac{1}{2}(f_1+f_2))}{\Gamma(\frac{1}{2}f_1)\Gamma(\frac{1}{2}f_2)} F^{\frac{1}{2}f_1-1}\left(1+\frac{f_1}{f_2}F\right)^{-\frac{1}{2}(f_1+f_2)} \quad . \tag{8.2.3}$$

This is shown in Fig. 8.2 for fixed values of f_1 and f_2. The distribution is reminiscent of the χ^2-distribution; it is only non-vanishing for $F \geq 0$, and has a long tail for $F \to \infty$. Therefore it cannot be symmetric. One can easily show that for $f_2 > 2$ the expectation value is simply

$$E(F) = f_2/(f_2-2) \quad .$$

We can now determine a limit F'_α with the requirement

$$P\left(\frac{\mathsf{s}_1^2}{\mathsf{s}_2^2} > F'_\alpha\right) = \alpha \quad . \tag{8.2.4}$$

[†]This is also often called the ν^2-distribution, ω^2-distribution, or Snedecor distribution.

This expression means that the limit F'_α is equal to the *quantile* $F_{1-\alpha}$ of the F-distribution (see Sect. 3.3) since

$$P\left(\frac{s_1^2}{s_2^2} < F'_\alpha\right) = P\left(\frac{s_1^2}{s_2^2} < F_{1-\alpha}\right) = 1-\alpha \quad . \tag{8.2.5}$$

If this limit is exceeded, then we say that $\sigma_1^2 > \sigma_2^2$ with the significance level α. The quantiles $F_{1-\alpha}$ for various pairs of values (f_1, f_2) are given in Table I.8. In general one applies a two-sided test, i.e., one tests whether the ratio F is between two limits F''_α and F'''_α, which are determined by

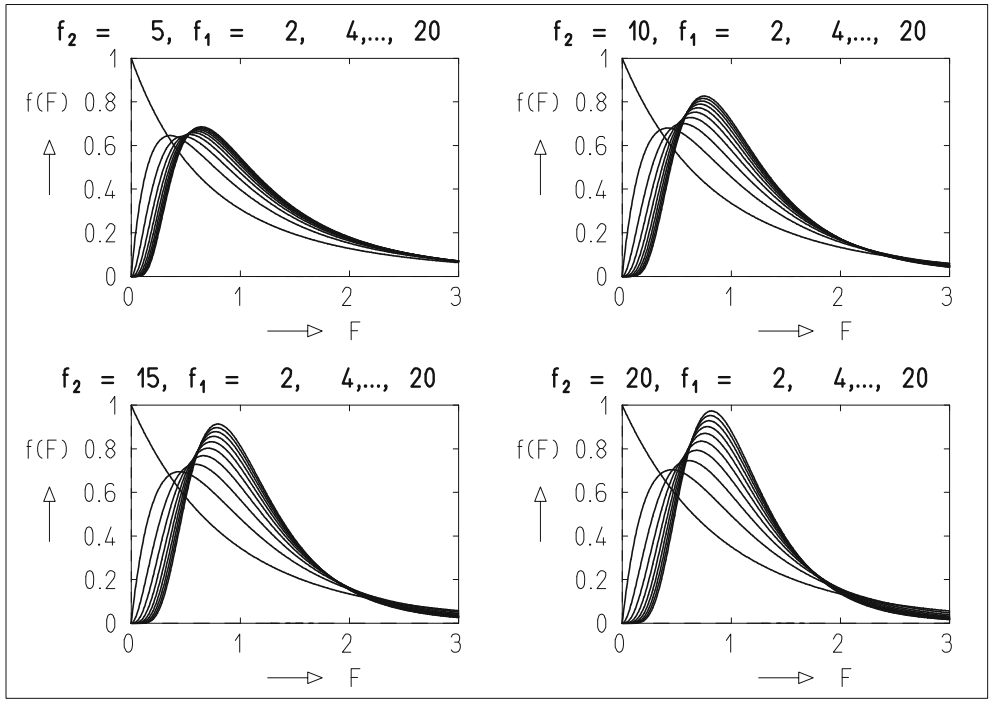

Fig. 8.2: Probability density of the F-distribution for fixed values of $f_1 = 2, 4, \ldots, 20$. For $f_1 = 2$ one has $f(F) = \mathrm{e}^{-F}$. For $f_1 > 2$ the function has a maximum which increases for increasing f_1.

$$P\left(\frac{s_1^2}{s_2^2} > F''_\alpha\right) = \frac{1}{2}\alpha, \qquad P\left(\frac{s_1^2}{s_2^2} < F'''_\alpha\right) = \frac{1}{2}\alpha \quad . \tag{8.2.6}$$

Because of the definition of F as a ratio, the inequality

$$s_1^2/s_2^2 < F'''_\alpha(f_1, f_2)$$

clearly has the same meaning as

$$\mathsf{s}_2^2/\mathsf{s}_1^2 > F_\alpha'''(f_2, f_1) \quad .$$

Here the first argument gives the number of degrees of freedom in the numerator, and the second in the denominator. The requirement (8.2.6) can also be written:

$$P\left(\frac{\mathsf{s}_1^2}{\mathsf{s}_2^2} > F_\alpha''(f_1, f_2)\right) = \frac{1}{2}\alpha, \qquad P\left(\frac{\mathsf{s}_2^2}{\mathsf{s}_1^2} > F_\alpha''(f_2, f_1)\right) = \frac{1}{2}\alpha \quad . \tag{8.2.7}$$

Table I.8 can therefore be used for the one-sided as well as the two-sided F-test.

A glance at Table I.8 also shows that $F_{1-\alpha/2} > 1$ for all reasonable values of α. Therefore one needs only to find the limit for the ratio

$$\mathsf{s}_g^2/\mathsf{s}_k^2 > F_{1-\frac{1}{2}\alpha}(f_g, f_k) \quad . \tag{8.2.8}$$

Here the indices g and k give the larger and smaller values of the two variances, i.e., $s_g^2 > s_k^2$. If the inequality (8.2.8) is satisfied, then the hypothesis of equal variances must be rejected.

Example 8.1: F-test of the hypothesis of equal variance of two series of measurements

A standard length (100 μm) is measured using two traveling microscopes. The measurements and computations are summarized in Table 8.1. From Table I.8 we find for the two-sided F-test with a significance level of 10%,

$$F_{0.1}''(6, 9) = F_{0.95}(6, 9) = 3.37 \quad .$$

The hypothesis of equal variances cannot be rejected. ■

8.3 Student's Test: Comparison of Means

We now consider a population that follows a standard Gaussian distribution. Let $\bar{\mathsf{x}}$ be the arithmetic mean of a sample of size N. According to (6.2.3) the variance of $\bar{\mathsf{x}}$ is related to the population variance by

$$\sigma^2(\bar{\mathsf{x}}) = \sigma^2(\mathsf{x})/N \quad . \tag{8.3.1}$$

If N is sufficiently large, then from the Central Limit Theorem, $\bar{\mathsf{x}}$ will be normally distributed with mean $\hat{x}$ and variance $\sigma^2(\bar{\mathsf{x}})$. That is,

$$y = (\bar{x} - \hat{x})/\sigma(\bar{x}) \tag{8.3.2}$$

will be described by a standard normal distribution. The quantity $\sigma(x)$ is, however, not known. We only know the estimate for $\sigma^2(x)$,

$$s_x^2 = \frac{1}{N-1}\sum_{j=1}^{N}(x_j - \widehat{x})^2 \quad . \tag{8.3.3}$$

Then with (8.3.1) we can also estimate $\sigma^2(\bar{x})$ to be

$$s_{\bar{x}}^2 = \frac{1}{N(N-1)}\sum_{j=1}^{N}(x_j - \bar{x})^2 \quad . \tag{8.3.4}$$

We now ask to what extent (8.3.2) differs from the standard Gaussian distribution if $\sigma(\bar{x})$ is replaced by $s_{\bar{x}}$. By means of a simple translation of coordinates we can always have $\hat{x} = 0$. We therefore only consider the distribution of

$$t = \bar{x}/s_{\bar{x}} = \bar{x}\sqrt{N}/s_x \quad . \tag{8.3.5}$$

Since $(N-1)\,s_x^2 = f\,s_x^2$ follows a χ^2-distribution with $f = N-1$ degrees of freedom, we can write

Table 8.1: F-test on the equality of variances. Data from Example 8.1.

	Measurement with	
Measurement number	Instrument 1 [μm]	Instrument 2 [μm]
1	100	97
2	101	102
3	103	103
4	98	96
5	97	100
6	98	101
7	102	100
8	101	
9	99	
10	101	
Mean	100	99.8
Degrees of freedom	9	6
s^2	$34/9 = 3.7$	$39/6 = 6.5$
$F = 6.5/3.7 = 1.8$		

$$\mathsf{t} = \bar{\mathsf{x}}\sqrt{N}\sqrt{f}/\chi \quad .$$

The distribution function of t is given by

$$F(t) = P(\mathsf{t} < t) = P\left(\frac{\bar{\mathsf{x}}\sqrt{N}\sqrt{f}}{\chi} < t\right) \quad . \tag{8.3.6}$$

After a somewhat lengthy calculation one finds

$$F(t) = \frac{\Gamma(\frac{1}{2}(f+1))}{\Gamma(\frac{1}{2}f)\sqrt{\pi}\sqrt{f}} \int_{-\infty}^{t} \left(1 + \frac{t^2}{f}\right)^{-\frac{1}{2}(f+1)} \mathrm{d}t \quad .$$

The corresponding probability density is

$$f(t) = \frac{\Gamma(\frac{1}{2}(f+1))}{\Gamma(\frac{1}{2}f)\sqrt{\pi}\sqrt{f}} \left(1 + \frac{t^2}{f}\right)^{-\frac{1}{2}(f+1)} \quad . \tag{8.3.7}$$

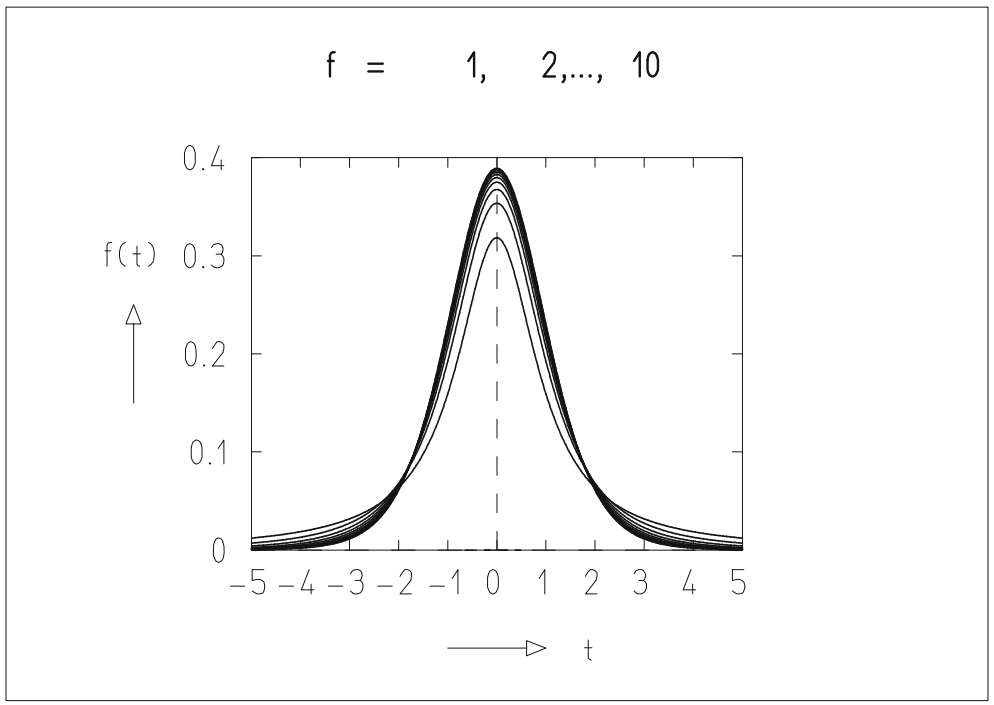

Fig. 8.3: Student's distribution $f(t)$ for $f = 1, 2, \ldots, 10$ degrees of freedom. For $f = 1$ the maximum is lowest and the tails are especially prominent.

Figure 8.3 shows the function $f(t)$ for various degrees of freedom $f = N - 1$. A comparison with Fig. 5.7 shows that for $f \to \infty$, the distribution (8.3.7) becomes the standard normal distribution $\phi_0(t)$, as expected.

Like $\phi_0(t)$, $f(t)$ is symmetric about 0 and has a bell shape. Corresponding to (5.8.3) one has

$$P(|t| \le t) = 2F(|t|) - 1 \quad . \tag{8.3.8}$$

By requiring

$$\int_0^{t'_\alpha} f(t)\,\mathrm{d}t = \frac{1}{2}(1-\alpha) \tag{8.3.9}$$

we can again determine limits $\pm t'_\alpha$ at a given significance level α, where

$$t'_\alpha = t_{1-\frac{1}{2}\alpha} \quad .$$

The quantiles $t_{1-\frac{1}{2}\alpha}$ are given in Table I.9 for various values of α and f.

The application of Student's test[‡] can be described in the following way: One has a hypothesis λ_0 for the population mean of a normal distribution. A sample of size N yields the sample mean $\bar{\mathsf{x}}$ and sample variance s_x^2. If the inequality

$$|t| = \frac{|\bar{\mathsf{x}} - \lambda_0|\sqrt{N}}{\mathsf{s}_\mathsf{x}} > t'_\alpha = t_{1-\frac{1}{2}\alpha} \tag{8.3.10}$$

is fulfilled for a given significance level α, then the hypothesis must be rejected.

This is clearly a two-sided test. If deviations only in one direction are important, then the corresponding test at the significance level α is

$$t = \frac{(\bar{\mathsf{x}} \pm \lambda_0)\sqrt{N}}{\mathsf{s}_\mathsf{x}} > t'_{2\alpha} = t_{1-\alpha} \quad . \tag{8.3.11}$$

We can make the test more general and apply it to the problem of comparing two mean values. Suppose samples of size N_1 and N_2 have been taken from two populations X and Y. We wish to find a measure of correctness for the hypothesis that the expectation values are equal,

$$\hat{x} = \hat{y} \quad .$$

Because of the Central Limit Theorem, the mean values are almost normally distributed. Their variances are

$$\sigma^2(\bar{\mathsf{x}}) = \frac{1}{N_1}\sigma^2(\mathsf{x}), \qquad \sigma^2(\bar{\mathsf{y}}) = \frac{1}{N_2}\sigma^2(\mathsf{y}) \tag{8.3.12}$$

and the estimates for these quantities are

[‡]The t-distribution was introduced by W. S. Gosset and published under the pseudonym "Student".

$$\begin{aligned} \mathsf{s}^2_{\bar{\mathsf{x}}} &= \frac{1}{N_1(N_1-1)} \sum_{j=1}^{N_1} (\mathsf{x}-\bar{\mathsf{x}})^2 \quad , \\ \mathsf{s}^2_{\bar{\mathsf{y}}} &= \frac{1}{N_2(N_2-1)} \sum_{j=1}^{N_2} (\mathsf{y}-\bar{\mathsf{y}})^2 \quad . \end{aligned} \tag{8.3.13}$$

According to the discussion in Example 5.10, the difference

$$\Delta = \bar{\mathsf{x}} - \bar{\mathsf{y}} \tag{8.3.14}$$

also has an approximate normal distribution with

$$\sigma^2(\Delta) = \sigma^2(\bar{\mathsf{x}}) + \sigma^2(\bar{\mathsf{y}}) \quad . \tag{8.3.15}$$

If the hypothesis of equal means is true, i.e., $\hat{\Delta} = 0$, then the ratio

$$\Delta / \sigma(\Delta) \tag{8.3.16}$$

follows the standard normal distribution. If $\sigma(\Delta)$ were known one could immediately give the probability according to (5.8.2) for the hypothesis to be fulfilled. But only s_Δ is known. The corresponding ratio

$$\Delta / \mathsf{s}_\Delta \tag{8.3.17}$$

will in general be somewhat larger.

Usually the hypothesis $\hat{x} = \hat{y}$ implies that $\bar{\mathsf{x}}$ and $\bar{\mathsf{y}}$ come from the same population. Then $\sigma^2(\mathsf{x})$ and $\sigma^2(\mathsf{y})$ are equal, and we can use the weighted mean of s^2_x and s^2_y as the corresponding estimator. The weights are given by $(N_1 - 1)$ and $(N_2 - 1)$:

$$\mathsf{s}^2 = \frac{(N_1-1)\mathsf{s}^2_\mathsf{x} + (N_2-1)\mathsf{s}^2_\mathsf{y}}{(N_1-1)+(N_2-1)} \quad . \tag{8.3.18}$$

From this we construct

$$\mathsf{s}^2_{\bar{\mathsf{x}}} = \frac{\mathsf{s}^2}{N_1}, \qquad \mathsf{s}^2_{\bar{\mathsf{y}}} = \frac{\mathsf{s}^2}{N_2} \quad ,$$

and

$$\mathsf{s}^2_\Delta = \mathsf{s}^2_{\bar{\mathsf{x}}} + \mathsf{s}^2_{\bar{\mathsf{y}}} = \frac{N_1+N_2}{N_1 N_2} \mathsf{s}^2 \quad . \tag{8.3.19}$$

It can be shown (see [8]) that the ratio (8.3.17) follows the Student's t-distribution with $f = N_1 + N_2 - 2$ degrees of freedom. With this one can now perform *Student's difference test*:

The quantity (8.3.17) is computed from the results of two samples. This value is compared to a quantile of Student's distribution with $f = N_1 + N_2 - 2$ degrees of freedom with a significance level α. If

$$|t| = \frac{|\Delta|}{\mathsf{s}_\Delta} = \frac{|\bar{\mathsf{x}} - \bar{\mathsf{y}}|}{\mathsf{s}_\Delta} \geq t'_\alpha = t_{1-\frac{1}{2}\alpha} \quad , \tag{8.3.20}$$

then the hypothesis of equal means must be rejected. Instead one would assume $\hat{x} > \hat{y}$ or $\hat{x} < \hat{y}$, depending on whether one has $\bar{\mathsf{x}} > \bar{\mathsf{y}}$ or $\bar{\mathsf{x}} < \bar{\mathsf{y}}$.

Example 8.2: Student's test of the hypothesis of equal means of two series of measurements

Column x of Table 8.2 contains measured values (in arbitrary units) of the concentration of neuraminic acid in the red blood cells of patients suffering from a certain blood disease. Column y gives the measured values for a group of healthy persons. From the mean values and variances of the two samples one finds

$$\begin{aligned} |\Delta| &= |\bar{\mathsf{x}} - \bar{\mathsf{y}}| = 1.3 \quad , \\ \mathsf{s}^2 &= \frac{15\mathsf{s}_\mathsf{x}^2 + 6\mathsf{s}_\mathsf{y}^2}{21} = 9.15 \quad , \\ \mathsf{s}_\Delta^2 &= \frac{23}{112}\mathsf{s}^2 = 1.88 \quad . \end{aligned}$$

For $\alpha = 5\%$ and $f = 21$ we find $t_{1-\alpha/2} = 2.08$. We must therefore conclude that the experimental data is not sufficient to determine an influence of the disease on the concentration. ■

8.4 Concepts of the General Theory of Tests

The test procedures discussed so far have been obtained more or less intuitively and without rigorous justification. In particular we have not given any specific reasons for the choice of the critical region. We now want to deal with the theory of statistical tests in a somewhat more critical way. A complete treatment of this topic would, however, go beyond the scope of this book.

Each sample of size N can be characterized by N points in the sample space of Sect. 2.1. For simplicity we will limit ourselves to a continuous random variable x, so that the sample can be described by N points $(\mathsf{x}^{(1)}, \mathsf{x}^{(2)}, \ldots, \mathsf{x}^{(N)})$ on the x axis. In the case of r random variables we would have N points in an r-dimensional space. The result of such a sample, however, could also be specified by a single point in a space of dimension rN. A sample of size 2 with a single variable could, for example, be depicted as a point in

Table 8.2: Student's difference test on the equality of means. Data from Example 8.2.

x	y
21	16
24	20
18	22
19	19
25	18
17	19
18	19
22	
21	
23	
18	
13	
16	
23	
22	
24	
$N_1 = 16$	$N_2 = 7$
$\bar{\mathsf{x}} = 20.3$	$\bar{\mathsf{y}} = 19.0$
$\mathsf{s}_\mathsf{x}^2 = 171.8/15$	$\mathsf{s}_\mathsf{y}^2 = 20/6$

a two-dimensional plane, spanned by the axes $\mathsf{x}^{(1)}, \mathsf{x}^{(2)}$. We will call such a space the E space. Every *hypothesis* H consists of an assumption about the probability density

$$f(x; \lambda_1, \lambda_2, \ldots, \lambda_p) = f(x; \boldsymbol{\lambda}) \quad . \tag{8.4.1}$$

The hypothesis is said to be *simple* if the function f is completely specified, i.e., if the hypothesis gives the values of all of the parameters λ_i. It is said to be *composite* if the general mathematical form of f is known, but the exact value of at least one parameter remains undetermined. A simple hypothesis could, for example, specify a standard Gaussian distribution. A Gaussian distribution with a mean of zero but an undetermined variance, however, is a composite hypothesis. The hypothesis H_0 is called the *null hypothesis*. Sometimes we will write explicitly

$$H_0(\boldsymbol{\lambda} = \boldsymbol{\lambda}_0) = H_0(\lambda_1 = \lambda_{10}, \lambda_2 = \lambda_{20}, \ldots, \lambda_p = \lambda_{p0}) \quad . \tag{8.4.2}$$

Other possible hypotheses are called *alternative hypotheses*, e.g.,

$$H_1(\boldsymbol{\lambda} = \boldsymbol{\lambda}_1) = H_1(\lambda_1 = \lambda_{11}, \lambda_2 = \lambda_{21}, \ldots, \lambda_p = \lambda_{p1}) \quad . \tag{8.4.3}$$

Often one wants to test a null hypothesis of the type (8.4.2) against a composite alternative hypothesis

$$H_1(\lambda \neq \lambda_0) = H_1(\lambda_1 \neq \lambda_{10}, \lambda_2 \neq \lambda_{20}, \ldots, \lambda_p \neq \lambda_{p0}) \quad . \tag{8.4.4}$$

Since the null hypothesis makes a statement about the probability density in the sample space, it also predicts the probability for observing a point $X = (\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N)})$ in the E space.[§] We now define a *critical region* S_c with the significance level α by the requirement

$$P(X \in S_c | H_0) = \alpha \quad , \tag{8.4.5}$$

i.e., we determine S_c such that the probability to observe a point X within S_c is α, under the assumption that H_0 is true. If the point X from the sample actually falls into the region S_c, then the hypothesis H_0 is rejected. One must note that the requirement (8.4.5) does not necessarily determine the critical region S_c uniquely.

Although using the E space is conceptually elegant, it is usually not very convenient for carrying out tests. Instead one constructs a *test statistic*

$$T = T(X) = T(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N)}) \tag{8.4.6}$$

and determines a region U of the variable T such that it corresponds to the critical region S_c, i.e., one performs the mapping

$$X \to T(X), \qquad S_c(X) \to U(X) \quad . \tag{8.4.7}$$

The null hypothesis is rejected if $T \in U$.

Because of the statistical nature of the sample, it is clearly possible that the null hypothesis could be true, even though it was rejected since $X \in S_c$. The probability for such an error, an *error of the first kind*, is equal to α. There is in addition another possibility to make a wrong decision, if one does not reject the hypothesis H_0 because X was not in the critical region S_c, even though the hypothesis was actually false and an alternative hypothesis was true. This is an *error of the second kind*. The probability for this,

$$P(X \notin S_c | H_1) = \beta \quad , \tag{8.4.8}$$

depends of course on the particular alternative hypotheses H_1. This connection provides us with a method to specify the critical region S_c. A test is clearly most reasonable if for a given significance level α the critical region is chosen such that the probability β for an error of the second kind is a

[§] Although X and the function $T(x)$ introduced below are random variables, we do not use for them a special character type.

minimum. The critical region and therefore the test itself naturally depend on the alternative hypothesis under consideration.

Once the critical region has been determined, we can consider the probability for rejecting the null hypothesis as a function of the "true" hypothesis, or rather as a function of the parameters that describe it. In analogy to (8.4.5), this is

$$M(S_c, \boldsymbol{\lambda}) = P(X \in S_c|H) = P(X \in S_c|\boldsymbol{\lambda}) \quad . \tag{8.4.9}$$

This probability is a function of S_c and of the parameters $\boldsymbol{\lambda}$. It is called the *power function* of a test. The complementary probability

$$L(S_c, \boldsymbol{\lambda}) = 1 - M(S_c, \boldsymbol{\lambda}) \tag{8.4.10}$$

is called the *acceptance probability* or the *operating characteristic function* of the test. It gives the probability to accept[¶] the null hypothesis. One clearly has

$$\begin{aligned} M(S_c, \boldsymbol{\lambda}_0) &= \alpha, & M(S_c, \boldsymbol{\lambda}_1) &= 1 - \beta, \\ L(S_c, \boldsymbol{\lambda}_0) &= 1 - \alpha, & L(S_c, \boldsymbol{\lambda}_1) &= \beta \quad . \end{aligned} \tag{8.4.11}$$

The *most powerful test* of a simple hypothesis H_0 with respect to the simple alternative hypothesis is defined by the requirement

$$M(S_c, \boldsymbol{\lambda}_1) = 1 - \beta = \max \quad . \tag{8.4.12}$$

Sometimes there exists a *uniformly most powerful test*, for which the requirement (8.4.12) holds for all possible alternative hypotheses.

A test is said to be *unbiased* if its power function is greater than or equal to α for all alternative hypotheses:

$$M(S_c, \boldsymbol{\lambda}_1) \geq \alpha \quad . \tag{8.4.13}$$

This definition is reasonable, since the probability to reject the null hypothesis is then a minimum if the null hypothesis is true. An *unbiased most powerful test* is the most powerful of all the unbiased tests. Correspondingly one can define a *unbiased uniformly most powerful test*. In the next sections we will learn the rules which sometimes allow one to construct tests with such desirable properties. Before turning to this task, we will first give an example to illustrate the definitions just introduced.

[¶]We use here the word "acceptance" of a hypothesis, although more precisely one should say, "There is no evidence to reject the hypothesis."

Example 8.3: Test of the hypothesis that a normal distribution with given variance σ^2 has the mean $\lambda = \lambda_0$

We wish to test the hypothesis $H_0(\lambda = \lambda_0)$. As a test statistic we use the arithmetic mean $\bar{\mathsf{x}} = \frac{1}{n}(\mathsf{x}_1 + \mathsf{x}_2 + \ldots + \mathsf{x}_n)$. (We will see in Example 8.4 that this is the most appropriate test statistic for our purposes.) From Sect. 6.2 we know that $\bar{\mathsf{x}}$ is normally distributed with mean λ and variance σ^2/n, i.e., that the probability density for $\bar{\mathsf{x}}$ for the case $\lambda = \lambda_0$ is given by

$$f(\bar{x}; \lambda_0) = \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \exp\left(-\frac{n}{2\sigma^2}(\bar{x} - \lambda_0)^2\right) \quad . \tag{8.4.14}$$

This is shown in Fig. 8.4 together with four different critical regions, all of which have the same significance level α.

These are the regions

$$\begin{aligned}
&U_1 : \bar{x} < \lambda^{\mathrm{I}} \text{ and } \bar{x} > \lambda^{\mathrm{II}} && \text{with } \textstyle\int_{-\infty}^{\lambda^{\mathrm{I}}} f(\bar{x})\,\mathrm{d}\bar{x} = \int_{\lambda^{\mathrm{II}}}^{\infty} f(\bar{x})\,\mathrm{d}\bar{x} = \frac{1}{2}\alpha\ , \\
&U_2 : \bar{x} > \lambda^{\mathrm{III}} && \text{with } \textstyle\int_{\lambda^{\mathrm{III}}}^{\infty} f(\bar{x})\,\mathrm{d}\bar{x} = \alpha\ , \\
&U_3 : \bar{x} < \lambda^{\mathrm{IV}} && \text{with } \textstyle\int_{-\infty}^{\lambda^{\mathrm{IV}}} f(\bar{x})\,\mathrm{d}\bar{x} = \alpha\ , \\
&U_4 : \lambda^{\mathrm{V}} \le \bar{x} < \lambda^{\mathrm{VI}} && \text{with } \textstyle\int_{\lambda^{\mathrm{V}}}^{\lambda_0} f(\bar{x})\,\mathrm{d}\bar{x} = \int_{\lambda_0}^{\lambda^{\mathrm{VI}}} f(\bar{x})\,\mathrm{d}\bar{x} = \frac{1}{2}\alpha\ .
\end{aligned}$$

In order to obtain the power functions for each of these regions, we must vary the mean value λ. The probability density of $\bar{\mathsf{x}}$ for an arbitrary value of λ is, in analogy to (8.4.14), given by

$$f(\bar{x}; \lambda) = \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \exp\left[-\frac{n}{2\sigma^2}(\bar{x} - \lambda)^2\right] \quad . \tag{8.4.15}$$

The dashed curve in Fig. 8.4b represents the probability density for $\lambda = \lambda_1 = \lambda_0 + 1$. The power function (8.4.9) is now simply

$$P(\bar{\mathsf{x}} \in U | \lambda) = \int_U f(\bar{x}; \lambda)\,\mathrm{d}\bar{x} \quad . \tag{8.4.16}$$

The power functions obtained in this way for the critical regions U_1, U_2, U_3, U_4 are shown in Fig. 8.4c for $n = 2$ (solid curve) and $n = 10$ (dashed curve).

We can now compare the effects of the four tests corresponding to the various critical regions. From Fig. 8.4c we immediately see that U_1 corresponds to an unbiased test, since the requirement (8.4.13) is clearly fulfilled. On the other hand, the test with the critical region U_2 is more powerful for the alternative hypothesis $H_1(\lambda_1 > \lambda_0)$, but is not good for $H_1(\lambda_1 < \lambda_0)$. For the test with U_3, the situation is exactly the opposite. Finally, the region U_4 provides a test for which the rejection probability is a maximum if the null hypothesis is true. Clearly this is very undesirable. The test was only constructed for demonstration purposes. If we compare the first three tests, we see that none of them are more powerful than the other two for all values of λ_1. Thus

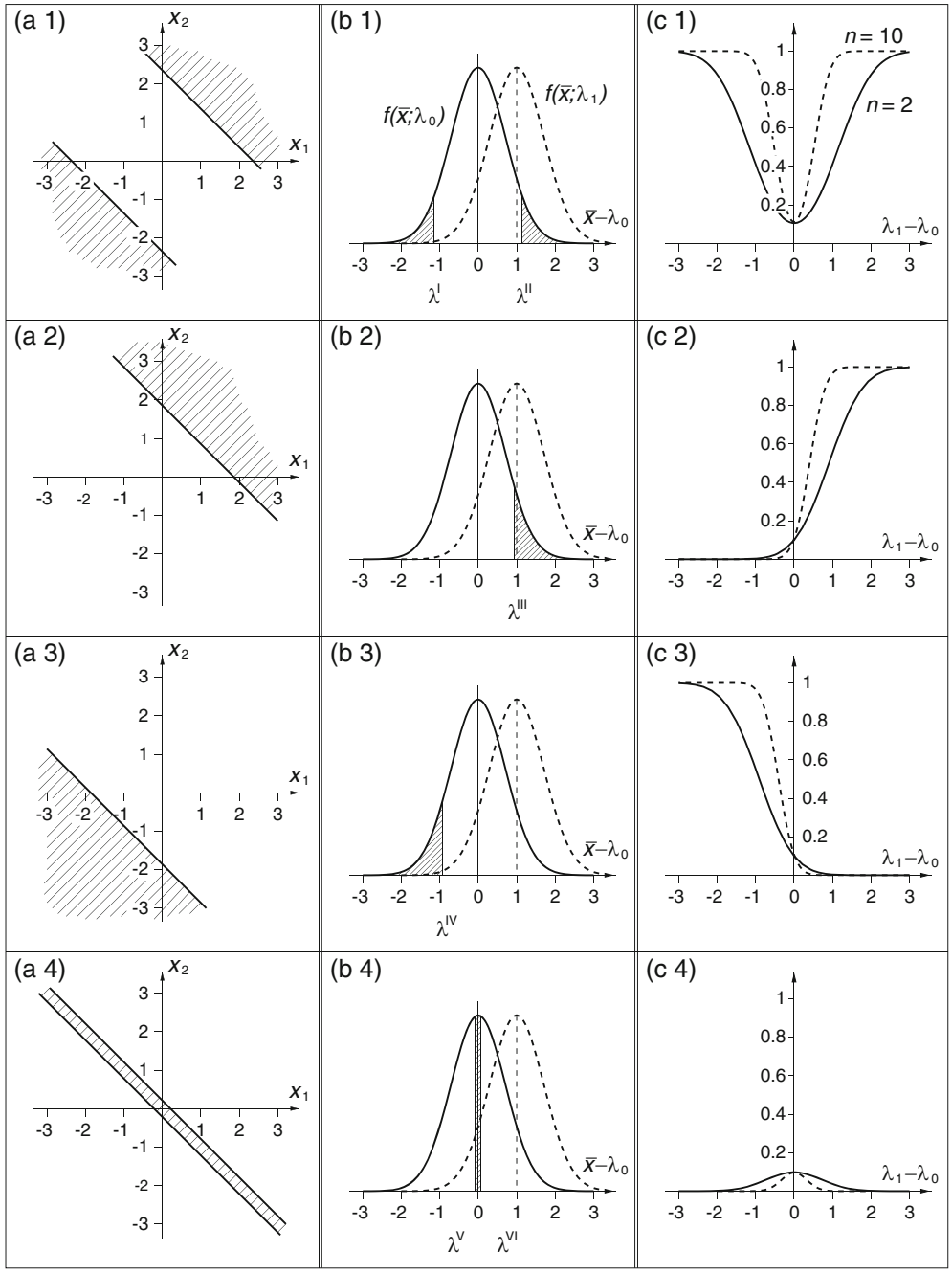

Fig. 8.4: (**a**) Critical regions in E space, (**b**) critical region of the test function, and (**c**) power function of the test from Example 8.3.

we have not succeeded in finding a uniformly most powerful test. In Example 8.4, where we will continue the discussion of the present example, we will determine that for this problem there does not exist a uniformly most powerful test. ■

8.5 The Neyman–Pearson Lemma and Applications

In the last section we introduced the E space, in which a sample is represented by a single point X. The probability to observe a point X within the critical region S_c – providing the null hypothesis H_0 is true – was defined in (8.4.5),

$$P(X \in S_c|H_0) = \alpha \quad . \tag{8.5.1}$$

We now define a conditional probability in E space,

$$f(X|H_0).$$

One clearly has

$$\int_{S_c} f(X|H_0)\,\mathrm{d}X = P(X \in S_c|H_0) = \alpha \quad . \tag{8.5.2}$$

The NEYMAN–PEARSON lemma states the following:

> A test of the simple hypothesis H_0 with respect to the simple alternative hypothesis H_1 is a most powerful test if the critical region S_c in E space is chosen such that
>
> $$\frac{f(X|H_0)}{f(X|H_1)} \begin{cases} \leq c \text{ for all } X \in S_c & , \\ \geq c \text{ for all } X \notin S_c & . \end{cases} \tag{8.5.3}$$
>
> Here c is a constant which depends on the significance level.

We will prove this by considering another region S along with S_c. It may partially overlap with S_c, as sketched in Fig. 8.5. We choose the size of the region S such that it corresponds to the same significance level, i.e.,

$$\int_S f(X|H_0)\,\mathrm{d}X = \int_{S_c} f(X|H_0)\,\mathrm{d}X = \alpha \quad .$$

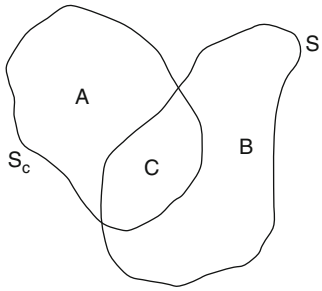

Fig. 8.5: The regions S and S_c.

Using the notation of Fig. 8.5, we can write

$$\begin{aligned}\int_A f(X|H_0)\,\mathrm{d}X &= \int_{S_c} f(X|H_0)\,\mathrm{d}X - \int_C f(X|H_0)\,\mathrm{d}X \\ &= \int_S f(X|H_0)\,\mathrm{d}X - \int_C f(X|H_0)\,\mathrm{d}X \\ &= \int_B f(X|H_0)\,\mathrm{d}X \quad .\end{aligned}$$

Since A is contained in S_c, we can use (8.5.3), i.e.,

$$\int_A f(X|H_0)\,\mathrm{d}X \leq c \int_A f(X|H_1)\,\mathrm{d}X \quad .$$

Correspondingly, since B is outside of S_c, one has

$$\int_B f(X|H_0)\,\mathrm{d}X \geq c \int_B f(X|H_1)\,\mathrm{d}X \quad .$$

We can now express the power function (8.4.9) using these integrals:

$$\begin{aligned}M(S_c, \boldsymbol{\lambda}_1) &= \int_{S_c} f(X|H_1)\,\mathrm{d}X = \int_A f(X|H_1)\,\mathrm{d}X + \int_C f(X|H_1)\,\mathrm{d}X \\ &\geq \frac{1}{c}\int_A f(X|H_0)\,\mathrm{d}X + \int_C f(X|H_1)\,\mathrm{d}X \\ &\geq \int_B f(X|H_1)\,\mathrm{d}X + \int_C f(X|H_1)\,\mathrm{d}X \\ &\geq \int_S f(X|H_1)\,\mathrm{d}X = M(S, \boldsymbol{\lambda}_1)\end{aligned}$$

or directly,

$$M(S_c, \boldsymbol{\lambda}_1) \geq M(S, \boldsymbol{\lambda}_1) \quad . \tag{8.5.4}$$

This is exactly the condition (8.4.12) for a uniformly most powerful test. Since we have not made any assumptions about the alternative hypothesis $H_1(\lambda = \lambda_1)$ or the region S, we have proven that the requirement (8.5.3) provides a uniformly most powerful test when it is fulfilled by the alternative hypothesis.

Example 8.4: Most powerful test for the problem of Example 8.3

We now continue with the ideas from Example 8.3, i.e., we consider tests with a sample of size N, obtained from a normal distribution with known variance σ^2 and unknown mean λ. The conditional probability density of a point $X = (\mathsf{x}^{(1)}, \mathsf{x}^{(2)}, \ldots, \mathsf{x}^{(N)})$ in E space is the joint probability density of the $\mathsf{x}^{(j)}$ for given values of λ, i.e.,

$$f(X|H_0) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{N}(\mathsf{x}^{(j)} - \lambda_0)^2\right] \tag{8.5.5}$$

for the null hypothesis and

$$f(X|H_1) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{N}(\mathsf{x}^{(j)} - \lambda_1)^2\right] \tag{8.5.6}$$

for the alternative hypothesis. The ratio (8.5.3) takes on the form

$$\begin{aligned} Q = \frac{f(X|H_0)}{f(X|H_1)} &= \exp\left[-\frac{1}{2\sigma^2}\left\{\sum_{j=1}^{N}(\mathsf{x}^{(j)} - \lambda_0)^2 - \sum_{j=1}^{N}(\mathsf{x}^{(j)} - \lambda_1)^2\right\}\right] \\ &= \exp\left[-\frac{1}{2\sigma^2}\left\{N(\lambda_0^2 - \lambda_1^2) - 2(\lambda_0 - \lambda_1)\sum_{j=1}^{N}\mathsf{x}^{(j)}\right\}\right] . \end{aligned}$$

The expression

$$\exp\left[-\frac{N}{2\sigma^2}(\lambda_0^2 - \lambda_1^2)\right] = k \geq 0$$

is a non-negative constant. The condition (8.5.3) thus has the form

$$k\exp\left[\frac{\lambda_0 - \lambda_1}{\sigma^2}\sum_{j=1}^{N}\mathsf{x}^{(j)}\right] \begin{cases} \leq c, & X \in S_c \quad , \\ \geq c, & X \notin S_c \quad . \end{cases}$$

This is the same as

$$(\lambda_0 - \lambda_1)\bar{\mathsf{x}} \begin{cases} \leq c', & X \in S_c \quad , \\ \geq c', & X \notin S_c \quad . \end{cases} \tag{8.5.7}$$

Here c' is a constant different from c. Equation (8.5.7) is, however, not only a condition for S_c, but also specifies directly that $\bar{\mathsf{x}}$ should be used as the test variable. For each given λ_1, i.e., for every simple alternative hypothesis $H_1(\lambda = \lambda_1)$, (8.5.7) gives a clear prescription for the choice of S_c or U, i.e., for the critical region and the test variable $\bar{\mathsf{x}}$.

For the case $\lambda_1 < \lambda_0$, the relation (8.5.7) becomes

$$\bar{\mathsf{x}} \begin{cases} \leq c'', & X \in S_c \quad , \\ \geq c'', & X \notin S_c \quad . \end{cases}$$

This corresponds to the situation in Fig. 8.4b3 with $c'' = \lambda^{IV}$. Similarly, for every alternative hypothesis with $\lambda_1 > \lambda_0$, the critical region of the most powerful test is given by

$$\bar{\mathsf{x}} \geq c'''$$

(see Fig. 8.4b2 with $c''' = \lambda'''$). There does not exist a uniformly most powerful test, since the factor $(\lambda_0 - \lambda_1)$ in Eq. (8.5.7) changes sign at $\lambda_1 = \lambda_0$. ■

8.6 The Likelihood-Ratio Method

The Neyman–Pearson lemma gave the condition for a uniformly most powerful test. Such a test did not exist if the alternative hypothesis included parameter values that could be both greater and less than that of the null hypothesis. We determined this in Example 8.4; it can be shown, however, that it is true in general. The question thus arises as to what test should be used when no uniformly most powerful test exists. Clearly this question is not formulated precisely enough to allow a unique answer. We would like in the following to give a prescription that allows us to construct tests that have desirable properties and that have the advantage of being relatively easy to use.

We consider from the beginning the general case with p parameters $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_p)$. The result of a sample, i.e., the point $X = (\mathsf{x}^{(1)}, \mathsf{x}^{(2)}, \ldots, \mathsf{x}^{(N)})$ in E space is to be used to test a given hypothesis. The (composite) null hypothesis is characterized by a given region for each parameter. We can use a p-dimensional space, with the $\lambda_1, \lambda_2, \ldots, \lambda_p$ as coordinate axes, and consider the region allowed by the null hypothesis as a region in this parameter space, called ω. We denote the region in this space representing all possible values of the parameters by Ω. The most general alternative hypothesis is then the part of Ω that does not contain ω. We denote this by $\Omega - \omega$. Recall now from Chap. 7 the maximum-likelihood estimator $\tilde{\lambda}$ for a parameter λ. It is that value of λ for which the likelihood function is a maximum. In Chap. 7 we tacitly assumed that one searched for the maximum in the entire allowable parameter space. In the following we will consider maxima in a restricted region (e.g., in ω). We write in this case $\tilde{\lambda}^{(\omega)}$. The *likelihood-ratio test* defines a test statistic

$$T = \frac{f(\mathsf{x}^{(1)}, \mathsf{x}^{(2)}, \ldots, \mathsf{x}^{(N)}; \tilde{\lambda}_1^{(\Omega)}, \tilde{\lambda}_2^{(\Omega)}, \ldots, \tilde{\lambda}_p^{(\Omega)})}{f(\mathsf{x}^{(1)}, \mathsf{x}^{(2)}, \ldots, \mathsf{x}^{(N)}; \tilde{\lambda}_1^{(\omega)}, \tilde{\lambda}_2^{(\omega)}, \ldots, \tilde{\lambda}_p^{(\omega)})} \quad . \tag{8.6.1}$$

Here $f(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N)}; \lambda_1, \lambda_2, \ldots, \lambda_p)$ is the joint probability density of the $\mathbf{x}^{(j)}$ $(j = 1, 2, \ldots, N)$, i.e., the likelihood function (7.1.5). The procedure of the likelihood ratio test prescribes that we reject the null hypothesis if

$$T > T_{1-\alpha} \quad . \tag{8.6.2}$$

Here $T_{1-\alpha}$ is defined by

$$P(T > T_{1-\alpha}|H_0) = \int_{T_{1-\alpha}}^{\infty} g(T|H_0)\,\mathrm{d}T \quad , \tag{8.6.3}$$

and $g(T|H_0)$ is the conditional probability density for the test statistic T. The following theorem by WILKS [9] concerns the distribution function of T (or actually $-2\ln T$) in the limiting case of very large samples:

> If a population is described by the probability density $f(x; \lambda_1, \lambda_2, \ldots, \lambda_p)$ that satisfies reasonable requirements of continuity, and if $p - r$ of the p parameters are fixed by the null hypothesis, while r parameters remain free, then the statistic $-2\ln T$ follows a χ^2-distribution with $p - r$ degrees of freedom for very large samples, i.e., for $N \to \infty$.

We now apply this method to the problem of Examples 8.3 and 8.4, i.e., we consider tests with samples from a normally distributed population with known variance and unknown mean.

Example 8.5: Power function for the test from Example 8.3

For the simple hypothesis $H_0(\lambda = \lambda_0)$, the region ω of the parameter space is reduced to the point $\lambda = \lambda_0$. We have thus

$$\tilde{\lambda}^{(\omega)} = \lambda_0 \quad . \tag{8.6.4}$$

If we consider the most general alternative hypothesis $H_1(\lambda = \lambda_1 \neq \lambda_0)$, then we obtain as the maximum-likelihood estimator of λ the sample mean $\bar{\mathbf{x}}$. The likelihood ratio (8.6.1) thus becomes

$$T = \frac{f(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N)}; \bar{\mathbf{x}})}{f(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N)}; \lambda_0)} \quad . \tag{8.6.5}$$

The joint probability density is given by (7.2.6),

$$f(x^{(1)}, x^{(2)}, \ldots, x^{(N)}) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^N \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{N}(x^{(j)} - \lambda)^2\right] \quad . \tag{8.6.6}$$

Therefore,

$$
\begin{aligned}
T &= \exp\left[\frac{1}{2\sigma^2}\left\{-\sum_{j=1}^{N}(\mathsf{x}^{(j)}-\bar{\mathsf{x}})^2+\sum_{j=1}^{N}(\mathsf{x}^{(j)}-\lambda_0)^2\right\}\right] \\
&= \exp\left[\frac{1}{2\sigma^2}\sum_{j=1}^{N}(\bar{\mathsf{x}}-\lambda_0)^2\right] \\
&= \exp\left[\frac{N}{2\sigma^2}(\bar{\mathsf{x}}-\lambda_0)^2\right] \quad .
\end{aligned}
$$

We must now calculate $T_{1-\alpha}$ and reject the hypothesis H_0 if the inequality (8.6.2) is fulfilled. Since the logarithm of T is a monotonic function of T, we can use

$$T' = 2\ln T = \frac{N}{\sigma^2}(\bar{\mathsf{x}}-\lambda_0)^2 \tag{8.6.7}$$

as the test statistic and reject H_0 if

$$T' > T'_{1-\alpha}$$

with

$$\int_{T'_{1-\alpha}}^{\infty} h(T'|H_0)\,\mathrm{d}T' = \alpha \quad .$$

In order to calculate the probability density $h(T'|H_0)$ of T', we start with the density $f(\bar{x})$ for the sample mean with the condition $\lambda = \lambda_0$,

$$f(\bar{x}|H_0) = \sqrt{\frac{N}{2\pi\sigma^2}}\exp\left(-\frac{N}{2\sigma^2}(\bar{x}-\lambda_0)^2\right) \quad .$$

In order to carry out the transformation of variables (3.7.1), we need in addition the derivative,

$$\left|\frac{\mathrm{d}\bar{x}}{\mathrm{d}T'}\right| = \frac{1}{2}\sqrt{\frac{\sigma^2}{N}}T'^{-1/2} \quad ,$$

which can be easily obtained from (8.6.7). One then has

$$h(T'|H_0) = \left|\frac{\mathrm{d}\bar{x}}{\mathrm{d}T'}\right| f(\bar{x}|H_0) = \frac{1}{\sqrt{2\pi}}T'^{-1/2}\mathrm{e}^{-T'/2} \quad . \tag{8.6.8}$$

This is indeed a χ^2-distribution for one degree of freedom. Thus in our example, WILKS' theorem holds even for finite N. We see, therefore, that the likelihood-ratio test yields the unbiased test of Fig. 8.4b1. The test

$$T' = \frac{N}{\sigma^2}(\bar{x}-\lambda_0)^2 > T'_{1-\alpha}$$

is equivalent to

$$\left(\frac{N}{\sigma^2}\right)^{1/2} |\bar{x} - \lambda_0| < \lambda', \qquad \left(\frac{N}{\sigma^2}\right)^{1/2} |\bar{x} - \lambda_0| > \lambda'' \tag{8.6.9}$$

with

$$-\lambda' = \lambda'' = (T'_{1-\alpha})^{1/2} = (\chi^2_{1-\alpha})^{1/2} = \chi_{1-\alpha} \quad .$$

We can use this result to compute explicitly the power function of our test. For a given value of the population mean λ, the probability density for the sample mean is

$$f(\bar{x};\lambda) = \left(\frac{N}{2\pi\sigma^2}\right)^{1/2} \exp\left[-\frac{N(\bar{x}-\lambda)^2}{2\sigma^2}\right] = \phi_0\left(\frac{\bar{x}-\lambda}{\sigma/\sqrt{N}}\right) \quad .$$

Using (8.4.9) and (8.6.9) we obtain

$$\begin{aligned} M(S_c;\lambda) &= \int_{-\infty}^{A} f(\bar{x};\lambda)\,\mathrm{d}\bar{x} + \int_{B}^{\infty} f(\bar{x};\lambda)\,\mathrm{d}\bar{x} \qquad (8.6.10)\\ &= \psi_0\left(\chi_{1-\alpha} - \frac{\lambda-\lambda_0}{\sigma/\sqrt{N}}\right) + \psi_0\left(\chi_{1-\alpha} + \frac{\lambda-\lambda_0}{\sigma/\sqrt{N}}\right) \quad ,\\ A &= -\chi_{1-\alpha}\sigma/\sqrt{N} - \lambda_0, \qquad B = \chi_{1-\alpha}\sigma/\sqrt{N} - \lambda_0 \quad . \end{aligned}$$

Here ϕ_0 and ψ_0 are the probability density and distribution function of the standard normal distribution. The power function (8.6.10) is shown in Fig. 8.6 for $\alpha = 0.05$ and various values of N/σ^2. ■

Example 8.6: Test of the hypothesis that a normal distribution of unknown variance has the mean value $\lambda = \lambda_0$

In this case the null hypothesis $H_0(\lambda = \lambda_0)$ is composite, i.e., it makes no statement about σ^2. From Example 7.8 we know the maximum-likelihood estimator in the full parameter space,

$$\tilde{\lambda}^{(\Omega)} = \bar{\mathsf{x}} \quad ,$$
$$\tilde{\sigma}^{2(\Omega)} = \frac{1}{N}\sum_{j=1}^{N}(\mathsf{x}^{(j)} - \bar{\mathsf{x}})^2 = \mathsf{s}'^2 \quad .$$

In the parameter space of the null hypothesis we have

$$\tilde{\lambda}^{(\omega)} = \lambda_0,$$
$$\tilde{\sigma}^{2(\omega)} = \frac{1}{N}\sum_{j=1}^{N}(\mathsf{x}^{(j)} - \lambda_0)^2 \quad .$$

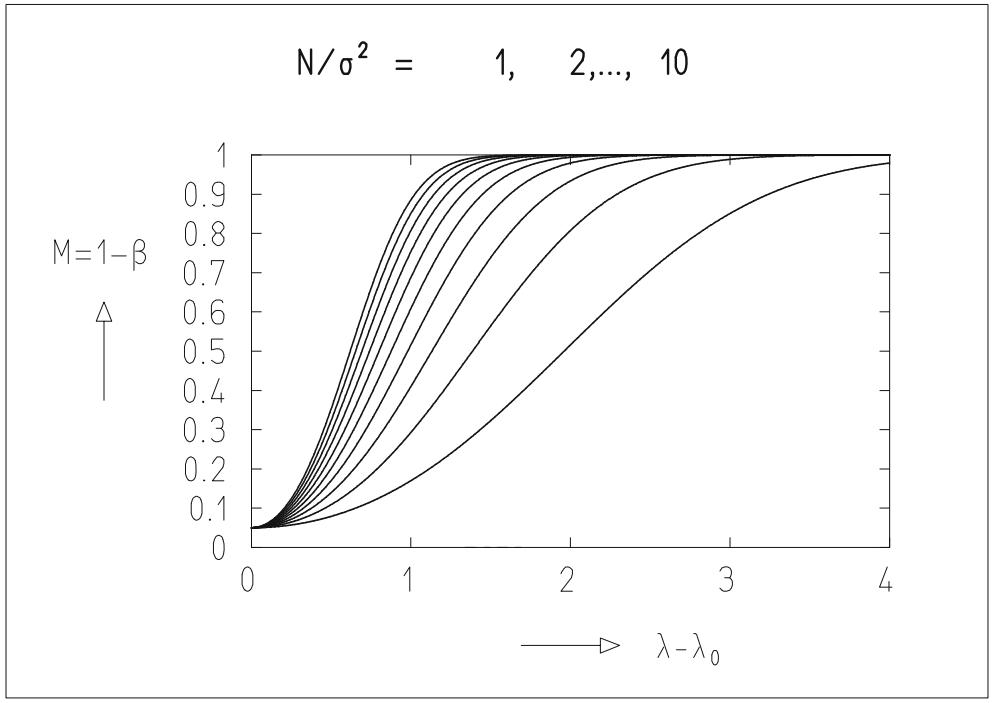

Fig. 8.6: Power function of the test from Example 8.5. The right-most curve corresponds to $N/\sigma^2 = 1$.

The likelihood ratio (8.6.1) is then

$$\begin{aligned} T &= \left(\frac{\sum(\mathsf{x}^{(j)}-\lambda_0)^2}{\sum(\mathsf{x}^{(j)}-\bar{\mathsf{x}})^2}\right)^{N/2} \exp\left(-\frac{N}{2}\frac{\sum(\mathsf{x}^{(j)}-\bar{\mathsf{x}})^2}{\sum(\mathsf{x}^{(j)}-\bar{\mathsf{x}})^2}+\frac{N}{2}\frac{\sum(\mathsf{x}^{(j)}-\lambda_0)^2}{\sum(\mathsf{x}^{(j)}-\lambda_0)^2}\right) \\ &= \left(\frac{\sum(\mathsf{x}^{(j)}-\lambda_0)^2}{\sum(\mathsf{x}^{(j)}-\bar{\mathsf{x}})^2}\right)^{N/2} \quad . \end{aligned}$$

We transform again to a different test statistic T' that is a monotonic function of T,

$$\begin{aligned} T' &= T^{2/N} = \frac{\sum(\mathsf{x}^{(j)}-\lambda_0)^2}{\sum(\mathsf{x}^{(j)}-\bar{\mathsf{x}})^2} = \frac{\sum(\mathsf{x}^{(j)}-\bar{\mathsf{x}})^2+N(\bar{\mathsf{x}}-\lambda_0)^2}{\sum(\mathsf{x}^{(j)}-\bar{\mathsf{x}})^2} \quad , (8.6.11) \\ T' &= 1+\frac{t^2}{N-1} \quad , \end{aligned}$$

where

$$t = \sqrt{N}\frac{\bar{\mathsf{x}}-\lambda_0}{\left(\frac{\sum(\mathsf{x}^{(j)}-\bar{\mathsf{x}})^2}{N-1}\right)^{1/2}} = \sqrt{N}\frac{\bar{\mathsf{x}}-\lambda_0}{\mathsf{s}_x} = \frac{\bar{\mathsf{x}}-\lambda_0}{\mathsf{s}_{\bar{\mathsf{x}}}} \qquad (8.6.12)$$

is Student's test variable introduced in Sect. 8.3. From (8.6.11) we can compute a value of t for a given sample and reject the null hypothesis if

$$|t| > t_{1-\frac{1}{2}\alpha} \quad .$$

The very generally formulated method of the likelihood ratio has led us to Student's test, which was originally constructed for tests with samples from a normal distribution with known mean and unknown variance. ■

8.7 The χ^2-Test for Goodness-of-Fit

8.7.1 χ^2-Test with Maximal Number of Degrees of Freedom

Suppose one has N measured values g_i, $i = 1, 2, \ldots, N$, each with a known measurement error σ_i. The meaning of the measurement error is the following: g_i is a measurement of the (unknown) true quantity h_i. One has

$$g_i = h_i + \varepsilon_i, \qquad i = 1, 2, \ldots, N \quad . \tag{8.7.1}$$

Here the deviation ε_i is a random variable that follows a normal distribution with mean 0 and standard deviation σ_i.

We now want to test the hypothesis specifying the values h_i on which the measurement is based,

$$h_i = f_i, \qquad i = 1, 2, \ldots, N \quad . \tag{8.7.2}$$

If this hypothesis is true, then all of the quantities

$$u_i = \frac{g_i - f_i}{\sigma_i}, \qquad i = 1, 2, \ldots, N \quad , \tag{8.7.3}$$

follow the standard Gaussian distribution. Therefore,

$$T = \sum_{i=1}^{N} u_i^2 = \sum_{i=1}^{N} \left(\frac{g_i - f_i}{\sigma_i} \right)^2 \tag{8.7.4}$$

follows a χ^2-distribution for N degrees of freedom. If the hypothesis (8.7.2) is false, then the individual deviations of the measured values g_i from the values predicted by the hypothesis f_i, normalized by the errors σ_i, (8.7.3), will be greater. For a given significance level α, the hypothesis (8.7.2) is rejected if

$$T > \chi^2_{1-\alpha} \quad , \tag{8.7.5}$$

i.e., if the quantity (8.7.4) is greater than the quantile $\chi^2_{1-\alpha}$ of the χ^2-distribution for N degrees of freedom.

8.7.2 χ^2-Test with Reduced Number of Degrees of Freedom

The number of degrees of freedom is reduced when the hypothesis to be tested is less explicit than (8.7.2). For this case we consider the following example. Suppose a quantity g can be measured as a function of an independent *controlled* variable t, which itself can be set without error,

$$g = g(t) \quad .$$

The individual measurements g_i correspond to given fixed values t_i of the independent variable. The corresponding true quantities h_i are given by some function

$$h_i = h(t_i) \quad .$$

A particularly simple hypothesis for this function is the linear equation

$$f(t) = h(t) = at + b \quad . \tag{8.7.6}$$

The hypothesis can in fact include specifying the numerical values for the parameters a and b. In this case, all values f_i in (8.7.2) are exactly known, and the quantity (8.7.4) follows – if the hypothesis is true – a χ^2-distribution for N degrees of freedom.

The hypothesis may, however, only state: There exists a linear relationship (8.7.6) between the controlled variable t and the variable h. The numerical values of the parameters a and b are, however, unknown. In this case one constructs estimators $\tilde{a}$, $\tilde{b}$ for the parameters, which are functions of the measurements g_i and the errors σ_i. The hypothesis (8.7.2) is then

$$h_i = h(t_i) = f_i = \tilde{a}t_i + \tilde{b} \quad .$$

Since, however, $\tilde{a}$ and $\tilde{b}$ are functions of the measurements g_i, the normalized deviations u_i in (8.7.3) are no longer all independent. Therefore the number of degrees of freedom of the χ^2-distribution for the sum of squares (8.7.4) is reduced by 2 to $N-2$, since the determination of the two quantities $\tilde{a}$, $\tilde{b}$ introduces two equations of constraint between the quantities u_i.

8.7.3 χ^2-Test and Empirical Frequency Distribution

Suppose we have a distribution function $F(x)$ and its probability density $f(x)$. The full region of the random variable x can be divided into r intervals

$$\xi_1, \xi_2, \ldots, \xi_i, \ldots, \xi_r \quad ,$$

as shown in Fig. 8.7. By integrating $f(x)$ over the individual intervals we obtain the probability to observe x in ξ_i,

$$p_i = P(\mathsf{x} \in \xi_i) = \int_{\xi_i} f(x)\,\mathrm{d}x; \qquad \sum_{i=1}^{r} p_i = 1 \quad . \tag{8.7.7}$$

We now take a sample of size n and denote by n_i the number of elements of the sample that fall into the interval ξ_i. An appropriate graphical representation of the sample is a histogram, as described in Sect. 6.3.

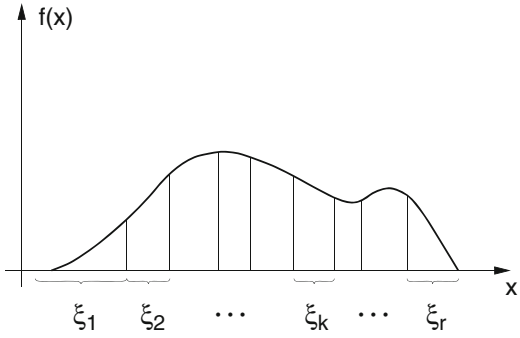

Fig. 8.7: Dividing the range of the variable x into the intervals ξ_k.

One clearly has

$$\sum_{i=1}^{r} n_i = n \quad . \tag{8.7.8}$$

From the (hypothetical) probability density for the population we would have expected the value

$$n\,p_i$$

for n_i. For large values of n_i, the variance of n_i is equal to n_i (Sect. 6.8), and the distribution of the quantity u_i with

$$u_i^2 = \frac{(n_i - n\,p_i)^2}{n_i} \tag{8.7.9}$$

becomes approximately – if the hypothesis is true – a standard Gaussian distribution. This holds also if one uses the expected variances $n\,p_i$ instead of the observed quantities n_i in the denominator of (8.7.9),

$$u_i^2 = \frac{(n_i - n\,p_i)^2}{n\,p_i} \quad . \tag{8.7.10}$$

If we now construct the sum of squares of the u_i for all intervals,

$$X^2 = \sum_{i=1}^{r} u_i^2 \quad , \tag{8.7.11}$$

then we expect (for large n) that this follows a χ^2-distribution if the hypothesis is true. The number of degrees of freedom is $r-1$, since the u_i are not independent because of (8.7.8). The number of degrees of freedom is reduced to $r-1-p$ if, in addition, p parameters are determined from the observations.

Example 8.7: χ^2-test for the fit of a Poisson distribution to an empirical frequency distribution

In an experiment investigating photon-proton interactions, a beam of high energy photons (γ-quanta) impinge on a hydrogen bubble chamber. The processes by which a photon materializes in the chamber, electron-positron pair conversion, are counted in order to obtain a measure of the intensity of the photon beam. The frequency of cases in which 0,1,2,... pairs are observed simultaneously, i.e., in the same bubble-chamber photograph, follows a Poisson distribution (see Example 5.3). Deviations from the Poisson distribution provide information about measurement losses, which are important for uncovering systematic errors. The results of observing $n = 355$ photographs are given in column 2 of Table 8.3 and in Fig. 8.8. From Example 7.4, we know that the maximum-likelihood estimator of the parameter of the Poisson distribution is given by $\tilde{\lambda} = \sum_k k\, n_k / \sum_k n_k$. We find $\tilde{\lambda} = 2.33$. The values p_k of the Poisson distribution with this parameter multiplied by n are given in column 3. By summing the squared terms in column 4 one obtains the value $X^2 = 10.44$. The problem has six degrees of freedom, since $r = 8$, $p = 1$. We chose $\alpha = 1\%$ and find $\chi^2_{0.99} = 16.81$ from Table I.7. We therefore have no reason to reject the hypothesis of a Poisson distribution. ■

Table 8.3: Data for the χ^2-test from Example 8.7.

Number of electron pairs per photograph k	Number of photographs with k electron pairs n_k	Prediction of Poisson distribution $n\,p_k$	$\frac{(n_k - n\,p_k)^2}{n\,p_k}$
0	47	34.4	4.61
1	69	80.2	1.56
2	84	93.7	1.00
3	76	72.8	0.14
4	49	42.6	0.96
5	16	19.9	0.76
6	11	7.8	1.31
7	3	2.5	0.10
8	—	(0.7)	
	$n = \sum n_k = 355$		$X^2 = 10.44$

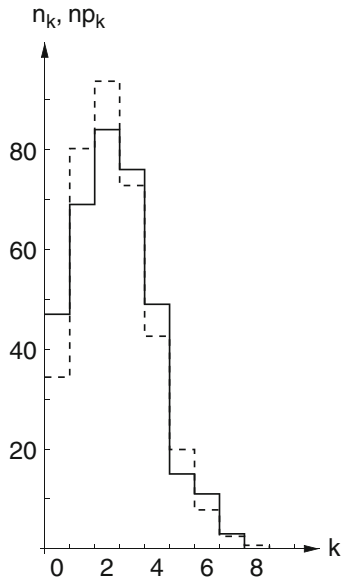

Fig. 8.8: Comparison of the experimental distribution n_k (histogram with *solid line*) from Example 8.7 with the Poisson distribution np_k (*dashed line*).

8.8 Contingency Tables

Suppose n experiments have been carried out whose results are characterized by the values of two random variables x and y. We consider the two variables as discrete, being able to take on the values $x_1, x_2, \ldots, x_k$; $y_1, y_2, \ldots, y_\ell$. Continuous variables can be approximated by discrete ones by dividing their range into intervals, as shown in Fig. 8.7. Let the number of times the result $\mathsf{x} = x_i$ and $\mathsf{y} = y_j$ is observed be n_{ij}. One can arrange the numbers n_{ij} in a matrix, called a *contingency table* (Table 8.4).

Table 8.4: Contingency table.

	y_1	y_2	$\ldots$	y_ℓ
x_1	n_{11}	n_{12}	$\ldots$	$n_{1\ell}$
x_2	n_{21}	n_{22}	$\ldots$	$n_{2\ell}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
x_k	n_{k1}	n_{k2}	$\ldots$	$n_{k\ell}$

We denote by p_i the probability for $\mathsf{x} = x_i$ to occur, and by q_j the probability for $\mathsf{y} = y_j$. If the variables are independent, then the probability to simultaneously observe $\mathsf{x} = x_i$ and $\mathsf{y} = y_j$ is equal to the product $p_i q_j$. The maximum-likelihood estimators for p and q are

$$\tilde{p}_i = \frac{1}{n}\sum_{j=1}^{\ell} n_{ij}, \qquad \tilde{q}_j = \frac{1}{n}\sum_{i=1}^{k} n_{ij} \quad .$$

Since

$$\sum_{j=1}^{\ell} \tilde{q}_j = \sum_{i=1}^{k} \tilde{p}_i = \frac{1}{n} \sum_{j=1}^{\ell} \sum_{i=1}^{k} n_{ij} = 1 \quad ,$$

one has $k + \ell - 2$ independent estimators $\tilde{p}_i$, $\tilde{q}_j$. We can now organize the elements of the contingency table into a single line,

$$n_{11}, n_{12}, \ldots, n_{1\ell}, n_{21}, n_{22}, \ldots, n_{2\ell}, \ldots, n_{k\ell} \quad ,$$

and carry out a χ^2-test. For this we must compute the quantity

$$X^2 = \sum_{i=1}^{k} \sum_{j=1}^{\ell} \frac{(n_{ij} - n\,\tilde{p}_i\,\tilde{q}_j)^2}{n\,\tilde{p}_i\,\tilde{q}_j} \tag{8.8.1}$$

and compare it to the quantile $\chi^2_{1-\alpha}$ of the χ^2-distribution corresponding to a given significance level α. The number of degrees of freedom is still obtained from the number of intervals minus the number of estimated parameters minus one,

$$f = k\ell - 1 - (k + \ell - 2) = (k-1)(\ell - 1) \quad .$$

If the variables are not independent, then n_{ij} will not, in general, be near $n\,\tilde{p}_i\,\tilde{q}_j$, i.e., one will find

$$X^2 > \chi^2_{1-\alpha} \tag{8.8.2}$$

and the hypothesis will be rejected.

8.9 2 × 2 Table Test

The simplest nontrivial contingency table has only two rows and two columns, and is called a 2×2 table, as shown in Table 8.5. It is often used in medical studies. (The variables x_1 and x_2 could represent, for example, two different treatment methods, and y_1 and y_2 could represent success and failure of the treatment. One wishes to determine whether success is independent of the treatment.)

Table 8.5: 2×2 table.

	y_1	y_2
x_1	$n_{11} = a$	$n_{12} = b$
x_2	$n_{21} = c$	$n_{22} = d$

One computes the quantity X^2 either according to (8.8.1) or from the formula

$$X^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \quad ,$$

which is obtained by rearranging (8.8.1). If the variables x and y are independent, then X^2 follows a χ^2-distribution with one degree of freedom. One rejects the hypothesis of independence at the significance level α if

$$X^2 > \chi^2_{1-\alpha} \quad .$$

In order for the quantity X^2 to actually follow a χ^2-distribution is again necessary that the individual n_{ij} are sufficiently large, (and the hypothesis of independence must be true).

8.10 Example Programs

Example Program 8.1: The class E1Test generates samples and tests the equality of their variances

The program performs a total of n_{exp} simulated experiments. Each experiment consists of the simulation of two samples of sizes N_1 and N_2 from normal distributions with standard deviations σ_1 and σ_2. The variance of each of the samples is computed using the class Sample. The sample variances are called s^2_g and s^2_k, so that

$$\mathsf{s}^2_g > \mathsf{s}^2_k \quad .$$

From the corresponding sample sizes the numbers of degrees of freedom $f_g = N_g - 1$ and $f_k = N_k - 1$ are computed. Finally, the ratio $\mathsf{s}^2_g/\mathsf{s}^2_k$ is compared with the quantile $F_{1-\alpha/2}(f_g, f_k)$ at a given confidence level $\beta = 1-\alpha$. If the ratio is larger than the quantile, then the hypothesis of equal variances has to be rejected. The program asks for the quantities n_{exp}, N_1, N_2, σ_1, σ_2, and β. For each simulated experiment one line of output is displayed.

Suggestions: Choose $n_{\text{exp}} = 20$ and $\beta = 0.9$. (a) For $\sigma_1 = \sigma_2$ you would expect the hypothesis to be rejected in 2 out of 20 cases because of an error of the first kind. Note the large statistical fluctuations, which obviously depend on N_1 and N_2, and choose different pairs of values N_1, N_2 for $\sigma_1 = \sigma_2$. (b) Check the power of the test for different variances $\sigma_1 \neq \sigma_2$.

Example Program 8.2: The class E2Test generates samples and tests the equality of their means with a given value using Student's Test

This short program performs n_{exp} simulation experiments. In each experiment a sample of size N is drawn from a normal distribution with mean x_0 and width σ. Using the class Sample the sample mean $\bar{\mathsf{x}}$ and the sample variance s^2_x are determined. If λ_0 is the population mean specified by the hypothesis, then the quantity

$$|t| = \frac{|\bar{\mathsf{x}} - \lambda_0|\sqrt{N}}{\mathsf{s_x}}$$

can be used to test the hypothesis. At a given confidence level $\beta = 1 - \alpha$ the hypothesis is rejected if

$$|t| > t_{1-\alpha/2} \quad .$$

Here $t_{1-\alpha/2}$ is the quantile of Student's distribution with $f = N - 1$ degrees of freedom. The program asks for the quantities n_{exp}, N, x_0, σ, λ_0, and β. For each simulated experiment one line of output is displayed.

Suggestion: Modify the suggestions at the end of Sect. 8.1 to apply to Student's test.

Example Program 8.3: The class E3Test generates samples and computes the test statistic χ^2 for the hypothesis that the samples are taken from a normal distribution with known parameters

For samples of size N the hypothesis H_0 that they stem from a normal distribution with mean a_0 and standard deviation σ_0 is tested. A total of n_{exp} samples are drawn in simulated experiments from a normally distributed population with mean a and standard deviation σ. For each sample the quantity

$$\mathsf{X}^2 = \sum_{i=1}^{N} \left(\frac{\mathsf{x}_i - a_0}{\sigma_0} \right)^2 \tag{8.10.1}$$

is computed. Here x_i are the elements of the sample. If $a = a_0$ and $\sigma = \sigma_0$, then the quantity X^2 follows a χ^2-distribution for N degrees of freedom. This quantity can therefore be used to perform a χ^2-test on the hypothesis H_0. The program does not, however, perform the χ^2-test, but rather it displays a histogram of the quantity X^2 together with the χ^2-distribution. One observes that for $a = a_0$ and $\sigma = \sigma_0$ the histogram and χ^2-distribution indeed coincide within statistical fluctuations. If, however, $a \neq a_0$ and/or $\sigma \neq \sigma_0$, then deviations appear. These deviations become particularly clear if instead of X^2 the quantity

$$P(\mathsf{X}^2) = 1 - F(\mathsf{X}^2; N) \tag{8.10.2}$$

is displayed. Here $F(\mathsf{X}^2, N)$ is the distribution function (C.5.2) of the χ^2-distribution for N degrees of freedom. $F(\mathsf{X}^2, N)$ is equal to the probability that a random variable drawn from a χ^2-distribution is smaller than X^2. Thus, P is the probability that it is greater than or equal to X^2. If the hypothesis H_0 is true, then F and therefore also P follow uniform distributions between 0 and 1. If, however, H_0 is false, then the distribution of the X^2 is not a χ^2-distribution, and the distribution of the P is not a uniform distribution. The test statistic X^2 often (not completely correctly) is simply called "χ^2" and the quantity P is then called the "χ^2-probability". Large values of X^2 obviously signify that the terms in the sum (8.10.1) are on the average large compared to unity, i.e., that the x_i are significantly different from a_0. For large values of X^2, however, P becomes small, cf. (8.10.2). Large values of "χ^2" therefore correspond to small values of the "χ^2-probability". The hypothesis H_0 is rejected at the confidence level $\beta = 1 - \alpha$ if $\mathsf{X}^2 > \chi^2_{1-\alpha}(N)$. That is equivalent to $F(\mathsf{X}^2, N) > \beta$ or $P < \alpha$.

The program allows for interactive input of the quantities n_{exp}, N, a, a_0, σ, σ_0 and displays the distributions of both X^2 and $P(\mathsf{X}^2)$ in the form of histograms.

Suggestions: (a) Choose $n_{\text{exp}} = 1000$; $a = a_0 = 0$, $\sigma = \sigma_0 = 1$ and for $N = 1$, $N = 2$, and $N = 10$ display both X^2 and $P(\mathsf{X}^2)$. (b) Repeat (a), keeping $a = 0$ and choosing $a_0 = 1$ and $a_0 = 5$. Explain the shift of the histogram for $P(\mathsf{X}^2)$. (c) Repeat (a), keeping $\sigma = 1$ fixed and choosing $\sigma_0 = 0.5$ and $\sigma_0 = 2$. Discuss the results. (d) Modify the program so that instead of a_0 and σ_0^2, the sample mean $\bar{\mathsf{x}}$ sample variance s^2 are used for the computation of X^2. The quantity X^2 can be used for a χ^2-test of the hypothesis that the samples were drawn from a normal distribution. Display histograms of X^2 and $P(\mathsf{X}^2)$ and show that X^2 follows a χ^2-distribution with $N - 2$ degrees of freedom.