6. Samples

In the last chapter we discussed a number of distributions, but we have not specified how they are realized in a particular case. We have only given the probability that a random variable x will lie within an interval with boundaries x and $x + \mathrm{d}x$. This probability depends on certain parameters describing its distribution (like λ in the case of the Poisson distribution) which are usually unknown. We therefore have no direct knowledge of the probability distribution and have to approximate it by a *frequency distribution* obtained experimentally. The number of measurements performed for this purpose, called a *sample*, is necessarily finite. To discuss the elements of sampling theory we first have to introduce a number of new definitions.

6.1 Random Samples. Distribution of a Sample. Estimators

Every sample is taken from a set of elements which correspond to the possible results of an individual observation. Such a set, which usually has infinitely many elements, is called a *population*. If a sample of n elements is taken from it, then we say that the sample has the *size* n. Let the distribution of the random variable x in the population be given by the probability density $f(x)$. We are interested in the values of x assumed by the individual elements of the sample. Suppose that we take ℓ samples of size n and find the following values for x:

$$
\begin{aligned}
&1\text{st sample:} \quad \mathsf{x}_1^{(1)}, \mathsf{x}_2^{(1)}, \ldots, \mathsf{x}_n^{(1)}, \\
&\vdots \\
&j\text{th sample:} \quad \mathsf{x}_1^{(j)}, \mathsf{x}_2^{(j)}, \ldots, \mathsf{x}_n^{(j)}, \\
&\vdots \\
&\ell\text{th sample:} \quad \mathsf{x}_1^{(\ell)}, \mathsf{x}_2^{(\ell)}, \ldots, \mathsf{x}_n^{(\ell)}.
\end{aligned}
$$

S. Brandt, *Data Analysis: Statistical and Computational Methods for Scientists and Engineers*,
DOI 10.1007/978-3-319-03762-2_6, © Springer International Publishing Switzerland 2014

We group the result of one sample into an n-dimensional vector

$$\mathbf{x}^{(j)} = (\mathsf{x}_1^{(j)}, \mathsf{x}_2^{(j)}, \ldots, \mathsf{x}_n^{(j)}) \quad , \tag{6.1.1}$$

which can be considered the position vector in an n-dimensional sample space (Sect. 2.1). Its probability density is

$$g(\mathbf{x}) = g(x_1, x_2, \ldots, x_n) \quad . \tag{6.1.2}$$

This function must fulfill two conditions in order for the sample to be *random*:

(a) The individual x_i must be independent, e.g., one must have

$$g(\mathbf{x}) = g_1(x_1) g_2(x_2) \ldots g_n(x_n) \quad . \tag{6.1.3}$$

(b) The individual marginal distributions must be identical and equal to the probability density $f(x)$ of the population,

$$g_1(x) = g_2(x) = \ldots = g_n(x) = f(x) \quad . \tag{6.1.4}$$

Comparing with (6.1.2) it is clear that there is a simple relation between a population and a sample only if these conditions are fulfilled. In the following we will mean by the word sample a random sample unless otherwise stated.

It should be emphasized that in the actual process of sampling it is often quite difficult to ensure randomness. Because of the large variety of applications, a general prescription cannot be given. In order to obtain reliable results from sampling, we have to take the utmost precautions to meet the requirements (6.1.3) and (6.1.4). Independence (6.1.3) can be checked to a certain extent by comparing the frequency distributions of the first, second, ... elements of a large number of samples. It is very difficult, however, to ensure that the samples in fact come from a population with the probability density $f(x)$. If the elements of the population can be numbered, it is often useful to use random numbers to select the elements for the sample.

We now suppose that the n elements of a sample are ordered according to the value of the variable, e.g., marked on the x axis, and we ask for the number of elements of the sample n_x for which $\mathsf{x} < x$, for arbitrary x. The function

$$W_n(x) = n_x / n \tag{6.1.5}$$

takes on the role of an empirical distribution function. It is a step function that increases by $1/n$ as soon as x is equal to one of the values x of an element of the sample. It is called the *sample distribution function*. It is clearly an approximation for $F(x)$, the distribution function of the population, which it approaches in the limit $n \to \infty$.

A function of the elements of a sample (6.1.1) is called a *statistic*. Since x is a random variable, a statistic is itself a random variable. The most important example is the *sample mean*,

$$\bar{\mathsf{x}} = \frac{1}{n}(\mathsf{x}_1 + \mathsf{x}_2 + \ldots + \mathsf{x}_n) \quad . \tag{6.1.6}$$

A typical problem of data analysis is the following. The general mathematical form of the probability density of the population is known. In radioactive decay, for example, the number of nuclei which decay before the time $t = \tau$ is $N_\tau = N_0(1 - \exp(-\lambda\tau))$, if N_0 nuclei existed at time $t = 0$. Here, however, the decay constant λ is, in general, not known. By taking a finite sample (measuring a finite number of decay times of individual nuclei) we want to determine the parameter λ as accurately as possible. Since such a task cannot be exactly solved, because the sample is finite, one speaks of the *estimation of parameters*. To estimate a parameter λ of a distribution function one uses an *estimator*

$$\mathsf{S} = \mathsf{S}(\mathsf{x}_1, \mathsf{x}_2, \ldots, \mathsf{x}_n) \quad . \tag{6.1.7}$$

An estimator is said to be *unbiased* if for arbitrary sample size the expectation value of the (random) quantity S is equal to the parameter to be estimated:

$$E\{\mathsf{S}(\mathsf{x}_1, \mathsf{x}_2, \ldots, \mathsf{x}_n)\} = \lambda \qquad \text{for all } n. \tag{6.1.8}$$

An estimator is said to be *consistent* if its variance vanishes for arbitrarily large sample size, i.e., if

$$\lim_{n\to\infty} \sigma(\mathsf{S}) = 0 \quad . \tag{6.1.9}$$

Often one can give a lower limit for the variance of an estimator of a parameter. If one finds an estimator S_0 whose variance is equal to this limit, then one apparently has the "best possible" estimator. S_0 is then said to be an *efficient estimator* for λ.

6.2 Samples from Continuous Populations: Mean and Variance of a Sample

The case of greatest interest in applications concerns a sample from an infinitely large continuous population described by the probability density $f(x)$. The sample mean (6.1.6) is a random variable, as are all statistics. Let us consider its expectation value

$$E(\bar{\mathsf{x}}) = \frac{1}{n}\{E(\mathsf{x}_1) + E(\mathsf{x}_2) + \ldots + E(\mathsf{x}_n)\} = \hat{x} \quad . \tag{6.2.1}$$

This expectation value is equal to the expectation value of x. Since Eq. (6.2.1) holds for all values of n, the arithmetic mean of a sample is, as one would expect, an unbiased estimator for the mean of the population. The characteristic function of the random variable $\bar{\mathsf{x}}$ is

$$\varphi_{\bar{\mathsf{x}}}(t) = \left\{\varphi_{\frac{\mathsf{x}}{n}}(t)\right\}^n = \left\{\varphi_{\mathsf{x}}\left(\frac{t}{n}\right)\right\}^n \quad . \tag{6.2.2}$$

Next we are interested in the variance of $\bar{\mathsf{x}}$,

$$\begin{aligned}\sigma^2(\bar{\mathsf{x}}) &= E\{(\bar{\mathsf{x}} - E(\bar{\mathsf{x}}))^2\} = E\left\{\left(\frac{\mathsf{x}_1 + \mathsf{x}_2 + \ldots + \mathsf{x}_n}{n} - \hat{x}\right)^2\right\} \\ &= \frac{1}{n^2} E\{[(\mathsf{x}_1 - \hat{x}) + (\mathsf{x}_2 - \hat{x}) + \ldots + (\mathsf{x}_n - \hat{x})]^2\} \quad .\end{aligned}$$

Since all of the x_i are independent, all of the cross terms of the type $E\{(\mathsf{x}_i - \hat{x})(\mathsf{x}_j - \hat{x})\}$, $i \neq j$ (i.e., all of the covariances) vanish, and we obtain

$$\sigma^2(\bar{\mathsf{x}}) = \frac{1}{n}\sigma^2(\mathsf{x}) \quad . \tag{6.2.3}$$

One thus shows that $\bar{\mathsf{x}}$ is a consistent estimator for $\hat{x}$. The variance (6.2.3) is itself, however, not a random variable, and is therefore not directly obtainable by experiment. As a definition for the *sample variance* we could try using the arithmetic mean of squared differences

$$\mathsf{s}'^2 = \frac{1}{n}\{(\mathsf{x}_1 - \bar{\mathsf{x}})^2 + (\mathsf{x}_2 - \bar{\mathsf{x}})^2 + \ldots + (\mathsf{x}_n - \bar{\mathsf{x}})^2\} \quad . \tag{6.2.4}$$

Its expectation value is

$$\begin{aligned}E(\mathsf{s}'^2) &= \frac{1}{n} E\left\{\sum_{i=1}^{n} (\mathsf{x}_i - \bar{\mathsf{x}})^2\right\} \\ &= \frac{1}{n} E\left\{\sum_{i=1}^{n} (\mathsf{x}_i - \hat{x} + \hat{x} - \bar{\mathsf{x}})^2\right\} \\ &= \frac{1}{n} E\left\{\sum_{i=1}^{n} (\mathsf{x}_i - \hat{x})^2 + \sum_{i=1}^{n} (\hat{x} - \bar{\mathsf{x}})^2 + 2\sum_{i=1}^{n} (\mathsf{x}_i - \hat{x})(\hat{x} - \bar{\mathsf{x}})\right\} \\ &= \frac{1}{n} \sum_{i=1}^{n} \{E((\mathsf{x}_i - \hat{x})^2) - E((\bar{\mathsf{x}} - \hat{x})^2)\} \\ &= \frac{1}{n}\left\{n\sigma^2(\mathsf{x}) - n\left(\frac{1}{n}\sigma^2(\mathsf{x})\right)\right\} \quad , \\ E(\mathsf{s}'^2) &= \frac{n-1}{n}\sigma^2(\mathsf{x}) \quad . \end{aligned} \tag{6.2.5}$$

Hence one sees that the sample variance defined in this way is a biased estimator for the population variance, having an expectation value smaller than $\sigma^2(\mathsf{x})$. We can see directly from (6.2.5), however, the size of the bias. We therefore change our definition (6.2.4) and write for the sample variance

$$\mathsf{s}^2 = \frac{1}{n-1}\{(\mathsf{x}_1 - \bar{\mathsf{x}})^2 + (\mathsf{x}_2 - \bar{\mathsf{x}})^2 + \ldots + (\mathsf{x}_n - \bar{\mathsf{x}})^2\} \quad . \tag{6.2.6}$$

This is now an unbiased estimator for $\sigma^2(\mathsf{x})$. The value $(n-1)$ in the denominator appears at first to be somewhat strange. One must consider, however, that for $n = 1$ the sample mean is equal to the value x of the sole element of the sample ($\mathsf{x} = \bar{\mathsf{x}}$) and that therefore the quantity (6.2.4) would vanish. That is related to the fact that in (6.2.4) – and also in (6.2.6) – the sample mean $\bar{\mathsf{x}}$ was used instead of the population mean $\hat{x}$, since the latter was not known. Part of the information contained in the sample first had to be used and was not available for the calculation of the variance. The effective number of elements available for calculating the variance is therefore reduced. This is taken into consideration by reducing the denominator of the arithmetic mean (6.2.4). The same line of reasoning is repeated quantitatively in Sect. 6.5.

If we substitute the estimator for the population variance (6.2.6) into (6.2.3), we obtain an estimator for the *variance of the mean*

$$\mathsf{s}^2(\bar{\mathsf{x}}) = \frac{1}{n}\mathsf{s}^2(\mathsf{x}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\mathsf{x}_i - \bar{\mathsf{x}})^2 \quad . \tag{6.2.7}$$

The corresponding standard deviation can be considered to be the *error of the mean*

$$\Delta\bar{\mathsf{x}} = \sqrt{\mathsf{s}^2(\bar{\mathsf{x}})} = \mathsf{s}(\bar{\mathsf{x}}) = \frac{1}{\sqrt{n}}\mathsf{s}(\mathsf{x}) \quad . \tag{6.2.8}$$

Of course we are also interested in the error of the sample variance (6.2.6). In Sect. 6.6 we will show that this quantity can be determined under the assumption that the population follows a normal distribution. We will use the result here ahead of time. The variance of s^2 is

$$\mathrm{var}(\mathsf{s}^2) = \left(\frac{\sigma^2}{n-1}\right)^2 2(n-1) \quad . \tag{6.2.9}$$

If we substitute the estimator (6.2.6) into the right-hand side for σ^2 and take the square root, we obtain for the error of the sample variance

$$\Delta\mathsf{s}^2 = \mathsf{s}^2\sqrt{\frac{2}{(n-1)}} \quad . \tag{6.2.10}$$

Finally we give explicit expressions for estimators of the *sample standard deviation* and its error. The first is simply the square root of the sample variance

$$\mathsf{s} = \sqrt{\mathsf{s}^2} = \frac{1}{\sqrt{n-1}} \sqrt{\sum (\mathsf{x}_i - \bar{\mathsf{x}})^2} \quad . \tag{6.2.11}$$

The error of the sample standard deviation is obtained from (6.2.10) by error propagation, which gives

$$\Delta \mathsf{s} = \frac{\mathsf{s}}{\sqrt{2(n-1)}} \quad . \tag{6.2.12}$$

Example 6.1: Computation of the sample mean and variance from data

Suppose one has $n = 7$ measurements of a certain quantity (e.g., the length of an object). Their values are 10.5, 10.9, 9.2, 9.8, 9.0, 10.4, 10.7. The computation is made easier if one uses the fact that all of the measured values are near $a = 10$, i.e., they are of the form $\mathsf{x}_i = a + \delta_i$. The relation (6.1.6) then gives

$$\bar{\mathsf{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathsf{x}_i = \frac{1}{n} \sum_{i=1}^{n} (a + \delta_i) = a + \frac{1}{n} \sum_{i=1}^{n} \delta_i = a + \Delta$$

with

$$\begin{aligned} \Delta &= \frac{1}{n} \sum_{i=1}^{n} \delta_i = \frac{1}{7}(0.5 + 0.9 - 0.8 - 0.2 - 1.0 + 0.4 + 0.7) \\ &= 0.5/7 = 0.07 \quad . \end{aligned}$$

We thus have $\bar{\mathsf{x}} = 10 + \Delta = 10.07$.

The sample variance is computed according to (6.2.6) to be

$$\begin{aligned} \mathsf{s}^2 &= \frac{1}{n-1} \sum_{i=1}^{n} (\mathsf{x}_i - \bar{\mathsf{x}})^2 \\ &= \frac{1}{n-1} \sum_{i=1}^{n} (\mathsf{x}_i^2 - 2\mathsf{x}_i \bar{\mathsf{x}} + \bar{\mathsf{x}}^2) \\ &= \frac{1}{n-1} \left\{ \sum_{i=1}^{n} \mathsf{x}_i^2 - n\bar{\mathsf{x}}^2 \right\} \quad . \end{aligned}$$

The result can be obtained either by the first or last line of the relation above. In the last line only one difference is computed, not n. The numbers to be squared, however, are usually considerably larger, and one must consider the problem of rounding errors. We therefore use the original expression

$$\begin{aligned} \mathsf{s}^2 &= \frac{1}{6}\{0.43^2 + 0.83^2 + 0.87^2 + 0.27^2 + 1.07^2 + 0.33^2 + 0.63^2\} \\ &= \frac{1}{6}\{0.1849 + 0.6889 + 0.7569 + 0.0729 + 1.1449 + 0.1089 \\ &\quad + 0.3969\} \\ &= 3.3543/6 \approx 0.56 \quad . \end{aligned}$$

The sample standard deviation is $\mathsf{s} \approx 0.75$. From (6.2.8), (6.2.10), and (6.2.12) we obtain finally $\Delta\bar{\mathsf{x}} = 0.28$, $\Delta\mathsf{s}^2 = 0.32$, and $\Delta\mathsf{s} = 0.21$. ■

Naturally one does not usually compute the sample mean and variance by hand, but rather by the class Sample and its methods.

6.3 Graphical Representation of Samples: Histograms and Scatter Plots

After the theoretical considerations of the last sections we now turn to some simple practical aspects of the analysis of sample data. An important tool for this is the representation of the data in graphical form.

A sample

$$\mathsf{x}_1, \mathsf{x}_2, \ldots, \mathsf{x}_n \quad ,$$

which depends on a single variable x can be represented simply by means of tick marks on an x axis. We will call such a representation a *one-dimensional*

Table 6.1: Values of resistance R of 100 individual resistors of nominal value 200 kΩ. The data are graphically represented in Fig. 6.1.

193.199	195.673	195.757	196.051	196.092
196.596	196.679	196.763	196.847	197.267
197.392	197.477	198.189	198.650	198.944
199.070	199.111	199.153	199.237	199.698
199.572	199.614	199.824	199.908	200.118
200.160	200.243	200.285	200.453	200.704
200.746	200.830	200.872	200.914	200.956
200.998	200.998	201.123	201.208	201.333
201.375	201.543	201.543	201.584	201.711
201.878	201.919	202.004	202.004	202.088
202.172	202.172	202.297	202.339	202.381
202.507	202.591	202.633	202.716	202.884
203.051	203.052	203.094	203.094	203.177
203.178	203.219	203.764	203.765	203.848
203.890	203.974	204.184	204.267	204.352
204.352	204.729	205.106	205.148	205.231
205.357	205.400	205.483	206.070	206.112
206.154	206.155	206.615	206.657	206.993
207.243	207.621	208.124	208.375	208.502
208.628	208.670	208.711	210.012	211.394

scatter plot. It contains all the information about the sample. Table 6.1 contains the values $\mathsf{x}_1, \mathsf{x}_2, \ldots, \mathsf{x}_n$ of a sample of size 100, obtained from measuring the resistance R of 100 individual resistors of nominal value 200 kΩ. After obtaining the sample the measurements were ordered.

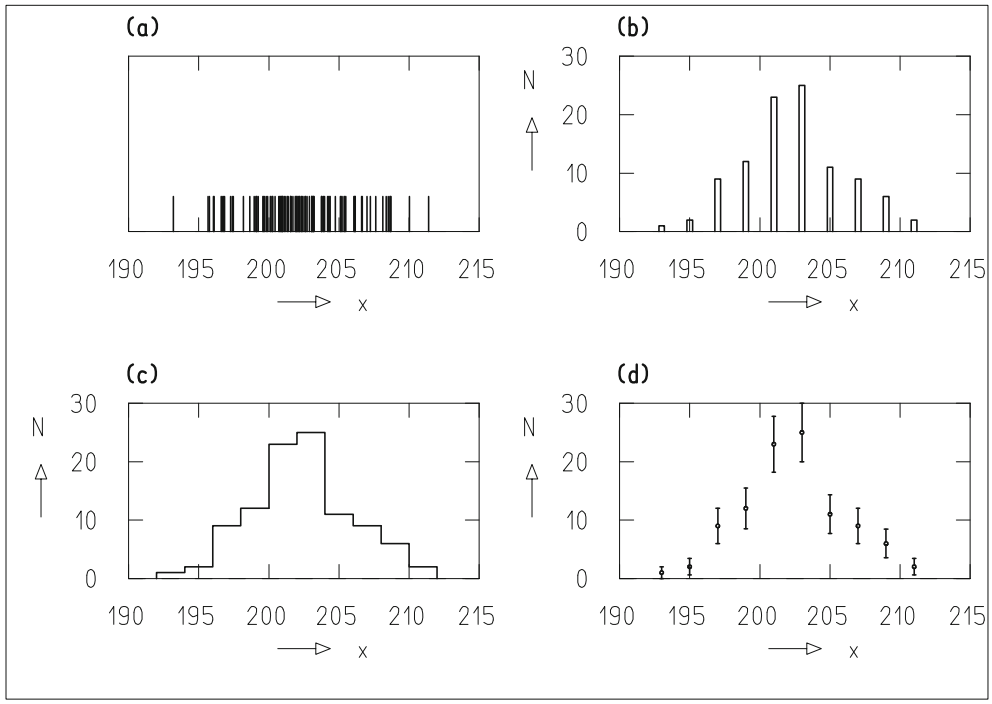

Fig. 6.1: Representation of the data from Table 6.1 as (**a**) a one-dimensional scatter plot, (**b**) a bar diagram, (**c**) a step diagram, and (**d**) a diagram of measured points with error bars.

Figure 6.1a shows the corresponding scatter plot. Qualitatively one can estimate the mean and variance from the position and width of the clustering of tick marks.

Another graphical representation is usually better suited to visualize the sample by using the second dimension available on the paper. The x axis is used as abscissa and divided into r intervals

$$\xi_1, \xi_2, \ldots, \xi_r$$

of equal width Δx. These intervals are called *bins*. The centers of the bins have the x-values

$$x_1, x_2, \ldots, x_r \quad .$$

On the vertical axis one plots the corresponding numbers of sample elements

$$n_1, n_2, \ldots, n_r$$

that fall into bins $\xi_1, \xi_2, \ldots, \xi_r$. The diagram obtained in this way is called a *histogram* of the sample. This can be interpreted as a frequency distribution, since $h_k = n_k/n$ is a frequency, i.e., a measure of the probability p_k to observe a sample element in the interval ξ_k. For the graphical form of histograms, various methods are used. In a *bar diagram* the values n_k are represented as bars perpendicular to the x axis on the x_k values (Fig. 6.1b). In a *step diagram* the n_k are represented as horizontal lines that cover the entire width ξ_k of the interval. Neighboring horizontal lines are connected by perpendicular lines (Fig. 6.1c). The fraction of the area covering each interval ξ_k of the x axis is then proportional to the number n_k of the sample elements in the interval. (If one uses the area in the interval for the graphical representation of n_k, then the bins can also have different widths.) In economics bar diagrams are most commonly used. (Sometimes one also sees diagrams in which, instead of bars, line segments are used to connect the tips of the bars. In contrast to the step diagram, the resulting figure does not have an area proportional to the sample size n.) In the natural sciences, step diagrams are more common.

In Sect. 6.8 we will determine that as long as the values n_k are not too small, their statistical errors are given by $\Delta n_k = \sqrt{n_k}$. In order to plot them on a graph, the observed values n_k can be drawn as points with vertical *error bars* ending at the points $n_k \pm \sqrt{n_k}$ (Fig. 6.1d).

It is clear that the relative errors $\Delta n_k/n_k = 1/\sqrt{n_k}$ decrease for increasing n_k, i.e., for a sample of fixed size n they decrease for increasing bin width of the histogram. On the other hand, by choosing a larger interval width, one loses any finer structure of the data with respect to the variable x. The ability of a histogram to convey information therefore depends crucially on the appropriate choice of the bin width, usually found only after several attempts.

Example 6.2: Histograms of the same sample with various choices of bin width

In Fig. 6.2 four histograms of the same sample are shown. The population is a Gaussian distribution, which is also shown as a continuous curve. This was scaled in such a way that the area under the histogram is equal to the area under the Gaussian curve. Although the information contained in the plot is greater for a smaller bin width – for vanishing bin width the histogram becomes a one dimensional scatter plot – one notices the similarity between the histogram and Gaussian curve much more easily for the larger bin width. This is because for the larger bin width the relative statistical fluctuations of the contents of individual bins are smaller. The individual steps of the histogram differ less from the curve. ■

Constructing a histogram from a sample is a simple programming task. Suppose the histogram has n_x bins of width Δx, with the first interval extending from $x = x_0$ to $x = x_0 + \Delta x$. The contents of the histogram is put into an

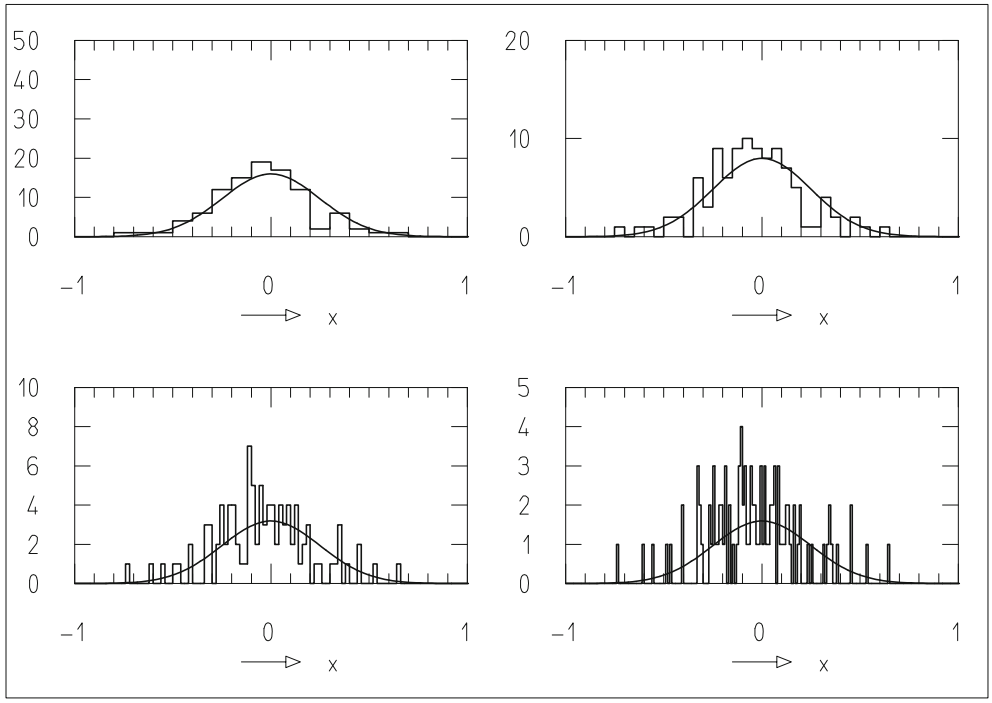

Fig. 6.2: Histogram of the same sample from a Gaussian distribution represented with four different bin widths.

array hist, with the first bin in hist[0] the second bin in hist[1], etc.
The histogram is then specified in the computer by the array hist and the three values x_0, Δx, and n_x. The class Histogram permits construction and administration of a histogram.

Graphical display of histograms can be accomplished using methods of the class DatanGraphics (Appendix F). With them the user can freely adjust all of the parameters that determine the appearance of the plot, such as the page format, scale factors, colors, line thickness, etc. Often it is convenient to use the class GraphicsWithHistogram which does not allow for this freedom but by a single call gives rise to the graphical output of a histogram, stored in the computer.

A histogram allows a first direct look at the nature of the data. It answers questions such as “Are the data more or less distributed according to a Gaussian?” or “Do there exist points exceptionally far away from the average value?”. If the histogram leads one to conclude that the population is distributed according to a Gaussian, then the mean and standard deviation can be estimated directly from the plot. The mean is the center of gravity of the histogram. The standard deviation is obtained as in the following example.

Example 6.3: Full width at half maximum (FWHM)
If the form of histogram allows one to assume that the represented sample originates from a Gaussian distribution, then one can draw by hand a Gaussian bell curve that follows the histogram as closely as possible. The position of the maximum is a good estimator for the mean of the sample. One then draws a horizontal line at half of the height of the maximum. This crosses the bell curve at the points x_a and x_b. The quantity

$$f = x_b - x_a$$

is called the *full width at half maximum* (FWHM). One can easily determine for a Gaussian distribution the simple relation

$$\sigma = \frac{f}{\sqrt{-8 \ln \frac{1}{2}}} \approx 0.4247\ f \tag{6.3.1}$$

between the standard deviation and FWHM. This expression can be used to estimate the standard deviation of a sample when f is obtained from a histogram. ■

We now use the Monte Carlo method (Chap. 4) together with histograms in order to illustrate the concepts of mean, standard deviation, and variance of a sample, and their errors, as introduced in Sect. 6.2.

Example 6.4: Investigation of characteristic quantities of samples from a Gaussian distribution with the Monte Carlo method
We generate successively 1000 samples of size $N = 100$ from the standard normal distribution, e.g., compute the mean $\bar{\mathsf{x}}$, variance s^2, and standard deviation s for each sample as well as the errors $\Delta\bar{\mathsf{x}}$, $\Delta\mathsf{s}^2$, and $\Delta\mathsf{s}$, with the methods of the classSample. We then produce for each of the six quantities a histogram (Fig. 6.3), containing 1000 entries. Since each of the quantities is defined as the sum of many random quantities, we expect in all cases that the histograms should resemble Gaussian distributions. From the histogram for $\bar{\mathsf{x}}$, we obtain a full width at half maximum of about 0.25, and hence a standard deviation of approximately 0.1. Indeed the histogram for $\Delta\bar{\mathsf{x}}$ shows an approximately Gaussian distribution with mean value $\Delta\bar{\mathsf{x}} = 0.1$. (From the width of this histogram one could determine the error of the error $\Delta\bar{\mathsf{x}}$ of the mean $\bar{\mathsf{x}}$!) From both histograms one obtains a very clear impression of the meaning of the error $\Delta\bar{\mathsf{x}}$ of the mean $\bar{\mathsf{x}}$ for a single sample, as computed according to (6.2.8). It gives (within its error) the standard deviation of the population, from which the sample mean $\bar{\mathsf{x}}$ comes. If many samples are successively taken (i.e., if the experiment is repeated many times) then the frequency distribution of the values $\bar{\mathsf{x}}$ follows a Gaussian distribution about the population mean with standard deviation $\Delta\bar{\mathsf{x}}$. The corresponding considerations also hold for the quantities s^2, s, and their errors $\Delta\mathsf{s}$ and $\Delta\mathsf{s}^2$. ■

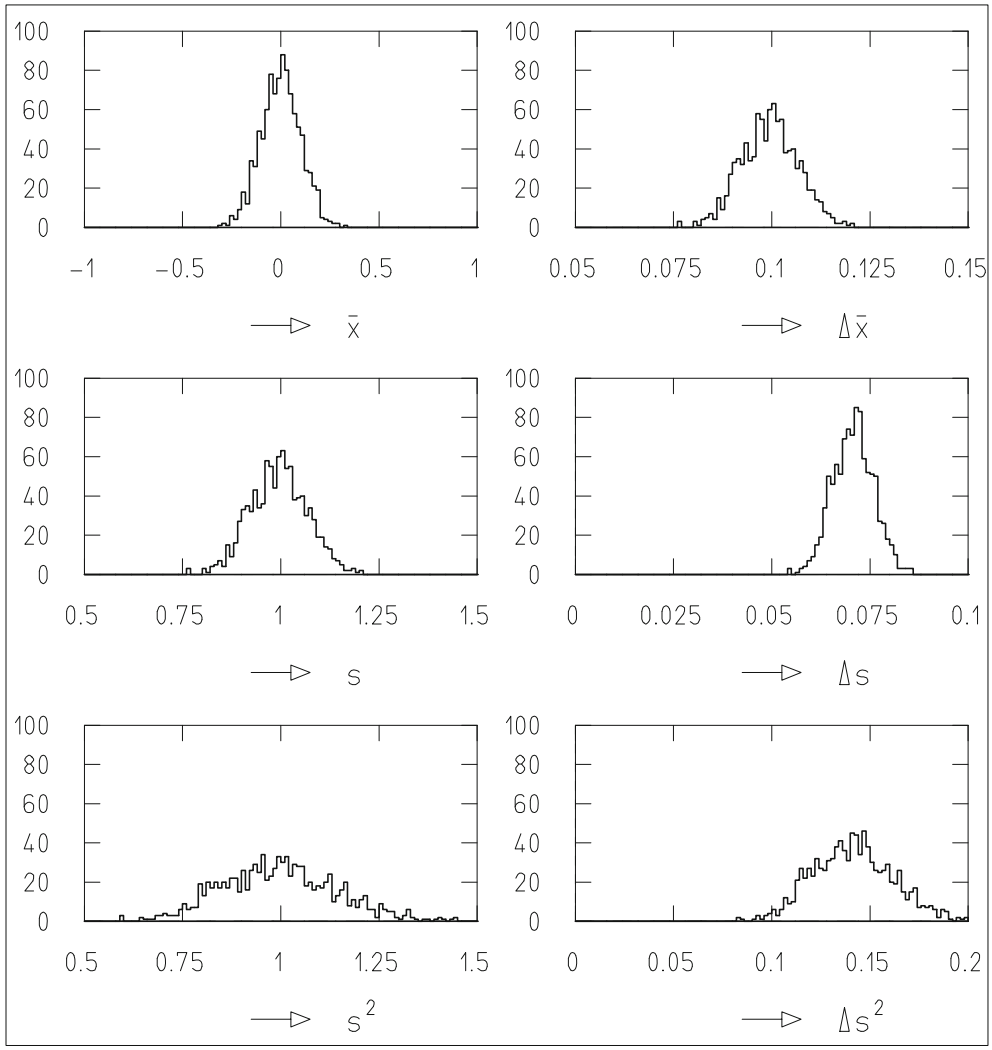

Fig. 6.3: Histograms of the quantities $\bar{x}$, $\Delta\bar{x}$, s, Δs, s^2, and Δs^2 from 1000 samples of size 100 from the standard normal distribution.

If the elements of the sample depend on two random variables x and y, then one can construct a scatter plot, where each element is represented as a point in a Cartesian coordinate system for the variables x and y. Such a *two-dimensional scatter plot* provides useful qualitative information about the relationship between the two variables.

The class GraphicsWith2DScatterDiagram generates such a diagram by a single call. (A plot in the format A5 landscape is generated, into which the scatter diagram, itself in square format, is fitted. If another plot format or edge ratio of the diagram is desired, the class has to be adapted accordingly.)

Example 6.5: Two-dimensional scatter plot: Dividend versus price for industrial stocks

Table 6.2 contains a list of the first 10 of 226 data sets, in which the dividend in 1967 (first column) and share price on December 31, 1967 (second column) are given for a number of industrial stocks. The third column shows the company name for all German corporations worth more than 10 million marks. The scatter plot of the number pairs (share price, dividend) is shown in Fig. 6.4.

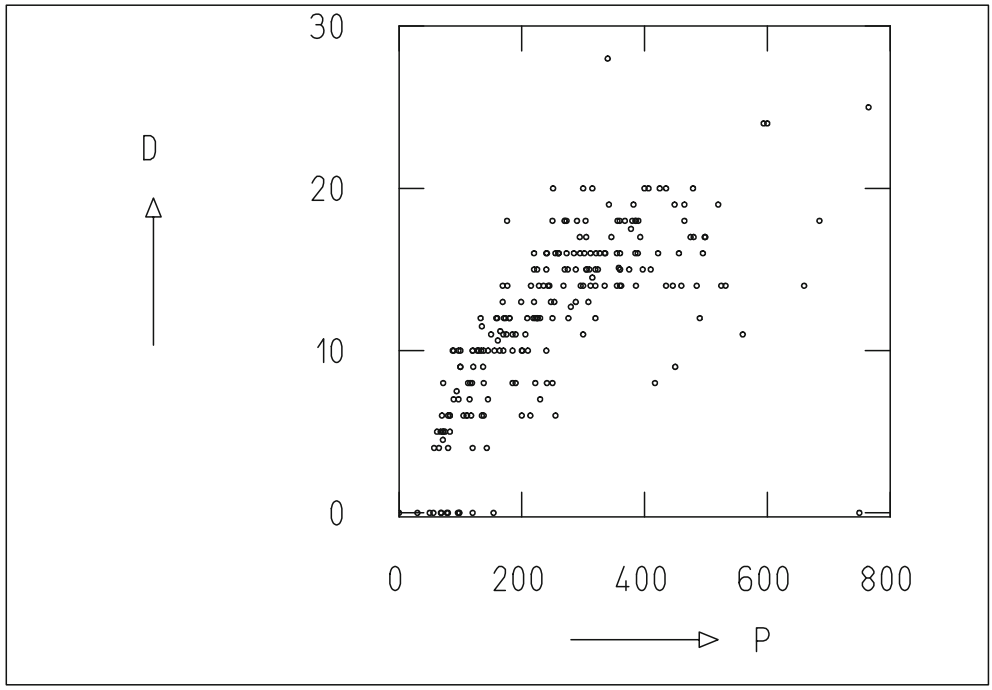

Fig. 6.4: Scatter plot of dividend D versus share price P for industrial stocks.

As expected we see a strong correlation between dividend and share price. One can see, however, that the dividend does not grow linearly with the price. It appears that factors other than immediate profit determine the price of a stock.

Also shown are histograms for the share price (Fig. 6.5) and dividend (Fig. 6.6) which can be obtained as projections of the scatter plot onto the abscissa and ordinate. One clearly observes a non-statistical behavior for the dividends. It is given as a percent of the nominal value and is therefore almost always integer. One sees that even numbers are considerably more frequent than odd numbers. ■

Table 6.2: Dividend, share price of a stock, and company name.

12.	133.	ACKERMANN-GOEGGINGEN
08.	417.	ADLERWERKE KLEYER
17.	346.	AGROB AG FUER GROB U. FEINKERAMIK
25.	765.	AG.F.ENERGIEWIRTSCHAFT
16.	355.	AG F. LICHT- U. KRAFTVERS.,MCHN.
20.	315.	AG.F. IND.U.VERKEHRSW.
08.	138.	AG. WESER
16.	295.	AEG ALLG.ELEKTR.-GES.
20.	479.	ANDREAE-NORIS ZAHN
10.	201.	ANKERWERKE

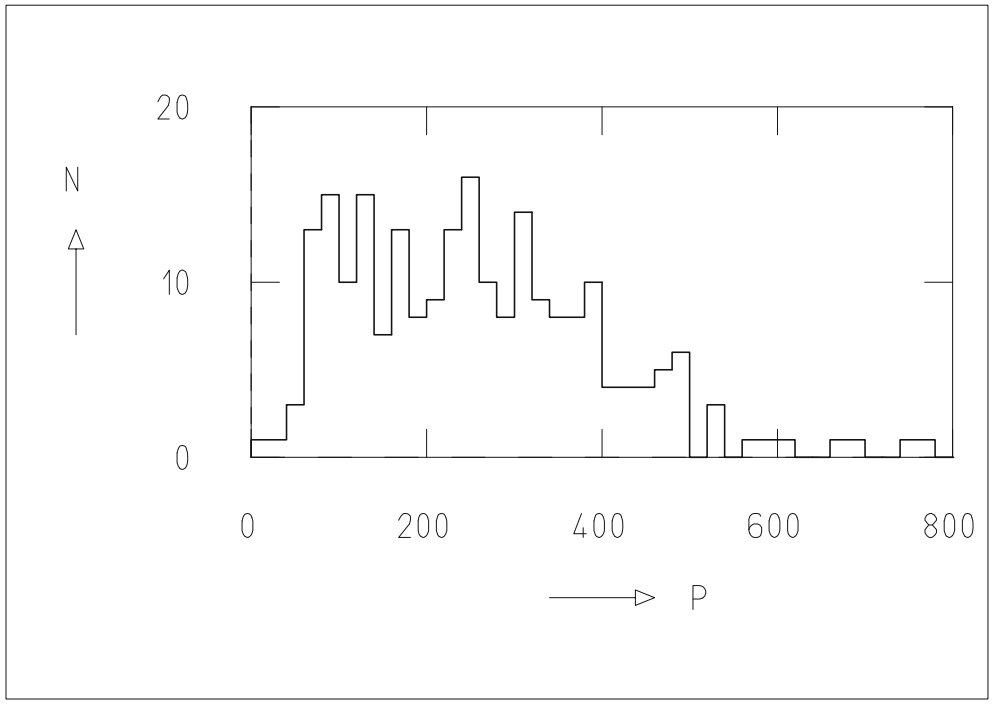

Fig. 6.5: Histogram of the price of industrial stocks.

6.4 Samples from Partitioned Populations

It is often advantageous to divide a population G (e.g., all of the students in Europe) into various *subpopulations* G_1, G_2, ..., G_t (students at university $1, 2, \ldots, t$). Suppose a quantity of interest x follows in the various subpopulations the probability densities $f_1(x)$, $f_2(x), \ldots, f_t(x)$. The distribution function corresponding to $f_i(x)$ is then

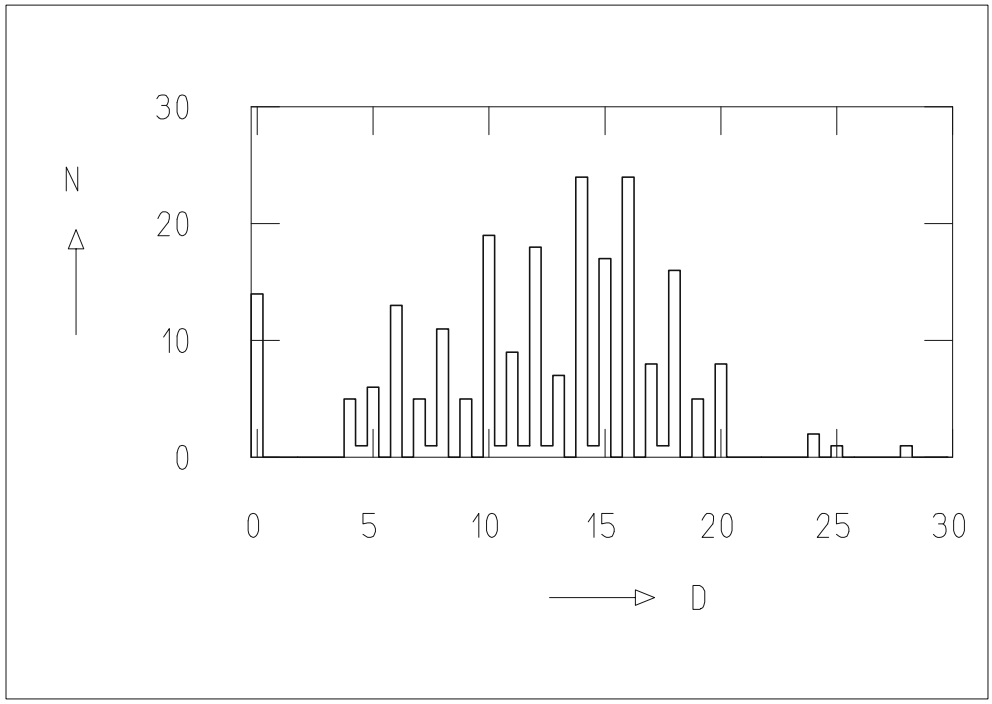

Fig. 6.6: Histogram of the dividend of industrial stocks.

$$F_i(x) = \int_{-\infty}^{x} f_i(x)\,\mathrm{d}x = P(\mathsf{x} < x | \mathsf{x} \in G_i) \quad . \tag{6.4.1}$$

This is equal to the conditional probability for $\mathsf{x} < x$ given that x is contained in the subpopulation G_i. The rule of total probability (2.3.4) provides the relationship between the various $F_i(x)$ and the distribution function $F(x)$ for G,

$$F(x) = P(\mathsf{x} < x | \mathsf{x} \in G) = \sum_{i=1}^{t} P(\mathsf{x} < x | \mathsf{x} \in G_i) P(\mathsf{x} \in G_i) \quad ,$$

i.e.,

$$F(x) = \sum_{i=1}^{t} P(\mathsf{x} \in G_i) F_i(x) \quad . \tag{6.4.2}$$

Correspondingly one has for the probability density

$$f(x) = \sum_{i=1}^{t} P(\mathsf{x} \in G_i) f_i(x) \quad . \tag{6.4.3}$$

If we now abbreviate $P(\mathsf{x} \in G_i)$ by p_i, then one has

$$\hat{x} = E(\mathsf{x}) = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x = \sum_{i=1}^{t} p_i \int_{-\infty}^{\infty} x f_i(x)\,\mathrm{d}x \quad ,$$

$$\hat{x} = \sum_{i=1}^{t} p_i \hat{x}_i \quad . \tag{6.4.4}$$

The population mean is thus the mean of the individual means of the subpopulations, each weighted by probability of its subpopulation. For the population variance one obtains

$$\begin{aligned} \sigma^2(\mathsf{x}) &= \int_{-\infty}^{\infty} (x-\hat{x})^2 f(x)\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} (x-\hat{x})^2 \sum_{i=1}^{t} p_i f_i(x)\,\mathrm{d}x \\ &= \sum_{i=1}^{t} p_i \int_{-\infty}^{\infty} \{(x-\hat{x}_i)+(\hat{x}_i-\hat{x})\}^2 f_i(x)\,\mathrm{d}x \quad . \end{aligned}$$

All cross terms vanish since the x_i are independent, leading to

$$\sigma^2(\mathsf{x}) = \sum_{i=1}^{t} p_i \left\{ \int_{-\infty}^{\infty} (x-\hat{x}_i)^2 f_i(x)\,\mathrm{d}x + (\hat{x}_i-\hat{x})^2 \int_{-\infty}^{\infty} f_i(x)\,\mathrm{d}x \right\}$$

or

$$\sigma^2(\mathsf{x}) = \sum_{i=1}^{t} p_i \{\sigma_i^2 + (\hat{x}_i-\hat{x})^2\} \quad . \tag{6.4.5}$$

One thus obtains the weighted mean of a sum of two terms. The first gives the dispersion of a subpopulation, the second gives the quadratic deviation of the mean of this subpopulation from the mean of the whole population.

Having discussed separating a population into parts, we now take from each subpopulation G_i a sample of size n_i (with $\sum_{i=1}^{t} n_i = n$) and examine the arithmetic mean of the total partitioned sample

$$\bar{\mathsf{x}}_p = \frac{1}{n} \sum_{i=1}^{t} \sum_{j=1}^{n_i} \mathsf{x}_{ij} = \frac{1}{n} \sum_{i=1}^{t} n_i \bar{\mathsf{x}}_i \tag{6.4.6}$$

with the expectation value and variance

$$
\begin{aligned}
E(\bar{\mathsf{x}}_p) &= \frac{1}{n}\sum_{i=1}^{n} n_i \hat{x}_i \quad , \qquad (6.4.7)\\
\sigma^2(\bar{\mathsf{x}}_p) &= E\{(\bar{\mathsf{x}}_p - E(\bar{\mathsf{x}}_p))^2\}\\
&= E\left\{\left(\sum_{i=1}^{t} \frac{n_i}{n}(\bar{\mathsf{x}}_i - \hat{x}_i)\right)^2\right\}\\
&= \frac{1}{n^2}\sum_{i=1}^{t} n_i^2 E\{(\bar{\mathsf{x}}_i - \hat{x}_i)^2\} \quad ,\\
\sigma^2(\bar{\mathsf{x}}_p) &= \frac{1}{n^2}\sum_{i=1}^{t} n_i^2 \sigma^2(\bar{\mathsf{x}}_i) \quad . \qquad (6.4.8)
\end{aligned}
$$

Using (6.2.3) this is finally

$$\sigma^2(\bar{\mathsf{x}}_p) = \frac{1}{n}\sum_{i=1}^{t} \frac{n_i}{n}\sigma_i^2 \quad . \qquad (6.4.9)$$

One would obtain the same result by application of the law of error propagation (3.8.7) to Eq. (6.4.6).

It is clear that the arithmetic mean $\bar{\mathsf{x}}_p$ cannot in general be an estimator for the sample mean $\hat{x}$, since it depends on the arbitrary choice of the size n_i of the samples from the subpopulations. A comparison of (6.4.7) with (6.4.4) shows that this is only true for the special case $p_i = n_i/n$.

The population mean $\hat{x}$ can be estimated in the following way. One first determines the means $\bar{\mathsf{x}}_i$ for the subpopulations, and constructs then the expression

$$\tilde{\mathsf{x}} = \sum_{i=1}^{t} p_i \bar{\mathsf{x}}_i \quad , \qquad (6.4.10)$$

in analogy to Eq. (6.4.4). By error propagation one obtains for the variance of $\tilde{\mathsf{x}}$

$$\sigma^2(\tilde{\mathsf{x}}) = \sum_{i=1}^{t} p_i^2 \sigma^2(\bar{\mathsf{x}}_i) = \sum_{i=1}^{t} \frac{p_i^2}{n_i}\sigma_i^2 \quad . \qquad (6.4.11)$$

Example 6.6: Optimal choice of the sample size for subpopulations

In order to minimize the variance $\sigma^2(\tilde{\mathsf{x}})$, we cannot simply differentiate the relation (6.4.11) with respect to all n_i, since the n_i must satisfy a constraint, namely

$$\sum_{i=1}^{t} n_i - n = 0 \quad . \tag{6.4.12}$$

We must therefore use the method of *Lagrange multipliers*, by multiplying Eq. (6.4.12) with a factor μ, adding this to Eq. (6.4.11), and finally setting the partial derivatives of the n_i with respect to μ equal to zero:

$$L = \sigma^2(\tilde{x}) + \mu(\sum n_i - n) = \sum (p_i^2/n_i)\sigma_i^2 + \mu(\sum n_i - n) \quad ,$$

$$\frac{\partial L}{\partial n_i} = -\frac{p_i^2 \sigma_i^2}{n_i^2} + \mu = 0 \quad , \tag{6.4.13}$$

$$\frac{\partial L}{\partial \mu} = \sum n_i - n = 0 \quad . \tag{6.4.14}$$

From (6.4.13) we obtain

$$n_i = p_i \sigma_i / \sqrt{\mu} \quad .$$

Together with (6.4.14) this gives

$$1/\sqrt{\mu} = n / \sum p_i \sigma_i$$

and therefore

$$n_i = n p_i \sigma_i / \sum p_i \sigma_i \quad . \tag{6.4.15}$$

The result (6.4.15) states that the sizes n_i of the samples from the subpopulations i should be chosen in such a way that they are proportional to the probability p_i of subpopulation i, weighted with the corresponding standard deviation.

As an example assume that a scientific publishing company wants to estimate the total amount spent for scientific books by two subpopulations: (1) students and (2) scientific libraries. Further, we will assume that there are 1000 libraries and 10^6 students in the population and that the standard deviation of the money spent by students is \$ 100, and for libraries (which are of greatly differing sizes) \$ $3 \cdot 10^5$. We then have

$$p_1 \approx 1 \quad , \qquad p_2 \approx 10^{-3} \quad , \qquad \sigma_1 = 100 \quad , \qquad \sigma_2 = 3 \times 10^5$$

and from (6.4.15)

$$n_1 = \text{const} \cdot 100 \quad , \qquad n_2 = \text{const} \cdot 300 \quad , \qquad n_2 = 3 n_1 \quad .$$

Note that the result does not depend on the means of the partial populations. The quantities p_i, x_i, and σ_i are in general unknown. They must first be estimated from preliminary samples. ■

The discussion of subpopulations will be taken up again in Chap. 11.

6.5 Samples Without Replacement from Finite Discrete Populations. Mean Square Deviation. Degrees of Freedom

We first encountered the concept a sample in connection with the hypergeometric distribution (Sect. 5.3). There we determined that the independence of the individual sample elements was lost by the process of taking elements without replacing them from a finite (and hence discrete) population. We are therefore no longer dealing with genuine random sampling, even if no particular choice among the remaining elements is made.

To discuss this further let us introduce the following notation. Suppose the population consists of N elements $y_1, y_2, \ldots, y_N$. From it we take a sample of size n with the elements $\mathsf{x}_1, \mathsf{x}_2, \ldots, \mathsf{x}_n$. (In the hypergeometric distribution, the y_j and hence the x_i could only take on the values 0 and 1.)

Since it is equally probable for each of the remaining elements y_j to be chosen, we obtain for the expectation value

$$E(y) = \hat{y} = \bar{y} = \frac{1}{N}\sum_{j=1}^{N} y_j \quad . \tag{6.5.1}$$

Although $\hat{y}$ is not a random variable, this expression is the arithmetic mean of a finite number of elements of the population. A definition of the population variance encounters the difficulties discussed at the end of Sect. 6.2. We define it in analogy to (6.2.6) as

$$\begin{aligned} \sigma^2(y) &= \frac{1}{N-1}\sum_{j=1}^{N}(y_j - \bar{y})^2 \\ &= \frac{1}{N-1}\left\{\sum_{j=1}^{N} y_j^2 - \frac{1}{N}\left(\sum_{j=1}^{N} y_j\right)^2\right\} \quad . \end{aligned} \tag{6.5.2}$$

Let us now consider the *sum of squares*,

$$\sum_{j=1}^{N}(y_j - \bar{y})^2 \quad . \tag{6.5.3}$$

Since we have not constrained the population in any way, the y_j can take on all possible values. Therefore the first element in the sum in (6.5.2) can also take on any of the possible values. The same holds for the 2nd, 3rd, ..., $(N-1)$th terms summed. The Nth term in the sum is then, however, fixed, since

$$\sum_{j=1}^{N}(y_j - \bar{y}) = 0 \quad . \tag{6.5.4}$$

We say that the *number of degrees of freedom* of the sum of squares (6.5.3) is $N-1$. One can illustrate this connection geometrically. We consider the case $\bar{y} = 0$ and construct an N-dimensional vector space with the y_j. The quadratic sum (6.5.3) is then the square of the absolute value of the position vector in this space. Because of the equation of constraint (6.5.4) the tip of the position vector can only move in a space of dimension $(N-1)$. In mechanics the dimension of such a constrained space is called the number of degrees of freedom. This is sketched in Fig. 6.7 for the case $N = 2$. Here the position vector is constrained to lie on the line $y_2 = -y_1$.

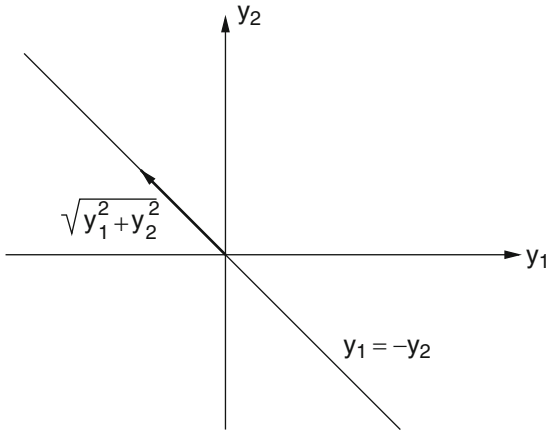

Fig. 6.7: A sample of size two gives a sum of squares with one degree of freedom.

A sum of squares divided by the number of degrees of freedom, i.e., an expression of the form (6.5.2) is called a *mean square* or to be more complete, since we are dealing with the differences of the individual values from the expectation or mean value, *mean square deviation*. The square root of this expression, which is then a measure of the dispersion, is called the root mean square (RMS) deviation.

We now return to the sample $\mathsf{x}_1, \mathsf{x}_2, \ldots, \mathsf{x}_n$. For simplicity of notation we will introduce the *Kronecker symbol*, which describes the selection procedure for the sample. It is defined as

$$\delta_i^j = \begin{cases} 1, \text{ if } \mathsf{x}_i \text{ is the element } y_j, \\ 0 \text{ otherwise.} \end{cases} \tag{6.5.5}$$

In particular one has

$$\mathsf{x}_i = \sum_{j=1}^{N} \delta_i^j y_j \quad . \tag{6.5.6}$$

Since the selection of any of the y_j as the ith element is equally probable, one has

$$P(\delta_i^j = 1) = 1/N \quad . \tag{6.5.7}$$

Since δ_i^j describes a random procedure, it is clearly a random variable itself. Its expectation value is found from Eq. (3.3.2) (where $n = 2$, $x_1 = 0$, $x_2 = 1$) to be

$$E(\delta_i^j) = P(\delta_i^j = 1) = 1/N \quad . \tag{6.5.8}$$

If now one element x_i of the sample is determined, one then has only $(N-1)$ selection possibilities out of the population for a further element, e.g., x_k. That is,

$$P(\delta_i^j \delta_k^\ell = 1) = \frac{1}{N}\frac{1}{N-1} = E(\delta_i^j \delta_k^\ell) \quad . \tag{6.5.9}$$

Since the sample is taken without replacement, one has $j \neq \ell$, i.e.,

$$\delta_i^j \delta_k^j = 0 \quad . \tag{6.5.10}$$

Similarly one has

$$\delta_i^j \delta_i^\ell = 0 \quad , \tag{6.5.11}$$

since two different elements of the population cannot simultaneously occur as the ith element of the sample.

We consider now the expectation value of x_1,

$$E(\mathsf{x}_1) = E\left\{\sum_{j=1}^{N} \delta_1^j y_j\right\} = \sum_{j=1}^{N} y_j E(\delta_1^j) = \frac{1}{N}\sum_{j=1}^{N} y_j = \bar{y} \quad . \tag{6.5.12}$$

Since x_1 is in not in any way special, the expectation values of all elements of the sample, and thus also of their arithmetic mean, have the same value

$$E(\bar{\mathsf{x}}) = \frac{1}{n}\sum_{i=1}^{n} E(\mathsf{x}_i) = \bar{y} \quad . \tag{6.5.13}$$

The arithmetic mean of the sample is thus an unbiased estimator for the population mean.

Next we consider the *sample variance*

$$\mathsf{s}_x^2 = \frac{1}{n-1}\sum_{i=1}^{n} (\mathsf{x}_i - \bar{\mathsf{x}})^2 \quad . \tag{6.5.14}$$

By means of a somewhat longer calculation it can be shown that the expectation value is

$$E(\mathsf{s}_x^2) = \sigma^2(y) \quad . \tag{6.5.15}$$

The sample variance is thus an unbiased estimator for the population variance.

The *variance of the mean* is also of interest:

$$\sigma^2(\bar{\mathsf{x}}) = E\{(\bar{\mathsf{x}} - E(\bar{\mathsf{x}}))^2\} \quad .$$

$E(\bar{\mathsf{x}}) = \bar{y}$ is, however, a fixed, *not* a random quantity, whereas $\bar{\mathsf{x}}$ depends on the individual sample, and is hence a random variable. One therefore has

$$\begin{aligned} \sigma^2(\bar{\mathsf{x}}) &= E(\bar{\mathsf{x}}^2) - \bar{y}^2 = \frac{1}{n}\left\{\left(1 - \frac{n}{N}\right)\sigma^2(y) + n\bar{y}^2\right\} - \bar{y}^2 \quad , \\ \sigma^2(\bar{\mathsf{x}}) &= \frac{\sigma^2(y)}{n}\left(1 - \frac{n}{N}\right) \quad . \end{aligned} \tag{6.5.16}$$

Comparing with the case of an infinite continuous sample (6.2.3) one sees the additional factor $(1 - n/N)$. This corresponds to the fact that the variance of $\bar{\mathsf{x}}$ vanishes in the case $n = N$, where the "sample" contains the entire population and where one has exactly $\bar{\mathsf{x}} = \bar{y}$.

6.6 Samples from Gaussian Distributions: χ^2-Distribution

We return now to continuously distributed populations and consider in particular a Gaussian distribution with mean a and variance σ^2. According to (5.7.7), the characteristic function of such a Gaussian distribution is

$$\varphi_{\mathsf{x}}(t) = \exp(\mathrm{i}ta)\exp(-\frac{1}{2}\sigma^2 t^2) \quad . \tag{6.6.1}$$

We now take a sample of size n from the population. The characteristic function of the sample mean was given in (6.2.2) in terms of the characteristic function of the population. From this we have

$$\varphi_{\bar{\mathsf{x}}}(t) = \left\{\exp\left(\mathrm{i}\frac{t}{n}a - \frac{\sigma^2}{2}\left(\frac{t}{n}\right)^2\right)\right\}^n \quad . \tag{6.6.2}$$

If we consider $(\bar{\mathsf{x}} - a) = (\bar{\mathsf{x}} - \hat{x})$ in place of x, then one obtains

$$\varphi_{\bar{\mathsf{x}}-a}(t) = \exp\left(-\frac{\sigma^2 t^2}{2n}\right) \quad . \tag{6.6.3}$$

This is again the characteristic function of a normal distribution, but with a different variance,

$$\sigma^2(\bar{\mathsf{x}}) = \sigma^2(\mathsf{x})/n \quad . \tag{6.6.4}$$

For the simple case of a standard Gaussian distribution ($a = 0$, $\sigma^2 = 1$) we have

$$\varphi_{\bar{\mathsf{x}}}(t) = \exp(-t^2/2n) \quad . \tag{6.6.5}$$

We take a sample from this distribution

$$\mathsf{x}_1, \mathsf{x}_2, \ldots, \mathsf{x}_n \quad ,$$

but we are interested in particular in the sum of the squares of the sample elements,

$$\mathsf{x}^2 = \mathsf{x}_1^2 + \mathsf{x}_2^2 + \ldots + \mathsf{x}_n^2 \quad . \tag{6.6.6}$$

We want to show that the quantity x^2 follows the distribution function*

$$F(\chi^2) = \frac{1}{\Gamma(\lambda)2^\lambda} \int_0^{\chi^2} u^{\lambda-1} \mathrm{e}^{-\frac{1}{2}u} \, \mathrm{d}u \quad , \tag{6.6.7}$$

where

$$\lambda = \frac{1}{2}n \quad . \tag{6.6.8}$$

The quantity n is called the *number of degrees of freedom*.

We first introduce the abbreviation

$$\frac{1}{\Gamma(\lambda)2^\lambda} = k \tag{6.6.9}$$

and determine the probability density to be

$$f(\chi^2) = k(\chi^2)^{\lambda-1} \mathrm{e}^{-\frac{1}{2}\chi^2} \quad . \tag{6.6.10}$$

For two degrees of freedom, the probability density is clearly an exponential function. First we want to prove what was claimed by (6.6.7) for one degree of freedom ($\lambda = 1/2$). Thus we ask for the probability that $\mathsf{x}^2 < \chi^2$, or rather, that $-\sqrt{\chi^2} < \mathsf{x} < +\sqrt{\chi^2}$. This is

$$\begin{aligned} F(\chi^2) &= P(\mathsf{x}^2 < \chi^2) = P(-\sqrt{\chi^2} < \mathsf{x} < +\sqrt{\chi^2}) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\sqrt{\chi^2}}^{\sqrt{\chi^2}} \mathrm{e}^{-\frac{1}{2}x^2} \, \mathrm{d}x = \frac{2}{\sqrt{2\pi}} \int_0^{\sqrt{\chi^2}} \mathrm{e}^{-\frac{1}{2}x^2} \, \mathrm{d}x \quad . \end{aligned}$$

*The symbol χ^2 (chi squared) was introduced by K. Pearson. Although it is written as something squared, which reminds one of its origin as a sum of squares, it is treated as a usual random variable.

Setting $x^2 = u$, $\mathrm{d}u = 2x\,\mathrm{d}x$, we obtain directly

$$F(\chi^2) = \frac{1}{\sqrt{2\pi}} \int_0^{\chi^2} u^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2}u}\,\mathrm{d}u \quad . \tag{6.6.11}$$

To prove the general case we first find the characteristic function of the χ^2-distribution to be

$$\varphi_{\chi^2}(t) = \int_0^{\infty} k(\chi^2)^{\lambda-1} \exp(-\frac{1}{2}\chi^2 + \mathrm{i}t\chi^2)\,\mathrm{d}\chi^2 \tag{6.6.12}$$

or with $(1/2 - \mathrm{i}t)\chi^2 = \nu$,

$$\varphi_{\chi^2}(t) = 2^{\lambda}(1-2\mathrm{i}t)^{-\lambda} k \int_0^{\infty} \nu^{\lambda-1}\mathrm{e}^{-\nu}\,\mathrm{d}\nu \quad .$$

The integral on the right side is, according to (D.1.1), equal to $\Gamma(\lambda)$. One therefore has

$$\varphi_{\chi^2}(t) = (1-2\mathrm{i}t)^{-\lambda} \quad . \tag{6.6.13}$$

If we now consider the case of a second distribution with λ', then one has

$$\varphi'_{\chi^2}(t) = (1-2\mathrm{i}t)^{-\lambda'} \quad .$$

Since the characteristic function of a sum is equal to the product of the characteristic functions, one has the following important theorem:

> The sum of two independent χ^2 variables with
> n_1, n_2 degrees of freedom follows itself a χ^2-distribution
> with $n = n_1 + n_2$ degrees of freedom.

This theorem can now be used to easily generalize the claim (6.6.7), proven up to now only for $n = 1$. The proof follows from the fact that the individual terms of the sum of squares are independent and therefore (6.6.6) can be treated as the sum of n different χ^2 variables, each with one degree of freedom.

In order to obtain the expectation value and variance of the χ^2-distribution, we use the characteristic function, whose derivatives (5.5.7) give the central moments. We obtain

$$\begin{aligned} E(\mathsf{x}^2) &= -\mathrm{i}\varphi'(0) = 2\lambda \quad , \\ E(\mathsf{x}^2) &= n \end{aligned} \tag{6.6.14}$$

and

$$\begin{aligned} E\{(\mathsf{x}^2)^2\} &= -\varphi''(0) = 4\lambda^2 + 4\lambda \quad , \\ \sigma^2(\mathsf{x}^2) &= E\{(\mathsf{x}^2)^2\} - \{E(\mathsf{x}^2)\}^2 = 4\lambda \quad , \\ \sigma^2(\mathsf{x}^2) &= 2n \quad . \end{aligned} \tag{6.6.15}$$

The expectation value of the χ^2-distribution is thus equal to the number of degrees of freedom, and the variance is two times larger. Figure 6.8 shows the probability density of the χ^2 distribution for various values of n. One sees [as can be directly seen also from (6.6.10)] that for $\chi^2 = 0$, the function diverges for $n = 1$, is equal to $1/2$ for $n = 2$, and vanishes for $n \geq 3$. A table of the χ^2-distribution is provided in the appendix (Table I.6).

The χ^2-distribution is of great significance in many applications, where the quantity χ^2 is used as a measure of confidence in a certain result. The smaller the value of χ^2, the greater is the confidence in the result. (After all, χ^2 was defined as the sum of squares of deviations of elements of a sample from the population mean. See Sect. 8.7.) The distribution function

$$F(\chi^2) = P(\mathsf{x}^2 < \chi^2) \tag{6.6.16}$$

gives the probability that the random variable x^2 is not larger than χ^2. In practice, one frequently uses the quantity

$$W(\chi^2) = 1 - F(\chi^2) \tag{6.6.17}$$

as a measure of confidence in a result. $W(\chi^2)$ is often called the *confidence level*. $W(\chi^2)$ is large for small values of χ^2 and falls with increasing χ^2. The

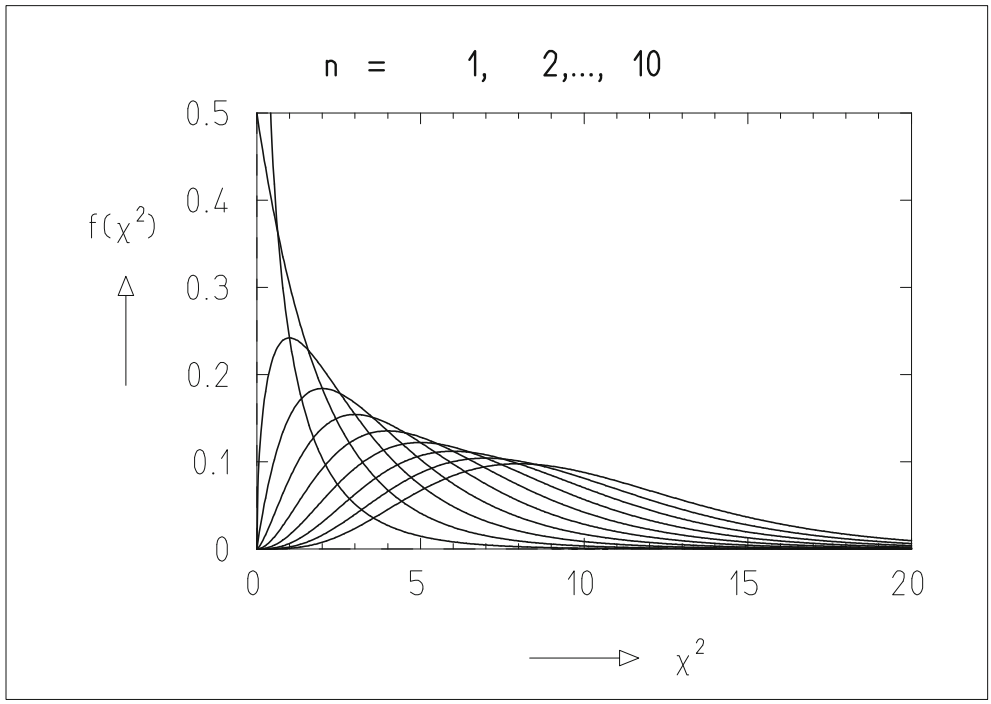

Fig. 6.8: Probability density of χ^2 for the number of degrees of freedom $n = 1, 2, \ldots, 10$. The expectation value $E(\chi^2) = n$ moves to the right as n increases.

distribution function (6.6.16) is shown in Fig. 6.9 for various numbers of degrees of freedom n. The inverse function, which gives the quantiles of the χ^2 distribution,

$$\chi_F^2 = \chi^2(F) = \chi^2 \tag{6.6.18}$$

is used especially often in "hypothesis testing" (see Sect. 8.7). It is tabulated in the appendix (Table I.7).

Up to now we have restricted ourselves to the case where the population is described by the standard normal distribution. Usually, however, one has a normal distribution in general form with mean a and variance σ^2. Then the sum of squares (6.6.6) is clearly no longer distributed according to the χ^2-distribution. One immediately obtains, however, a χ^2-distribution by considering the quantity

$$\mathsf{x}^2 = \frac{(\mathsf{x}_1 - a)^2 + (\mathsf{x}_2 - a)^2 + \cdots + (\mathsf{x}_n - a)^2}{\sigma^2} \quad . \tag{6.6.19}$$

This result follows directly from Eq. (5.8.4).

If the expectation values a_i and variances σ_i of the individual variables are different, then one has

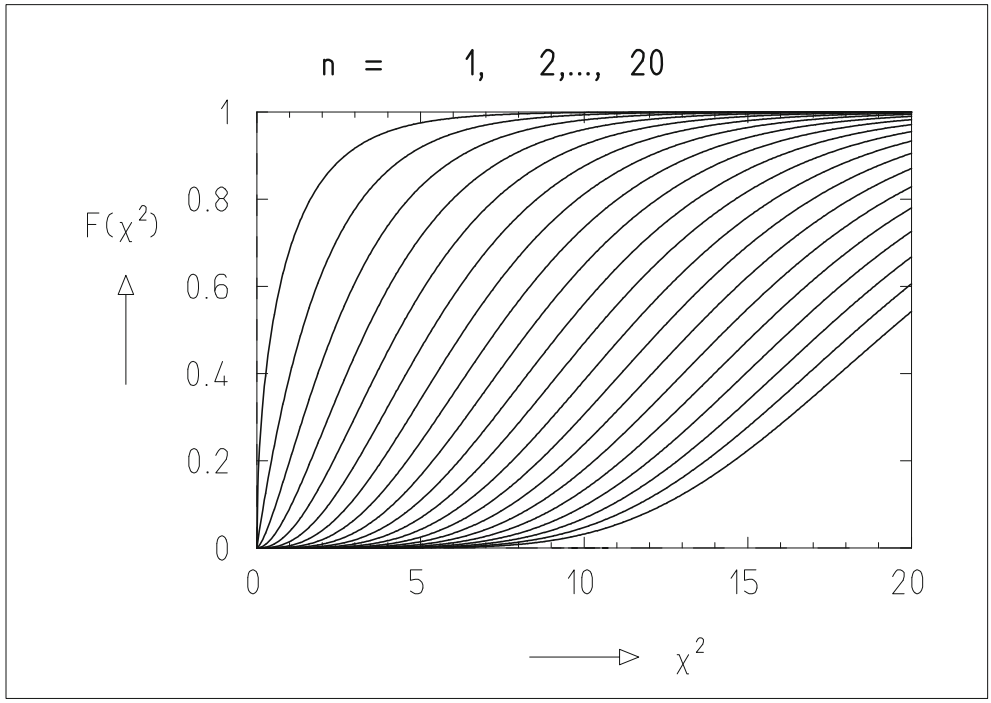

Fig. 6.9: Distribution function for χ^2 for the number of degrees of freedom $n = 1, 2, \ldots, 20$. The function for $n = 1$ is the curve at the far left, and the function for $n = 20$ is at the far right.

$$\mathsf{x}^2 = \frac{(\mathsf{x}_1 - a_1)^2}{\sigma_1^2} + \frac{(\mathsf{x}_2 - a_2)^2}{\sigma_2^2} + \cdots + \frac{(\mathsf{x}_n - a_n)^2}{\sigma_n^2} \quad . \tag{6.6.20}$$

Finally if the n variables are not independent, but are described by a joint normal distribution (5.10.1) with the expectation values given by the vector $\mathbf{a}$ and the covariance matrix $C = B^{-1}$, then one has

$$\mathsf{x}^2 = (\mathbf{x} - \mathbf{a})^{\mathrm{T}} B (\mathbf{x} - \mathbf{a}) \quad . \tag{6.6.21}$$

6.7 χ^2 and Empirical Variance

In Eq. (6.2.6) we found that

$$\mathsf{s}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\mathsf{x}_i - \bar{\mathsf{x}})^2 \tag{6.7.1}$$

is a consistent, unbiased estimator for the variance σ^2 of a population. Let the x_i be independent and normally distributed with standard deviation σ. We want to show that the quantity

$$\frac{n-1}{\sigma^2} \mathsf{s}^2 \tag{6.7.2}$$

follows the χ^2-distribution with $f = n - 1$ degrees of freedom. We first carry out an orthogonal transformation of the n variables x_i (see Sect. 3.8):

$$\begin{aligned}
\mathsf{y}_1 &= \frac{1}{\sqrt{1 \cdot 2}} (\mathsf{x}_1 - \mathsf{x}_2) \quad , \\
\mathsf{y}_2 &= \frac{1}{\sqrt{2 \cdot 3}} (\mathsf{x}_1 + \mathsf{x}_2 - 2\mathsf{x}_3) \quad , \\
\mathsf{y}_3 &= \frac{1}{\sqrt{3 \cdot 4}} (\mathsf{x}_1 + \mathsf{x}_2 + \mathsf{x}_3 - 3\mathsf{x}_4) \quad , \\
&\vdots \\
\mathsf{y}_{n-1} &= \frac{1}{\sqrt{(n-1)n}} (\mathsf{x}_1 + \mathsf{x}_2 + \cdots + \mathsf{x}_{n-1} - (n-1)\mathsf{x}_n) \quad , \\
\mathsf{y}_n &= \frac{1}{\sqrt{n}} (\mathsf{x}_1 + \mathsf{x}_2 + \cdots + \mathsf{x}_n) = \sqrt{n} \bar{\mathsf{x}} \quad .
\end{aligned} \tag{6.7.3}$$

One can verify that this transformation is in fact orthogonal, i.e., that

$$\sum_{i=1}^{n} \mathsf{x}_i^2 = \sum_{i=1}^{n} \mathsf{y}_i^2 \quad . \tag{6.7.4}$$

Since a sum or difference of independent normally distributed quantities is again itself normally distributed, all of the y_i are normally distributed. The factors in (6.7.3) ensure that the y_i have mean values of zero and standard deviations σ.

From (6.7.1) and (6.7.2) one then has

$$\begin{aligned}(n-1)\mathsf{s}^2 &= \sum_{i=1}^{n}(\mathsf{x}_i-\bar{\mathsf{x}})^2 = \sum_{i=1}^{n}\mathsf{x}_i^2 - 2\bar{\mathsf{x}}\sum_{i=1}^{n}\mathsf{x}_i + n\bar{\mathsf{x}}^2 \\ &= \sum_{i=1}^{n}\mathsf{x}_i^2 - n\bar{\mathsf{x}}^2 = \sum_{i=1}^{n}\mathsf{y}_i^2 - \mathsf{y}_n^2 = \sum_{i=1}^{n-1}\mathsf{y}_i^2 \quad .\end{aligned}$$

This expression is a sum of only $(n-1)$ independent squared terms. A comparison with (6.6.19) shows that the quantity (6.7.2) in fact follows a χ^2-distribution with $(n-1)$ degrees of freedom.

The squared terms $(\mathsf{x}_i - \bar{\mathsf{x}})^2$ are not linearly independent. One has the following relation between them:

$$\sum_{i=1}^{n}(\mathsf{x}_i-\bar{\mathsf{x}}) = 0 \quad .$$

One can show that every additional relation between the squared terms reduces the number of degrees of freedom by one. Later we will make frequent use of this result, which we only state here without proof.

6.8 Sampling by Counting: Small Samples

Samples are often obtained in the following way. One draws n elements from a population, checks if they possess a given characteristic and accepts only those k elements into the sample that have the characteristic. The remaining $n - \mathsf{k}$ elements are rejected, i.e., their properties are not recorded. This approach thus becomes the counting of k out of n elements drawn.

This approach corresponds exactly to selecting a sample according to a binomial distribution. The parameters p and q of this distribution correspond then to the occurrence or non-occurrence of the property in question. As will be shown in Example 7.5,

$$\mathsf{S}(p) = \frac{\mathsf{k}}{n} \tag{6.8.1}$$

is the maximum likelihood estimator of the parameter p. The variance of S is

$$\sigma^2(\mathsf{S}(p)) = \frac{p(1-p)}{n} \quad . \tag{6.8.2}$$

By using (6.8.1) it can be estimated from the sample by

$$\mathsf{s}^2(\mathsf{S}(p)) = \frac{1}{n}\frac{\mathsf{k}}{n}\left(1-\frac{\mathsf{k}}{n}\right) \quad . \tag{6.8.3}$$

We define the error $\Delta\mathsf{k}$ as

$$\Delta\mathsf{k} = \sqrt{[\mathsf{s}^2(\mathsf{S}(np))]} \quad . \tag{6.8.4}$$

By using (6.8.3) we obtain

$$\Delta\mathsf{k} = \sqrt{\left[\mathsf{k}\left(1-\frac{\mathsf{k}}{n}\right)\right]} \quad . \tag{6.8.5}$$

The error $\Delta\mathsf{k}$ only depends on the number of elements counted and on the size of the sample. It is called the *statistical error*. A particularly important case is that of small k, or more precisely, the case $\mathsf{k} \ll n$. In this limit we can define $\lambda = np$ and following Sect. 5.4 consider the counted number k as a single element of a sample taken from a Poisson distributed population with parameter λ. From (6.8.1) and (6.8.5) we obtain

$$\mathsf{S}(\lambda) = \mathsf{S}(np) = \mathsf{k} \quad , \tag{6.8.6}$$

$$\Delta\lambda = \sqrt{\mathsf{k}} \quad . \tag{6.8.7}$$

(This can be derived by using the result of Example 7.4 with $N = 1$.) The result (6.8.7) is often written in an actually incorrect but easy to remember form,

$$\Delta\mathsf{k} = \sqrt{\mathsf{k}} \quad ,$$

which is read: The *statistical error* of the *counted number* k is $\sqrt{\mathsf{k}}$.

In order to interpret the statistical error $\Delta\lambda = \sqrt{\mathsf{k}}$ we must examine the Poisson distribution somewhat more closely. Let us begin with the case where k is not too small (say, $\mathsf{k} > 20$). For large values of λ the Poisson distribution becomes a Gaussian distribution with mean λ and variance $\sigma^2 = \lambda$. This can be seen qualitatively from Fig. 5.6. As long as k is not too small, i.e., $\mathsf{k} \gg 1$, we can then treat the Poisson distribution in k with parameter λ as a normal distribution in x with mean λ and variance $\sigma^2 = \lambda$. The discrete variable k is then replaced by the continuous variable x. The probability density of x is

$$f(x;\lambda) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(x-\lambda)^2}{2\sigma^2}\right\} = \frac{1}{\sqrt{2\pi\lambda}}\exp\left\{-\frac{(x-\lambda)^2}{2\lambda}\right\} \quad . \tag{6.8.8}$$

The observation of k events corresponds to observing once the value of the random variable $\mathsf{x} = \mathsf{k}$.

With the help of the probability density (6.8.8) we now want to determine the *confidence limits* at a given *confidence level* $\beta = 1-\alpha$ in such a way that

$$P(\lambda_- \le \lambda \le \lambda_+) = 1-\alpha \quad . \tag{6.8.9}$$

That is, one requires that the probability that the true value of λ is contained within the confidence limits λ_- and λ_+ be equal to the confidence level $1-\alpha$. The limiting cases $\lambda = \lambda_-$ and $\lambda = \lambda_+$ are depicted in Fig. 6.10. They are determined such that

$$P(x > \mathsf{k}|\lambda = \lambda_+) = 1-\alpha/2 \quad , \qquad P(x < \mathsf{k}|\lambda = \lambda_-) = 1-\alpha/2 \quad . \tag{6.8.10}$$

One clearly has

$$\begin{aligned} \alpha/2 &= \int_{x=-\infty}^{x=\mathsf{k}} f(x;\lambda_+)\,\mathrm{d}x = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x=\mathsf{k}} \exp\left\{-\frac{(x-\lambda_+)^2}{2\sigma^2}\right\} \\ &= \int_{u=-\infty}^{u=(\mathsf{k}-\lambda_+)/\sigma} \phi_0(u)\,\mathrm{d}u = \psi_0\left(\frac{\mathsf{k}-\lambda_+}{\sigma}\right) \quad , \end{aligned} \tag{6.8.11}$$

and correspondingly

$$1-\alpha/2 = \int_{-\infty}^{x=\mathsf{k}} f(x;\lambda_-)\,\mathrm{d}x = \psi_0\left(\frac{\mathsf{k}-\lambda_-}{\sigma}\right) \quad . \tag{6.8.12}$$

Here ϕ_0 and ψ_0 are the probability density and distribution function of the standard normal distribution introduced in Sect. 5.8. By using the inverse function Ω of the distribution function ψ_0 [see Eq. (5.8.8)], one obtains

$$\frac{\mathsf{k}-\lambda_-}{\sigma} = \Omega(1-\alpha/2) \quad , \qquad \frac{\mathsf{k}-\lambda_+}{\sigma} = \Omega(\alpha/2) \quad . \tag{6.8.13}$$

Because of (5.8.10) one has $\Omega(1-\alpha/2) = \Omega'(1-\alpha)$ and because of the symmetry of the function Ω, $\Omega(1-\alpha/2) = -\Omega(\alpha/2)$. Further, since $\alpha < 1$, one has $\Omega(1-\alpha/2) > 0$, $\Omega(\alpha/2) < 0$. From this we finally obtain

$$\lambda_- = \mathsf{k} - \sigma\Omega'(1-\alpha) \quad , \qquad \lambda_+ = \mathsf{k} + \sigma\Omega'(1-\alpha) \quad . \tag{6.8.14}$$

According to (6.8.6), k is the best estimator for λ. Since $\sigma^2 = \lambda$, the best estimator for σ is given by $\mathsf{s} = \sqrt{\mathsf{k}}$. Since we have assumed that $\mathsf{k} \gg 1$, the uncertainty in s is significantly smaller than the uncertainty in k. We can therefore substitute $x = k$ and $\mathsf{s} = \sqrt{\mathsf{k}}$ in (6.8.9) and obtain for the confidence interval with confidence level $1-\alpha$

$$\lambda_- = \mathsf{k} - \sqrt{\mathsf{k}}\Omega'(1-\alpha) \le \lambda \le \mathsf{k} + \sqrt{\mathsf{k}}\Omega'(1-\alpha) = \lambda_+ \quad . \tag{6.8.15}$$

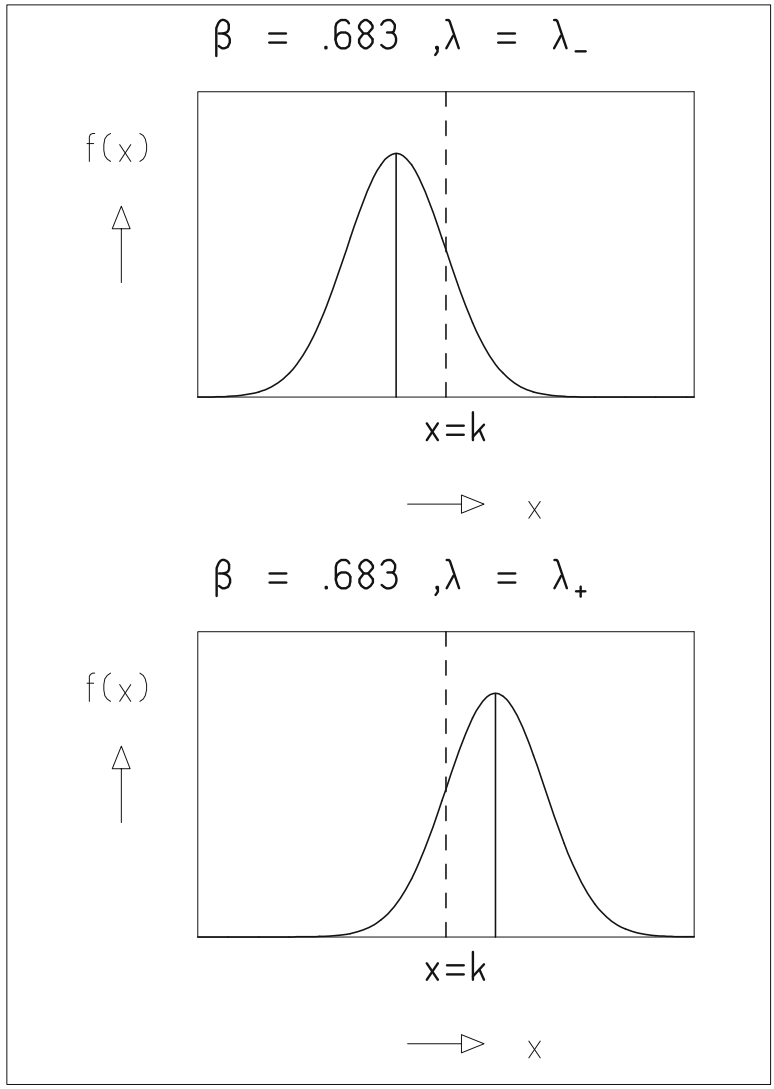

Fig. 6.10: Normal distribution with mean λ and standard deviation σ for $\lambda = \lambda_-$ and $\lambda = \lambda_+$.

For $1-\alpha = 68.3\,\%$ we find from Sect. 5.8 or Table I.5 that $\Omega'(\alpha) = 1$. What is usually reported,

$$\lambda = \mathsf{k} \pm \sqrt{\mathsf{k}} \quad ,$$

which was already the result of (6.8.6) and (6.8.7), thus gives the confidence limits at the confidence level of 68.3 %, but only for the case $\mathsf{k} \gg 1$. For the confidence level of 90 %, i.e., for $\alpha = 0.1$, we find $\Omega'(0.1) = 1.65$ and for the confidence level 99 % one has $\Omega'(0.01) = 2.57$.

For very small values of k, one can no longer replace the Poisson distribution with the normal distribution. We follow therefore reference [25]. We start again from Eq. (6.8.10), but use, instead of the probability density (6.8.8) for

the continuous random variable x with fixed parameter λ, the Poisson probability for observing the discrete random variable n for a given λ,

$$f(n;\lambda) = \frac{\lambda^n}{n!}\mathrm{e}^{-\lambda} \quad . \tag{6.8.16}$$

For the observation k we now determine the confidence limits λ_- and λ_+, which fulfill (6.8.10) with $x = n$ (Fig. 6.11) and obtain in analogy to (6.8.11) and (6.8.12)

$$1-\alpha/2 \;=\; \sum_{n=\mathsf{k}+1}^{\infty} f(n;\lambda_+) = 1 - \sum_{n=0}^{\mathsf{k}} f(n;\lambda_+) = 1 - F(\mathsf{k}+1;\lambda_+) \quad ,$$

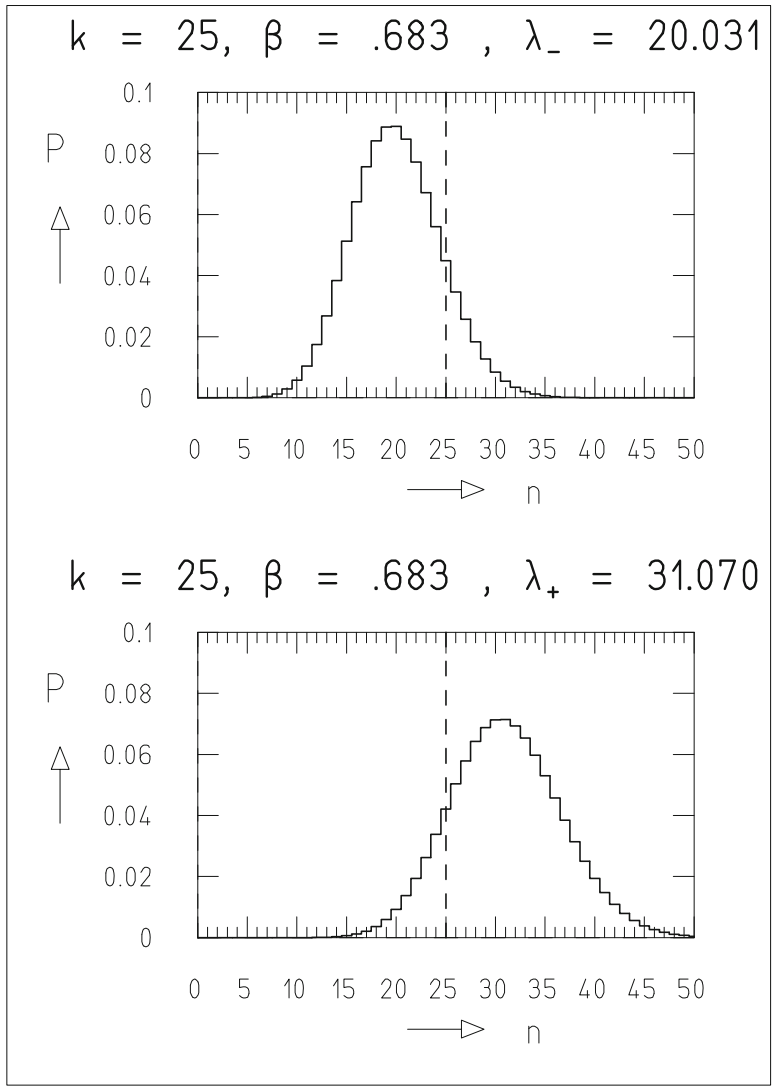

Fig. 6.11: Poisson distribution with parameter λ for $\lambda = \lambda_-$ and $\lambda = \lambda_+$.

$$1-\alpha/2 \;=\; \sum_{n=0}^{\mathsf{k}-1} f(n;\lambda_-) = F(\mathsf{k};\lambda_-)$$

or

$$\alpha/2 \;=\; F(\mathsf{k}+1;\lambda_+) \quad , \tag{6.8.17}$$

$$1-\alpha/2 \;=\; F(\mathsf{k};\lambda_-) \quad . \tag{6.8.18}$$

Here,

$$F(k;\lambda) = \sum_{n=0}^{k-1} f(n;\lambda) = P(\mathsf{k} < k)$$

is the distribution function of the Poisson distribution.

In order to obtain numerical values for the confidence limits λ_+ and λ_-, we solve Eqs. (6.8.17) and (6.8.18). That is, we must construct the inverse function of the Poisson distribution for fixed k and given probability P (in our case $\alpha/2$ and $1-\alpha/2$),

$$\lambda = \lambda_P(k) \quad . \tag{6.8.19}$$

This is done numerically with the method StatFunct.quantile Poisson. The function (6.8.19) is given in Table I.1 for frequently occurring values of P.

For extremely small samples one is often only interested in an *upper confidence limit* at confidence level $\beta = 1-\alpha$. This is obtained by requiring

$$P(n > \mathsf{k}|\lambda = \lambda^{(\mathrm{up})}) = \beta = 1-\alpha \tag{6.8.20}$$

instead of (6.8.10). Thus one has

$$\alpha = \sum_{n=0}^{\mathsf{k}} f(n;\lambda^{(\mathrm{up})}) = F(\mathsf{k}+1;\lambda^{(\mathrm{up})}) \quad . \tag{6.8.21}$$

For the extreme case $\mathsf{k} = 0$, i.e., for a sample in which no event was observed, one obtains the upper limit $\lambda^{(\mathrm{up})}$ by inverting $\alpha = F(1;\lambda^{(\mathrm{up})})$. The upper limit then has the following meaning. If the true value of the parameter were in fact $\lambda = \lambda^{(\mathrm{up})}$ and if one were to repeat the experiment many times, then the probability of observing at least one event is β. The observation $\mathsf{k} = 0$ is then expressed in the following way: One has $\lambda < \lambda^{(\mathrm{up})}$ with a confidence level of $1-\alpha$. From Table I.1 one finds that $k = 0$ corresponds to $\lambda < 2.996 \approx 3$ at a confidence level of 95 %.

Example 6.7: Determination of a lower limit for the lifetime of the proton from the observation of no decays

As already mentioned, the probability for the decay of a radioactive nucleus with the time t is

$$P(t) = \frac{1}{\tau} \int_0^t \mathrm{e}^{-x/\tau}\, \mathrm{d}x \quad .$$

Here τ is the mean lifetime of the nucleus. For $t \ll \tau$ the expression simplifies to

$$P(t) = t/\tau \quad .$$

For a total of N nuclei one expects that

$$k = N\,P(t) = N \cdot t/\tau$$

nuclei will decay within the time t. The mean lifetime τ is obtained by counting the number of such decays. If one observes k decays from a total of N nuclei in a time t, then one obtains as the measured value of τ

$$\tilde{\tau} = \frac{N}{\mathsf{k}} t \quad .$$

Of particular interest is the mean lifetime of the proton, one of the primary building blocks of matter. In experiments recently carried out with great effort, one observes large numbers of protons with detectors capable of detecting each individual decay. Up to now, not a single decay has been seen. According to Table I.9, the true expected number of decays λ does not exceed three (at a confidence level of 95 %). One has therefore

$$\tau > \frac{N}{3} t$$

at this confidence level. Typical experimental values are $t = 0.3$ years, $N = 10^{33}$, i.e.,

$$\tau > 10^{32}\,\text{years} \quad .$$

The proton can therefore be considered as stable even over cosmological time scales, if one considers that the age of the universe is estimated to be only around 10^{10} years. ■

6.9 Small Samples with Background

In many experiments one is faced with the following situation. For the detected events one cannot determine whether they belong to the type that is actually of interest (*signal events*) or to another type (*background events*).

For the expected number of events in the experiment one then has a Poisson distribution with the parameter $\lambda = \lambda_S + \lambda_B$. Here λ_S is the sought after parameter of the number of signal events, and λ_B is the parameter for the background events, which must of course be known if one wants to obtain information about λ_S. (In an experiment as in Example 6.7, one might have, for example, an admixture of radioactive nuclei whose decays cannot be distinguished from those of the proton. If the number of such nuclei and their lifetime is known, then λ_B can be computed.)

We are now tempted to simply take the results of the last section, to determine the confidence limits $\lambda_\pm$ and the upper limit $\lambda^{(\mathrm{up})}$ and to set $\lambda_{S\pm} = \lambda_\pm - \lambda_B$, $\lambda_S^{(\mathrm{up})} = \lambda^{(\mathrm{up})} - \lambda_B$. This procedure can, however, lead to nonsensical results. (As seen in Example 6.7, one has $\lambda^{(\mathrm{up})} = 3$ at a confidence level of 95 %, for $k = 0$. For $\lambda_B = 4$, $k = 0$ we would obtain $\lambda_S^{(\mathrm{up})} = -1$, although a value $\lambda_S < 0$ has no meaning.)

The considerations up to now are based on the following. The probability for observing n events, $n = n_S + n_B$, is

$$f(n; \lambda_S + \lambda_B) = \frac{1}{n!} \mathrm{e}^{-(\lambda_S+\lambda_B)} (\lambda_S + \lambda_B)^n \quad , \tag{6.9.1}$$

and the probabilities to observe n_S signal events, and n_B background events are

$$f(n_S; \lambda_S) = \frac{1}{n_S!} \mathrm{e}^{-\lambda_S} \lambda_S^{n_S} \quad , \tag{6.9.2}$$

$$f(n_B; \lambda_B) = \frac{1}{n_B!} \mathrm{e}^{-\lambda_B} \lambda_B^{n_B} \quad . \tag{6.9.3}$$

The validity of (6.9.1) was shown in Example 5.5 with the help of the characteristic function, starting from the independence of the two Poisson distributions (6.9.2) and (6.9.3). One can also obtain them directly by summation of all products of the probabilities (6.9.2) and (6.9.3) that lead to $n = n_S + n_B$, by application of (B.4) and (B.6),

$$\begin{aligned}
&\sum_{n_S=0}^{n} f(n_S; \lambda_S) f(n - n_S; \lambda_B) \\
&= \mathrm{e}^{-(\lambda_S+\lambda_B)} \sum_{n_S=0}^{n} \frac{1}{n_S!(n-n_S)!} \lambda_S^{n_S} \lambda_B^{n-n_S} \\
&= \frac{1}{n!} \mathrm{e}^{-(\lambda_S+\lambda_B)} \sum_{n_S=0}^{n} \binom{n}{n_S} \lambda_S^{n_S} \lambda_B^{n-n_S} \\
&= \frac{1}{n!} \mathrm{e}^{-(\lambda_S+\lambda_B)} (\lambda_S + \lambda_B)^n \\
&= f(n; \lambda_S + \lambda_B) \quad .
\end{aligned}$$

The difficulties explained above are overcome with a method developed by ZECH [26]. In an experiment in which k events are recorded, the number of background events cannot simply be given by (6.9.3), since from the result of the experiment it is known that $n_B \leq \mathsf{k}$. One must therefore replace (6.9.3) by

$$f'(n_B;\lambda) = f(n_B;\lambda_B) \Bigg/ \sum_{n_B=0}^{\mathsf{k}} f(n_B;\lambda_B) \quad , \qquad n_B \leq \mathsf{k} \quad . \tag{6.9.4}$$

This distribution is normalized to unity in the region $0 \leq n_B \leq k$. In a corresponding way the distribution

$$f'(n;\lambda_S+\lambda_B) = f(n;\lambda_S+\lambda_B) \Bigg/ \sum_{n_B=0}^{\mathsf{k}} f(n_B;\lambda_B) \tag{6.9.5}$$

takes the place of (6.9.1).

In this way one obtains in analogy to (6.8.17) and (6.8.18) for the limits of the confidence interval $\lambda_{S-} \leq \lambda_S \leq \lambda_{S+}$ at a confidence level of $1-\alpha$

$$\alpha/2 = F'(\mathsf{k}+1,\lambda_{S+}+\lambda_B) \quad , \tag{6.9.6}$$

$$1-\alpha/2 = F'(\mathsf{k},\lambda_{S-}+\lambda_B) \quad . \tag{6.9.7}$$

Here,

$$F'(\mathsf{k};\lambda_S+\lambda_B) = \sum_{n=0}^{\mathsf{k}-1} f'(n;\lambda_S+\lambda_B) = P(\mathsf{k}<k) \tag{6.9.8}$$

is the distribution function of the renormalized distribution (6.9.4). If only an upper limit at confidence level $1-\alpha$ is desired, then one clearly has in analogy to (6.8.21)

$$\alpha = F'(\mathsf{k}+1,\lambda_S^{(\mathrm{up})}+\lambda_B) \quad . \tag{6.9.9}$$

Table 6.3 gives some numerical values computed with the methods of the class `SmallSample`. Note that for $\mathsf{k}=0$, Eq. (6.9.7) has no meaning, so that λ_{S-} cannot be defined. In this case Eq. (6.9.6) and also λ_{S+} are not meaningful. (In the table, however, the values for λ_{S-} and λ_{S+} are shown as computed by the program. This sets $\lambda_{S-}=0$ and computes λ_{S+} according to (6.9.6).)

6.10 Determining a Ratio of Small Numbers of Events

Often a number of *signal* events k is measured and compared to a number of *reference* events d. One is interested in the true value r of the ratio of the

Table 6.3: Limits λ_{S-} and λ_{S+} of the confidence interval and upper confidence limits $\lambda_S^{(\mathrm{up})}$, for various values of λ_B and various very small sample sizes k for a fixed confidence level of 90%.

$\beta = 0.90$				
k	λ_B	λ_{S-}	λ_{S+}	$\lambda_S^{(\mathrm{up})}$
0	0.0	0.000	2.996	2.303
0	1.0	0.000	2.996	2.303
0	2.0	0.000	2.996	2.303
1	0.0	0.051	4.744	3.890
1	1.0	0.051	4.113	3.272
1	2.0	0.051	3.816	2.995
2	0.0	0.355	6.296	5.322
2	1.0	0.100	5.410	4.443
2	2.0	0.076	4.824	3.877
3	0.0	0.818	7.754	6.681
3	1.0	0.226	6.782	5.711
3	2.0	0.125	5.983	4.926
4	0.0	1.366	9.154	7.994
4	1.0	0.519	8.159	7.000
4	2.0	0.226	7.241	6.087
5	0.0	1.970	10.513	9.275
5	1.0	1.009	9.514	8.276
5	2.0	0.433	8.542	7.306

number of signal events to the number of reference events, or more precisely, the ratio of the probability to observe a signal event to that of a reference event. As an estimator for this ratio one clearly uses

$$\tilde{r} = \mathsf{k}/\mathsf{d} \quad .$$

We now ask for the confidence limits of r. If k and d are sufficiently large, then they may be approximated as Gaussian variables with standard deviations $\sigma_k = \sqrt{\mathsf{k}}$, $\sigma_d = \sqrt{\mathsf{d}}$. Then according to the law of error propagation one has

$$\Delta r = \sqrt{\left(\frac{\partial r}{\partial \mathsf{k}}\right)^2 \mathsf{k} + \left(\frac{\partial r}{\partial \mathsf{d}}\right)^2 \mathsf{d}} = r\sqrt{\frac{1}{\mathsf{k}} + \frac{1}{\mathsf{d}}} \quad . \tag{6.10.1}$$

If in addition one has $\mathsf{d} \gg \mathsf{k}$, then Δr is simply

$$\Delta r = \frac{r}{\sqrt{\mathsf{k}}} \quad . \tag{6.10.2}$$

If the requirements for the validity of (6.10.1) or even of (6.10.2) are not fulfilled, i.e., if k and d are small numbers, then one must use considerations developed by JAMES and ROOS [23]. Clearly when one observes an individual event, the probability that it is a signal event is given by

$$p = \frac{r}{1+r} \quad , \tag{6.10.3}$$

and the probability that it is a reference event is

$$q = 1 - p = \frac{1}{1+r} \quad . \tag{6.10.4}$$

In an experiment in which a total of $\mathsf{N} = \mathsf{k} + \mathsf{d}$ are observed, the probability that exactly n signal events are present is given by a binomial distribution (5.1.3). This is

$$f(n;r) = \binom{\mathsf{N}}{n} p^n q^{N-n} = \binom{\mathsf{N}}{n} \left(\frac{r}{1+r}\right)^n \left(\frac{1}{1+r}\right)^{\mathsf{N}-n} \quad . \tag{6.10.5}$$

The probability to have $n < \mathsf{k}$ is then

$$P(n < \mathsf{k}) = \sum_{n=0}^{\mathsf{k}-1} f(n;r) = F(\mathsf{k};r) \tag{6.10.6}$$

with

$$F(\mathsf{k};r) = \sum_{n=0}^{\mathsf{k}-1} \binom{\mathsf{N}}{n} \left(\frac{r}{1+r}\right)^n \left(\frac{1}{1+r}\right)^{\mathsf{N}-n} \quad , \tag{6.10.7}$$

i.e., the distribution function of the binomial distribution. To determine the limits r_- and r_+ of the confidence interval at the confidence level $\beta = 1-\alpha$, we use in analogy to (6.8.17) and (6.8.18)

$$\alpha/2 = F(\mathsf{k}+1;r_+) \quad , \tag{6.10.8}$$

$$1-\alpha/2 = F(\mathsf{k};r_-) \quad . \tag{6.10.9}$$

If one only seeks an upper limit at the confidence level $\beta = 1-\alpha$, it can be obtained from [see Eq. (6.8.21)]

$$\alpha = F(\mathsf{k}+1;r^{(\mathrm{up})}) \quad . \tag{6.10.10}$$

The quantities r_+, r_-, and $r^{(\mathrm{up})}$ can be computed for given values of k, d, and β with the class SmallSample.

6.11 Ratio of Small Numbers of Events with Background

By combining as done by SWARTZ [24] the ideas of Sects. 6.9 and 6.10, one can deal with the following situation. In an experiment one has three types of events: *signal*, *background*, and *reference*. Signal and background events cannot be distinguished from each other. Suppose in the experiment one has detected a total of k signal and background events and d reference events. Let us label with r_S and r_B the true values (in the sense of the definition at the beginning of the previous section) of the ratios of the numbers of signal to reference and background to reference events. Then the probabilities that a randomly selected event is signal or background, p_S and p_B, are

$$p_S = \frac{r_S}{1+r_S+r_B} \quad , \qquad p_B = \frac{r_B}{1+r_S+r_B} \quad . \tag{6.11.1}$$

The probability that it is a reference event is then

$$p_R = 1 - p_S - p_B = \frac{1}{1+r_S+r_B} \quad . \tag{6.11.2}$$

If one has a total of $\mathsf{N} = \mathsf{k}+\mathsf{d}$ events in the experiment, then the individual probabilities that one has exactly n_S signal events, n_B background events, and $n_R = \mathsf{N} - n_S - n_B$ reference events are

$$f_S(n_S; p_S) = \binom{\mathsf{N}}{n_S} p_S^{n_S} (1-p_S)^{\mathsf{N}-n_S} \quad , \tag{6.11.3}$$

$$f_B(n_B; p_B) = \binom{\mathsf{N}}{n_B} p_B^{n_B} (1-p_B)^{\mathsf{N}-n_B} \quad , \tag{6.11.4}$$

$$f_R(n_R; p_R) = \binom{\mathsf{N}}{n_R} p_R^{n_R} (1-p_R)^{\mathsf{N}-n_R} \quad . \tag{6.11.5}$$

Since there are now three mutually exclusive types of events, one has instead of a binomial distribution (6.10.5) a trinomial distribution, i.e., a multinomial distribution (5.1.10) with $\ell = 3$. The probability that in an experiment with a total of N events one has exactly n_S signal, n_B background, and $\mathsf{N} - n_S - n_B$ reference events is therefore

$$f(n_S, n_B; r_S, r_B) = \frac{\mathsf{N}!}{n_S! n_B! (\mathsf{N}-n_S-n_B)!} p_S^{n_S} p_B^{n_B} (1-p_S-p_B)^{\mathsf{N}-n_S-n_B} \quad . \tag{6.11.6}$$

Here, however, we have not taken into consideration that the number of background events cannot be greater than k. In a manner similar to (6.9.4) one uses

$$f'_B(n_B;p_B) = f_B(n_B;p_B) \Big/ \sum_{n_B=0}^{\mathsf{k}} f(n_B;p_B) \quad , \qquad n_B \le \mathsf{k} \quad , \quad (6.11.7)$$

in place of f_B. This replacement must also be made in (6.11.6), which gives

$$f'(n_S,n_B;r_S,r_B) = f(n_S,n_B;r_S,r_B) \Big/ \sum_{n_B=0}^{\mathsf{k}} f(n_B;p_B) \quad . \quad (6.11.8)$$

The probability to have n_S signal events regardless of the number of background events is

$$f'(n_S;r_S,r_B) = \sum_{n_B=0}^{\mathsf{k}} f'(n_S,n_B;r_S,r_B)$$

and finally the probability to have $n_S \le \mathsf{k}$ is

$$\begin{aligned} F'(\mathsf{k};r_S,r_B) &= \sum_{n_S=0}^{\mathsf{k}-n_B-1} f'(n_S;r_S,r_B) \\ &= \frac{\displaystyle\sum_{n_S=0}^{\mathsf{k}-n_B-1}\sum_{n_B=0}^{\mathsf{k}-1} \frac{\mathsf{N}!}{n_S!n_B!(\mathsf{N}-n_S-n_B)!} p_S^{n_S} p_B^{n_B} (1-p_S-p_B)^{\mathsf{N}-n_S-n_B}}{\displaystyle\sum_{n_B=0}^{\mathsf{k}} \frac{\mathsf{N}!}{n_B!(N-n_B)!} p_B^{n_B}(1-p_B)^{\mathsf{N}-n_B}} \quad . \end{aligned}$$

Since r_B was assumed to be known, the quantity F' for a given k depends only on r_S. Similar to (6.9.6) and (6.9.7) one can determine the limits r_{S+} and r_{S-} of the confidence region for r_S with confidence level $\beta = 1-\alpha$ from the following requirement:

$$\alpha/2 = F'(\mathsf{k}+1;r_{S+},r_B) \quad , \quad (6.11.9)$$

$$1-\alpha/2 = F'(\mathsf{k};r_{S-},r_B) \quad . \quad (6.11.10)$$

If one only wants an upper limit with confidence level $\beta = 1-\alpha$, this can be found according to (6.9.9) from

$$\alpha = F'(\mathsf{k}+1;r_S^{(\mathrm{up})},r_B) \quad . \quad (6.11.11)$$

Table 6.4 contains some numerical values computed with methods of the class SmallSample. For $\mathsf{k}=0$, however, (6.11.10) and hence also r_{S-} have no meaning. Similarly for (6.11.9), r_{S+} for $\mathsf{k}=0$ is not meaningful. (In the table, however, values for r_{S-} and r_{S+} are given for $\mathsf{k}=0$ as computed by the program. For $\mathsf{k}=0$ this sets $r_{S-}=0$ and determines r_{S+} according to (6.11.9).)

Table 6.4: Limits r_{S-} and r_{S+} of the confidence interval and upper confidence limit $r_S^{(\mathrm{up})}$ for various values of r_B and various very small values of k for a fixed number of reference events d and fixed confidence level of 90 %.

$\beta = 0.90, \mathsf{d} = 10$				
k	r_B	r_{S-}	r_{S+}	$r_S^{(up)}$
0	0.0	0.000	0.349	0.259
0	0.1	0.000	0.349	0.259
0	0.2	0.000	0.349	0.259
1	0.0	0.005	0.573	0.450
1	0.1	0.005	0.502	0.382
1	0.2	0.005	0.464	0.348
2	0.0	0.034	0.780	0.627
2	0.1	0.010	0.686	0.535
2	0.2	0.007	0.613	0.467
3	0.0	0.077	0.979	0.799
3	0.1	0.020	0.880	0.701
3	0.2	0.012	0.788	0.612
4	0.0	0.127	1.174	0.968
4	0.1	0.044	1.074	0.869
4	0.2	0.019	0.976	0.771
5	0.0	0.180	1.367	1.135
5	0.1	0.085	1.267	1.035
5	0.2	0.034	1.167	0.936

6.12 Java Classes and Example Programs

Java Classes Referring to Samples

Sample contains methods computing characteristic parameters of a sample: mean, variance and standard deviation as well as the errors of these quantities.

SmallSamle contains methods computing the confidence limits for small samples.

Histograms allows the construction and administration of a histogram.

Example Program 6.1: The class E1Sample demonstrates the use of the class `Sample`

This short program generates a sample of size N taken from the standard normal distribution. It computes the six quantities: mean, error of the mean, variance, error of the variance, standard deviation, and error of the standard deviation, and outputs each quantity in a single line.

Example Program 6.2: The class E2Sample demonstrates the use of the classes Histogram and GraphicsWithHistogarm

Initialization, filling and graphical representation of a histogram are demonstrated for a sample of N elements from the standardized normal distribution. Interactive input is provided for N as well as for the lower boundary x_0, bin width Δx and the number of bins n_x of the histogram. The histogram is initialized, the sample elements are generated and entered into the histogram. Finally the histogram graphics is produced.

Example Program 6.3: The class E3Sample demonstrates the use of class GraphicsWith2DScatterDiagrams

A scatter plot is created and later displayed graphically. The coordinates of the points making up the scatter plot are given as pairs of random numbers from a bivariate normal distribution (cf. Sect. 4.10). The program asks for the parameters of the normal distribution (means a_1, a_2, standard deviations σ_1, σ_2, correlation coefficient ρ) and for the number of random number pairs to be generated. It generates the pairs and prepares the caption and the labeling of axes and scales and displays the plot (Fig. 6.12).

Example Program 6.4: The class E4Sample demonstrates using the methods of the class SmallSample to compute confidence limits

The program computes the limits λ_{S-}, λ_{S+}, and $\lambda_S^{(\mathrm{up})}$ for the Poisson parameter of a signal. The user enters interactively the number of observed events k, the confidence level $\beta = 1-\alpha$, and the Poisson parameter λ_B of the background. **Suggestions:** (a) Verify a few lines from Table 6.3.
(b) Choose $\beta = 0.683$, $\lambda_B = 0$ and compare for different values k the values λ_{S-} and λ_{S+} with the naive statement $\lambda = \mathsf{k} \pm \sqrt{\mathsf{k}}$.

Example Program 6.5: The class E5Sample demonstrates the use of methods of the class SmallSample to compute confidence limits of ratios

The program computes the limits r_{S-}, r_{S+}, and $r_S^{(\mathrm{up})}$ for the ratio r of the number of signal to reference events in the limit of a large number of events. More precisely, r is the ratio of the Poisson parameters λ_S to λ_R of the signal to reference events. The program asks interactively for the number k of observed (signal plus reference) events, for the number d of reference events, for the confidence level $\beta = 1-\alpha$, and for the expected ratio $r_B = \lambda_B/\lambda_R$ for background events.

Suggestion: Verify a few lines from Table 6.4.

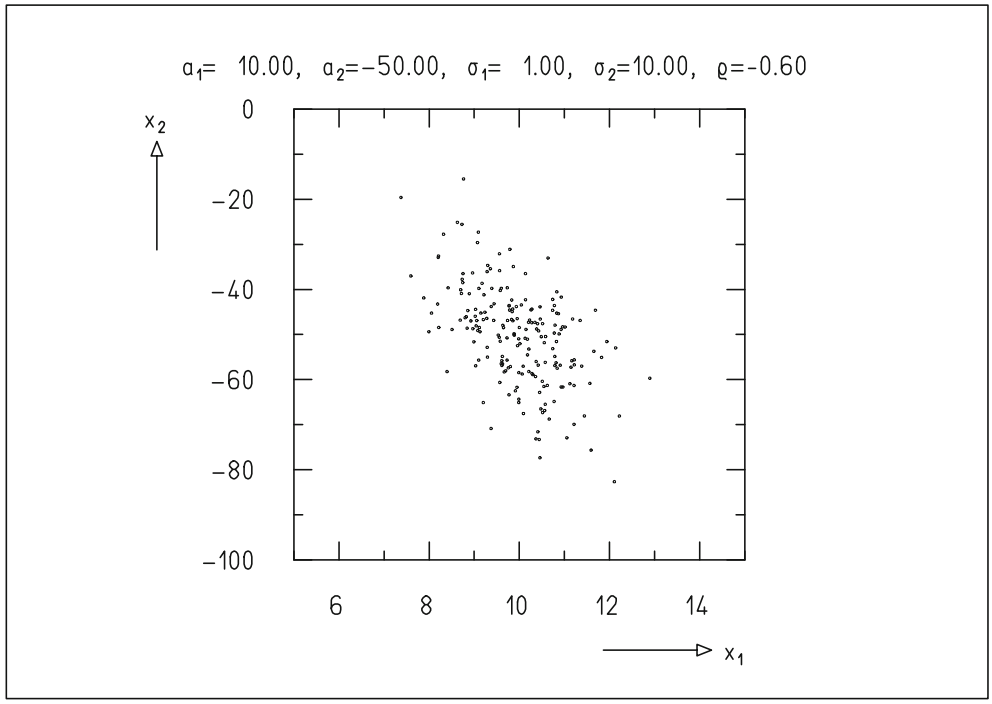

Fig. 6.12: Pairs of random numbers taken from a bivariate normal distribution.

Example Program 6.6: The class E6Sample simulates experiments with few events and background

A total of n_{exp} experiments are simulated. In each experiment N objects are analyzed. Each object yields with probability $p_S = \lambda_S/N$ a signal event and with probability $p_B = \lambda_B/N$ a background event. The numbers of events found in the simulated experiment are k_S, k_B, and $\mathsf{k} = \mathsf{k}_S + \mathsf{k}_B$. In the real experiment only k is known. The limits λ_{S-}, λ_{S+}, and $\lambda_S^{(\text{up})}$ for k are computed for a given confidence level $\beta = 1-\alpha$ and a given value of λ_B and are displayed for each experiment.

Suggestion: Choose, e.g., $n_{\text{exp}} = 20$, $N = 1000$, $\lambda_S = 5$, $\lambda_B = 2$, $\beta = 0.9$ and find out whether, as expected, this simulation yields for 10 % of the experiments an interval $(\lambda_{S-}, \lambda_{S+})$, which does not contain the value λ_S. Keep in mind the meaning of the statistical error of your observation when using only 20 experiments.

Example Program 6.7: The class E7Sample simulates experiments with few signal events and with reference events

The program asks interactively for the quantities n_{exp}, N, λ_S, λ_B, λ_R, and the confidence level $\beta = 1-\alpha$. It computes the probabilities $p_S = \lambda_S/N$, $p_B = \lambda_B/N$, $p_R = \lambda_R/N$, as well as the ratios $r_S = p_S/p_R$ and $r_B = p_B/p_R$ of a total of n_{exp} simulated experiments, in each of which N objects are analyzed. Each object is taken to be a signal event with probability p_S, a background event with probability p_B, and a reference event with probability p_R. (Here $p_S + p_B + p_R \ll 1$ is assumed.) The simulation yields the numbers k_S, k_B, and d for signal, background, and refer-

ence events, respectively. In a real experiment only the numbers $\mathsf{k} = \mathsf{k}_S + \mathsf{k}_B$ and d are known, since signal and background events cannot be distinguished. For the given values of β and r_B and for the quantities k and d which were found in the simulated experiments, the limits r_{S-}, r_{S+}, and $r_S^{(\mathrm{up})}$ are computed and displayed.

Suggestion: Modify the suggestion accompanying Example Program 6.6 by choosing an additional input parameter $\lambda_S = 20$. Find out in how many cases the true value r_S used in the simulation is not contained in the interval (r_{S-}, r_{S+}).