5 Processing of experimental data

This chapter is about the processing of data in its simplest form: given a number of similar observations $x_i = \mu + \epsilon_i$ of an unknown quantity μ, yielding values that only differ in their random fluctuations ϵ_i, how can you make the best estimate $\hat{\mu}$ of the true μ? And how can you best estimate the *accuracy* of $\hat{\mu}$, i.e., how large do you expect the deviation of $\hat{\mu}$ from the true μ to be? Each observation is a sample from an underlying distribution; how can you characterize that distribution? If you have reasons to assume that the underlying distribution is normal, how do you estimate its mean and variance and how do you assess the relative accuracy of those parameters? And how do you proceed if you don't wish to make any assumptions about the underlying distribution?

Suppose you have a number of similar observations x_i, differing only in deviations of random character. *Assume* for the time being that the probability distribution of those random deviations is a normal distribution, characterized by a mean μ and a standard deviation σ or variance σ^2. Although you don't know the real distribution function because your data set is limited (and these data are only samples from the distribution), it is possible to make *estimates* of μ and σ. Such estimates are often indicated by a *hat* over the symbol: $\hat{\mu}, \hat{\sigma}$. What you really want to know is the best estimate for the true value (e.g. the mean) and the uncertainty in that estimate. In practice you can use the estimated variance of the distribution to derive the uncertainty in the mean.

In this chapter we shall first look at the distribution function of the data (Section 5.1) and then indicate how the properties of the data (Section 5.2) lead to estimation of the properties of the distribution function (Section 5.3). The uncertainties in the estimates are discussed for the mean in Section 5.4 and for the variance in Section 5.5. Section 5.6 considers the case that individual data have different statistical weights. Finally, Section 5.7 treats some methods that are robust against dependency on the exact shape of the underlying distribution.

5.1 The distribution function of a data series

In order to get an impression of the distribution of the data it is useful to plot the data in a *histogram*. This is done by first sorting the data in increasing order and subsequently grouping the data in predetermined intervals. A plot of the number of observations in each interval, e.g. as bars, versus the central values of the intervals is called a histogram.

Be careful with computer programs that generate fancy histograms. For example, if you display a perspective view using three-dimensional bars, the point of view may be chosen such that certain bars appear relatively larger than they are. Horizontal lines may appear as having positive or negative slopes, and the reader may be misled by the graph. The same happens when icons are used instead of lines or bars, e.g. an oil barrel to indicate the volume of oil production: a barrel that is twice as large gives the impression of an increase much larger than a factor of two. It is naive to use fancy displays for esthetical reasons if they produce misleading results; it is a scientific crime to purposely construct misleading displays.[1]

Let us use the example "Thirty Observations" given in Chapter 2 on page 6. The data, already sorted, are given in Table 2.1 on page 6 and a histogram is shown in Fig. 2.3.

A histogram is an approximation to the probability density function that is sampled by the data. When the number of observations is limited, as in the present example of thirty observations, a histogram is quite noisy and it is difficult to judge the probability density function from a fit to the histogram. It is then much better to display the *cumulative distribution* of the data. This is quite similar to the cumulative distribution function $F(x)$ of a continuous probability distribution, as described in the previous chapter (see page 31), except that the data are now discrete. For a set of n values $x_1, \ldots, x_n$, the cumulative distribution $F_n(x)$ is defined as

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leq x), \tag{5.1}$$

where $I(condition)$ is the *indicator function*, defined as equal to 1 when *condition* is true and equal to 0 otherwise. Thus $F_n(x)$ is equal to the fraction of all samples x_i for which $x_i \leq x$. So, between x_{i-1} and x_i the function is equal to $(i-1)/n$, but it jumps to i/n for $x = x_i$. See Fig. 5.1.

For a series of measurements with equal weight, the cumulative distribution is constructed by plotting the sequence number in an *ordered* series of data $x_1 \leq x_2 \leq \ldots \leq x_n$ versus the value of x.

The definition (5.1) assumes that all data points have equal statistical weights. This is often not the case. For example, the data may have been

[1] There is nothing new in misleading your readers. For examples, see Huff (1973).

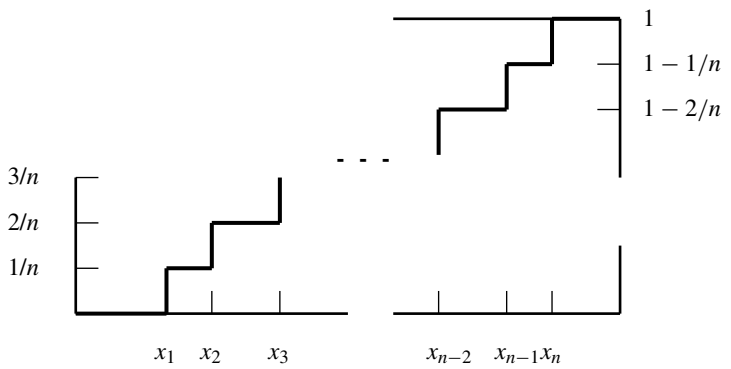

Figure 5.1 Detailed aspects of the cumulative distribution function of a set of discrete data $x_1, \ldots, x_n$.

gathered in bins before analysis (resulting in a histogram) and the individual original data are not available anymore. In that case we have, instead of n points each with statistical weight $1/n$, n bins each with a given statistical weight w_i. The latter is the number of observations within the i-th bin, preferably relative to the total number of observations, so that the total weight equals 1.

Figure 5.2 is an example of such a histogram. The data are the distribution of the height of men and women in the Netherlands in the age group 20–29, averaged over the years 1998, 1999 and 2000. The data are available from official statistical sources[2] but only in the form of percentages in bins of 5 cm width. The bin with midpoint 180 cm accumulates the rounded heights 178–182, i.e., all heights between 177.5 and 182.5 cm. The corresponding bar in the histogram should be centered at the midpoint value.

The definition of the cumulative distribution of the data is now slightly different from (5.1), as each point must be scaled according to its weight w_i:

$$F_n(x) = \frac{\sum_{i=1}^{n} w_i I(x_i \leq x)}{\sum_{i=1}^{n} w_i}. \tag{5.2}$$

The data should be plotted with the "jumps" located at the midpoints of the bins. The left panel of Fig. 5.3 plots the population–height data of Fig. 5.2 as a cumulative distribution. The staircase curve, of course, is an approximation to the exact cumulative length distribution. The dots in this figure denote the points at which this approximate curve coincides with the exact

[2] http://statline.cbs.nl/StatWeb/publications.

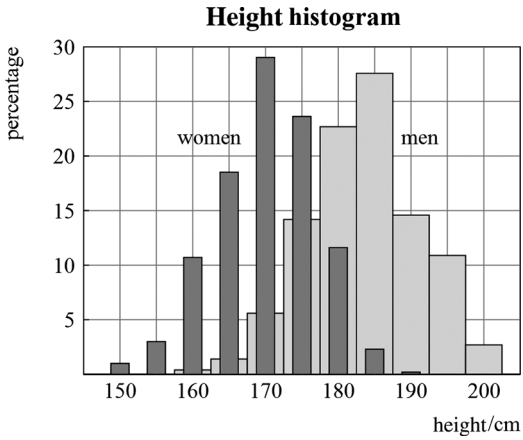

Figure 5.2 Histograms of the height distribution of men (light gray) and women (dark gray) in the age group 20–29 in the Netherlands, averaged over the years 1998, 1999 and 2000. The data have been gathered in bins of 5 cm width.

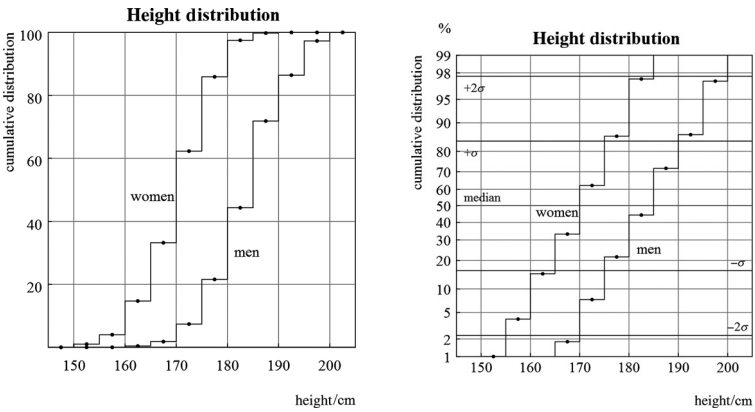

Figure 5.3 The cumulative probability distribution of the data of Fig. 5.2. Left: linear scale; right: probability scale. The dots indicate the values where the cumulative function is exact.

cumulative distribution. These points are located at the boundaries between bins. Thus, if you want to fit a theoretical distribution function to the experimental data, the theoretical curve should match these points as closely as possible.

The right panel of Fig. 5.3 plots the same data on a probability scale (see page 39). A normal distribution should give a straight line. It is obvious from this plot that the distribution of the data is very nearly normal.

5.2 The average and the mean squared deviation of a data series

In this book we denote *averages* over a data series with $\langle \ldots \rangle$ (e.g. $\langle x \rangle$).[3] In order to estimate the properties of the probability distribution from which the data are samples, the following averages are needed:

(i) The *average* $\langle x \rangle$ of a series of equivalent (i.e., equally probable) independent samples $x_i, i = 1, \ldots, n$ is given by

$$\langle x \rangle = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{5.3}$$

See Section 5.6 for the handling of data series with unequal statistical weights.

(ii) The *mean squared deviation* (msd) from the average is defined as

$$\langle (\Delta x)^2 \rangle = \frac{1}{n} \sum_{i=1}^{n} (\Delta x_i)^2, \tag{5.4}$$

where Δx_i is the deviation of the average:

$$\Delta x_i = x_i - \langle x \rangle. \tag{5.5}$$

The root of the msd, which is naturally called the *root-mean-squared deviation* (rms deviation or rmsd), is a measure for the spread of the data around the average.

In order to determine the msd, you must pass through the data twice: first to determine $\langle x \rangle$ and subsequently to determine $\langle (\Delta x)^2 \rangle$. This can be avoided by using the following identity (see Exercise 5.2):

$$\langle (\Delta x)^2 \rangle = \langle x^2 \rangle - \langle x \rangle^2, \tag{5.6}$$

where

$$\langle x^2 \rangle = \frac{1}{n} \sum_{i=1}^{n} x_i^2. \tag{5.7}$$

[3] Often averages are denoted with a bar over the variable, e.g. $\bar{x}$; this symbol we shall reserve for *averages over time*. The *expectation* (see page 29) is also an average, e.g. over a probability density function; this kind of average is usually named the *mean*. In the literature the term *mean* is often also employed for averages over data series.

Note: If the x_i's are large numbers with a relatively small spread, (5.6) could give inaccurate results by truncation errors, especially on a computer with single-precision arithmetic. Therefore the general use of (5.6) is not recommended. The remedy is to subtract from all x values a constant which is close to $\langle x \rangle$, e.g. the first value of the series. The computed average must of course be corrected for this shift.

5.3 Estimates for mean and variance

The averages $\langle x \rangle$ and $\langle (\Delta x)^2 \rangle$ are simple properties of the data set. We wish to use those to *estimate* the mean and variance (and hence also the standard deviation) of the underlying probability distribution of which the data are supposed to be *independent* random samples.

For the mean μ the answer is simple: the best estimate $\hat{\mu}$ for the mean of the underlying distribution is the average of the data themselves:

$$\hat{\mu} = \langle x \rangle. \tag{5.8}$$

It is easy to show that this choice for $\hat{\mu}$ minimizes the total squared deviation from $\hat{\mu}$:

$$\sum_{i=1}^{n} (x_i - \hat{\mu})^2 \text{minimal}. \tag{5.9}$$

For the variance the choice is less straightforward. The best estimate $\hat{\sigma}^2$ for the variance of the underlying distribution is slightly larger than the mean squared deviation of the average of the data:

$$\hat{\sigma}^2 = \frac{n}{n-1} \langle (\Delta x)^2 \rangle = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \langle x \rangle)^2. \tag{5.10}$$

The best estimate for the standard deviation (s.d.) of the underlying distribution is the square root of $\hat{\sigma}^2$:

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2}. \tag{5.11}$$

The reason that the factor $n/(n-1)$ figures in (5.10) is that $\langle x \rangle$ is not exactly equal to the mean of the distribution, but is itself correlated with the data. One could loosely say that one data point has been "used" to compute the average, so that only $n-1$ points provide new data to compute the variance. For a derivation of this term see Appendix A6 on page 151. The equation for $\hat{\sigma}^2$ is only valid when the data are independent samples (which we assumed

to be the case). When the data are correlated, $\hat{\sigma}^2$ is even larger. As you can see, the factor $n/(n-1)$ is not very important when n is large.[4]

5.4 Accuracy of mean and Student's t-distribution

The accuracy of the mean does not equal σ, but it does follow from the value of σ. The more data points are available, the more accurately the average of the measured values will represent the true mean of the underlying distribution. The average $\langle x \rangle$ is itself also a sample from a probability distribution; we could recover that distribution if we could repeat the whole series of measurements a larger number of times. When many series of n independent measurements had been performed, the variance of the average would be given by

$$\sigma^2_{\langle x \rangle} = \sigma^2/n. \tag{5.12}$$

See Appendix A7 for the derivation of this equation. Thus the *estimate* $\hat{\sigma}_{\langle x \rangle}$ of the standard deviation of the average $\langle x \rangle$ (also called the *standard error* or *rms error* of $\langle x \rangle$) is:

$$\hat{\sigma}_{\langle x \rangle} = \frac{\hat{\sigma}}{\sqrt{n}} = \sqrt{\frac{\langle (\Delta x)^2 \rangle}{n-1}}. \tag{5.13}$$

Also this equation is only valid when the statistical deviations in all measured values are independent. If they are not independent, the individual fluctuations will not add quadratically and the standard error will become *larger*. It is as if the number of independent points is less than n. For the common case that dependencies in a series of measurements result from correlation between successive points it is possible to define a *correlation length* n_c. The equations then remain valid, but the number of data points n must sometimes be replaced by the *effective number* n/n_c. For example, in (5.10) $n/(n-1)$ must be replaced by $n/(n-n_c)$, making the estimate of the variation somewhat larger. But the standard inaccuracy in the sample mean becomes $\sqrt{n_c}$ times larger, as n in (5.12) must be replaced by n/n_c. See Appendices A6 on page 151 and A7 on page 154 for more details.

When the measurements are samples from a normal distribution, one might well expect that the quantity

$$t = \frac{\langle x \rangle - \mu}{\hat{\sigma}/\sqrt{n}} \tag{5.14}$$

[4] Note that calculators with statistical functions often let you choose between a σ based on n and a σ based on $n-1$. The former gives the rmsd of the data set and the latter gives the best estimate of the standard deviation of the underlying probability distribution.

will be a sample from the standard normal distribution $N(0,1)$. This, however, is not the case because $\hat{\sigma}$ is not exactly equal to the true σ of the distribution; there is also a spread in $\hat{\sigma}$ itself. If this is taken into account, then one finds that t is a sample from a distribution called the *Student's t-distribution*.[5] For details see the data sheet STUDENT'S T-DISTRIBUTION on page 213. For a derivation in a Bayesian context see the second example in Section 8.4 on page 115.

In the limit of large numbers of data points the t-distribution equals a normal distribution, but for small numbers the t-distribution is broader. The t-distribution has as parameter the number of *degrees of freedom* $\nu = n - 1$, one less than the number of (independent) data points. One data point has already been "used" to determine the average, just as in the case of the estimation of σ, see (5.10). It is clear that it is only possible to say anything about the accuracy of the mean when at least two data points are available.

When the t-distribution is used, one can best give a *confidence interval*, e.g. the lower and upper limits between which the true mean is expected to lie with a probability of 50% (or 80%, 90%, 95%, 99%, ..., your choice!).

5.5 Accuracy of variance

Finally we give an indication for the accuracy of $\hat{\sigma}$: if the measurements are independent and the deviations are random samples from a normal distribution, then the *relative* standard inaccuracy of $\hat{\sigma}$ equals $1/\sqrt{2(n-1)}$. Appendix A7 on page 154 gives more details. The same applies to the relative standard inaccuracy in the computed standard error of the mean. For example, if you find for the estimated mean of a series of 10 measurements and its estimated inaccuracy 5.367 ± 0.253 than you should report this as 5.4 ± 0.3 because the relative inaccuracy of the number 0.253 equals $1/\sqrt{18}$ or 24% (=0.06) (insufficiently accurate for two significant digits). Had these numbers been the result of 100 independent measurements, then the proper report would have been 5.37 ± 0.25. Table 5.1 gives the relative s.d. as a percentage of $\hat{\sigma}$ for various numbers n of independent data points. The same relative inaccuracy also applies to the s.d. of the mean, as calculated by (5.13).

While the accuracy of the standard deviation is usually not very large, the estimated *skewness* or *excess* is often hardly significant. For near-Gaussian distributions these estimates with their s.d. are

[5] See Gosset (1908). "Student" was the pseudonym of the English statistician W. S. Gosset (b. 1876).

Table 5.1 *Relative inaccuracy (s.d.) of the estimated standard deviation $\hat{\sigma}$ of a distribution based on a series of n independent samples.*

n	s.d.($\hat{\sigma}$) %	n	s.d.($\hat{\sigma}$) %	n	s.d.($\hat{\sigma}$) %
2	70	10	24	50	10.1
3	50	15	19	60	9.2
4	41	20	16	70	8.5
5	35	25	14	80	8.0
6	32	30	13	90	7.5
7	29	35	12	100	7.1
8	27	40	11	150	5.8
9	25	45	11	200	5.0

$$skewness = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i}{\hat{\sigma}}\right)^3 \pm \sqrt{\frac{15}{n}}, \tag{5.15}$$

$$excess = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i}{\hat{\sigma}}\right)^4 - 3 \pm \sqrt{\frac{96}{n}}. \tag{5.16}$$

5.6 Handling data with unequal weights

Until this point we have assumed that all data points have the same *statistical weight*, i.e., that they are all samples from the same probability distribution. But it is quite common that one measurement is more accurate than another; in such cases the more accurate measurement must get a larger weight in the statistical analysis (e.g. in the determination of the mean) than a less accurate measurement. This may happen when the same quantity is determined in different ways, yielding several values with their individual uncertainty estimates, and the best estimate for the mean is required. Unequal weights must also be given to histogram data that result from adding observations in bins: it is obvious that each bin (central) value x_i must be multiplied by the number of observations in that bin n_i in order to obtain the proper mean over all observations:

$$\langle x \rangle = \frac{\sum_i n_i x_i}{\sum_i n_i}. \tag{5.17}$$

In general, the best estimate $\hat{\mu}$ for the mean of the underlying distribution is the *weighted average* defined as

$$\langle x\rangle = \frac{1}{w}\sum_{i=1}^{n} w_i x_i; \quad w = \sum_{i=1}^{n} w_i, \tag{5.18}$$

where the *weight factors* w_i are proportional to $1/\sigma_i^2$. Only proportionality is needed because the sum is divided by the total weight. Why this is the correct way of averaging is explained in Appendix A8 on page 158.

This type of averaging does not only apply to x but to any quantity that is to be averaged, e.g.

$$\langle x^2\rangle = \frac{1}{w}\sum_{i=1}^{n} w_i x_i^2; \quad w = \sum_{i=1}^{n} w_i, \tag{5.19}$$

or, in general,

$$\langle f(x)\rangle = \frac{1}{w}\sum_{i=1}^{n} w_i f(x_i); \quad w = \sum_{i=1}^{n} w_i. \tag{5.20}$$

Accuracy of the estimated mean

When the mean of a data series has been estimated by weighted averaging of $x_i \pm \sigma_i$, then the estimate for the standard inaccuracy of the estimated mean is given by

$$\hat{\sigma}_{\langle x\rangle} = \left(\sum_{i=1}^{n} \frac{1}{\sigma_i^2}\right)^{-1/2}. \tag{5.21}$$

Why this is so is also explained in Appendix A8. Using this formula we assume that the values of σ_i^2 are reliable; we have not used the value of $\langle(\Delta x)^2\rangle$ for the estimation of $\hat{\sigma}_{\langle x\rangle}$. Whether the observed spread in the measured values will be statistically acceptable (i.e., compatible with the known σ_i^2), can be tested with a *chi-squared test*. The chi-squared test will be fully treated in Section 7.4 on page 95 (see also the data sheet chi-squared distribution on page 199), but here we already make superficial use of it. For this case the number of degrees of freedom equals $n-1$ and χ^2 is defined as

$$\chi^2 = \sum_{i=1}^{n} \frac{(x_i - \langle x\rangle)^2}{\sigma_i^2} = \frac{\langle(\Delta x)^2\rangle}{\hat{\sigma}^2_{\langle x\rangle}}. \tag{5.22}$$

Note that $\langle(\Delta x)^2\rangle$ must have been determined by weighted averaging according to (5.20). In the last term we have used (5.21). The value of χ^2 should

be in the neighborhood of the number of degrees of freedom $n - 1$. How much it can reasonably deviate from this value is given by the cumulative chi-squared distribution (see page 2 of the data sheet chi-squared distribution on page 199).

If the σ_i's are *not* accurately known *and* the number of observations is sufficiently large, it is possible to use $\langle(\Delta x)^2\rangle$ for the determination of $\hat{\sigma}_{\langle x\rangle}$. In that case assume that $\chi^2 = n - 1$, so that

$$\hat{\sigma}^2_{\langle x\rangle} = \frac{\langle(\Delta x)^2\rangle}{n-1}. \tag{5.23}$$

This equation – as expected – also applies to the case of independent samples of equal weight, and is therefore equivalent to (5.13) on page 59.

Which choice of method you make is up to you. When your individual variance estimations are unreliable, choose the latter method. To be on the safe side, you may also choose the largest of the two uncertainties from the two methods.

5.7 Robust estimates

Estimates of parameters like standard deviation and standard error, as have been treated in the previous sections, are quite sensitive to outliers in the data. The reason for this is the use of squared deviations; an outlier contributes rather heavily to a sum of squares. When a deviation is so large that its occurrence in the data set is rather unlikely, one may eliminate such an observation (see below). Some of the methods treated in the previous sections are only valid for normally distributed data, such as the confidence intervals determined by Student's t method. In modern statistics *robust* methods have been developed to handle data series in such a way that outliers play a lesser role and the results depend less strongly on the type of distribution function of the data. These robust methods are based on the *ranking order* of the data ('rank-based methods'). In this book we give only a brief summary of these methods and refer for further details to the literature (Petruccelli *et al.*, 1999; Birkes and Dodge, 1993; Huber and Ronchetti, 2009).

Elimination of outliers

It may happen that a particular measured value falls outside the expected range of values. This may be due to a random fluctuation, but it can also be the result of an experimental error or mistake. It is warranted to eliminate such a data point from the data series before further processing. A reasonable, and often used, criterion is that the deviation exceeds 2.5 σ. Don't apply such elimination more than once in a data series. Prudence is required, because the choice whether or not to eliminate a data point may be influenced

by subjective considerations if the particular measurement does or does not suit your purposes. Of course, rather than elimination, it is much better to repeat the measurement: a possible error or mistake can then be identified. If repeated measurements also show significant deviations from the expected value, you may be on the track of an interesting phenomenon worth further investigation.

The 2.5 σ criterion is rather arbitrary and many researchers prefer a limit of 3 σ. The criterion should be chosen such that the random occurrence of a value beyond the chosen limit is an unlikely event, e.g. with a probability of less than 5%. But with such a criterion the limiting value depends on the *number* of data points in the data series: the probability – given a normal distribution – that a *single* point deviates more than 2.5 σ is a bit over 1%; the probability that *at least one* point in a series of 20 data points deviates more than 2.5 σ exceeds 20%. The first case is unlikely, but the second case may easily occur by random sampling. In the table on page 2 of the data sheet NORMAL DISTRIBUTION on page 205 the probability is tabulated that *at least one* data point out of n points falls outside the range $(\mu - d, \mu + d)$ (a *two-sided* criterion), for various values of d/σ; on page 4 the probability is tabulated that at least one data point exceeds the value $\mu + d$ (a *one-sided* criterion). If you choose a 5% limit for this one-sided probability, you see that for less than 10 data points 2.5 σ is a good choice; for 10 to 50 points 3 σ is better and for 50 to several hundred points 3.5 σ is the best choice.

Rank-based estimates

The estimated mean of a distribution is usually taken as equal to the average of the measured values. When you have a good reason to assume a symmetric underlying probability distribution, but no good reason to assume a normal distribution, you can also take the *median* of the measured values. For a large number of points the result is the same, but for a small number of points the median is less sensitive to outliers than the average. The median has the property that the number of positive and negative deviations are equal; in order to obtain the median only the *sign* of the deviations is used.

Sign-based confidence intervals

A sign-based estimate of a confidence interval is obtained from the binomial distribution of the number of positive signs of the possible deviations. Suppose you have five measurements, sorted in ascending order: x_1, x_2, x_3, x_4, x_5. As estimate for the mean $\hat{\mu}$ you take the median x_3. Now consider the probability that the $\mu < x_1$. In that case the deviations would have the signs $+++++$ and the binomial probability of obtaining five pluses when each sign has a 50% chance of being plus, is

$$p(\mu < x_1) = 2^{-5}\binom{5}{5} = 1/32 \tag{5.24}$$

(see Section 4.3). The same probability is obtained when $\mu > x_5$, so the interval (x_1, x_5) has a confidence level of $30/32 = 94\%$. When μ lies between x_2 and x_4, the deviations have the sign $--+++$ or $---++$; the binomial probability that this happens is:

$$p(x_2 < \mu < x_4) = 2^{-5}\binom{5}{3} + 2^{-5}\binom{5}{2} = 20/32 = 62\%. \tag{5.25}$$

Because only a small number of discrete values are available, the interval for a preset confidence level of, say, 90%, cannot be given. The method is robust, but also quite inaccurate. If there is an indication that the data points are samples from a normal distribution, the "classical" parameter estimates are much better. In order to keep in line with the classical report of the standard deviation, an alternative robust estimate of the "standard deviation" can be obtained by abstracting the 68% confidence interval from an analysis of the cumulative distribution function (see Section 5.1 on page 54).[6]

The bootstrap method

Finally a few words about another *distribution-free* method, the *bootstrap*, designed to obtain an approximate probability distribution (a "sampling distribution") of an estimated mean on the basis of a data series, without any assumption about the probability distribution from which the data are sampled. The method originates from 1979, see Efron and Tibshirani (1993). The method is simple but can only be realized by using a computer.

Assume you have a data series of n independent samples of equal weight from an unknown distribution. The average of this series is a good estimate for the mean of the distribution. You wish to generate a number of such averages to produce a sampling distribution of the mean; this gives you details on the accuracy of the estimated mean. Unfortunately, this is only possible if you could produce many new sets of measurements, which would each freshly sample the data distribution. But you don't have more new data, so you must rely on the n samples you already have. Now generate a large number (say, 3000) of series, each consisting of n "measurements," each drawn randomly from the n original data, but with *replacement*, i.e., without changing the probability of drawing a particular value. From each series determine the average. The collection of all 3000 averages so obtained approximates

[6] The deviation d for which the interval $(\mu - d, \mu + d)$ equals the 68 percent confidence interval is strictly not a standard deviation and it does not imply the validity of other confidence intervals derived from normal distributions. It should be checked if this value equals the best estimate of the standard deviation within its error limits.

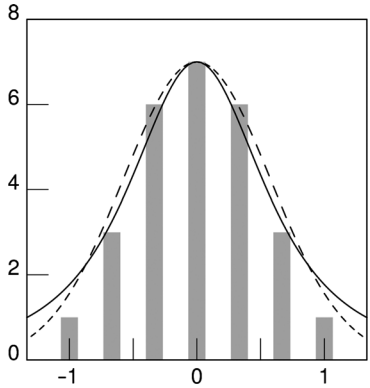

Figure 5.4 Histogram of the bootstrap distribution of the mean of three data values -1, 0 and 1 (plotted ordinates $\times 1/27$). Drawn line: Student's t-distribution for two degrees of freedom; dashed line: normal distribution with "classical" standard deviation. All distributions have been scaled to yield the same maximum.

the true sampling distribution you would have obtained from 3000 sets of fresh measurements.

For a small number of measurements it is possible to generate *all* possible series (there are n^n of those), but for more than five data points this runs out of hand. For illustrative purposes Fig. 5.4 gives the bootstrap distribution for three data points with values -1, 0 and 1: there are seven possible averages. In the same figure the Student's t-distribution is given for the same three measured values (i.e., for two degrees of freedom), for which $\hat{\mu} = 0$ and $\hat{\sigma} = 1$. The "classical" standard uncertainty of the mean equals $\hat{\sigma}/\sqrt{3} = 0.577$; the s.d. of the bootstrap distribution is $\sqrt{2}/3 = 0.471$. The latter is also the standard uncertainty of the mean using the biased estimator: $\sqrt{\langle(\Delta x)^2\rangle}/\sqrt{3}$. Also the normal distribution with $\sigma = 0.577$ is given. We see that for this symmetric case the normal distribution and the bootstrap distribution agree well; the t-distribution has broader flanks. If you only have three values and no good reason to assume normality of the underlying distribution, there is no good reason to apply the Student's t-distribution.

The term "bootstrap" now becomes meaningful: A bootstrap method is a method to obtain something new from nothing, which in principle is impossible, such as lifting yourself off the ground by pulling your bootstraps. Have we gained anything new by applying the bootstrap method? No! The bootstrap produces an array of averages of n samples taken from a given distribution: a sum of n δ-functions at the original data points. The

distribution function of these averages can be computed by methods treated in Appendix A5; its mean and standard deviation are completely determined by the original data. In fact, the mean of the bootstrap distribution equals the mean of the original data and the s.d. of the bootstrap distribution equals the rmsd $\sqrt{\langle(\Delta x)^2\rangle}$ of the original data divided by $\sqrt{n}$. This equals the *biased* estimate of the standard uncertainty in the mean; we know that the *unbiased* estimate rmsd divided by $\sqrt{n-1}$ is better. A bootstrap distribution with unbiased s.d. can be obtained by adding $n-1$ rather than n samples from the original data.

So it seems that the bootstrap method is rather meaningless. It is meaningless, indeed, for obtaining a best estimate for the mean and its standard uncertainty. It is not meaningless, however, for obtaining confidence intervals for a given confidence level. But you should always be aware of the fact that the bootstrap distribution does not extend beyond the minimum and maximum data value, while the underlying probability distribution may well have tails extending (far) beyond those values. Confidence limits derived from the tails of the bootstrap distribution may well be unrealistically narrow and could lead to erroneous conclusions. See Exercise 5.6 for a comparison of various estimates.

A program that will generate an array with averages from random samples taken from a given dataset is **Python code** 5.1 on page 175.

A program `report` to analyze a data set is **Python code** 5.2 on page 176. Given a set of independent data, it produces a graph of the cumulative distribution (on a probability scale) and a graph with data points and standard deviations (if given); it prints properties of the data (including skewness and excess) and identifies outliers. In addition it performs functions explained in Chapter 7: drift analysis with significance tests and a chi-squared analysis if standard deviations are given. Look for updates on www.hjcb.nl.

Summary *You are now able to make a clear distinction between the distribution of measured data x_i and the (unknown) underlying distribution from which the data are supposed to be random samples. Properties of your measured data are number n, average $\langle x\rangle$, mean squared deviation (msd) $\langle(\Delta x)^2\rangle$, and root-mean-squared deviation (rmsd) $\sqrt{\langle(\Delta x)^2\rangle}$, but also rank-based properties such as range, median and various percentiles. From these properties you can derive best estimates $\hat{\mu}$, $\hat{\sigma}$ for the parameters of the underlying distribution: mean and standard deviation. An important quantity is the inaccuracy of the estimated mean $\sigma_{\langle x\rangle}$ (the s.d. of the sampling mean), which equals*

$\hat{\sigma}/\sqrt{n}$. You are aware of the fact that all these formulas are valid for a set of n independent samples; if samples are correlated, the estimated variance becomes somewhat $(((n-1)/(n-n_c))\times)$ larger and the standard inaccuracy in the sample mean becomes considerably $(n_c\times)$ larger, where n_c is a correlation length. You know how to handle your data if the data points have unequal statistical weights: in all kinds of averaging, the values to be averaged are multiplied by their weight w_i/w, where w is the total weight.

You can express results in terms of confidence intervals. These can be one-sided or double-sided. For example, a 90 percent double-sided confidence interval gives the estimated range from the 5th to the 95th percentile of the underlying distribution. For normally distributed variables, the confidence intervals follow from the normal distribution if you know σ beforehand, or from Student's t-distribution if you don't. An alternative determination of confidence intervals for the sampling mean is to construct a bootstrap distribution based on your data itself. You are aware of the pitfalls of this "distribution-free" method.

Finally, if you have a set of data and a good prior estimate of the inaccuracy of each data point, you can use the chi-squared distribution to assess whether the spread in the measured data is compatible with the a-priori inaccuracies. If the spread is improbably large, there is probably an error source that you overlooked.

Exercises

5.1 Could the data given in Table 2.1 on page 6 be sampled from a normal distribution? If so, estimate $\hat{\mu}$ and $\hat{\sigma}$ by drawing a straight line through the cumulative distribution function of Fig. 2.1.

5.2 Prove (5.6).

5.3 If you subtract a constant from all values of x and then compute the msd using (5.6), is a further correction still required?

5.4 Generate 1000 normally distributed variables with mean c and s.d. 1. Compare the rmsd computed by both (5.4) and (5.6). Vary the constant c (e.g. 1.e6, 1.e7, 1.e8, 1.e9).

5.5 (refer to Table 5.1 on page 61)
A series of n independent measurements of a physical quantity yields an average of 75.325 78 and a mean squared deviation of 25.643 06. Report,

with the correct number of digits, your best estimates of the mean and standard deviation of the underlying probability distribution, for two cases: (a) $n = 15$, (b) $n = 200$.

5.6 You live in Germany and want to calibrate the speedometer of your car. On a quiet, mostly straight and level Autobahn section you keep your speed as accurate as possible at 130 km/hr on your speedometer. Your companion measures with a stopwatch the time between passing two kilometer marks that are exactly 1 km apart. She finds the following nine intervals (in s):[7]
29.04, 29.02, 29.24, 28.89, 29.33, 29.35, 29.00, 29.25, 29.43

1. Compute the following properties of the measured set of time intervals:
   (a) the average,
   (b) the average squared deviation from the average,
   (c) the root-mean-squared average deviation from the average,
   (d) the range, median and the first and third quartiles.
2. Compute the best estimates for the following properties of the underlying distribution function:
   (a) the mean $\hat{\mu}$,
   (b) the variance $\hat{\sigma}^2$,
   (c) the standard deviation $\hat{\sigma}$,
   (d) the standard uncertainty of the estimated mean,
   (e) the uncertainty of the last three values.
3. What is (the best estimate for) your car's real velocity? What is the standard uncertainty of this value? How large is the speedometer's deviation and what is the relative accuracy of that deviation? Give all values with the correct number of significant digits.
4. If you as driver assert that you have kept the speed within a deviation of ± 0.5 km/hr, does this knowledge influence your conclusions in any way?
5. Assuming that the (biased) bootstrap yields a reliable sampling distribution of the mean, generate a bootstrap distribution of 2000 samples and compute the 80%, 90% and 95% confidence limits for the time interval.
6. Using this bootstrap distribution, compute the 80%, 90% and 95% confidence limits for the velocity.
7. Assuming the underlying distribution to be normal $N(\hat{\mu}, \hat{\sigma})$, compute the 80%, 90% and 95% confidence limits for the velocity.

[7] These numbers are from a real experiment.

8. Assuming the underlying distribution to be normal with unknown s.d., compute the 80%, 90% and 95% confidence limits for the velocity according to Student's t-distribution.

5.7 You are a member of a CODATA committee with the task to update Avogadro's number. The following reliable data are at your disposal:

- the already known number (see data sheet PHYSICAL CONSTANTS on page 209)
- a series of measurements by scientist A with result: $6.022\,141\,48(75) \times 10^{23}$
- a series of measurements by scientist B with result: $6.022\,142\,05(30) \times 10^{23}$
- a series of measurements by scientist C with result: $6.022\,1420(12) \times 10^{23}$

Give the weighted mean and its standard uncertainty.

5.8 Plot the bootstrap distribution, the histogram of which is given in Fig. 5.4, on a probability scale. Is this distribution compatible with a normal distribution? Estimate graphically the mean and s.d. and compare to the values given in the text.

5.9 (*This advanced exercise requires reading of Appendix A3 and Appendix A5.*)
Determine – using the characteristic function – the distribution function of the sum of three samples, each randomly chosen with equal probability from the three values -1, 0 and 1. Note that the distribution function for the sum of three values equals the convolution of the distribution functions of each value. Determine its variance. Compare your result with Fig. 5.4.