4 Probability distributions

Every measurement is in fact a random sample from a probability distribution. In order to make a judgment on the accuracy of an experimental result we must know something about the underlying probability distribution. This chapter treats the properties of probability distributions and gives details about the most common distributions. The most important distribution of all is the normal distribution, not in the least because the central limit theorem tells us that it is the limiting distribution for the sum of many random disturbances.

4.1 Introduction

Every measurement x_i of a quantity x can be considered to be a *random sample* from a *probability distribution* $p(x)$ of x. In order to be able to analyze random deviations in measured quantities we must know something about the *underlying* probability distribution, from which the measurement is supposed to be a random sample.

If x can only assume discrete values $x = k, k = 1, \ldots, n$ then $p(k)$ forms a *discrete probability distribution* and $p(k)$ (often called the *probability mass function*, pmf) indicates the probability that an arbitrary sample has the value k. If x is a continuous variable, then $p(x)$ is a continuous function of x: the *probability density function*, pdf. The meaning of $p(x)$ is: *the probability that a sample x_i occurs in the interval $(x, x + dx)$ equals $p(x)\,dx$.*

Probability density functions (or probability mass functions) are defined on a *domain* of possible values the random variable can assume. The function value itself is a non-negative real number. The integral over the domain (or the sum in the case of a discrete distribution) equals 1, i.e., the pdf (or pmf) is normalized. In general pdf's can be *multidimensional*, i.e., a function of one, two or more variables. Thus the *joint pdf* $p(x, y)$ means that the probability of finding a sample x_i in the interval $(x, x + dx)$ *and* of finding a sample y_i in the interval $(y, y + dy)$ is given by $p(x, y)\,dx\,dy$. If a pdf $p(x, y)$ is integrated over one variable, say y, the resulting pdf is called the *marginal* pdf of x; multiplied by dx it is the probability of

finding x_i in the interval $(x, x + dx)$ *irrespective* of the value of y. Probabilities can also be defined under a restrictive condition, e.g. $p(x|y)$ is the *conditional* probability of finding x, *given* the value of y. The conditional probability makes sense only when x and y are somehow related to each other: if they are independent of each other, $p(x|y)$ obviously does not depend on y:

$$p(x|y) = p(x) \quad (x, y \text{ independent}). \tag{4.1}$$

The following relations hold:

$$p(x, y) = p(x)p(y|x) = p(y)p(x|y), \tag{4.2}$$

$$p(x, y) = p(x)p(y) \quad (x, y \text{ independent}), \tag{4.3}$$

where $p(x)$ and $p(y)$ are the marginal distributions:

$$p(x) = \int p(x, y)\, dy, \tag{4.4}$$

$$p(y) = \int p(x, y)\, dx. \tag{4.5}$$

The integrations are carried out over the full domains of the variables y and x.

A summary of the properties of one- and two-dimensional probability functions is given on the data sheet PROBABILITY DISTRIBUTIONS on page 211.

In this chapter we consider the properties of a few common one-dimensional probability distributions: the *binomial distribution*, the *Poisson distribution*, the *normal distribution* and a few others. The first two are discrete distributions, the latter is a continuous distribution. In the following chapter on page 53 we consider how, given a series of measured samples, we can derive the *best estimates* of properties of the underlying probability distribution. The real distribution can never be precisely determined because that would require an infinite number of samples.

We shall also change notation and denote the pdf's with $f(x)$ rather than $p(x)$. The reason is that the probability functions we consider in this chapter are based on counting the *frequencies of occurrences* of the possible outcomes given the statistical process that produces the samples. This is in contrast to more general interpretations of probabilities $p(x)$, which may include probabilities based on *beliefs* or *best estimates* considering all knowledge we have. Chapter 8 on page 111 elaborates on this point.

4.2 Properties of probability distributions

Normalization

Both continuous probability density functions $f(x)$ and discrete probability mass functions $f(k)$ are *normalized*, i.e., the sum of all probabilities (over the possible domain of sample values[1]) is equal to 1:

$$\int_{-\infty}^{\infty} f(x)\,dx = 1; \tag{4.6}$$

$$\sum_{k=1}^{n} f(k) = 1. \tag{4.7}$$

For the continuous density function $f(x)$ we have assumed that the domain of possible x-values comprises all real numbers, i.e., the interval $\langle -\infty, +\infty \rangle$, but there are also density functions with a different domain, such as $[0, 1]$ or $[0, +\infty\rangle$. Probabilities are never negative: $f(k) \geq 0$; $f(x) \geq 0$.

Expectation, mean and variance

The *expectation* of a function $g(x)$ of x over the probability density function $f(x)$ (sometimes called the *expected value*) $E[g]$ of $g(x)$ is defined as

$$E[g] = \int_{-\infty}^{\infty} g(x) f(x)\,dx, \tag{4.8}$$

or, in the discrete case:

$$E[g] = \sum_{k=1}^{n} g(k) f(k). \tag{4.9}$$

We use the notation $E[\ ]$ to indicate that E is a *functional*, i.e., a function of a function. Thus the *mean* of x, usually indicated by μ, is equal to the expectation of x itself over the density function:

$$\mu = E[x] = \int_{-\infty}^{\infty} x f(x)\,dx, \tag{4.10}$$

[1] The *domain* is the set of possible values of k or x; the *range* of a series of samples is the difference between the largest and the smallest value occurring in the data set. An *interval* is a set of values between a lower and an upper limit; one indicates the interval limits by [or] if the limit itself is included and by $\langle$ or $\rangle$ if the limit is not included. Normal brackets (or) may be used when the distinction is irrelevant.

or, in the discrete case:

$$\mu = E[k] = \sum_{k=1}^{n} kf(k). \tag{4.11}$$

The *variance* σ^2 of a probability distribution is the expectation of the squared deviation from the mean:

$$\sigma^2 = E[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx, \tag{4.12}$$

or, in the discrete case:

$$\sigma^2 = E[(k-\mu)^2] = \sum_{k=1}^{n} (k-\mu)^2 f(k). \tag{4.13}$$

The square root of σ^2 is called the *standard deviation* (s.d.) σ. Alternatively the s.d. is called the 'rms (root-mean-square) deviation'. The s.d. of the uncertainty distribution of an experimental result is called the *standard uncertainty* or *standard error* or *r.m.s. error*.

Moments and central moments

These are the most important averages over probability distributions. They are related to the first and second *moment* of the distribution. The n-th moment μ_n of a distribution is defined as

$$\mu_n = E[x^n]. \tag{4.14}$$

It is often more useful to employ the *central moments* which are defined with respect to the mean of the distribution. The n-th central moment is

$$\mu_n^c = E[(x-\mu)^n]. \tag{4.15}$$

The second central moment is the variance. The third central moment, expressed in units of σ^3, is called the *skewness* and the fourth central moment (in units σ^4) is the *kurtosis*. Since the kurtosis of a normal distribution equals 3 (see Section 4.5), the *excess* is defined as the deviation from the kurtosis of a normal distribution:[2]

$$skewness = E[(x-\mu)^3/\sigma^3] \tag{4.16}$$

$$kurtosis = E[(x-\mu)^4/\sigma^4] \tag{4.17}$$

$$excess = kurtosis - 3. \tag{4.18}$$

[2] Some books use the name *kurtosis* or *coefficient of kurtosis* for what we have defined as *excess*.

Cumulative distribution function

The cumulative distribution function (cdf) $F(x)$ gives the probability that a value x is not exceeded:

$$F(x) = \int_{-\infty}^{x} f(x')\,dx', \tag{4.19}$$

or, in the discrete case:

$$F(k) = \sum_{l=1}^{k} f(l). \tag{4.20}$$

Note that the value $f(k)$ is included in the cumulative sum $F(k)$. The function $1 - F(x)$ is called the *survival function* (sf), indicating the probability that x *is* exceeded:

$$sf(x) = 1 - F(x) = \int_{x}^{\infty} f(x')\,dx', \tag{4.21}$$

or, in the discrete case:

$$sf(k) = 1 - F(k) = \sum_{l=k+1}^{n} f(l). \tag{4.22}$$

From these definitions it is clear that

$$f(x) = \frac{dF(x)}{dx} \tag{4.23}$$

$$f(k) = F(k) - F(k-1). \tag{4.24}$$

The function F is monotonically increasing, with a value in the interval $[0, 1]$. Cumulative distribution functions F and their inverse functions F^{-1} are necessary to determine *confidence intervals* and *confidence limits*. For example, the probability that x lies between $x_1 = F^{-1}(0.25)$, i.e., $F(x_1) = 0.25$, and $x_2 = F^{-1}(0.75)$, i.e., $F(x_2) = 0.75$, is 50%. The values of x that will be exceeded with a probability of 1% is equal to $F^{-1}(0.99)$, i.e., $F(x) = 0.99$. The value $F^{-1}(0.5)$, i.e., the value of x for which $F(x) = 0.5$, is the *median* of the distribution; $F^{-1}(0.25)$ and $F^{-1}(0.75)$ are the first and third *quartiles* of the distributions. Similarly one may define *deciles* and *percentiles*. The q-th *quantile* equals x if $F(x) = q$.

Characteristic function

Every probability density function $f(x)$ has associated with it a *characteristic function* $\Phi(t)$, which is defined as

$$\Phi(t) \stackrel{\text{def}}{=} E[e^{itx}] = \int_{-\infty}^{\infty} e^{itx} f(x)\,dx. \tag{4.25}$$

The characteristic function is mathematically helpful to analyze probability functions. For example, its series expansion in t generates moments of the distribution. For the usual statistical data treatment you will not require the characteristic function. Interested readers, however, who are not unfamiliar with Fourier transforms, may consult Appendix A3 on page 141 for further details.

A word on nomenclature

The word *probability distribution* is often used in a general sense, meaning any kind of discrete or continuous probability function. However, sometimes the term *distribution function* is specifically meant to indicate the *cumulative distribution function* $F(x)$, in contrast to the continuous *probability density function* $f(x)$ or the discrete probability mass function $f(k)$. To avoid confusion, it is recommended that the modifier "cumulative" is included in this case. Instead of "probability distribution" you should use the term "probability density function" when the latter is meant.

Numerical values of distribution functions

Statistical tables generally give both the density functions and the cumulative functions. They can – among others – be found in Beyer (1991), in Abramowitz and Stegun (1964), and in the *Handbook of Chemistry and Physics* (CRC Handbook, each year). Values can more easily and more accurately be extracted from computer packages. The Python extension SciPy offers a package "stats" with more than 80 continuous and 12 discrete distributions; for each one can invoke the probability density function (pdf), the cumulative distribution function (cdf), the survival function (sf), the percent-point function (ppf, inverse of cdf) and the isf (inverse survival function). Random variates (rvs) and common statistical properties can be obtained from each distribution as well.

4.3 The binomial distribution

Definition and properties

Suppose you measure a *binary* quantity, i.e. a quantity that can assume either one of two values (e.g. 0 or 1, false or true) and every measurement has a probability p to be 1 (or true), then the probability $f(k;n)$ that out of n measurements exactly k have the outcome 1, equals

$$f(k;n) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{4.26}$$

Here

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{4.27}$$

is the *binomial coefficient* "n over k" indicating the number of ways k objects can be chosen from a set of n objects. The random process to choose one possibilities out of two, with probability p, is called a *Bernouilli trial*. Some important properties of the binomial distribution are:

$$\textit{mean}: \ \mu = E[k] = pn, \tag{4.28}$$

$$\textit{variance}: \ \sigma^2 = E[(k-\mu)^2] = p(1-p)n, \tag{4.29}$$

$$\textit{s.d.}: \ \sigma = \sqrt{p(1-p)n}. \tag{4.30}$$

Appendix A4 explains why.

Variance proportional to number

The variance is proportional to the total number of observations n (called the *sample size*). Therefore the *relative* standard uncertainty is *inversely* proportional to the *square root* of the sample size. This is an important rule of thumb to remember: for a 100 times larger sample size the relative uncertainty becomes 10 times smaller. You can buy accuracy by doing more experiments.

Note that for small p the standard deviation is approximately equal to the square root of the mean number of observed events pn. If you have observed 100 events that only seldom occur, the s.d. in the observed number is 10, or 10%; if you have observed 1000 events, the s.d. is 32 or 3.2%. If you want to gain a factor of 10 in accuracy, your observation time must be 100 times longer.

Examples

Here are a few examples of binomial distributions. Figure 4.1 shows the probability of obtaining k heads in 10 coin tosses, assuming the probability of obtaining a head in each throw is 0.5. Figure 4.2 shows the probability of obtaining k times a "six" in 60 throws of a perfect dice. You see that the distribution tends to become symmetrical for larger numbers, even if the probability for a single event is far from the symmetrical 0.5.

Figure 4.3 relates to "extra-sensory perception" (ESP) experiments parapsychologists used to perform to investigate the possibility of telepathy.[3]

[3] This is a case where scientists would demand a very high significance level in order to even consider positive experimental outcomes. Previous experimenters have fallen into all the statistical pitfalls you can think of. See Gardner (1957).

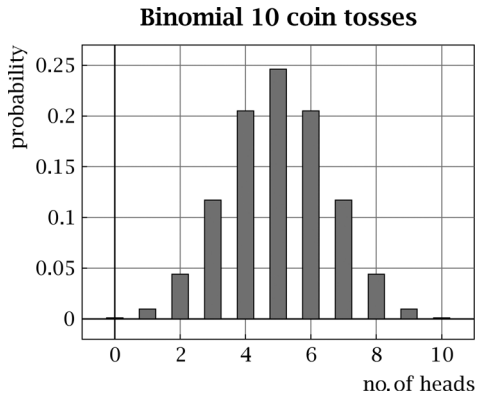

Figure 4.1 The probability of obtaining k heads in 10 coin tosses.

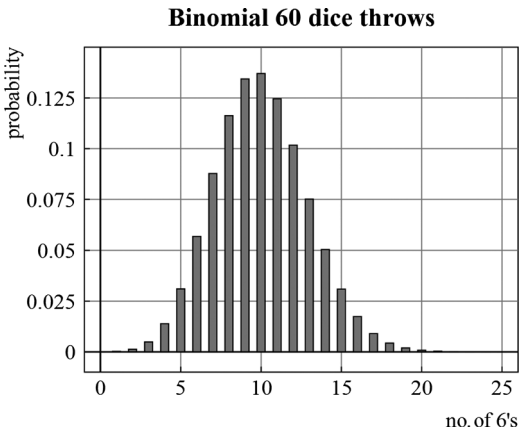

Figure 4.2 The probability of obtaining k faces "6" in 60 throws of a dice.

The "sender" sequentially selects cards from a well-mixed pack of "Zener cards," which contains an equal number of five types of card (each with a simple figure: square, circle, cross, star, wavy lines) and concentrates for a moment on the figure; the "receiver" notes which card he thinks has been drawn, without being able to see the card. One such experiment involves 25 cards. Assuming that telepathy does not exist, the probability of a correct guess is 0.2 and on average 5 cards will be guessed correctly. The probability of guessing *more than k* cards correctly is the binomial *survival function*

Table 4.1 *The binomial survival function* $1 - F(k)$*, giving the probability that more than k Zener cards are guessed correctly out of 25 trials.*

$\geq(k+1)$	$>k$	survival $1 - F(k)$
12	11	0.001 540
11	10	0.005 555
10	9	0.017 332
9	8	0.046 774
8	7	0.109 123
7	6	0.219 965

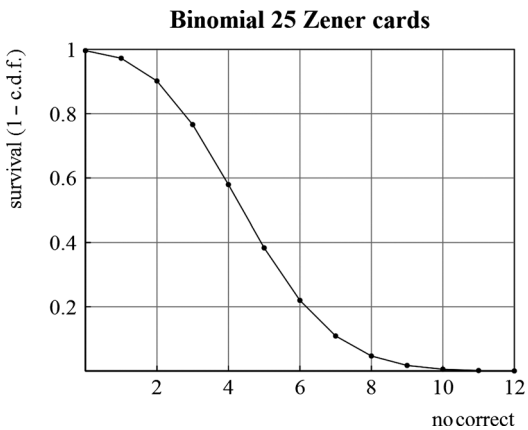

Figure 4.3 The "survival", i.e. the probability of guessing *more than k* cards correctly out of 25 trials. There are five different cards which are randomly presented.

(sf), which is 1 minus the cumulative distribution function (cdf). The exact meaning of the cdf and sf is:

$$\text{cdf}: \quad F(x): \quad \text{Prob}\{k \leq x\} = F(x); \tag{4.31}$$

$$\text{sf}: \quad 1 - F(x): \quad \text{Prob}\{k > x\} = 1 - F(x). \tag{4.32}$$

The survival function is given for some relevant values in Table 4.1 and in Fig. 4.3.

See **Python code** 4.1 on page 173 for codes to generate the functions and figures of this section.

From binomial to multinomial

When a random choice is made not among two possibilities, but among a number m possibilities, the statistics is that of a *multinomial* distribution. For example, an opinion poll asks which choice a voter will make among the five parties that figure in an election. Or, a certain sequence of amino acids in a protein can be classified as either $\alpha-helix$, $\beta-sheet$ or *random coil*. Or, you gather random variables in n distinct bins. The details of the multinomial distribution can be found in Appendix A4.

4.4 The Poisson distribution

You will encounter the Poisson distribution whenever you are *counting numbers*, such as a number of objects in a small volume of a homogeneous suspension (e.g. bacteria under a microscope or fish in a representative volume in a lake), or a number of photons detected in a given time interval Δt with a "single photon counter," or the number of gamma quanta counted in a given time interval originating from the radioactive decay of unstable nuclei.

If μ is the *average* number of events that can be expected, then the probability $f(k)$ of counting exactly k events is given by the Poisson distribution:

$$f(k) = \mu^k e^{-\mu}/k! \tag{4.33}$$

The Poisson distribution is a limiting case of the binomial distribution ($p \to 0$); for large k the Poisson distribution itself approaches a normal distribution. Details are given in Appendix A4.

The Poisson mass distribution is normalized. The mean and variance are given by:

$$E[k] = \mu, \tag{4.34}$$

$$\sigma^2 = E[(k-\mu)^2] = \mu. \tag{4.35}$$

The most important property of the Poisson distribution is that the standard deviation σ equals the square root of the mean μ. For example, a measurement counting 10 000 photons has a s.d. of 100, i.e. an uncertainty of 1 percent. When the number of events is sufficiently large (say, >20) then

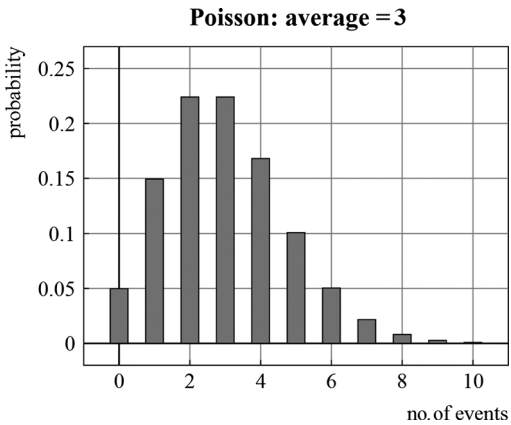

Figure 4.4 The probability that exactly k events are observed in a given time interval, when the events arrive randomly with an average of 3 per time interval.

the Poisson distribution is almost equal to the normal distribution with mean μ and s.d. $\sqrt{\mu}$.

Figure 4.4 shows the probability $f(k)$ of observing k events when the mean number $\mu = 3$. For example, a specialized hospital ward admits on the average 3 urgent patients per day; $f(k)$ is the probability that on a given day k patients arrive, assuming the patients arrive randomly. See Exercise 4.6.

4.5 The normal distribution

See data sheet NORMAL DISTRIBUTION on page 205.

The Gauss function

The pdf of the normal distribution is known mathematically as a *Gauss function*:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]. \tag{4.36}$$

The mean is μ, the variance is σ^2 and the s.d. is σ. The normal distribution is usually indicated by $N(\mu, \sigma)$. When we make the substitution

$$z = \frac{x-\mu}{\sigma}, \tag{4.37}$$

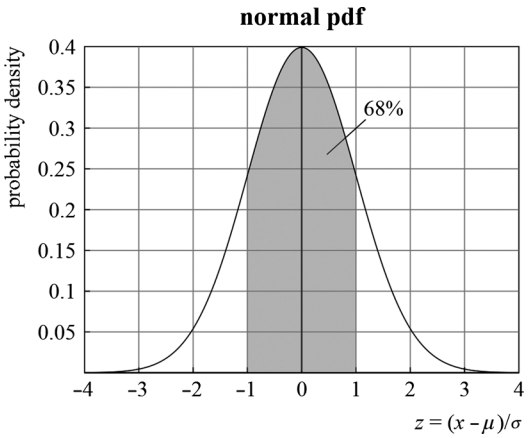

Figure 4.5 The standardized normal probability density function (pdf) $f(z); z = (x - \mu)/\sigma$, with μ being the mean and σ the standard deviation of the random variable x.

the *standardized normal distribution* is obtained, indicated by $N(0, 1)$. Its density distribution is

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{z^2}{2}\right]. \tag{4.38}$$

Figure 4.5 gives the standardized normal pdf. On the horizontal axis the reduced coordinate $(x - \mu)/\sigma$ is given. Thus the value 0 corresponds with $x = \mu$ and the value 1 with $x = \mu + \sigma$. The grey area gives the (integrated) probability that x lies between the values $\mu - \sigma$ and $\mu + \sigma$; this follows from the cumulative distribution function $F(z)$ and the probability equals $F(1) - F(-1) = 1 - 2F(-1) = 0.6826$ (68%).

Figure 4.6 gives the cumulative distribution function (cdf)

$$F(z) = \int_{-\infty}^{z} f(z')\, dz, \tag{4.39}$$

which expresses the probability that a sample from the normal distribution is not larger than z. The survival function (sf) $1 - F(z)$, expressing the probability that a normal variate exceeds the value z, is also given.

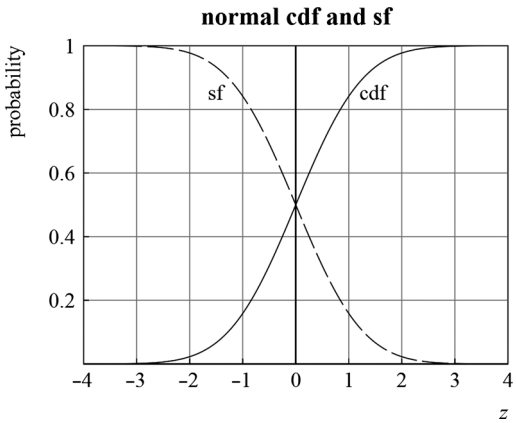

Figure 4.6 The standardized normal cumulative probability distribution function (pdf) $F(z); z = (x - \mu)/\sigma$, with μ being the mean and σ the standard deviation of the random variable x. The dashed curve is the survival function (sf) $1 - F(z)$.

Relation of cdf to error function

The function $F(z)$ can be expressed in terms of the *error function* $\text{erf}\,(z)$, which is a mathematical function defined as:[4]

$$\text{erf}\,(x) \stackrel{\text{def}}{=} \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2)\, dt. \tag{4.40}$$

Its complement is the complementary error function erfc:

$$\text{erfc}\,(x) = 1 - \text{erf}\,(x). \tag{4.41}$$

The relation is:

$$F(x) = \frac{1}{2}\,\text{erfc}\left(-x/\sqrt{2}\right) \quad \text{for } x < 0; \tag{4.42}$$

$$= \frac{1}{2}\left[1 + \text{erf}\,(x/\sqrt{2})\right] \quad \text{for } x \geq 0. \tag{4.43}$$

Probability scales

In order to judge whether a distribution is approximately normal, it is convenient to plot the cdf on a scale designed to produce a straight line in case of

[4] See, for example, Abramowitz and Stegun (1964).

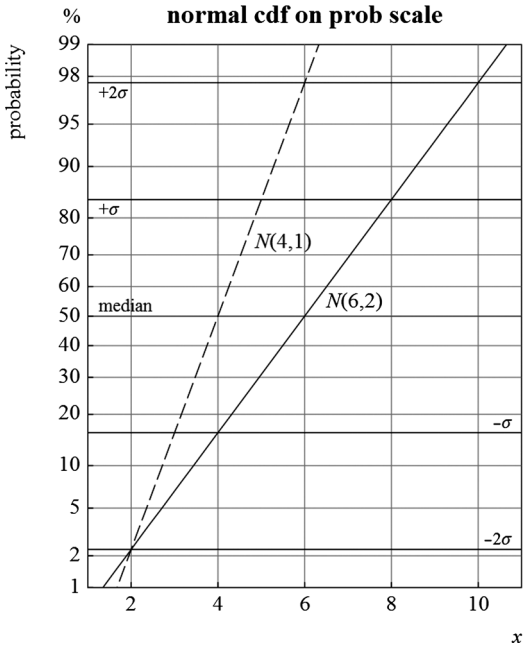

Figure 4.7 The cumulative distribution function (cdf) of a normal distribution $N(6,2)$, i.e. $\mu = 5; \sigma = 2$ is plotted on a "probability scale" (drawn line). Dashed line: $N(4,1)$.

normal distributions. Graph paper with appropriate divisions along the ordinate is commercially available (*probability paper*; print-yourself files can be downloaded from www.hjcb.nl/). With adequate computer software you can let the computer make the plots, rather than plotting by hand on paper. The plotting package `plotsvg` allows one to plot functions and cumulative distributions on a probability scale and such plots are often used in this book. Figure 4.7 plots two perfect normal distributions $N(6,2)$ and $N(4,1)$ on a probability scale: of course these are perfect straight lines. One can read the mean value and the standard deviation from such plots.

Significant deviations

Table 4.2 gives the probability that a sample x lies in a given interval and the probability that x exceeds a given value (the survival function $1 - F(z)$). You see that deviations of more than 2σ don't occur very often; deviations of more than 3σ are very rare. So if you find deviations of more than 3σ

Table 4.2 *Probability that a sample from a normal distribution occurs in the interval* $(\mu - \Delta, \mu + \Delta)$ *and the probability that a sample value exceeds* $\mu + \Delta$ *(or, equivalently, is smaller than* $\mu - \Delta$*), for various values of* Δ.

deviation Δ in units σ	Probability in $(\mu - \Delta, \mu + \Delta)$	Probability $> \mu + \Delta$
0.6745	50%	25%
1	68.3%	15.9%
1.5	86.6%	6.68%
2	95.45%	2.28%
2.5	98.76%	0.62%
3	99.73%	0.135%
4	99.993 66%	0.003 17%
5	99.999 943%	0.000 029%

in an experiment, you may safely conclude that is it improbable that such a deviation occurs by chance and designate the deviation as *significant*. Some researchers prefer to set the significance limit at 2.5σ or even at 2σ; what is best depends on the purpose (i.e. on the consequences of the decision taken on the basis of the measurement) and on the taste of the researcher. Of course the criterium used should always be made specific.

You should be especially careful when you consider the significance of one out of a *series* of experiments. It is not at all significant (on the contrary, it is quite likely with a probability of more than 70%) that at least one out of 100 independent measurements deviates more than 2.5σ; if you wish to maintain a significance level of e.g. 5% on the whole series of experiments, you should insist on a deviation of 3.5σ for at least 1 out of 100 results. Selecting the "significant" experiments and disregarding the "insignificant" ones, is a scientific crime. See pages 3 and 4 of the data sheet NORMAL DISTRIBUTION on page 205.

4.6 The central limit theorem

Of the various types of probability distributions the *normal distribution* is by far the most common in practice. The reason for this is that random fluctuations that are a result of the sum of many independent random components, tend to be distributed normally, independent of the type of distribution sampled by each component. This is the famous *central limit theorem*. The mean

or variance of the distribution of the sum equals the sum of the means or variances of the distribution of each of the contributing components: More precisely:

Let $x_i, i = 1 \ldots n$ be a set of random variables with arbitrary probability distribution with finite mean m_i and variance σ_i^2. Then for large n the random sum variable $x = x_1 + \cdots + x_n$ tends to sample a normal distribution $N(m, \sigma)$, with

$$m = \sum_{i=1}^{n} m_i, \tag{4.44}$$

$$\sigma^2 = \sum_{i=1}^{n} \sigma_i^2. \tag{4.45}$$

The theorem should be used with caution: if the distribution functions of contributing components have a non-existing (infinite) variance, the central limit theorem breaks down. Heavily skewed distributions may give problems as well. Appendix A5 on page 148 gives details.

Although the central limit theorem is very important and powerful, it is not a general justification for the assumption of normality of underlying probability distributions. Relatively small deviations are often normally distributed. This is not always true for larger deviations, for example in quantities like a concentration or an intensity that can only be positive. Be aware that in such cases non-normal, skewed, distributions are likely to occur.

4.7 Other distributions

There are many other probability distributions. Some are described shortly in this section; others we shall encounter later in this book: they are important for the assessment of confidence intervals for the properties derived from data series.

Log-normal distribution

The log-normal distribution is a normal distribution of $\log x$ instead of x. It is of course only defined for $x > 0$. This distribution is especially appropriate for variables that can never be negative, such as a concentration, a length, a volume, a time interval, etc.

The standard form for the density distribution function, as available in Python through the SciPy function `stats.lognorm.pdf`, is

$$f_{\mathrm{st}}(x, s) = \frac{1}{sx\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\ln x}{s}\right)^2\right], \tag{4.46}$$

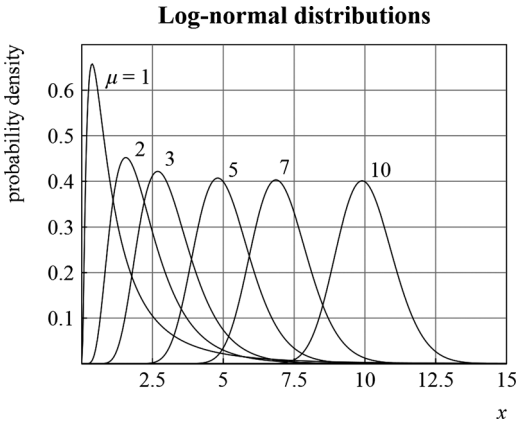

Figure 4.8 The probability density function (pdf) of log-normal distributions $f(x;\mu,\sigma)$, see (4.47), for various values of μ. The value of $\sigma = 1$ for all curves.

but a more convenient form is

$$f(x;\mu,\sigma) = \frac{1}{\mu} f_{\text{st}}\left(\frac{x}{\mu}, \frac{s}{\mu}\right). \tag{4.47}$$

In this form $f(x;\mu,\sigma)$ approaches the normal density function $N(\mu,\sigma)$ when μ/σ becomes larger. Figure 4.8 shows examples of the log-normal pdf with various μ, but all with the same $\sigma = 1$. For $\mu = 10\sigma$ the shape of the curve is virtually indistinguishable from the normal pdf.

The Lorentz distribution: undefined variance

A somewhat unusual distribution, but one with special interest, is the *Lorentz distribution*, also known by the name *Cauchy distribution*:

$$f(x;\mu,w) = \frac{1}{\pi w}\left[1 + \left(\frac{x-\mu}{w}\right)^2\right]^{-1}, \tag{4.48}$$

where μ is the mean and w is a width parameter. At $x = \mu \pm w$ the function is at half its maximum height. A measure for the width is the FWHH (full width at half height), equal to $2w$. This distribution may arise from spectroscopic experiments: the frequency distribution of emitted quanta from a sharp lifetime-limited excited state has a Lorentzian shape. The Lorentzian shape also arises in another context: Student's t-distribution for one degree of freedom (see data sheet STUDENT'S T-DISTRIBUTION on page 213).

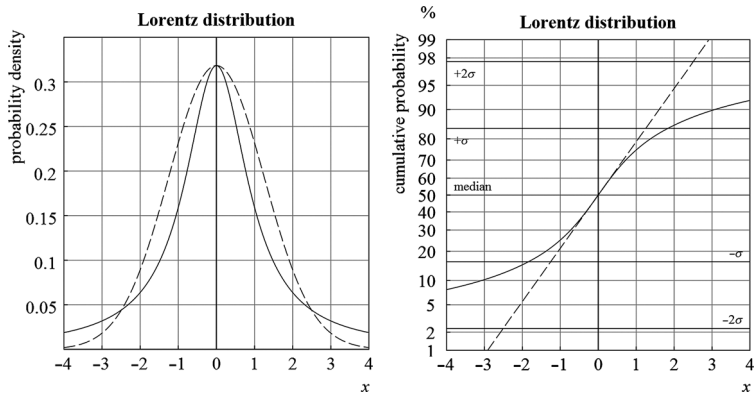

Figure 4.9 The probability density function (pdf) of the Lorentz distribution (drawn lines) $f(x; 0, 1)$, see (4.48), compared with a normal distribution (broken lines) at the same maximum height of the pdf, i.e. the same slope of the cdf at the median ($\sigma = \sqrt{\pi/2}$). Left: pdf, right: cdf on a probability scale.

The cumulative distribution is

$$F(x) = \frac{1}{2} + \frac{1}{\pi} \arctan \frac{x}{w}. \tag{4.49}$$

The problem with this distribution is that it has an infinite variance. Thus it makes no sense to estimate its variance from an actual data set. For distributions like this, including other distributions with wide tails, one should use *robust* methods (see Section 5.7 on page 63) to assess the accuracy of the mean of a series of measured samples.

Figure 4.9 depicts the Lorentz distribution, together with a normal distribution fitted with the same maximum of the pdf.

Lifetime and exponential distributions

Special types of distribution arise from considering lifetime distributions. For example, consider a large batch of incandescent lamps, all new from the factory. At time $t = 0$ you switch them all on and note the moment each lamp fails. The fraction of lamps that fails between t and $t + \Delta t$ (or equivalently, the fraction that has a lifetime between t and $t + \Delta t$) is $f(t)\Delta t$ (for small Δt), where $f(t)$ is the probability density function for the lifetime distribution. The cumulative distribution function $F(t) = \int_0^t f(t')\, dt'$ is the fraction that failed up to time t, and the survival function $1 - F(t)$ is the fraction that survives at time t (i.e., the fraction that has not (yet) failed). Another example is the

lifetime distribution of individuals in a population. Consider a large number of individuals and set for each $t = 0$ at birth; $f(t)\Delta t$ is the fraction with life span between t and $t + \Delta t$; $F(t)$ is the fraction with life span $\leq t$; $1 - F(t)$ is the fraction that survives at time t. An example from the molecular sciences is the time dependence of fluorescent intensity (emitted radiation quanta) after a fluorescent molecule has been excited by a short laser pulse at $t = 0$; $f(t)$ is the normalized time-dependent intensity.

The hazard function

The lifetime probability density or its cumulative distribution function describes the lifetime statistics, but does not describe the basic cause of death or failure. More basic is the *hazard function* (also called the *failure rate function*) $h(t)$. The hazard function is the probability density that a member of the population with age t will fail (die, drop out). In other words, the probability that a member fails in a small time interval Δt around t equals $h(t)\Delta t$. Because only a fraction $1 - F(t)$ is present in the population at time t, the following relation holds:

$$h(t) = \frac{f(t)}{1 - F(t)}. \tag{4.50}$$

From this relation, and using the fact that f is the derivative of F, we can solve for the lifetime density function:

$$f(t) = h(t) \exp\left[- \int_0^t h(t')\, dt' \right]. \tag{4.51}$$

The exponential distribution

Several distribution functions result from various choices of $h(t)$. By far the simplest choice, which describes quite common phenomena in physics or chemistry such as radioactive decay and first-order chemical reactions, is

$$h(t) = k \ \text{(constant)}, \tag{4.52}$$

called the *rate constant*. Its meaning is the relative fraction of the population members (e.g. number of radioactive nuclei n, concentration of reactant c, etc.) that disappear per unit of time

$$\frac{dn}{dt} = -kn, \tag{4.53}$$

$$\frac{dc}{dt} = -kc. \tag{4.54}$$

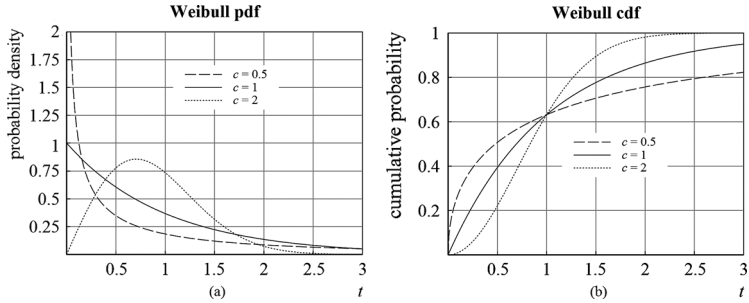

Figure 4.10 The distribution functions (pdf and cdf) of three Weibull distributions with $c = 0.5, 1, 2$. For $c = 1$ the exponential distribution is obtained.

It now follows from (4.51) that

$$f(t) = ke^{-kt} \tag{4.55}$$

and

$$F(t) = 1 - e^{-kt}. \tag{4.56}$$

This is an *exponential distribution*. The exponential distribution ($c = 1$) is depicted in Fig. 4.10

Population statistics

For the purpose of population statistics, e.g. for human population dynamics or for failure analysis, various general forms for the hazard functions have been proposed, leading to more general probability density functions for populations. The *Weibull* distribution[5] is a generalized form of the exponential distribution: the hazard function has the form

$$h(t) = ct^{c-1}. \tag{4.57}$$

Here c sets the time dependence of the failure rate; $c = 1$ recovers the exponential pdf, $c < 1$ means a higher initial rate (like a high infant mortality) and $c > 1$ means a higher failure rate at older age. The corresponding pdf is

$$f(t) = ct^{c-1} \exp[-t^c] \tag{4.58}$$

[5] A valuable source for information on distributions is the on-line NIST/SEMATECH e-Handbook of Statistical Methods on www.itl.nist.gov/div898/handbook.

and the cumulative distribution (cdf) is

$$F(t) = 1 - \exp[-t^c]. \tag{4.59}$$

Additional location (translating t) and scale (scaling t) parameters may be included. Figure 4.10 gives a few examples of Weibull distributions, including the exponential distribution.

See for the generation of Weibull distribution functions **Python code** 4.2 on page 174.

Chi-squared distribution

This is the distribution of the sum χ^2 of the squares of a number of normally distributed variables. The χ^2-distribution is used to obtain confidence intervals for predicted values when the s.d. of the data is known. See Section 7.4 on page 95 and the data sheet CHI-SQUARED DISTRIBUTION on page 199.

Student's t-distribution

This is the distribution of the ratio of a normally distributed variable and a χ^2-distributed variable. The t-distribution is used to assess confidence intervals for the mean, given a series of normally distributed data, when the s.d. of the distribution is not known beforehand. See Section 5.4 on page 59, the second example of Section 8.4 on page 115 and the data sheet STUDENT'S T-DISTRIBUTION on page 213.

F-distribution

This is the distribution of the ratio of two χ^2-distributed variables. The *F-ratio* is the ratio between two mean sum of squares (i.e., the sum of square deviations of a set of samples with respect to their average or with respect to a predicted value, divided by the number of degrees of freedom ν). The F-distribution (named by Snedecor) is the cumulative distribution function of the F-ratio F_{ν_1,ν_2} for the case that both sets of samples come from distributions with the same variance. It is usual to take the ratio as the largest value divided by the smallest value; if F_{ν_1,ν_2} exceeds the 99 percent level, the probability that both sets of samples come from the same distribution is less than 1 percent. The equation for F_{ν_1,ν_2} and a short table are given in data sheet F-DISTRIBUTION on page 201.

The F-distribution is useful in linear regression (see Chapter 7) in order to assess the relevance of the model that is fitted to the data. It compares the

variance in the data as explained by the model with the remaining variance of the data with respect to the model; the cumulative probability of the F-ratio then indicates whether the model contributes significantly to the explanation of the data variance.

The use in regression is a special case of the general "analysis of variance" (ANOVA), which is widely used in the assessments of the influence of external factors on a normally distributed variable. Such assessments belong to the statistical domain of *experimental* or *factorial design*: the analysis of the influence of a designed external factor. As this book concentrates on the processing of data to estimate probability distributions of parameters, the statistical treatment of experimental design falls outside of its scope.[6] However, in order to give some insight into the use of F-distributions, a simple one-way ANOVA example is given below.

A group of patients, randomly selected from a homogeneous population, is treated with a drug, while another group, randomly selected from the same population, is treated with a placebo. The groups are compared by measuring an objective test value and a statistical test is performed to assess the probability that the drug treatment has been effective. The assessment is phrased in terms of the probability that the *null hypothesis* H_0 = "the drug has no influence" is true. One computes two types of mean squared averaged deviations: first of the averages of each group with respect to the global average ("between-groups variance" or mean of the "regression sum of squares" SSR) and second of the values within each group with respect to the average of that group, added over all groups ("within-group variance" or mean of "error sum of squares" SSE). Each sum of squares is divided by the number of degrees of freedom ν, i.e., the number of samples minus the number of adjustable parameters. For the "between-groups variance" $\nu_1 = k - 1$ when there are k groups; for the "within-group variance" $\nu = n - k$. In this example there are two groups: $k = 2$, one control group with n_1 observations and average μ_1, and one treated group with n_2 observations and average μ_2. The overall average of $n = n_1 + n_2$ observations y_i is μ. The *F-ratio* is

$$F_{1,n-2} = \frac{\mathrm{SSR}/1}{\mathrm{SSE}/(n-2)}, \tag{4.60}$$

where

$$\mathrm{SSR} = n_1(\mu_1 - \mu)^2 + n_2(\mu_2 - \mu)^2; \tag{4.61}$$

$$\mathrm{SSE} = \sum_{i=1}^{n_1}(y_i - \mu_1)^2 + \sum_{i=n_1+1}^{n}(y_i - \mu_2)^2. \tag{4.62}$$

[6] There are many books covering factorial design, e.g. Walpole *et al.* (2007).

The F-test – in fact, the value $1 - F(F_{1,n-2})$ – now tells you what the probability is that *at least this ratio* would be found if your null-hypothesis were true. If this value is small (say, less than 0.01), you may conclude that the treatment has a significant effect.

Example

Imagine that you are a physician and you want to test a new drug for treating patients with high blood pressure. You select a group of ten patients with high blood pressure who do not (yet) receive treatment and who all have agreed to participate in your trial. You design a standard way to determine the blood pressure (e.g. the average of systolic pressures at 9 am on five consecutive days) and define the test value e.g. as the blood pressure after two weeks of treatment minus the value before treatment. Then you select five patients randomly to form the "treatment group"; the remaining five patients form the control group. The treatment group receives the drug treatment and the control group receives an indistinguishable placebo. You will accept the treatment as effective when the null hypothesis is rejected at a 95 percent confidence level.[7] The outcome of the experiment (the test values in mm Hg) is as follows:

treatment group: $-21, -2, -15, +3, -22$
control group: $-8, +2, +10, -1, -4$

The treatment group has an average of -11.5 and the control group has an average of -0.2. This looks like a positive result, but if you evaluate the appropriate sums you obtain:

$$\text{SSR} = 314;\ \text{SSE} = 698;\ F_{1,8} = [314/1]/[698/8] = 3.59;\ F(3.59) = 0.91.$$

This means that there is a 9 percent probability that the null hypothesis ("the treatment has no effect") is true and a 91 percent probability that the alternative hypothesis ("the treatment is effective") is true. So, although the result suggests that the treatment is effective, you cannot come to that conclusion

[7] It is important that you define all experimental details *and* the statistical methods to be used *before* you do the experiment without changing your methods during or after the experiment. The selection of patients and the performance of the measurements must be completely unbiased. In a serious experiment neither the patient nor the physician who performs the measurements is allowed to know which of the patients receive the treatment (a *double-blind* experiment). A serious experiment should involve a much larger group and include safeguards when intolerable side effects occur or when the treatment appears to be so effective that it would be ethically unacceptable to deprive the control group from the benefits of treatment. A serious hospital or research organization will set rules for such experiments on humans and establish an ethical approval committee. A serious journal will evaluate the quality of the experiment before publishing the results.

when you adhere to your preset 95 percent confidence level! What you have to do, of course, is to repeat your experiment with a (much) larger number of patients.

Summary *You now have distinguished probability density distributions, cumulative probability distributions and survival functions. You know what the expectation of a function over a given distribution is and you know how the mean, the variance, the standard deviation, the skewness and the kurtosis of a distribution are defined. The binomial distribution is the simplest discrete distribution; it is suitable to describe the random picking of one out of two unequal possibilities. Random picking of one out of several possibilities is described by the multinomial distribution. Random picking of an event on a continuous scale, as the time at which an impulse or photon is observed, leads to the Poisson distribution. In the limit of many events the latter leads to a continuous Gaussian or normal distribution. The normal distribution is quite common; it emerges when a deviation is composed of many independent random contributions, irrespective of the individual distributions for each of the contributions (the central limit theorem). Some other distributions play a role in special applications; lifetime distributions are an important subclass. Distributions that have infinite variance, such as the Lorentz distribution, may cause trouble because common rules do not apply. The chi-square, Student's t- and Snedecor's F-distributions play a role in the evaluation of data series.*

Exercises

4.1 In a lottery 5 percent of the tickets will produce a prize. If you buy ten tickets, what is the probability that you obtain no prize, 1 prize, 2 prizes, ...? Assume that there are so many tickets and prizes that the probability of obtaining a prize does not depend on the number of prizes you already have (this is called: a lottery with replacement).

4.2 When it is known that one measurement x has a probability of exceeding a given value x_m of 1 percent what then is the probability that *at least* one measurement in a series of 20 independent measurements will exceed x_m?

4.3 There will be elections where voters can elect one of two presidential candidates. You want to perform an opinion poll and predict the outcome

with a standard uncertainty of 1 percent. You expect roughly equal votes for either candidate. Assume that you are able to obtain the opinion of an unbiased random selection of voters, how many people do you have to select (what should be your sample size)?

4.4 You observe n independent events, each of which can have an outcome of 0 or 1. You count k_0 zeros and k_1 ones ($k_0 + k_1 = n$).
(a) What is your best estimate of the probability that a one appears?
(b) Give an estimate for the standard uncertainty in k_0.
(c) What is the standard uncertainty in k_1?
(d) You are finally interested in the ratio $r = k_1/k_0$. What is the standard uncertainty in r?

4.5 Show that the Poisson function 4.33 is normalized.

4.6 (a) With the hospital example of Fig. 4.4: assume each patient occupies a bed for one day and the ward has seven beds. When more than seven patients arrive, the excess is transported to another hospital. How many beds are occupied on average?
(b) How many patients per day are transported on average?
(c) If an unoccupied bed costs \$ 300 per day and transporting one patient costs \$ 1500, financially optimize the number of beds. How many patients per day are transported in the optimized case?

4.7 (a) A photosensitive device produces one electrical impulse for every absorbed photon, but also produces impulses when there is no light (the "dark current"). The number of impulses counted in 1 s is 100 without radiation and 900 with radiation. How large is the relative standard uncertainty in the measured radiation intensity?
(b) How large will be the relative standard uncertainty in the measured radiation intensity when the measurement (with and without radiation) is repeated 100 times?

4.8 What is the probability that a sample from a normally distributed quantity lies in the interval $[\mu - 0.1\sigma, \mu + 0.1\sigma]$?

4.9 (See the data sheet NORMAL DISTRIBUTION on page 205)
Using the approximation for large x mentioned on page 2 of the data sheet NORMAL DISTRIBUTION, determine the probability that the value $x = 6\sigma$ is exceeded. Is this approximation valid for this case?

4.10 The central limit theorem has a useful application: By adding 12 random numbers r, which are uniformly distributed over the interval $[0, 1\rangle$, and

subtracting 6 from the sum, you obtain in a good approximation a sample from a normal distribution with $\mu = 0$ and $\sigma = 1$:

$$x = \sum_{i=1}^{12} r_i - 6$$

(a) Show that $\langle x^2 \rangle = 1$.

(b) Generate a list of 100 normally distributed numbers by this method.

(c) Plot the cdf of this list on a "probability" scale.

4.11 Compute the mean and variance of the exponential distribution.

4.12 Refer to the example on page 49. In a similar trial the following results were obtained:

treatment group: $-6, 2, -8, -7, -12$
control group: $5, -1, 3, -4, 0$

Compute the F-ratio and the corresponding cumulative probability using the F-distribution. What conclusions would you derive from this F-test?