1
Fundamental Concepts

Roger Barlow

1.1
Introduction

Particle physics is all about random behaviour. When two particles collide, or even when a single particle decays, we can't predict with certainty what will happen, we can only give probabilities of the various different outcomes. Although we measure the lifetimes of unstable particles and quote them to high precision – for the τ lepton, for example, it is 0.290 ± 0.001 ps – we cannot say exactly when a particular τ will decay: it may well be shorter or longer. Although we know the probabilities (called, in this context, branching ratios) for the different decay channels, we can't predict how any particular τ will decay – to an electron, or a muon, or various hadrons.

Then, when particles travel through a detector system they excite electrons in random ways, in the gas molecules of a drift chamber or the valence band of semiconducting silicon, and these electrons will be collected and amplified in further random processes. Photons and phototubes are random at the most basic quantum level. The experiments with which we study the properties of the basic particles are random through and through, and a thorough knowledge of that fundamental randomness is essential for machine builders, for analysts, and for the understanding of the results they give.

It was not always like this. Classical physics was deterministic and predictable. Laplace could suggest a hypothetical demon who, aware of all the coordinates and velocities of all the particles in the Universe, could then predict all future events. But in today's physics the demon is handicapped not only by the uncertainties of quantum mechanics – the impossibility of knowing both coordinates and velocities – but also by the greater understanding we now have of chaotic systems. For predicting the flight of cannonballs or the trajectories of comets it was assumed, as a matter of common sense, that although our imperfect information about the initial conditions gave rise to increasing inaccuracy in the predicted motion, better information would give rise to more accurate predictions, and that this process could continue without limit, getting as close as one needed (and could afford) to

Data Analysis in High Energy Physics, First Edition. Edited by Olaf Behnke, Kevin Kröninger, Grégory Schott, and Thomas Schörner-Sadenius.

perfect prediction. We now know that this is not true even for some quite simple systems, such as the compound pendulum.

That is only one of the two ways that probability comes into our experiments. When a muon passes through a detector it may, with some probability, produce a signal in a drift chamber: the corresponding calculation is a *prediction*. Conversely a drift chamber signal may, with some probability, have been produced by a muon, or by some other particle, or just by random noise. To interpret such a signal is a process called *inference*. Prediction works forwards in time and inference works backwards. We use the same mathematical tool – *probability* – to cover both processes, and this causes occasional confusion. But the statistical processes of inference are, though less visibly dramatic, of vital concern for the analysis of experiments. Which is what this book is about.

1.2 Probability Density Functions

The outcomes of random processes may be described by a variable (or variables) which can be *discrete* or *continuous*, and a discrete variable can be *quantitative* or *qualitative*. For example, when a τ lepton decays it can produce a muon, an electron, or hadrons: that's a qualitative difference. It may produce one, three or five charged particles: that's quantitative and discrete. The visible energy (i.e. not counting neutrinos) may be between 0 and 1777 MeV: that's quantitative and continuous.

The probability prediction for a variable x is given by a function: we can call it $f(x)$. If x is discrete then $f(x)$ is itself a probability. If x is continuous then $f(x)$ has the dimensions of the inverse of x: it is $\int f(x)\mathrm{d}x$ that is the dimensionless probability, and $f(x)$ is called a *probability density function* or *pdf*.[1] There are clearly an infinite number of different pdfs and it is often convenient to summarise the properties of a particular pdf in a few numbers.

1.2.1 Expectation Values

If the variable x is quantitative then for any function $g(x)$ one can form the average

$$E[g] = \int g(x)\, f(x)\mathrm{d}x \quad \text{or, as appropriate,} \quad E[g] = \sum g(x)\, f(x)\ , \qquad (1.1)$$

where the integral (for continuous x) or the sum (for discrete x) covers the whole range of possible values. This is called the *expectation value*. It is also sometimes written $\langle g \rangle$, as in quantum mechanics. It gives the mean, or average, value of g, which is not necessarily the most likely one – particularly if x is discrete.

1) The parton density functions of QCD, PDFs, share the abbreviation and are indeed pdfs in both senses.

1.2.2 Moments

For any pdf $f(x)$, the integer powers of x have expectation values. These are called the (algebraic) *moments* and are defined as

$$\alpha_n = E[x^n] \,. \tag{1.2}$$

The first moment, α_1, is called the *mean* or, more properly, *arithmetic mean* of the distribution; it is usually called μ and often written $\overline{x}$. It acts as a key measure of *location*, in cases where the variable x is distributed with some known shape about a particular point.

Conversely there are cases where the shape is what matters, and the absolute location of the distribution is of little interest. For these it is useful to use the *central moments*

$$m_n = E[(x-\mu)^n] \,. \tag{1.3}$$

1.2.2.1 Variance

The second central moment is also known as the *variance*, and its square root as the *standard deviation*:

$$V[x] = \sigma^2 = m_2 = E[(x-\mu)^2] \,. \tag{1.4}$$

The variance is a measure of the width of a distribution. It is often easier to deal with algebraically whereas the standard deviation σ has the same dimensions as the variable x; which to use is a matter of personal choice. Broadly speaking, statisticians tend to use the variance whereas physicists tend to use the standard deviation.

1.2.2.2 Skew and Kurtosis

The third and fourth central moments are used to build shape-describing quantities known as *skew* and *kurtosis* (or *curtosis*):

$$\gamma_1 = \frac{m_3}{\sigma^3} \,, \tag{1.5}$$

$$\gamma_2 = \frac{m_4}{\sigma^4} - 3 \,. \tag{1.6}$$

Division by the appropriate power of σ makes these quantities dimensionless and thus independent of the scale of the distribution, as well as of its location. Any symmetric distribution has zero skew: distributions with positive skew have a tail towards higher values, and conversely negative skew distributions have a tail towards lower values. The Poisson distribution has a positive skew, the energy recorded by a calorimeter has a negative skew. A Gaussian has a kurtosis of zero – by definition, that's why there is a '3' in the formula. Distributions with positive

kurtosis (which are called *leptokurtic*) have a wider tail than the equivalent Gaussian, more centralised or *platykurtic* distributions have negative kurtosis. The Breit–Wigner distribution is leptokurtic, as is Student's *t*. The uniform distribution is platykurtic.

1.2.2.3 Covariance and Correlation

Suppose you have a pdf $f(x, y)$ which is a function of two random variables, x and y. You can not only form moments for both x and y, but also for combinations, particularly the *covariance*

$$\mathrm{cov}[x, y] = E[xy] - E[x]E[y] \,. \tag{1.7}$$

If the joint pdf is factorisable: $f(x, y) = f_x(x) \cdot f_y(y)$, then x and y are independent, and the covariance is zero (although the converse is not necessarily true: a zero covariance is a necessary but not a sufficient condition for two variables to be independent).

A dimensionless version of the covariance is the *correlation* ρ:

$$\rho = \frac{\mathrm{cov}[x, y]}{\sigma_x \sigma_y} \,. \tag{1.8}$$

The magnitude of the correlation lies between 0 (uncorrelated) and 1 (completely correlated). The sign can be positive or negative: amongst a sample of students there will probably be a positive correlation between height and weight, and a negative correlation between academic performance and alcohol consumption.

If there are several (i.e. more than two) variables, $x_1, x_2, \ldots, x_N$, one can form the *covariance* and *correlation matrices*:

$$V_{ij} = \mathrm{cov}[x_i, x_j] = E[x_i x_j] - E[x_i]E[x_j] \,, \tag{1.9}$$

$$\rho_{ij} = \frac{V_{ij}}{\sigma_i \sigma_j} \,, \tag{1.10}$$

and V_{ii} is just σ_i^2.

1.2.2.4 Marginalisation and Projection

Mathematically, any pdf $f(x, y)$ is a function of two variables x and y. They can be similar in nature, for example the energies of the two electrons produced by a converting high energy photon, or they can be different, for example the position and direction of particles undergoing scattering in material.

Often we are really interested in one parameter (say x) while the other (say y) is just a *nuisance parameter*. We want to reject the extra information shown in the two-dimensional function (or scatter plot). This can be done in two ways: the *projection* of x, $f(x)|_y$ is obtained by choosing a particular value of y, the *marginal distribution* $f(x) = \int f(x, y)\mathrm{d}y$ is found by integrating over y.

Projections can be useful for illustration, otherwise to be meaningful you have to have a good reason for choosing that specific value of y. Marginalisation requires that the distribution in y, like that of x, is properly normalised.

1.2.2.5 Other Properties

There are many other properties that can be quoted, depending on the point we want to bring out, and on the established usage of the field.

The mean is not always the most helpful measure of location. The *mode* is the value of x at which the pdf $f(x)$ is maximum, and if you want a typical value to quote it serves well. The *median* is the midway point, in the sense that half the data lie above and half below. It is useful in describing very skewed distributions (particularly financial income) in which fluctuations in a small tail would give a big change in the mean.

We can also specify dispersion in ways that are particularly useful for non-Gaussian distributions by using *quantiles*: the upper and lower *quartiles* give the values above which, and below which, 25% of the data lie. *Deciles* and *percentiles* are also used.

1.2.3 Associated Functions

The *cumulative distribution function*

$$F(a) = \int_{-\infty}^{a} f(x)\mathrm{d}x \quad \text{or, as appropriate,} \quad F(a) = \sum f(x_i)\Theta(a - x_i)\ , \tag{1.11}$$

where Θ is the Heaviside or step function ($\Theta(x) = 1$ for $x \geq 0$ and 0 otherwise), giving the probability that a variable will take a value up to a, is occasionally useful.

The *characteristic function*

$$\phi(u) = E[\mathrm{e}^{\mathrm{i}ux}] = \int \mathrm{e}^{\mathrm{i}ux} f(x)\mathrm{d}x\ , \tag{1.12}$$

which is just (up to factors of 2π) the Fourier transform of the pdf, is also met with sometimes as it has useful properties.

1.3 Theoretical Distributions

A pdf is a mathematical function. It involves a variable (or variables) describing the random quantity concerned. This may be a discrete integer or a continuous real number. It also involves one or more parameters. In what follows we will denote a random variable by x for a real number and r for an integer. Parameters generally have their traditional symbols for particular pdfs: where we refer to a generic parameter we will call it θ. It is often helpful to write a function as $f(x;\theta)$ or $f(x|\theta)$, separating this way more clearly the random variable(s) from the adjustable parameter(s). The semicolon is preferred by some, the line has the advan-

tage that it matches the notation used for conditional probabilities, described in Section 1.4.4.1.

There are many pdfs in use to model the results of random processes. Some are based on physical motivations, some on mathematics, and some are just empirical forms that happen to work well in particular cases.

The overwhelmingly most useful form is the *Gaussian* or *normal* distribution. The *Poisson* distribution is also encountered very often, and the *binomial* distribution is not uncommon. So we describe these in some detail, and then some other distributions rather more briefly.

1.3.1 The Gaussian Distribution

The Gaussian, or normal, distribution for a continuous random variable x is given by

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad . \tag{1.13}$$

It has two parameters; the function is manifestly symmetrical about the location parameter μ, which is the mean (and mode, and median) of the distribution. The scale parameter σ is also the standard deviation of the distribution. So there is, in a sense, only one Gaussian, the *unit Gaussian* or *standard normal distribution* $f(x;0,1)$ shown in Figure 1.1. Any other Gaussian can be obtained from this by scaling by a factor σ and translating by an amount μ. The Gaussian distribution is sometimes denoted $\mathcal{N}(x;\mu,\sigma)$.

The Gaussian is ubiquitous (hence the name 'normal') because of the *central limit theorem*, which states that if any distribution is convoluted with itself a large number of times, the resulting distribution tends to a Gaussian form. For a proof, see for example Appendix 2 in [1].

Gaussian random numbers are much used in simulation, and a suitable random number generator is available on most systems. If it is not, then you can generate a unit Gaussian by taking two uniformly generated random numbers u_1, u_2, set

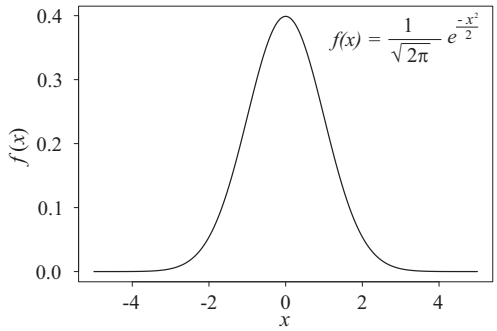

Figure 1.1 The unit Gaussian or standard normal distribution.

$\theta = 2\pi u_1, r = \sqrt{-2\ln u_2}$, and then $r\cos\theta$ and $r\sin\theta$ are independent samples from a unit Gaussian.

The product of two independent Gaussians gives a two-dimensional function

$$f(x, y; \mu_x, \mu_y, \sigma_x, \sigma_y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]}, \tag{1.14}$$

but the most general quadratic form in the exponent must include the cross term and can be written as

$$f(x, y; \mu_x, \mu_y, \sigma_x, \sigma_y, \rho) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{(1-\rho^2)}} e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]}, \tag{1.15}$$

where the parameter ρ is the correlation between x and y. For N variables, for which we will use the vector $\boldsymbol{x}$, the full form of the multivariate Gaussian can be compactly written using matrix notation:

$$f(\boldsymbol{x}; \boldsymbol{\mu}, \mathbf{V}) = \frac{1}{(2\pi)^{N/2}|\mathbf{V}|} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{V}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}. \tag{1.16}$$

Here, $\mathbf{V}$ is the covariance matrix described in Section 1.2.2.3.

The *error function* and the *complementary error function* are basically closely related to the cumulative Gaussian

$$\mathrm{erf}(y) = \frac{2}{\sqrt{\pi}} \int_0^y e^{-x^2} dx, \tag{1.17}$$

$$\mathrm{erfc}(y) = \frac{2}{\sqrt{\pi}} \int_y^\infty e^{-x^2} dx. \tag{1.18}$$

Their main use is in calculating Gaussian *p-values* (see Section 1.3.4.6). The probability that a Gaussian random variable will lie within one standard deviation, or '1 σ', of the mean is 68% obtained by calculating $\mathrm{erf}(y = 1)$. Conversely, the chance that a variable drawn from a Gaussian random process will lie outside 1 σ is 32%. Given such a process – say a mean of 10.2 and a standard deviation of 3.1 – then if you confront a particular measurement – say 13.3 – it is quite plausible that it was produced by the process. One says that its *p*-value, the probability that the process would produce a measurement this far, or further, from the ideal mean, is 32%. Conversely, if the number were 25.7 rather than 13.3, that would be 5 σ rather than 1 σ, for which the *p*-value is only $5.7 \cdot 10^{-7}$. In discussion of discoveries (or otherwise) of new particles and new effects this language is turned round, and a discovery with a *p*-value of $5.7 \cdot 10^{-7}$ is referred to as a '5 σ result'[2]. A translation is given in Table 1.1 – although for practical purposes it is easier to use functions such as `pnorm` and `qnorm` in the programming language R [2], or `TMath::Prob` in ROOT [3].

2) Note that there is a subtle difference between a one-sided and two-sided *p*-value. Details will be discussed in Chapter 3

Table 1.1 Two-sided Gaussian *p*-values for 1σ to 5σ deviations.

Deviation	*p*-value (%)
$1\,\sigma$	31.7
$2\,\sigma$	4.56
$3\,\sigma$	0.270
$4\,\sigma$	0.006 33
$5\,\sigma$	0.000 057 3

1.3.2 The Poisson Distribution

The Poisson distribution

$$f(n;\nu) = \frac{\nu^n}{n!}\mathrm{e}^{-\nu} \tag{1.19}$$

describes the probability of n events occurring when the mean expected number is ν; n is discrete and ν is continuous. Typical examples are the number of clicks produced by a Geiger counter in an interval of time, or, famously, the number of Prussian cavalrymen killed by horse-kicks [4]. Some examples are shown in Figure 1.2.

The Poisson distribution has a mean of ν and a standard deviation $\sigma = \sqrt{\nu}$. This property – that the standard deviation is the square root of the mean – is a key

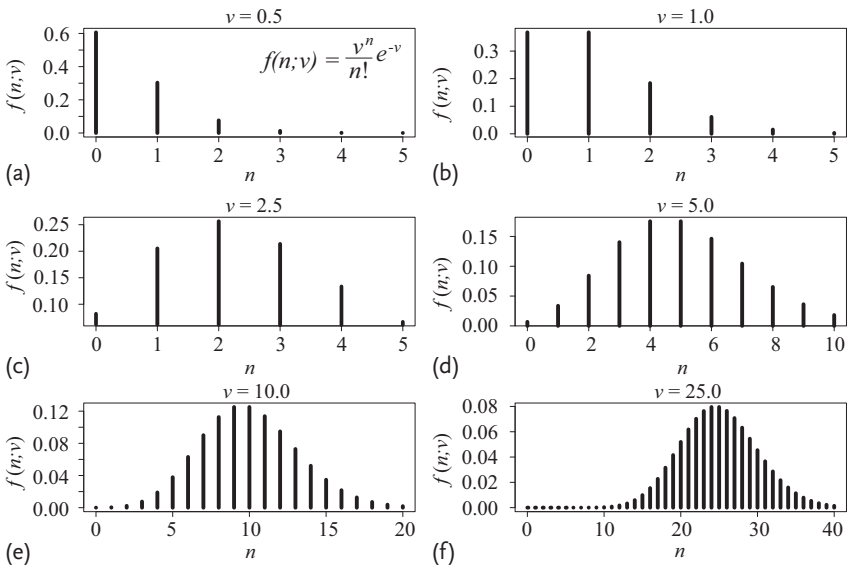

Figure 1.2 Poisson distributions for (a) $\nu = 0.5$, (b) $\nu = 1.0$, (c) $\nu = 2.5$, (d) $\nu = 5.0$, (e) $\nu = 10.0$, (f) $\nu = 25.0$.

fact about distributions generated by a Poisson process, which is important as this includes most cases where a number of samples is taken, including the contents of the bin of a histogram.

Example 1.1 Counting cosmic muons

In an experiment built to measure cosmic muons, the expected rate of muons in one run of the experiment is 0.45 events. This means that you have a 64% probability of observing no decays, a 29% probability of a single decay, 6% chance of seeing two and less than 1% of seeing three.

1.3.3 The Binomial Distribution

The binomial distribution describes a generalisation of the simple problem of the numbers of heads and tails that can arise from spinning a coin several times. The probability for getting r 'successes' from N 'trials' given an intrinsic probability of success p is

$$f(r; N, p) = \frac{N!}{r!(N-r)!} p^r (1-p)^{n-r} \,. \tag{1.20}$$

Sometimes one writes q instead of $1-p$, which makes the algebra prettier. The distribution has a mean of Np and a standard deviation $\sigma = \sqrt{Np(1-p)} = \sqrt{Npq}$. The factor $N!/[r!(N-r!)]$ is the number of ways that r objects may be chosen from N, and is often written $\binom{N}{r}$.

Example 1.2 Tracking chambers

A charged particle in an experiment goes through a set of six tracking chambers, which measure its position. Each of them is 95% efficient. If you require all six chambers to register a hit in order to define a reconstructed track, the efficiency of the system will clearly be $0.95^6 = 73.5\%$. If you are satisfied with five or more hits the efficiency is 96.7%. If at least four hits are enough, the track efficiency is 99.8%.

If p is small then the distribution can be approximated by a Poisson distribution[3] of mean Np. This is often used implicitly when analysing Monte Carlo samples: if you generate 1 000 000 Monte Carlo events, of which 100 end up in some particular histogram bin, then strictly speaking this is described by a binomial process rather than a Poisson. In practice you can take the error as the Poisson $\sqrt{100}$ rather than a binomial $\sqrt{1\,000\,000 \cdot 0.0001 \cdot 0.9999}$. This doesn't work if p is large. If 9 out of 10 events are accepted by the trigger, the error on the trigger efficiency of 90% is not

3) Indeed the Poisson can be derived as the limit of the binomial as $N \to \infty, p \to 0$ with Np constant.

$\sqrt{9}/10 = 30\%$ but $\sqrt{0.9 \cdot 0.1/10} = 9.5\%$ (in such a case the shortcut is to take the one lost event as approximately Poisson, giving the error as 10%, which is close).

If N is large and p is not small then the distribution is approximately a Gaussian.

If there are not just two possible outcomes but n, with probabilities $\{p_1, p_2, \ldots, p_n\}$, then the total probability of getting r_1 of the first outcome, r_2 of the second, and so on, is

$$f(r_1, r_2, \ldots, r_n; N, p_1, p_2, \ldots, p_n) = \frac{N!}{\prod r_i!} \prod p_i^{r_i} \,. \tag{1.21}$$

This is the *multinomial distribution*.

1.3.4 Other Distributions

There are many, many other possible distribution functions, and it is worth listing some of those more often met with.

1.3.4.1 The Uniform Distribution

The *uniform distribution*, also known as the *rectangular* or *top-hat* distribution, is constant inside some range – call this range $-a/2$ to $+a/2$, so the width is a; if the range is not central about zero but about some other value this is easily done by a translation. The mean, clearly, is zero, and the standard deviation is $a/\sqrt{12}$. This can be used in position measurements by a hodoscope: if a rectangular slab of scintillator gives a signal, you know that a track went through it but you do not know where. It is reasonable to assume a uniform distribution for the pdf of the hit position.

This can be relevant in considering some systematic uncertainties on the total result, as is also discussed in Section 8.4.1.2. For example, if you set up an experiment to run overnight, counting events with some efficiency E_1, and when you arrive in the morning you find a component has tripped so the efficiency is E_2, with no information about when this happened, your efficiency has to be quoted as $(E_1 + E_2)/2 \pm (|E_1 - E_2|)/\sqrt{12}$. It can also be applied to theoretical models: when two models give different predictions you are justified in using their mean as your prediction, with a (systematic) error which is the difference divided by $\sqrt{12}$, if (and only if) these two models represent absolute extremes and you really have no feeling as to where between the two extremes the truth may lie.

1.3.4.2 The Cauchy, or Breit–Wigner, or Lorentzian Distribution

In nuclear and particle physics the function

$$f(E; M, \Gamma) = \frac{1}{2\pi} \frac{\Gamma}{(E - M)^2 + (\Gamma/2)^2} \tag{1.22}$$

gives the variation with the energy E of a cross section produced by the formation of a state with mass M and width Γ. It can be written more neatly in dimensionless

form as

$$f(x) = \frac{1}{\pi}\frac{1}{1+x^2}\,, \tag{1.23}$$

where $x = (E - M)/(\Gamma/2)$. The mean is clearly M. It does not have a variance: the integral $\int x^2 f(x)\mathrm{d}x$ is divergent. If you have to compare this curve and with that of a Gaussian, the *full width at half maximum* (FWHM) is clearly Γ for this curve and for a Gaussian it is $2\sqrt{2\ln 2}\sigma = 2.35\sigma$.

This distribution is used in fitting resonance peaks (provided the width is much larger than the measurement error on E). It also has an empirical use in fitting a set of data which is almost Gaussian but has wider tails. This often arises in cases where a fraction of the data is not so well measured as the rest. A double Gaussian may give a good fit, but it often turns out that this form does an adequate job without the need to invoke extra parameters.

1.3.4.3 The Landau Distribution

When a charged particle passes an atom, its electrons experience a changing electromagnetic field and acquire energy. The amount of energy may be large; on rare occasions it will be large enough to create a delta ray. The probability distribution for the energy loss was computed by Landau [5] and is given by

$$f(\lambda) = \frac{1}{\pi}\int_0^\infty \mathrm{e}^{-u\ln u - \lambda u}\sin(\pi u)\mathrm{d}u\,, \tag{1.24}$$

where $\lambda = (\Delta - \Delta_0)/\xi$. Here, Δ is the actual energy loss, Δ_0 is a location parameter, and ξ is a scale, exact values for which depend on the material. This distribution has a peak at Δ_0, cuts off quickly below that, and has a very large long positive tail. The function is shown in Figure 1.3.

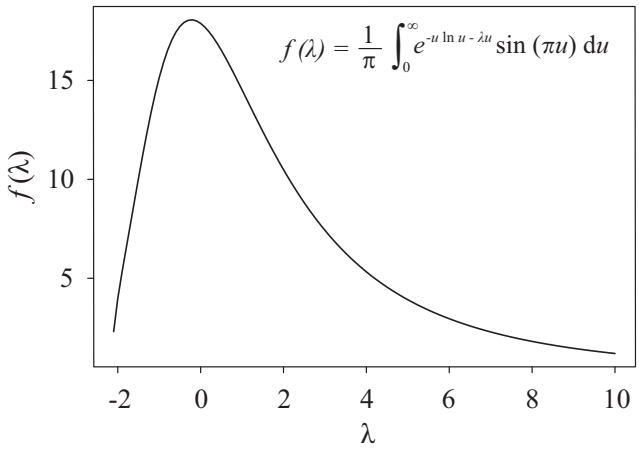

Figure 1.3 The Landau distribution.

The Landau distribution has very unpleasant mathematical properties. Some of its integrals diverge, for example it has no variance (like the Cauchy distribution), and, worse than that, it does not even have a mean. The ensuing complications can be avoided on a case-by-case basis by imposing an upper limit on the energy loss, as a particle cannot lose more than 100% of its energy.

There is a function $1/(\sqrt{2\pi})\mathrm{e}^{-1/2\lambda+\mathrm{e}^{-\lambda}}$ which is described in some places as 'the Landau distribution'. It is not. It is an approximation to the Landau distribution [6], and not a very good one at that.

1.3.4.4 The Negative Binomial Distribution

This considers the familiar binomial, but with a twist. As before, some process has a random probability p of success and $q \equiv 1 - p$ of failure, and is repeated for many trials. But now instead of asking the probability of r successes from a fixed number of trials n, we ask for the probability of r successes before encountering a fixed number k of failures. This is given by

$$f(r; k, p) = \frac{(k + r - 1)!}{r!(k - 1)!} q^k p^r \,. \tag{1.25}$$

It is the probability for r successes and $k - 1$ failures in any permutation, followed by a final kth failure. The combinatorial factor can also be written $(-1)^r \binom{-k}{r}$, hence the name 'negative binomial'. This can readily be extended to non-integer values by writing it as

$$f(r; k, p) = \frac{\Gamma(k + r)}{\Gamma(k) r!} q^k p^r \,, \tag{1.26}$$

although it is not clear what physical meaning this may have. Γ is the Gamma function, defined as

$$\Gamma(k) = \int_0^{+\infty} \mathrm{e}^{-t} t^{k-1} \mathrm{d}t \,. \tag{1.27}$$

The negative binomial distribution has a mean $\mu = (p/q)k$ and a variance $V = (p/q^2)k$. The negative binomial approaches the Poisson as k becomes large and p small with constant $p\,k \equiv \mu$.

1.3.4.5 Student's *t* Distribution

If you take a sample of n values, $\{x_1, \ldots, x_n\}$, from a Gaussian and histogram their differences from the true mean, divided by the standard deviation (a quantity often called the *pull distribution*), then this gives a unit Gaussian, that is a Gaussian with $\mu = 0$, $\sigma = 1$, which can be a useful check that you have your errors right. If, as often happens, the true mean is unknown, then the spread about the measured mean is slightly smaller than 1, by a factor $\sqrt{(n-1)/n}$.

If the standard deviation σ is also unknown, then you can use instead the estimated $\hat{\sigma} = \sqrt{\overline{(x-\mu)^2}}$ if μ is known or $\hat{\sigma} = \sqrt{n/(n-1)\overline{(x-\overline{x})^2}}$ if it is not. Now,

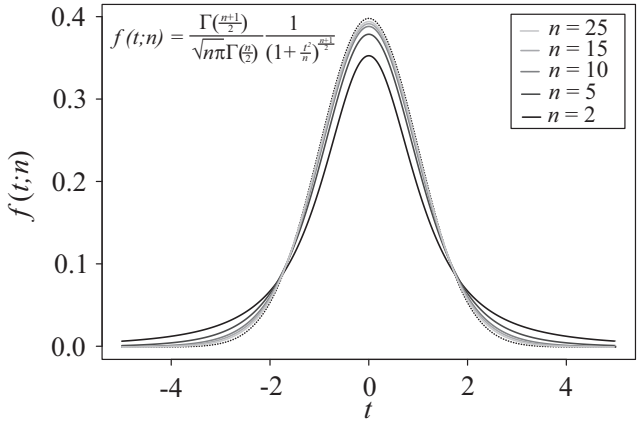

Figure 1.4 Student's t distribution for $n = 2, 5, 10, 15, 25$ with the Gaussian (dotted) for comparison.

for small n especially, this is not a very good estimator, and because you are dividing the differences from the mean by this bad estimate, the distribution for

$$t = \frac{x - \mu}{\hat{\sigma}} \tag{1.28}$$

is not given by a Gaussian, but by *Student's t distribution* for $n - 1$ degrees of freedom, where Student's t distribution is given by

$$f(t; n) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \frac{1}{\left(1 + \frac{t^2}{n}\right)^{\frac{n+1}{2}}} . \tag{1.29}$$

This tends to a unit Gaussian as n becomes large, but for small n it has tails which are significantly wider (see Figure 1.4): large t values can result if $\hat{\sigma}$ is an underestimate of the true value. The mean is clearly zero; the variance is not one, as it would be for a unit Gaussian, but $n/(n-2)$.

Example 1.3 Light yields in scintillators

You have five samples of scintillator from a manufacturer with light yields measured (in some units) as 1.23, 1.42, 1.35, 1.29 and 1.40. A second, cheaper, manufacturer provides a sample whose yield is 1.19. Does this give reason to believe that the cheaper sample has an inferior light yield?

The sample mean is 1.338 and the estimated standard deviation is 0.079, so the cheaper sample is 1.90 standard deviations below the mean. If this were a Gaussian distribution then the probability of a value lying this far below the mean is only 2.9% – so you would take this as strong evidence that the cheaper process was not so good. But for Student's t with four degrees of freedom the probability is (consulting the tables or evaluating a function) only 6.5%, so your evidence would be weaker (the calculations were done using the R function `pt(x,ndf)`).

1.3.4.6 The χ^2 Distribution

In describing the agreement between a predictive function $g(x)$ and a set of n measurements $\{(x_i, y_i)\}$, it is useful to form the total squared deviation

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{y_i - g(x_i)}{\sigma_i} \right]^2 , \tag{1.30}$$

where σ_i is the Gaussian error on measurement i: if these errors are the same for all measurements then the factor can, of course, be taken outside the summation.

Each term will clearly contribute an amount of order one to the sum, and it is no surprise that $E[f(\chi^2; n)] = n$. The distribution is given by

$$f(\chi^2; n) = \frac{2^{-\frac{n}{2}}}{\Gamma(\frac{n}{2})} \chi^{n-2} \mathrm{e}^{-\chi^2/2} . \tag{1.31}$$

Some examples for different n are shown in Figure 1.5.

The χ^2 distribution is used a great deal in considering the question of whether a particular set of measurements (with their errors) and a particular model are compatible. This is addressed through the *cumulative* χ^2 *distribution*. For a given value of χ^2, the complement of the cumulative distribution gives the *p*-value, the probability that, given that the model is indeed correct, a measurement would give a result with a χ^2 this large, or larger. If the value of χ^2 obtained is large compared to n then the *p*-value is small, that is the probability that a set of measurements truly described by this model would give such a large disagreement is small, and doubt is cast on the model, or the data (or both). The mean of $f(\chi^2, n)$ is just n, and the standard deviation is $\sqrt{2n}$. For large n the distribution converges to the Gaussian, as it must by the central limit theorem. However, the convergence is actually rather slow, and this approximation is not often used. Instead the *p*-value

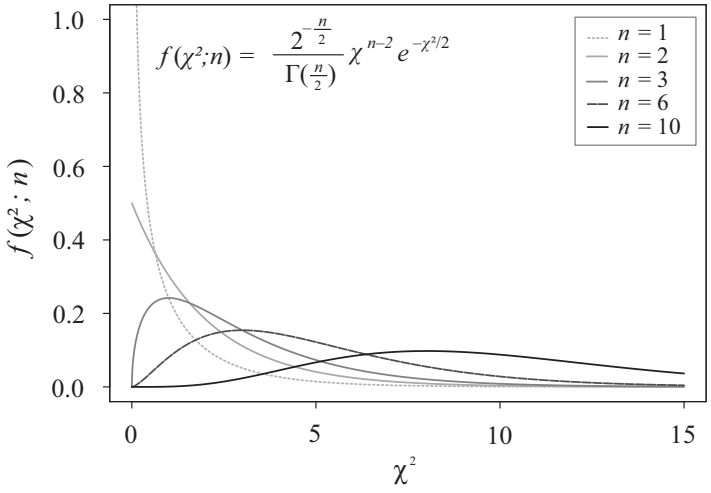

Figure 1.5 χ^2 distributions for $n = 1, 2, 3, 6$ and 10.

should be obtained accurately from functions such as `TMath::Prob` in ROOT or `pchisq` in R.

If the model has free parameters θ which are not given, but were found by fitting the data, then the same χ^2 test can be used, but for n one takes the number of data points minus the number of fitted parameters. This is called the number of degrees of freedom. Strictly speaking this is only true if the model is a linear one (i.e. linear in the parameters). This is often the case, either exactly or to a good approximation, but there are some instances where this condition does not hold, leading to the computation of deceptively small and inaccurate p-values.

Example 1.4 Resistance measurements

A series of ten measurements are made of resistance R as a function of temperature T. The temperature is controlled very accurately, but the resistance is only measured with an accuracy of $2\,\Omega$. A theoretical model predicts a value for R of $(10.3 + 0.047 \cdot T)\,\Omega$. The evaluation of χ^2 gives a value of 25.1. What can you say?

In this case one would use $n = 10$. Evaluating (using the R function `pchisq`) the probability of getting a value as large as 25.1 from $n = 10$ gives 0.5%. It seems very implausible that the model really describes this data. (This does not necessarily mean the model is wrong. It could be that the data are badly measured. Or that the measurement accuracy has been estimated too optimistically.)

You will occasionally obtain χ^2 values that seem very small: $\chi^2 \ll n$. There is no standard procedure for rejecting these, but you should treat them with some suspicion and consider whether the model may have been formulated after the data had been measured ('retrospective prediction'), or whether perhaps the errors have been over-generously estimated.

1.3.4.7 The Log-Normal Distribution

If the logarithm of the variable is given by a Gaussian distribution $f(\ln\, x; \mu, \sigma)$ then the distribution for x itself is the *log-normal distribution*

$$f(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \mathrm{e}^{-\left[\frac{(\ln x - \mu)^2}{2\sigma^2}\right]} . \tag{1.32}$$

Just as the central limit theorem dictates that any variable which is the sum of a large number of random components is described by a Gaussian distribution, any variable which is the product of a large number of random factors, none of which dominates the behaviour, is described by the log-normal. For instance, the signal registered by an electron in a calorimeter may be described by a log-normal distribution, as a certain fraction of the energy may be lost to dead material, a fraction to lost photons, a fraction to neutron production, and so on. The mean is given by $\mathrm{e}^{\mu+\sigma^2/2}$, and the standard deviation is $\mathrm{e}^{\mu+\sigma^2/2}\sqrt{\mathrm{e}^{\sigma^2} - 1}$.

1.3.4.8 The Weibull Distribution

The *Weibull distribution* is:

$$f(x; \alpha, \beta) = \alpha\beta(\alpha x)^{\beta-1}\mathrm{e}^{-(\alpha x)^{\beta}} . \tag{1.33}$$

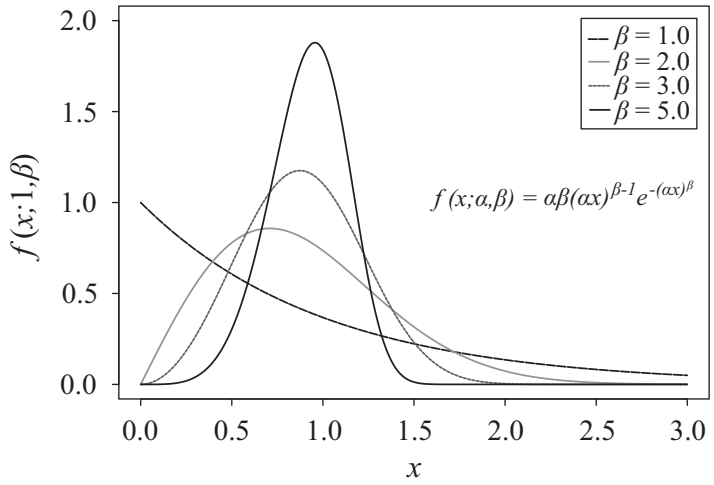

Figure 1.6 Weibull distributions for $\alpha = 1.0$, progressively more peaked for $\beta = 1.0$, 2.0, 3.0 and 5.0.

This gives a shape which rises from zero to a peak and then falls back to zero again. It was originally invented to describe the failure rates in aging light bulbs. There are no failures at small times (because they are new and fresh) or at long times (because they have all failed). It is a rather more realistic modelling of real-life 'lifetime' than the simple exponential decay law for which the failure probability is constant.

The parameter α is just a scale factor and β controls the shape. The case $\beta = 1$ corresponds to the simple exponential decay law, whereas $\beta > 1$ describes the behaviour when the failure probability increases with age, and gives successively sharper peaks. A case where the failure probability falls with time (perhaps because of initial burn-in) is described by $\beta < 1$. Examples are shown in Figure 1.6. The mean is $\frac{1}{\alpha}\Gamma[1+(1/\beta)]$ and the variance is $1/\alpha^2 \left\{\Gamma\left[1+(2/\beta)\right] - \Gamma\left[1+(1/\beta)\right]^2\right\}$. A location parameter x_0 may also be needed in some problems, replacing x by $x - x_0$.

1.4 Probability

We use probability every day, in both our work as physicists and our everyday lives. Sometimes this is a matter of precise calculation, when we buy an insurance policy or decide whether to publish a result, sometimes it is more intuitive, as when we decide to take an umbrella to work in the morning.

But although we are familiar with the concept of probability, on closer inspection it turns out that there are subtleties. When we get into technicalities there turn out to be different definitions of the concept which are not always compatible.

1.4.1
Mathematical Definition of Probability

Let A be an event. Then the probability $P(A)$ is a number obeying three conditions, the *Kolmogorov axioms* [7]:

1. $P(A) \geq 0$;
2. $P(U) = 1$, where U is the set of all A, the sample space;
3. $P(A \cup B) = P(A) + P(B)$ for any A, B which are exclusive, that is $A \cap B = 0$.

From these axioms a whole system of theorems and properties can be derived. However, the theory contains no statement as to what the numbers actually mean. For mathematicians this is, of course, not a problem, but it does not help us to apply the results.

1.4.2
Classical Definition of Probability

The probability of a coin landing heads or tails is clearly 1/2. Symmetry dictates that it cannot be anything else. Likewise the chance of drawing a particular card from a pack has to be 1/52. The original development of probability by Laplace, Pascal and their contemporaries, to aid the gambling fraternity, was founded on this equally likely construction. 'Probability' could be defined by taking fundamental symmetry where all cases were equally likely (say, the six sides on a dice), and extended to more complex cases (say, rolling two dice) by counting combinations.

Unfortunately this definition does not generalise to cases of continuous variables, where there is no fundamental symmetry: if you 'draw a line at random' from a given point, this could be done by taking coordinates of the endpoint from a uniform distribution, or by drawing an angle uniformly taken between 0 and 360°, the results are incompatibly different. This approach thus leads to a dead end.

1.4.3
Frequentist Definition of Probability

Problems with the classical definition led to the alternative definition of probability as the limit of frequency by Venn, von Mises [8] and others. If a selection is made N times under identical circumstances, then the fraction of cases resulting in a particular outcome A tends to a limit, and this limit is what is meant by the probability:

$$P(A) = \lim_{N \to \infty} \frac{N(A)}{N} \,. \tag{1.34}$$

This is the generally adopted definition, taught in most elementary courses and textbooks. It satisfies, of course, the Kolmogorov axioms.

Where the classical definition is valid it leads to the same results. But there is an important philosophical difference. The probability $P(A)$ is not some intrinsic property of A, it also depends on the way the sampling is done: on how the collective or ensemble of total possible outcomes has been constructed.

Thus, to use von Mises' example: the life insurance companies determine that the probability of one of their (male) clients dying between the ages of 40 and 41 is 1.1%. This is a hard and verifiable number, essential for the correct adjustment of the premium paid. However, it is not an intrinsic probability of the person concerned: you cannot say that a particular client has this number attached to them as a property in the same way that their height and weight are. The client belongs not just to this ensemble (insured 40-yr-old males) but to many others: 40-yr-old males, non-smoking 40-yr-old males, non-smoking professional lion tamers – and for each of these ensembles there will be a different number.

So there are cases with several possible ensembles, and the value of $P(A)$ is ambiguous until the ensemble is specified. There are also cases where there is no ensemble, as the event is unique. The Big Bang is an obvious example, but others can be found much nearer home. For example, what is the probability $P(\text{rain})$ that it will rain tomorrow? Now, there is only one tomorrow, and it will either rain or it will not, so $P(\text{rain})$ is either 0 or 1. Von Mises condemns any further discussion as 'unscientific' use of language. This is further discussed (and resolved) in Section 1.5.2.

1.4.4 Bayesian Definition of Probability

Another way of extending the unsatisfactory classical definition of probability was made by de Finetti [9] and others. De Finetti's starting point is the provocative 'Probability does not exist.' It has no objective status: it is something the human mind has constructed.

He shows that one can consistently define a personal probability (or *degree-of-belief*) $P(A)$ in A by establishing the odds of a bet whereby you lose €1 if A subsequently turns out to be false, and you receive €G if it turns out to be true. If $P(A) > 1/(1 + G)$ you will accept the bet; if $P(A) < 1/(1 + G)$ you will decline it.

Such personal probability is indeed something we use every day: when you decide whether or not to take an umbrella to work in the morning your decision is based on your personal probability of there being rain (and also the 'costs' involved in (a) getting wet and (b) having something extra to carry). However, there is no need for my personal probability to be the same as yours, or anyone else's. It is thus often referred to as a *subjective* probability. Subjective probability is also generally known as *Bayesian probability*, because of the great use it makes of Bayes' theorem [10]. This is a simple and fundamental result which is actually valid for any of the probability definitions being used.

1.4.4.1 Bayes' Theorem

Suppose A and B are two events, and introduce the conditional probability $P(A|B)$, the probability of event A given that B is true (for instance: the probability that a card is the six of spades, given that it is black, $P(\text{six of spades}|\text{black})$ is 1/26).

The probability of both A and B occurring, $P(A \cap B)$ is clearly $P(A|B)\,P(B)$. But it is also $P(B|A)\,P(A)$. Equating these two quantities gives

$$P(A|B) = \frac{P(B|A)}{P(B)}\,P(A)\,. \tag{1.35}$$

This is used in problems like the famous 'taxi colour' example.

Example 1.5 Taxi colour

In some city, 15% of taxi cabs are yellow, and 85% are green. A taxi is involved in a hit-and-run accident, and an eyewitness says it was a yellow cab. The police have established that such eyewitness statements get the colour correct in 80% of cases and wrong in 20%. What is the probability that the cab was yellow?

The arithmetic is simple: just plug the numbers into Bayes' theorem. Note that the $P(B)$ term in the denominator can be helpfully written as $P(B|A)\,P(A) + P(B|\bar{A})\,\big[1 - P(A)\big]$, where $\bar{A}$ denotes 'not A'. If the cab's true colour is denoted by Y or G, and the colour the witness says they saw by y or g, then

$$\begin{aligned} P(Y|y) &= \frac{P(y|Y)}{P(y|Y)\,P(Y) + P(y|G)\,P(G)}\,P(Y) \\ &= \frac{0.8}{0.8 \cdot 0.15 + 0.2 \cdot 0.8} \cdot 0.15 = 0.4\,. \end{aligned}$$

So the cab was more likely (60% probability) to have been green – despite the witness saying exactly the opposite.

The 'Bayesian' use of Bayes' theorem uses the same algebra but applies it to cases where B represents some experimental result and A some theory. $P(B|A)$ is the probability of the result occurring if the theory is true, and $P(A)$ is the personal probability you ascribe to the theory being true before the experiment is done – the *prior probability*. $P(A|B)$ is the probability you ascribe to the theory in the light of the experiment – the *posterior probability* (the prior $P(A)$ and the posterior $P(A|B)$ are meaningless in the frequentist definition).

This all works neatly. If a result is forbidden in some theory, $P(B|A) = 0$, then its observation must lead to the theory being discarded. If a result is favoured in some theory, then observation of that result increases our degree-of-belief in that theory, although this increase is tempered by the probability of observing the result in any case.

1.5 Inference and Measurement

Standard probability calculations are all about getting from the theory to the data. They address questions like: under such-and-such conditions, what is the probability that a specified random event will happen?

Inference is the reverse process: getting from the data to the theory. There is a theoretical model, containing some parameter (or parameters) θ, which predicts the probability of getting a certain result (or set of results) x. What does the observation of a particular value of x tell you about θ? A simple example would be a particle of true energy $E_{\text{true}} \equiv \theta$ giving a measured energy of $E_{\text{meas}} \equiv x$ in the calorimeter. A less simple example would be the existence of a Higgs particle with mass m_{H} giving a set of events with particular characteristics in different channels.

1.5.1 Likelihood

When you make a measurement, then $f(x, \theta)$, the probability of obtaining a result x given the value of a model parameter θ, can also be written as the *likelihood* $L(x; \theta)$. This change is purely cosmetic: the actual algebra of the function is the same. Taking the Poisson as an example, and contemplating the observation of five events and a prediction of 3.4, one can write $f(5; 3.4) \equiv L(5; 3.4) \equiv (3.4^5/5!)\mathrm{e}^{-3.4}$.

Given a result x, the value of $L(x; \theta)$ tells you the probability that θ would lead to x, which in turn tells you something about the plausibility that a particular value of θ is the true one. The latter statement is purposefully made vague: it will be considered in proper detail later.

For practical purposes one often uses the logarithm of the likelihood, as, if you had a set of independent results $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, then $\ln L(\boldsymbol{x}; \theta) = \sum_i \ln f(x_i; \theta)$, and sums are easier to handle than products.

The *likelihood principle* states that if you have a result x then the likelihood function $L(x; \theta)$ contains all the information relevant to your measurement of θ. This principle is regarded by some as an irrefutable axiom, and by others as an irrelevance. Bayesian inference generally satisfies this, whereas frequentist inference generally violates it as the frequentist also has to consider the ensemble of experimental results that might have been obtained.

1.5.2 Frequentist Inference

As von Mises points out, the probability of rain tomorrow is either 0 or 1, and no more can be said. However, you can construct an ensemble for something that looks very similar. Suppose that the pressure is falling and the clouds are gathering. A local weather forecast (perhaps made by a professional meteorologist, perhaps by the ache in your grandmother's left elbow) predicts rain. If you consider the track record of this particular prediction and count the number of times it has proved

correct, that gives a probability which is valid in the frequentist sense. So although you cannot say 'It will probably rain tomorrow', you can say 'The statement "It will rain tomorrow." is probably true.'

Indeed, if your weather prophet has been correct nine times out of ten, you can say 'The statement "It will rain tomorrow." has a 90% probability of being true.' Again notice that the number is a property not just of the event (rain) but of the ensemble, in the form of the weather forecaster.

Now apply this approach to the interpretation of a measurement. Suppose your measurement process is known to give a result x which differs from the true value μ with a probability distribution which is Gaussian with some known σ. You quote the result, whether it is the mass of the top quark determined from years of collider data, or a measurement of a resistance on a lab bench, as

$$x \pm \sigma \,. \tag{1.36}$$

This seems to say that μ lies in the range $[x - \sigma, x + \sigma]$ with 68% probability. But it can't. The top mass, m_t, for which we currently quote $173.2 \pm 0.9\,\text{GeV}$, either lies in the range $[172.3,\ 174.1]\,\text{GeV}$ or it does not. It is our measurement which is random, not the true value. So, as a frequentist, you make a statement about statements. 'The statement "$172.3 < m_t\,[\text{GeV}] < 174.1$" has a 68% chance of being true.' Or, to put it another way, you make the statement '$172.3 < m_t\,[\text{GeV}] < 174.1$' with 68% confidence. There is a trade-off between the accuracy of the statement and the confidence you have in it. You could have played safer, and said with 95% confidence '$171.4 < m_t\,[\text{GeV}] < 175.0$'. In other cases one-sided (upper or lower) limits may be appropriate.

1.5.3
Bayesian Inference

The Bayesian has no need of such mental gymnastics:

- $\pi(\theta)$ is the pdf describing my prior belief in the value of θ.
- After a result x, Bayes' theorem then gives my posterior belief $f(\theta|x)$ as $f(x|\theta)/f(x)\pi(\theta)$ where $f(x) = \int f(x|\theta)\pi(\theta)\mathrm{d}\theta$.
- The denominator does not contain θ, so we can write $f(\theta|x) \propto f(x|\theta)\pi(\theta)$.
- If you also decide that $\pi(\theta)$ is uniform, just a constant, then $f(\theta|x) \propto f(x|\theta)$.
- The proportionality constant can be fixed by the normalisation.

In particular, the Bayesian interpretation of a Gaussian measurement, assuming a flat prior, equates the likelihood with the posterior probability: $f(\mu|x) = 1/(\sigma\sqrt{2\pi})\mathrm{e}^{-(x-\mu)^2/2\sigma^2}$. This interpretation of the likelihood $L(x;\mu)$ as a pdf in the parameter μ looks especially plausible in the case of a Gaussian measurement: one has to remember that it is only valid for Bayesians and not for frequentists. Actually, the depiction of Bayesians and frequentists as different and rival schools of thought is not really correct. Yes, some statisticians can be fairly described as one or the other, but most of us adopt the approach most appropriate for a particular

problem. But care must be taken not to use concepts that are inapplicable in the framework chosen.

The uniform prior $\pi(\theta) = \text{const}$ has the problem that, if the range of θ is infinite, then to preserve $\int \pi(\theta)\mathrm{d}\theta = 1$, the constant must vanish. However, one normally just works with priors which do not integrate to unity – so-called *improper priors*[4] – relying on the final normalisation of the posterior.

Information from further measurements can be neatly incorporated into this framework. The posterior distribution from the first experiment is taken as the prior for the second, and its posterior forms the prior for the third (the order in which the combination is done is irrelevant).

1.5.3.1 Use of Different Priors

The simplicity of a uniform prior is misleading. In the first place, it probably does not represent your true personal belief. Consider some hypothetical X particle predicted by some far-beyond-the-Standard-Model theory, and suppose you are convinced that it does exist. If you say $\pi(m_X) = \text{const}$, then that implies that your prior expectation that m_X lies between 1 and 2 GeV is the same as your prior expectation that it lies between 100 001 and 100 002 GeV, which frankly is not credible (given the choice, would you rather work on an experiment that could detect an X between 1 and 2 GeV, or between 100 001 and 100 002 GeV?).

Secondly, uniformity does not survive reparameterisation. If an angle has a uniform pdf in θ, then the distribution in $\cos\theta$ is very non-uniform, and in $\sin\theta$ and $\tan\theta$ it is different again. You cannot claim to have an objective analysis through having a uniform prior, as the choice of which variable to make uniform will affect the result.

Once the data have had a chance to constrain the results, then the effects of the choice of prior are reduced. Suppose we change from a prior uniform in m_X to one uniform in $\ln m_X$ (this corresponds to a prior for m_X proportional to $1/m_X$). If your measurements cover a range of m_X from 100 to 200 GeV, then this is a big change, but if it is only from 171 to 173 GeV then the difference is small – and we gloss over such differences in everyday statistics when we write expressions like $\sigma_{\ln x} = \sigma_x/x$.

So a sound measurement does not depend (much) on the choice of prior. This is called a *robust measurement*. In presenting a Bayesian result you may justify it by any of the following:

- showing that the result is robust: that the arbitrary choice of prior makes no great difference;
- justifying the prior in some way as being correct – or, perhaps, showing that the uniform prior is uniform in the correct variable;
- saying that you have chosen this prior and it represents your personal belief.

But just saying 'We took a uniform prior.' is not doing a proper job.

4) All possible jokes have already been made. Don't go there, OK?

1.5.3.2 Jeffreys Priors

One attempt to systematise the choice of priors was made by Jeffreys [11]. His argument is based on the idea that an impartial prior should be 'uninformative' – it should not prefer any particular value or values.

Still speaking loosely: if the log-likelihood $\ln L(x;\theta)$ has a nice sharp peak, then the data are telling you something about θ, and if it is just a broad spread then it's not being much help. The 'peakiness' of a distribution can be expressed using the second differential (with a helpful minus sign). On, or near, a sharp peak, $-\left(\partial^2 \ln L\right) / \left(\partial\theta^2\right)$ will be large and positive.

Now we take a step back and forget any measurements made, and ask: given some value of θ, what would we expect, on average, from a measurement? This quantity is called the *Fisher information*:

$$I(\theta) = -E\left(\frac{\partial^2 \ln L}{\partial\theta^2}\right) , \tag{1.37}$$

where the expectation value is evaluated by multiplying by $f(x;\theta)$ and integrating over all possible results x for this particular value of θ. A large value of $I(\theta)$ means that if you make a measurement it will (probably) provide useful knowledge about the true value of θ, and a small value of $I(\theta)$ tells you that the measurement will not tell you much and is hardly worth doing. It can easily be shown (see e.g. Eq. (5.8) in [1]) that

$$I(\theta) = E\left[\left(\frac{\partial \ln L}{\partial\theta}\right)^2\right] . \tag{1.38}$$

Jeffreys answers the question 'Should we use θ or $\ln\theta$ or $\sqrt{\theta}$ as our fundamental variable?' by saying that we should choose a form such that no particular value will yield more informative results than any other. He prescribes a parameterisation $\theta'(\theta)$ for which $-(\partial^2 \ln L)/(\partial\theta'^2)$ is constant. This is a variable in which all values are (from the Fisher information viewpoint) equal, and if we make the prior in this variable flat we are clearly being fair and even-handed.

In practice one does not have to find θ' explicitly. If $\pi(\theta')$ is the prior for θ' and is constant, and as $I(\theta')$ is constant by construction, then

$$\begin{aligned}
\pi(\theta) = \pi(\theta')\left|\frac{\partial\theta'}{\partial\theta}\right| \propto & \sqrt{E\left[\left(\frac{\partial \ln L}{\partial\theta'}\right)^2\right]}\left|\frac{\partial\theta'}{\partial\theta}\right| \\
= & \sqrt{E\left[\left(\frac{\partial \ln L}{\partial\theta'}\frac{\partial\theta'}{\partial\theta}\right)^2\right]} \\
= & \sqrt{E\left[\left(\frac{\partial \ln L}{\partial\theta}\right)^2\right]} \\
= & \sqrt{I(\theta)} .
\end{aligned} \tag{1.39}$$

That means that for any θ, one should take the prior as $\pi(\theta) = \sqrt{I(\theta)}$. For a location parameter, the *Jeffreys prior* is indeed just uniform in $[-\infty, \infty]$ but for a scale parameter the prior is proportional to $1/\theta$ – equivalently, the prior is uniform if you use $\ln\theta$ as the fundamental form. For a Poisson mean it is $1/\sqrt{\theta}$ where $\theta = \nu$.

Jeffreys' method rests on the idea that the prior should not prejudge the result: that it should be as 'uninformative' as possible. But it also provides a structure for giving a unique answer. Whether you choose the fundamental parameter to be ν or $\sqrt{\nu}$ does not change your final quoted result, thanks to the different priors you would have to use. This is why such priors are often termed 'objective', in that the dependence on your personal choice is removed.

Extension to more than one parameter is difficult but not impossible through a technique called *reference priors* [12].

Although the Jeffreys prior offers a way to getting unambiguous results, it has not been universally taken up. Partly because some are too lazy to consider anything other than a uniform prior in their favourite variable. Partly because of the difficulty of applying it to more than one parameter. Partly because it violates the likelihood principle. Partly because the prior does depend on the likelihood function and thus on the experimental technique, so you would invoke a different prior for (say) the Higgs mass, as determined with ATLAS through $H \to \gamma\gamma$ than you would for the Higgs mass, as determined by CMS through $H \to W^+W^-$.

1.5.3.3 The Correct Prior?

So what prior, or what collection of priors, should you use in a Bayesian analysis, if you are forced to do so? The answer, clearly, is: whatever happens to be your personal belief. But although a prior is subjective, it should not be arbitrary. Other data, measurements of this quantity or similar ones, can be used for guidance. Asking (theorist) colleagues can be useful, but if you do that be sure to ask for a wide range.

The 'quest for the correct prior' is an issue for physicists, who are conditioned to expect problems for which there is a unique correct answer, rather than statisticians, who know better. There is no unique correct prior for a problem. There is a range of sensible priors, and you should use these to check the robustness of your result. If it is stable, then the choice does not matter. If it is unstable, then the measurements cannot tell you anything honestly useful.

1.6 Exercises

Exercise 1.1 Uniform distributions

Show, by integration, that the standard deviation of a uniform distribution of width w is $w/\sqrt{12}$.

Exercise 1.2 Poisson distributions 1

Show that the characteristic function of the Poisson distribution is $\phi(u) = e^{\lambda(e^{iu}-1)}$. Using the fact that the characteristic function of a convolution is the product of the individual characteristic functions, show that the convolution of two Poisson distributions of means λ_1 and λ_2 is also a Poisson, of mean $\lambda_1 + \lambda_2$. Now prove the result without using characteristic functions.

Exercise 1.3 Poisson distributions 2

A Poisson distribution has a mean of 3.7. Calculate 'by hand' the probability that it will give two events or less. Then calculate the same result using `ppois` in R or `poisson_cdf` in ROOT, or the equivalent in your favourite math program.

Exercise 1.4 Bayes' theorem

If some object X exists, it may be found with equal probability in one of N locations. Show, using Bayes' theorem, that if P is your prior belief in the existence of X, then after a location is unsuccessfully investigated your belief changes to $P' = (N-1)/(N-P)P$. If $N = 10$ and there are nine unsuccessful searches, calculate and contrast the final posteriors for a prior of 0.9 and of 0.99.

Exercise 1.5 *p*-values

Find the number of standard deviations corresponding to p-values of 10%, 5% and 1% for a Gaussian distribution. Consider both one-sided and two-sided p-values.

Exercise 1.6 Jeffreys prior

Prove the result given for the Jeffreys prior of a Poisson distribution of mean ν. Do this by writing down the log-likelihood, differentiating twice and negating, taking the expectation value and then taking the square root.

References

1 Barlow, R.J. (1989) *Statistics: a Guide to the Use of Statistical Methods in the Physical Sciences*, John Wiley & Sons.

2 R Development Core Team (2011) *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing.

3 Antcheva, I. *et al.* (2009) ROOT – A C++ framework for petabyte data storage, statistical analysis and visualization, *Comput. Phys. Commun.*, **180**, 2499.

4 von Bortkewitsch, L. (1898) Das Gesetz der kleinen Zahlen. *Monatsh. Math.*, **9**, 39.

5 Landau, L.D. (1944) On the energy loss of fast particles by ionization. *J. Phys. (USSR)*, **8**, 201.

6 Kolbig, K.S. and Schorr, B. (1984) A program package for the Landau distribution. *Comp. Phys. Commun.*, **31**, 97. Erratum: (2008) *Comp. Phys. Commun.*, **178**, 972.

7 Kolmogorov, A.N. (1950) *Foundations of the Theory of Probability*, Chelsea Publishing Company.

8 von Mises, R. (1957) *Probability, Statistics and Truth*, Dover Publications.

9 de Finetti, B. (1974) *Theory of Probability*, John Wiley & Sons.

10 Bayes, T. (1763) An essay towards solving a problem in the doctrine of chances. *Philos. Trans. R. Soc.*, **53**, 370.

11 Jeffreys, H. (1966) *Theory of Probability*, Oxford University Press.

12 Berger, J.O., Bernardo, J.M., and Sun, D. (2009) The formal definition of reference priors. *Ann. Stat.*, **37**, 905.