Chapter 2

Mathematical Foundations

In this chapter we will outline the basic mathematical foundations of probability and statistical theory. So far we have discussed the actual data but now we are going to start to understand the theory behind the distributions that can be used to describe these data. One of the key parts of data science involves *statistical inference* which is to *infer* the properties, or parameters, of the underlying distribution based on the data we observe.

2.1 The two philosophies of statistics

A rather awkward starting point is that statisticians do not actually agree on the most basic underlying principles of statistical theory, namely the definition of probability. There are two predominant schools of thought, known as *Bayesian* and *frequentist* (or *classical*). There is certainly historical tension between these two schools, with individuals proudly labelling themselves as belonging to one school or another. I have never really understood this. In my view, both are equally valid and they should converge to the same inference in the *asymptotic nirvana* (*i.e.* as the sample size tends to infinity). The cause of the argument is that they are founded on different principles, but for applications this is perhaps rather unimportant. There is also a history of mud-slinging between the two camps. A rather unfortunate example of this can be seen between the `CKMfitter` (frequentist) and `UTfit` (Bayesian) groups, which I will not cover the details of here but you can follow in Refs. [12, 13, 14].

Different fields of science tend to favour one approach over another and it is reasonable to chose the statistical approach depending on the problem. Astrophysicists are often Bayesians, they only have one universe so perhaps the idea of averaging over many universes is not suitable. Particle physicists are often frequentists, facilties like the LHC produce millions of events so computing the frequency of different outcomes seems reasonable.

2.2 Probability foundations and Random Variables

Let's now formally write down some definitions of probability and we will see what the differences are between Bayesians and frequentists.

2.2.1 Definitions of probability

Mathematical probability

It was Kolmogorov [15] who developed the first mathematical theory of probability. The *Kolmogorov axioms* state that for a set Ω of all possible *exclusive*[1] outcomes, X_i, the probability of a given occurence, $P(X_i)$ must satisfy:

$$\begin{aligned} &\text{(i)} \quad P(X_i) \geq 0 \;\; \text{for all } i && \text{probability must be zero or positive} \\ &\text{(ii)} \quad P(X_i \text{ or } X_j) = P(X_i) + P(X_j) && \text{exclusive probabilities sum} \\ &\text{(iii)} \quad \sum_{\Omega} P(X_i) = 1 && \text{all probabilities must sum to unity} \end{aligned} \tag{2.1}$$

These properties kind of seem obvious but infact all other properties of probabilities which are more complex can be derived from just these.

Frequentist probability

The *frequentist* definition of probability is perhaps a more scientific approach, where the probability is given by the relative frequency of a particular outcome. Imagine an experiment with a total number of trials, N, in which the outcome X is observed n times. The frequentist definition of probability that any given event will be of type X is then the limit of the frequency ratio as the number of trials tends to inifinty:

$$P(X) = \lim_{N \to \infty} \frac{n}{N}. \tag{2.2}$$

There are few things to note on the frequentist definition:

- The *true* probability in practise is never obtainable (because we cannot perform infinite experiments). We can only estimate the probability given the sample size we have. However, if it always possible to perform more experiments then we can keep doing so until we reach the desired accuracy (and any accuracy is in principle achievable).

- Frequentist probability can only be applied to repeatable experiments. For example I cannot use it to determine the probability it will rain the day after tomorrow. I need a system in which I can keep relevent conditions stable to perform repeatable experiments.

- One does not need to have any *prior* beliefs about outcomes, the probability is purely determined from observations.

Bayesian probability

The *Bayesian* approach is considered a bit more *intuitive* and more akin to everyday reasoning, where probability is interpreted as a degree of belief that a particular outcome will occur or a particular parameter will take a given value. To define a probability that can be applied to non-repeatable experiments we cannot use the concept of frequency so we have

[1] exclusive means that the occurence of one prohibits the occurence of any other, *e.g.* the day of the week is Monday.

to replace it with our prior degree of belief,

$$P(X) = \text{degree of belief that } X \text{ happens.} \tag{2.3}$$

We can base this on de Finetti's *coherent bet* [16]. The idea being that to determine someones prior belief on outcome X we can ask how much they would be willing to bet on its occurence. In that case,

$$P(X) = \frac{\text{largest amount they would be willing to bet}}{\text{amount they stand to gain if they win}}. \tag{2.4}$$

Clearly any prior probability we have will have to obey Kolmogorov axioms in (2.1).

Here are a few thoughts on Bayesian probability:

- It is a property of the observer and is therefore subjective, the amount *you* are willing to bet may be different to the amount *I* am willing to bet. The subjectivity of the Bayesian prior is one of the criticisms levelled at this approach by frequentists.
- It will depend on how much the observer *knows*, and will change if they know more. The ability to incorporate enhanced levels of knowledge into the Bayesian prior is one of the advantages of the Bayesian approach.

2.2.2 Properties of probability

Based on the *Kolmogorov axioms* we can now derive various other properties of probability that will apply to any probability that obeys the axioms.

Addition

For two non-exclusive sets A and B, containing elementary events X_i, the probability that an event is *either* in A *or* B is

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B). \tag{2.5}$$

Conditional probability

Now we can introduce an absolutely key concept, that of *conditional probability*, the probability of A given B, written as $P(A|B)$. This is the probability that an event *known* to be in set B is also in set A. It is defined via

$$P(A \text{ and } B) = P(A|B)P(B) = P(B|A)P(A). \tag{2.6}$$

Notice there is a very key difference between $P(A|B)$ and $P(B|A)$. They are not the same thing and often get confused. For example the probability you are pregnant *given* you are biologically female is about one in fifty. The probability you are biologically female given you are pregnant is one. This brings us to another important concept, that of *independence*. The sets A and B are independent if

$$P(A|B) = P(A), \tag{2.7}$$

which is to say that the occurence of B has absolutely no impact on the outcome A. Thus, using (2.6), if A and B are independent then,

$$P(A \text{ and } B) = P(A) \cdot P(B). \tag{2.8}$$

This definition of independence is important and useful. It means we can factorise the probability of A and B occuring into the product of the two independent probabilities. It is worth remembering that independence implies zero correlation but the converse is not true. Events can have zero linear correlation but still not be independent.

Bayes' theorem for discrete outcomes

There is an important theorem which links $P(A|B)$ to $P(B|A)$ and we have infact already seen it in (2.6). Bayes' theorem states that

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{2.9}$$

We can see a simple example by computing the probability it's a Monday given it's a weekday,

$$P(\text{Monday}|\text{weekday}) = \frac{P(\text{weekday}|\text{Monday})P(\text{Monday})}{P(\text{weekday})} = \frac{1 \cdot \frac{1}{7}}{\frac{5}{7}} = \frac{1}{5}. \tag{2.10}$$

We will come back to Bayes' theorem for continuous probability distributions later and we will discuss some of the debate over its use versus the classical approach.

The law of total probability

If we suppose there are many exclusive sets $A_i, \ldots, A_N$ which are also disjoint from B then we can write

$$\begin{aligned} P(B) &= P((B \text{ and } A_1) \text{ or } (B \text{ and } A_2) \text{ or } \ldots \text{ or } (B \text{ and } A_N)) && (2.11)\\ &= \sum_i P(B \text{ and } A_i) && (2.12)\\ &= \sum_i P(B|A_i)P(A_i). && (2.13) \end{aligned}$$

This is known as the *law of total probability* and can often be useful if we do not know or cannot calculate $P(B)$ easily. Substituting (2.13) into Bayes'theorem, (2.9), gives

$$P(A|B) = \frac{P(B|A)P(A)}{\sum_i P(B|A_i)P(A_i)}. \tag{2.14}$$

2.2.3 Random variables

A random event is an event with more than one possible outcome. We can associate a probability to each outcome. For some random event we cannot possibly know the outcome but only the probability of an outcome. We can associate the possible different outcomes to a *random variable*, X. For example, if I have a six-sided die, the random event is me rolling the die. The random outcomes can be associated to a random variable, X, which can take six different values, $X_1 = 1, X_2 = 2, ..., X_6 = 6$. The corresponding probabilities for each outcome, $P(X_1), P(X_2), ..., P(X_6)$ form a *probability distribution*, $P(X)$. In this case all the probabilities are 1/6 so the probability distribution is uniform and simply given by $P(X) = 1/6$. Suppose I now took my die and taped over one of the dots on the six face so it now looks like a five (and I have two sides with a five) and I also add a second dot to the face with a one (so I now have two sides with a two), my probability distribution is now different. It can be described by the probability distribution

$$P(X) = \begin{cases} 0 & \text{if } X = 1 \text{ or } X = 6 \\ \frac{1}{3} & \text{if } X = 2 \text{ or } X = 5 \\ \frac{1}{6} & \text{if } X = 3 \text{ or } X = 4 \end{cases} \tag{2.15}$$

which is plotted below in Fig. 2.1.

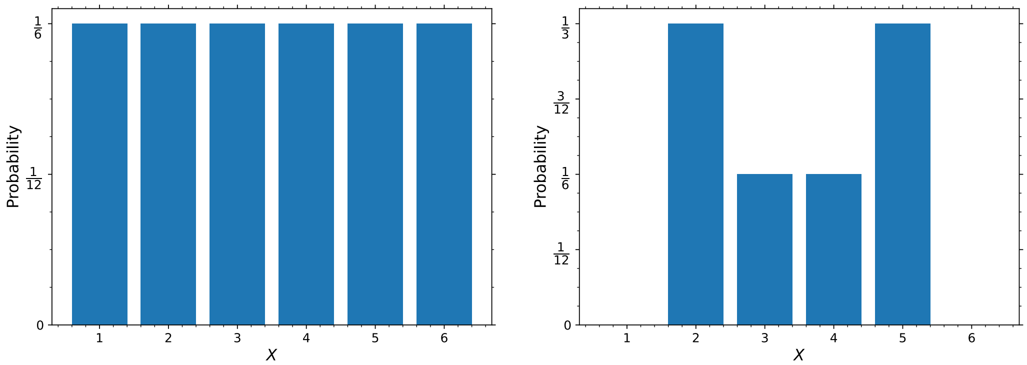

Figure 2.1: Probability distributions for a fair (left) and fixed (right) dice.

So far we have limited our discussion to discrete outcomes, such as is event X_i in set A. In the case of discrete outcomes probability distributions are called *probability mass functions* (p.m.f.s)[2] and they have the property that the sum of probabilities over all possible values must be unity (obeying the Kolmogorov axioms)

$$\sum_i P(X_i) = 1. \tag{2.16}$$

Discrete outcome probabilities are applicable to problems like coin flips and dice rolls, and questions like "is it going to rain today" or "what is the probability I am going to have a beer tonight given I had seven last night"? We will now turn our attention to continuous outcomes. These are a bit more exciting for scientists as we are often interested in measuring parameters. We don't just want to know the probability of a particular outcome but we want to know the *distribution of probabilities* as a function of some variable.

2.2.4 Probability density functions

For a continuous outcome our random variable, X, can take any real value. The *probability distribution*, $p(X)$, then gives the probability as a function of the variable. For a comparison to the discrete case, we can think about this as the probability for outcome X within some

[2] Perhaps a fun discussion topic would be *why* they are called this?

small interval ΔX, and see that the continuous outcome is obtained by taking the limit as the interval vanishes,

$$p(X) = \lim_{\Delta X \to 0} \frac{P(X - \Delta X/2 < X < X + \Delta X/2)}{\Delta X}. \tag{2.17}$$

From this we can see that the dimensions of $p(X)$ are inverse units of X and hence the probability distribution is called the *probability density function* (p.d.f.). Probability density functions are always normalised, to ensure the third Kolmogorov axiom that all probabilties sum to unity:

$$\int_\Omega p(X) dX = 1. \tag{2.18}$$

Similarly the probability that an observation of X lies within the range $a < X < b$ is given by

$$p(a < X < b) = \int_a^b p(X) dX. \tag{2.19}$$

We will discuss some of the most important and commonly used probability distributions later in Sec. 2.4.

Throughout these notes, I will normally use capital letters for random variables, *e.g.* X and Y, and I will often use a capital $P(X)$ for a discrete probability distribution (a *probability mass function*) and a lower case $p(X)$ for a continuous probability distribution (a *probability density function*).

2.2.5 Change of variables

Functions of random variables are themselves also random variables. It is useful to know how to change variables or understand the probability distribution of a change of variables. For example if the probability distribution of the random variable θ is uniformly distributed then it's clear the probability distribution of $f(\theta) = \cos(\theta)$ will not be.

If $f(X)$ is the probability distribution for a random variable X, we want to know the probability distribution $g(Y)$ where Y is related to X by some known transformation function, $Y = h(X)$, see Fig. 2.2. We know that the probability for X to occur between X and $X + dX$ must be equal to the probability that Y occurs between Y and $Y + dY$, so

$$g(Y) dY = f(X) dX. \tag{2.20}$$

If the function h is monotonic and invertible then we can write, $X = h^{-1}(Y)$, which inserting into (2.20), gives

$$g(Y) dY = f(h^{-1}(Y)) \frac{dX}{dY} dY. \tag{2.21}$$

If $Y = h(X)$ is monotonically decreasing then $dX/dY < 0$ and $f(h^{-1}(x)) > 0$ and thus the absolute value of the derivative is required to keep $g(Y) \geq 0$, which it has to be because it's a p.d.f. This transformation formula is often written

$$g(Y) = \frac{f(X)}{|h'(X)|} = f(h^{-1}(Y)) \left| \frac{dX}{dY} \right| = f(X) \left| \frac{dX}{dY} \right|, \tag{2.22}$$

where $|h'(X)|$ is the Jacobian of the transformation, *i.e.* the modulus of the transformation differential. This has important consequences for how we propagate errors as we will see

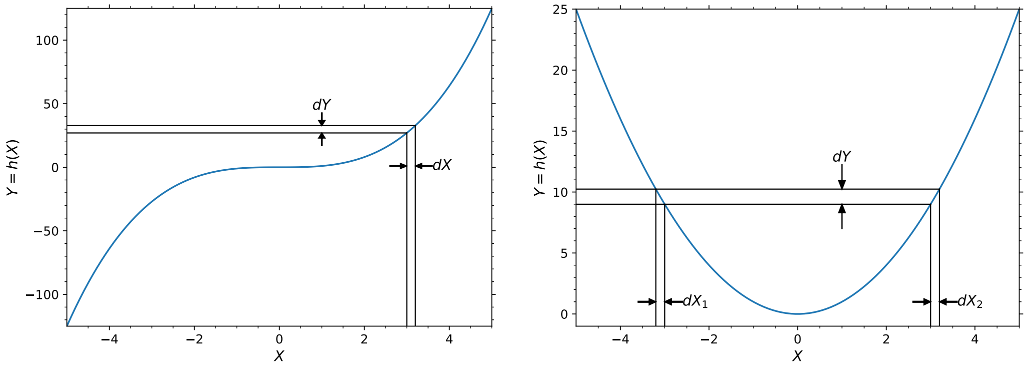

Figure 2.2: Visual demonstration of a change of variable

shortly.

If $h(X)$ is not a one-to-one mapping but instead the inverse has multiple possible solutions then the transformation must be summed over all disjoint regions,

$$g(Y) = \sum_i \frac{f_i(X)}{|h'(X)|}. \tag{2.23}$$

We can extend this to the case of two independent variables X and Y, distributed accordinging to the joint density $f(X,Y)$, which are transformed by functions, $U = U(X,Y)$ and $V = V(X,Y)$. The joint densitiy for U and V is then given by,

$$G(U,V) = F(X,Y)\left|J\left(\frac{X,Y}{U,V}\right)\right| = F(X,Y)\left|\begin{array}{cc} \frac{\partial X}{\partial U} & \frac{\partial Y}{\partial U} \\ \frac{\partial X}{\partial V} & \frac{\partial Y}{\partial V} \end{array}\right|. \tag{2.24}$$

Extending this to any number of variables, $\boldsymbol{X} = (X_1, X_2, X_3, ..., X_N)$, with transformation functions $\boldsymbol{Y} = (Y_1(\boldsymbol{X}), Y_2(\boldsymbol{X}), ..., Y_M(\boldsymbol{X}))$, then the probability density of the transformed variables is

$$g(\boldsymbol{Y}) = \left|J\left(\frac{\boldsymbol{X}}{\boldsymbol{Y}}\right)\right| f(\boldsymbol{X}). \tag{2.25}$$

2.2.6 The cumalative distribution

The integrated *probability density function* (p.d.f.) is known as the *cumalative distribution function* (c.d.f.)[3] which is defined as

$$F(X) = \int_{-\infty}^{X} f(X')dX'. \tag{2.26}$$

It is defined so that

$$F(X_{\text{min}}) = 0, \tag{2.27}$$

$$F(X_{\text{max}}) = 1, \tag{2.28}$$

(2.29)

[3] statisticians often refer to the c.d.f. as *the* distribution

with some very useful properties

$$P(X < X') = \int_{X_{\min}}^{X'} f(X)dX = F(X'), \tag{2.30}$$

$$P(X' < X < X'') = \int_{X'}^{X''} f(X)dX = F(X'') - F(X). \tag{2.31}$$

The snippet below shows a plot of the p.d.f. and c.d.f. of a unit Gaussian.

```python
import numpy as np
from scipy.stats import norm
import matplotlib.pyplot as plt

xlim = (-4,4)
x = np.linspace(*xlim,100)
fig, ax = plt.subplots(1,2,figsize=(12,4))
ax[0].plot(x, norm.pdf(x))
ax[1].plot(x, norm.cdf(x))
ax[0].set_xlabel('$X$')
ax[1].set_xlabel('$X$')
ax[0].set_ylabel('$f(X)$')
ax[1].set_ylabel('$F(X)$')
fig.tight_layout()
```

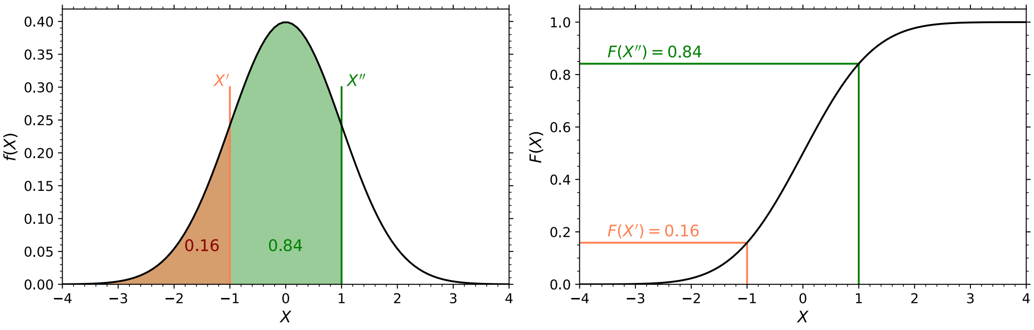

Figure 2.3: The p.d.f. (left) and c.d.f. (right) of a normal distribution.

2.2.7 The joint, marginal and conditional distributions

Let's now consider a so-called *joint* p.d.f., which is a probability distribution of two random variables, $f(X, Y)$. Recall our definition of independence above, (2.8), which we can write for continuous probabilities by requiring that the p.d.f. factorises into *independent* p.d.f.s for X and Y:

$$f(X, Y) = g(X)h(Y). \tag{2.32}$$

The functions, $g(X)$ and $h(Y)$ are the projections of the distribution over either X or Y, obtained when integrating over Y or X. These are known as the *marginal* distributions in

X and Y respectively:

$$g(X) = \int f(X,Y)dY \quad \text{the marginal distribution in } X, \tag{2.33}$$

$$h(Y) = \int f(X,Y)dX \quad \text{the marginal distribution in } Y. \tag{2.34}$$

We can then define the conditional probability, which is a slice through the *joint* density:

$$\boxed{g(X|Y) = \frac{f(X,Y)}{h(Y)}} = \frac{f(X,Y)}{\int f(X,Y)dX} \quad \text{the probability of } X \text{ given } Y, \tag{2.35}$$

$$\boxed{h(Y|X) = \frac{f(X,Y)}{g(X)}} = \frac{f(X,Y)}{\int f(X,Y)dY} \quad \text{the probability of } Y \text{ given } X. \tag{2.36}$$

A visualisation of joint and marginal densities is provided in Fig. 2.4.

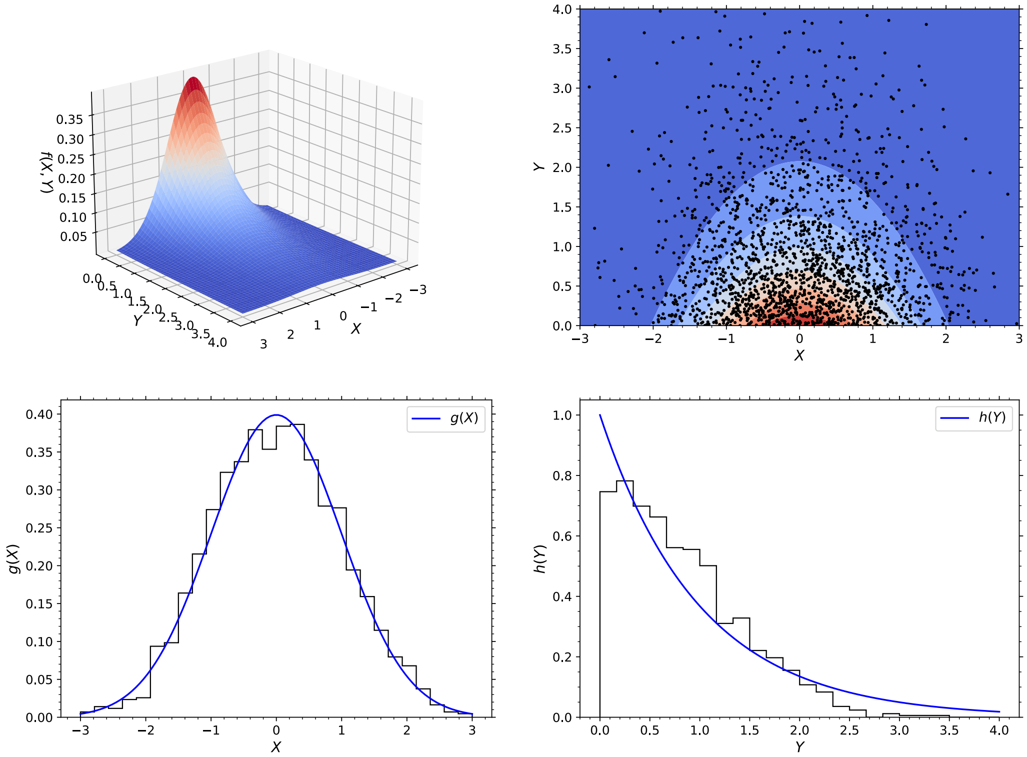

Figure 2.4: Top row: two different visualtions of some 2D joint p.d.f., $f(X,Y)$. The top right plot shows some scatter points for a dataset generated from this distribution. Bottom row: the marginal distributions $g(X)$ (left) and $h(Y)$ (right) along with histograms of the scatter data.

2.2.8 Bayes' theorem for continuous variables

Now that we have defined the conditional probability in (2.35) and 2.36 we can derive Bayes' theorem for a continuous random variable,

$$f(X,Y) = g(X|Y)h(Y) = h(Y|X)g(X), \tag{2.37}$$

$$\Rightarrow \boxed{g(X|Y) = \frac{h(Y|X)g(X)}{h(Y)}.} \tag{2.38}$$

This is one of the most important theorems in all of statistics. You will encounter it over and over again from both a machine learning context and a pure data analysis context. We will now take a few moments to discuss a few of the important consequences of Bayes' theorem but for most of the rest of the course we will focus on *classical* or *frequentist* statistics. More on Bayes' theorem and Bayesian inference will be presented in the Advanced Statistcs course.

In (2.9) we introduced Bayes' theorem for discrete outcomes, X_i, and discrete hypotheses, θ_i. Let's now suppose that θ is a continuous range of hypotheses and can represent a physical parameter which can take a range of different values (*e.g.* a particle's mass or lifetime). We can then consider N measurements of a continuous random variable, X, which is described by the p.d.f. $f(X|\theta)$, which will have a joint density

$$p(X|\theta) = \prod_{i=1}^{N} f(X_i|\theta). \tag{2.39}$$

This is an incredibly important quantity, known as the *likelihood function*. Sometimes it is written as $L(\theta)$. It tells us the likelihood that we observe the data or measurements, X_i, given a particular value of the parameter(s), θ. You ***should not*** confuse this with a p.d.f.. It is not a p.d.f.. It is a function of the parameters or hypotheses θ and not of the random variable X. In other words given a specific dataset, or set of measured values, it only varies with the value of θ. We will encounter the likelihood function so often that by the end of this lecture course it should feel like an old friend.

The big question that Bayes' theorem helps us to answer is, given N observations of X_i from the distribution $f(X|\theta)$, what can I say about the value of θ? In frequentist or classical statistics, the parameter θ has a true fixed value (which may be unknown). Later in the course (see Sec. 3) we will see how we can make estimates of that parameter using frequentist treatments. For a Bayesian, there is no such thing as a fixed parameter. All parameters are random variables and that includes θ. So what we can do is deploy Bayes' theorem to infer the probability distribution of θ given our prior beliefs and the data we have observed. Mathematically this looks like,

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)} = \frac{p(X|\theta)p(\theta)}{\int p(X|\theta)p(\theta)d\theta}. \tag{2.40}$$

The terms in this equation are:

- $p(\theta|X)$ - the *posterior distribution* - our probability distribution for the parameter θ given the data we have observed X.

- $p(X|\theta)$ - the *likelihood function* - the likelihood we observe the data X given a particular value of θ. This is of vital importance across all of statistics and machine learning. We will discuss the likelihood in much more detail in Sec. 3.2.

- $p(\theta)$ - the *prior distribution* - encompassing our prior beliefs about θ, this can be based on previous measurements, previous beliefs or indeed be flat (although note that a change of variables or basis will not necessarily maintain flatness). The prior influences the outcome of Bayesian inference, which can be seen as an advatange and a disadvantage. I will leave discussion of priors to the other stats course.

- $p(X)$ - the *evidence* - this is just a normalisation factor that ensures the *posterior* is a p.d.f.. For Bayesian inference the evidence can often be ignored as the posterior is proportional to the numerator of (2.40). However, the evidence can be quite a useful quantity for goodness of fit tests. I will leave discussion of the evidence to the other stats course.

Bayes' theorem is incredibly powerful. However, this course is on classical statistics and so we will now curtail our discussion of Bayesian methods. You will return to these in the other stats module. Please note for now that Bayes' theorem is still valid for multi-dimensional data observations, *e.g.* the height and weight of a population, and for any number of parameters, *e.g.* the mean and width of the distribution of heights and weights, so we could perhaps more accurately write them explictly as vectors:

$$p(\vec{\theta}|\boldsymbol{X}) = \frac{p(\boldsymbol{X}|\vec{\theta})p(\vec{\theta})}{p(\boldsymbol{X})} \tag{2.41}$$

2.3 Properties of Distributions

When discussing properties of datasets in Sec. 1 we saw how to compute sample estimates of the mean and variance. We will now extend this to apply to distributions using the concept of the expectation value.

2.3.1 Expectation, mean and variance

If we have a random variable X with probability density distributed as $f(X)$ (elsewhere in the notes I may use the notation, $X \sim f(X)$ to demonstrate that X is distributed as $f(X)$) and we have some function operating on that random variable $g(X)$, then the *expectation* of $g(X)$ is given by

$$E[g(X)] = \int g(X)f(X)dX, \tag{2.42}$$

where the integral is over the entire space of X. The expectation takes a value (*i.e.* it is a number) because it's integrated over X. It is not a function of X. Sometimes the notation for the expectation value is to put it inside angled brackets, $E[g(X)] =< g(X) >$. The expectation is a linear operator so that

$$E[ag(X) + bh(X)] = aE[g(X)] + bE[h(X)]. \tag{2.43}$$

For a discrete probability distribution the integral is instead a sum over all probabilities:

$$E[g(X)] = \sum_X g(X)P(X). \tag{2.44}$$

The expectation value of X itself is the mean and can be found by inserting $g(X) = X$ into (2.42). The mean density of a probability distribution is often labelled with a μ symbol and we will also adopt that convention, so the mean of a p.d.f. is defined as

$$\mu = \int Xf(X)dX. \tag{2.45}$$

By the same logic deployed in Sec. 1.2.2 we can quantifty the spread of the distribution using the expectation value of $(X-\mu)^2$ which is defined as the *variance* of the distribution. The *standard deviation* is the square root of the variance and conventionally labelled with a σ, a convention we also adopt, so that,

$$\begin{aligned} V(x) = \sigma^2 &= E[(X-\mu)^2] && (2.46)\\ &= E[X^2 - 2\mu X + \mu^2] && (2.47)\\ &= E[X^2] - \mu^2] && (2.48)\\ = \sigma^2 &= \int (X-\mu)^2 f(X)dX. && (2.49) \end{aligned}$$

We can also define the moments of the distribution using expectation values. The moments are defined as

$$\begin{aligned} \mu_\ell &= E[X^\ell] && \text{the } \ell^{\text{th}} \text{ algebraic moment} && (2.50)\\ \alpha_\ell &= E[(X-E[X])^\ell] && \text{the } \ell^{\text{th}} \text{ moment about the mean.} && (2.51) \end{aligned}$$

In our notation the mean is $\mu = \mu_1$ and the variance is $\sigma^2 = \alpha_2$. The skew is $\gamma_1 = \sqrt{\beta_1} = \mu_3/\mu_2^{3/2}$. The kurtosis is $\gamma_2 = \beta_2 - 3 = \mu_4/\mu_2^2 - 3$.

2.3.2 Covariance and correlation

In analogy with (1.23) we can also define the covariance between two random variables X and Y,

$$V(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = E[XY] - E[X]E[Y], \tag{2.52}$$

and then the correlation

$$\rho(X,Y) = \frac{V(X,Y)}{\sigma_X\sigma_Y}. \tag{2.53}$$

2.3.3 The characteristic function

For a random variable $X \sim f(X)$ with cumalative distribution $F(X)$ we defined the *characteristic function* as the Fourier transformation, such that,

$$\varphi(t) = E[e^{itX}] = \int_{-\infty}^{\infty} e^{itX} f(X) dX. \tag{2.54}$$

By performing the inverse Fourier transform of $\varphi(t)$ we see that the characteristic function can completely define the probability distribution,

$$f(X) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \varphi(t) e^{-iXt} dt. \tag{2.55}$$

This is a useful property that allows us to switch between the *probability distribution* and the *characteristic function.* If we now investigate the algebraic moments,

$$\mu_n = E[X^n] = \int_{-\infty}^{\infty} X^n f(X) dX, \tag{2.56}$$

we find that the n^{th} moment about the origin, μ_n, can be obtained by differentiating the characteristic function n times at the point $t = 0$:

$$\varphi_n(t) = \frac{d^n \varphi(t)}{dt^n} = i^n \int_{-\infty}^{\infty} x^n e^{itX} f(X) dX, \tag{2.57}$$

so that

$$\varphi_n(0) = i^n \mu_n. \tag{2.58}$$

If we also perform a simple coordinate transformation to the mean $Y = X - \mu$ and follow the same process we see that the n^{th} moment about the mean, α_n, can also be generated from the n^{th} derivative of the transformed characteristic function. We can find a relation between the characteristic function and the moments of a distribution by performing the power series expansion of the exponential

$$\begin{aligned}
\varphi(t) = E[e^{itX}] &= E\left[\sum_{n=0}^{\infty} \frac{(itX)^n}{n!}\right] && (2.59)\\
&= \sum_{n=0}^{\infty} \frac{(it)^n}{n!} E[X^r] && (2.60)\\
&= \sum_{n=0}^{\infty} \frac{(it)^n}{n!} \mu_n. && (2.61)
\end{aligned}$$

Thus we see that the characteristic function can be expanded as the sum of moments, μ_n, which appear as coefficients of $(it)^n/n!$ in the expansion. This will be highly useful when we come to proving the central limit theorem later in the lectures, Sec. 2.5.2.

The *characteristic function* is something we would rarely come across as applied scientists. It's really only something for the purists. However you can see here how incredibly useful it can be (and it has various other nice properties I will not cover here). The reason I introduce is that it's important for proving the Central Limit Theorem which I consider to be such an important concept underpinning the application of statistics that it would be negligent to not discuss it in detail. For that we will wait until Sec. 2.5.2. For now we are going to study some of the most commonly encountered probability distributions in nature.

2.4 Common Distributions

In this section we will discuss some of the most common probability distributions.

2.4.1 Binomial Distribution

The *binomial* distribution originates from an experiment with a fixed number of trials and only two outcomes, with probabilities p and $q = 1 - p$. Examples include tossing coins, quality control (pass or fail), medical treatments (cure or not) and computing efficiencies (sensor does register a hit or not). Imagine we have an experiment with n trials and the probability of success is p. The probability of failure is then $q = 1 - p$ and the total probability that the first k trials are successes is

$$P = p^k (1-p)^{n-k}. \tag{2.62}$$

The number of combinations, or ways, we can get k successes in n trials is given by

$$C(n,k) = \left(\begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{k!(n-k)!}. \tag{2.63}$$

Therefore the *binomial probability distribution* is described by

$$\boxed{P(k;p,n) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}}. \tag{2.64}$$

I leave it as an exercise for the problem sheets to show that the sum of all outcomes is unity.

The mean number of successes can be found using (2.44) (*proof is in the lectures*) and is given by

$$\mu = E[k] = np. \tag{2.65}$$

The variance is given by (*proof in the lectures*)

$$V(k) = np(1-p). \tag{2.66}$$

The standard deviation is therefore

$$\sigma = \sqrt{np(1-p)}. \tag{2.67}$$

A few different binomial distributions (for different values of n and p) are shown in Fig. 2.5. I made use of the `scipy.stats.binom` class to draw these.

2.4.2 Poisson Distribution

The Poisson distribution (note the capitalised “P” because it is named after Siméon Poisson) is used for so-called “counting” experiments. In this case we are still interested in certain

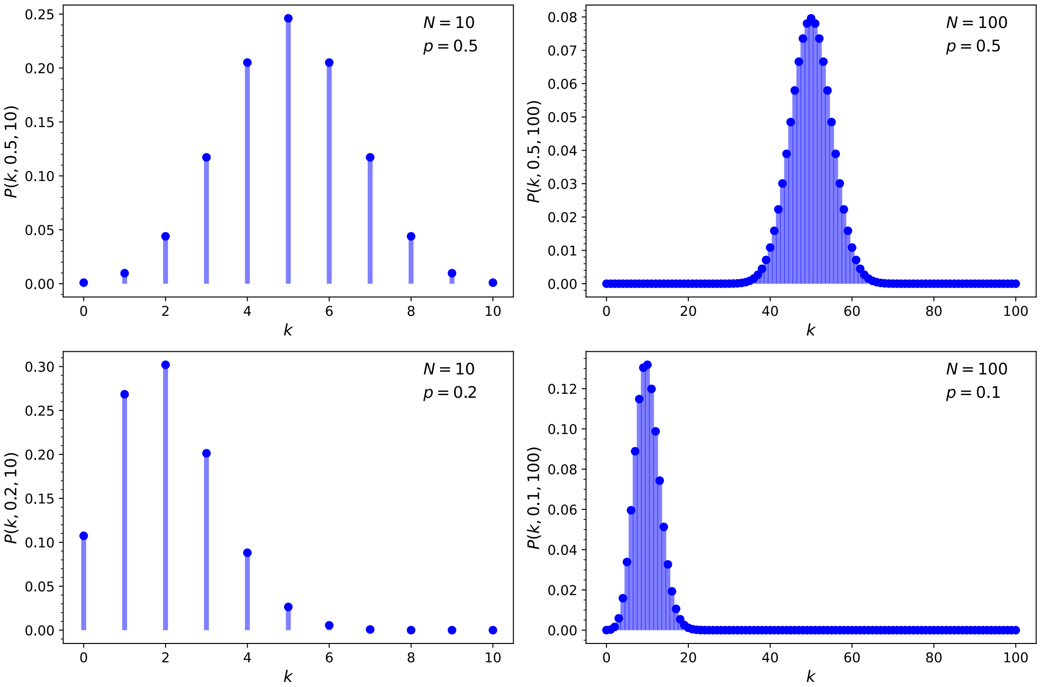

Figure 2.5: A few examples of the binomial distribution

outcomes but we do not have a concept of the number of trials. The Poisson distribution arises from the case where $n \to \infty$ but $np = \lambda$ stays fixed. Some examples of this might be the number of buses that go past you at a bus stop (within some time interval) where it makes no sense to ask how many buses did not go past you? When we plot data in a histogram we are counting the number of events in each bin of the histogram, so each bin has a Poisson distribution.

We can derive the Poisson p.d.f. in a few ways. One way is to take the limit of the binomial p.d.f. as $n \to \infty$ and $\lambda = np$ is kept fixed. The binomial p.d.f. in this case looks like,

$$P(k; \lambda/n, n) = \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \frac{n!}{k!(n-k)!}. \tag{2.68}$$

If we take the limit as $n \to \infty$ and k kept finite, the factorials in n give us,

$$\frac{n!}{(n-k)!} = n(n-1)(n-2)\dots(n-k+1) \to n^k. \tag{2.69}$$

The $1 - \lambda/n$ term gives us an exponential:

$$\left(1 - \frac{\lambda}{n}\right)^{n-k} \to \left(1 - \frac{\lambda}{n}\right)^{n} \to e^{-\lambda}. \tag{2.70}$$

Therefore the *Poisson probability distribution* is described by

$$\boxed{P(k, \lambda) = \frac{e^{-\lambda}\lambda^k}{k!}}. \tag{2.71}$$

The mean of the Poisson distribution is $\mu = \lambda$ and the variance is also $V(k) = \lambda$ so that the standard deviation is $\sigma = \sqrt{\lambda}$ (*the proofs are covered in the lectures*). A useful to remember property of the Poisson distribution (*the proof is one of the problem sheet questions*) is that the sum of two Poisson distributed variables will itself be a Poisson distribution whose expectation value (mean) is the sum of the two individual expectation values.

A few different Poisson distributions (for different values of λ) are shown in Fig. 2.6. I made use of the `scipy.stats.poisson` class to draw these.

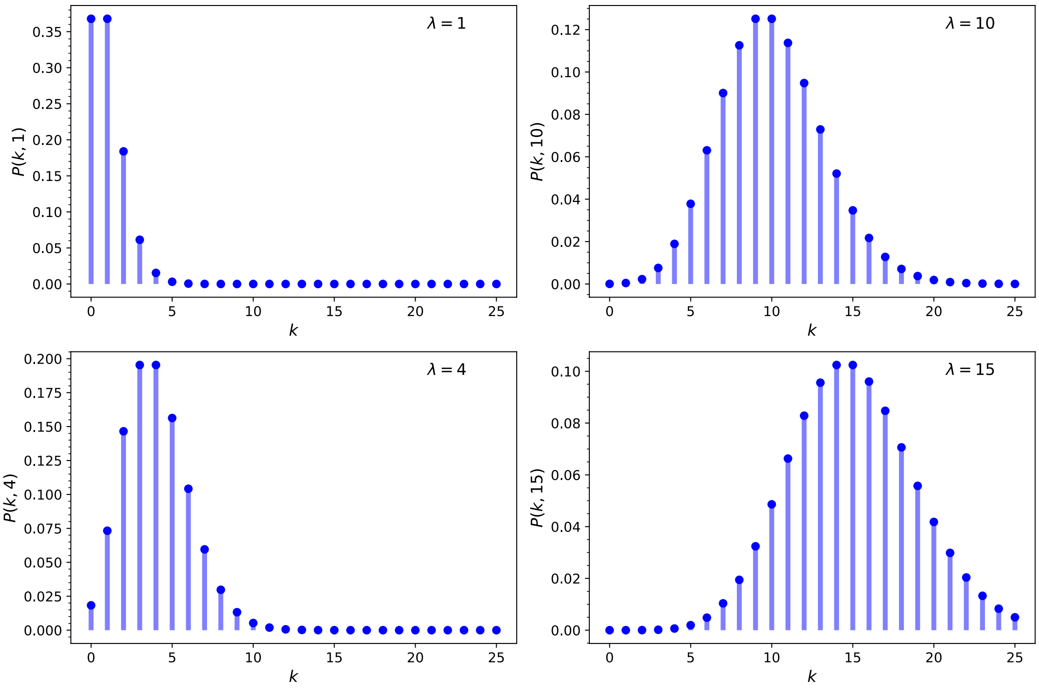

Figure 2.6: The Poisson distribution for different values of λ.

A note on errors on histogram entries

When we fill a histogram then (assuming the total number in each bin is not fixed) we have a counting experiment in each bin, so if the number of entries in each bin is n_i then the uncertainty (or error bar) on the entry in each bin is $\sqrt{n_i}$. When we visualise fits to data it is often instructive to show this error on the points. Below is a snippet of code showing a little comparison between plotting the same data as either a histogram or an error bar, the output is plotted on the left of Fig. 2.7.

```python
import numpy as np
import matplotlib.pyplot as plt

# set reproducible seed
np.random.seed(210187)

# generate some data
x = np.random.normal(size=500)

# make a histogram
nh, xe = np.histogram(x, bins='auto')

```

```
# get the centre of the bins
cx = (xe[1:] + xe[:-1])/2

fig, ax = plt.subplots(1,3,figsize=(18,4))

# plot as a hist and as errorbars
ax[0].hist(x, bins=xe, histtype='step', lw=2, label='Histogram')
ax[0].errorbar(cx, nh, yerr=nh**0.5, fmt='ko', capsize=2, label='Error Bars')

# now scale the histogram to the density and plot again (errors are wrong)
scf = 1/sum(nh)/(xe[1]-xe[0])
ax[1].hist(x, bins=xe, histtype='step', lw=2, density=True, label='Density Histogram')
ax[1].errorbar(cx, scf*nh, (scf*nh)**0.5, fmt='ko', capsize=2, label='Wrongly Scaled Errors')

# now scale appropriately
ax[2].hist(x, bins=xe, histtype='step', lw=2, density=True, label='Density Histogram')
ax[2].errorbar(cx, scf*nh, scf*nh**0.5, fmt='ko', capsize=2, label='Properly Scaled Errors')

plt.show()
```

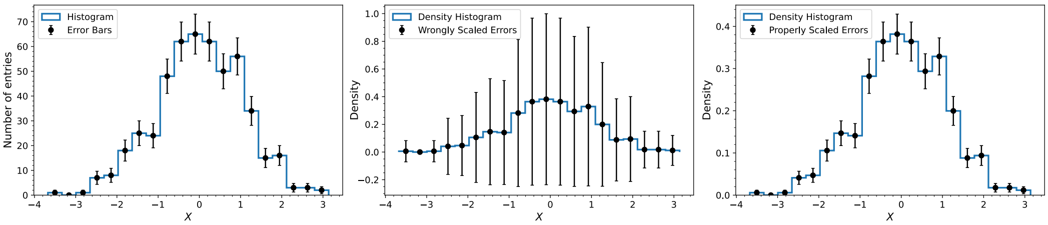

Figure 2.7: Examples of improper and proper scaling of histogram error bars

In other cases we may have more than an overall scale factor for our histogram but actually a scale factor per-event, known as a *weight*. For example, perhaps I line up the whole class at 30cm intervals away from a ball, and ask each person to independently estimate the radius of the ball. I could make the assumption that those closest to the ball would estimate it better so I may decide to weight each of my entries for your estimates by the inverse of the distance you are from the ball. In this case each entry in the histogram has its own weight, w_i. The count in each bin then becomes the sum of weights,

$$n_B = \sum_{i \in B} w_i. \tag{2.72}$$

Most histogram packages (like the ones from `numpy` and `matplotlib` we have been using) allow you to also pass weights. However the uncertainty is no longer the square-root of the bin count because uncertainties do not sum linearly. Instead we have to compute the variance in each bin, which is the sum of squared weights,

$$V(n_B) = \sum_{i \in B} w_i^2, \tag{2.73}$$

and then the uncertainty is the square-root of that, *i.e.*

$$\sigma(n_B) = \sqrt{\sum_{i \in B} w_i^2}. \qquad (2.74)$$

This means that if you want to keep track of the uncertainties for a weighted histogram you also have to keep track of the squared weights. Some histogram packages (*e.g.* `boost-histogram`) will do this for you but the deaulft `numpy.histogram` and `matplotlib.pyplot.hist` will not do so. Below is a short example of how you could do this:

```python
import numpy as np
import matplotlib.pyplot as plt

# set reproducible seed
np.random.seed(210187)

# generate some data
x = np.random.normal(size=500)

# generate some weights
ws = np.random.uniform(0,0.2,size=500)

# make a histogram with weights and squared weights
nh, xe = np.histogram(x, bins=25, weights=ws)
nh2, xe = np.histogram(x, bins=xe, weights=ws**2)
cx = 0.5*(xe[1:]+xe[:-1])

fig, ax = plt.subplots()

# plot weighted histogram
ax.hist(x, bins=xe, weights=ws, histtype='step', lw=2)

# plot weighted histogram with wrong errors (\ie assuming sqrt)
ax.errorbar(cx, nh, yerr=nh**0.5, fmt='ko', capsize=2, label='Naive $\sqrt{n}$
    ')

# plot weighted histogram with correct errors (using sqrt w^2)
ax.errorbar(cx, nh, yerr=nh2**0.5, fmt='r.', capsize=3, lw=2, label='$\sqrt{ \
    sum w^2}$')

plt.show()
```

and the corresponding figure in Fig. 2.8

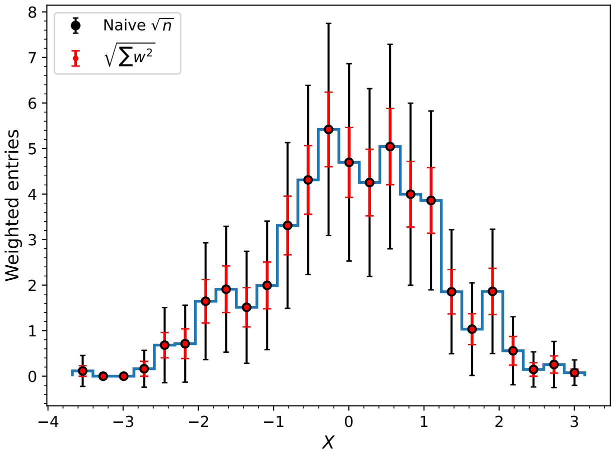

Figure 2.8: A comparison of properly and improperly weighted histograms.

2.4.3 Normal Distribution

Also called the Gaussian distributed (after Carl Gauss). Perhaps the most important distribution in all of statistics. We will see shortly that in the infinite sample limit all distributions will tend to a normal distribution. When we quote uncertainties of parameters in scientific papers we do so under the assumption that the parameter probability density is normally distributed and we are quoting the central value (the mean of the distribution) with an uncertainty corresponding to the width (the standard deviation of the distribution). When we propagate errors (which we will show shortly) we are assuming they are Gaussian distributed. Assumptions of Gaussianity are made all the time, especially in classical statistics, hence the huge importance of this distribution. Consequently, in this section we will not only show the p.d.f. of the normal distribution but also discuss some of it's other properties.

The p.d.f. for the special case of the *standard normal distribution*, which is centered on zero and has a width of one, is,

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}. \tag{2.75}$$

The factor of $\sqrt{2\pi}$ ensures the distribution is normalised and it is worth remembering that $\int_{-\infty}^{\infty} e^{-ax^2} dx = \sqrt{\pi/a}$. The mean is trivially computed because the expectation value of μ is the integral of an antisymmetric function and thus $\mu = 0$. Integration by parts allows us to compute that $\sigma = 1$.

A shifiting and scaling of x allows us to move the normal distribution to another location with another width. In other words making the subtitution $x \to (x-a)/b$ in (2.75) will produce a normal distribution with $\mu = a$ and $\sigma = b$. Consequently the normal (or Gaussian) p.d.f. is written

$$\boxed{f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}}\ . \tag{2.76}$$

The c.d.f. of the normal distribution, often denoted Φ, is also very useful. It is normally expressed in terms of the error function,

$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt, \tag{2.77}$$

where $z \in \mathbb{C}$, so that the Gaussian c.d.f. is

$$\Phi(x) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\right]. \tag{2.78}$$

Sometimes you see the c.d.f. written is a slightly different form:

$$F(X) = \phi\left(\frac{x-\mu}{\sigma}\right) \quad \text{where} \quad \phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2} dx. \tag{2.79}$$

The characteristic function has the nice form

$$phi(t) = e^{it\mu - t^2\sigma^2/2}. \tag{2.80}$$

A few nice properties of the normal distribution are

- Any linear combination of normally distributed random variables is also normally distributed
- The *sample mean* (1.1) and *sample variance* (1.6) are independent if the sample is drawn from the same normal distribution
- If X_i are standard normally distributed, the p.d.f. is constant on the hypersphere $\sum_i^N X_i^2 =$ constant which is a unique property of the normal distribution.

Given that we often quote errors as normally distributed the c.d.f. is highly useful because it tells us what fraction of a normal distribution is contained within each standard deviation. A few common values are shown in Table 2.1 below. Here's a little snippet of code that generates these values:

Table 2.1: The probability values, or p-values, for a few different standard deviations of a normal distribution.

σ	p	$1-p$
1	0.68268949	0.31731051
2	0.95449974	0.04550026
3	0.99730020	0.00269980
4	0.99993666	0.00006334
5	0.99999943	0.00000057

```python
import numpy as np
from scipy.stats import norm
from tabulate import tabulate

sigma = np.arange(1,6,1)
probs = norm.cdf(sigma) - norm.cdf(-sigma)
pvals = 1-probs

print(tabulate(zip(sigma,probs,pvals),floatfmt=[".0f",".8f",".8f"]))
```

which generates the following terminal output

```
-  ----------  ----------
1  0.68268949  0.31731051
2  0.95449974  0.04550026
3  0.99730020  0.00269980
4  0.99993666  0.00006334
5  0.99999943  0.00000057
-  ----------  ----------
```

This is graphically represented in Fig. 2.9 in which we see what fraction of a normal distribution is contained within each standard deviation.

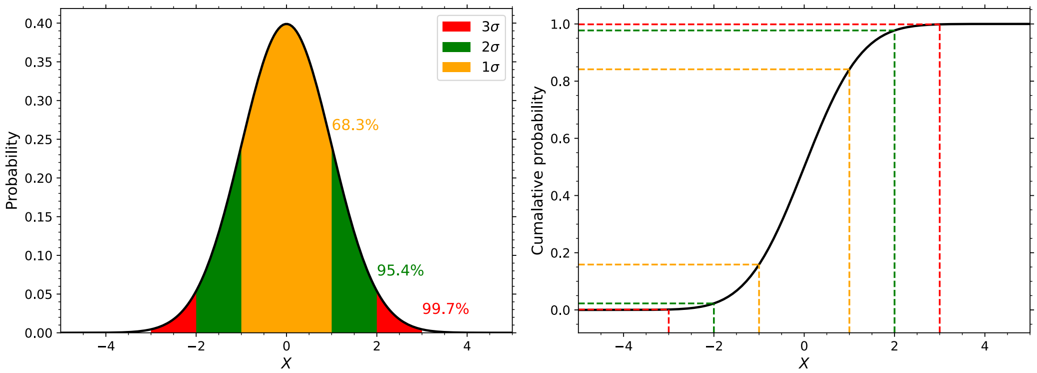

Figure 2.9: The p.d.f. (left) and c.d.f. (right) or the standard normal distribution with coloured contours showing what fraction is contained within a given standard deviation.

We are scientists. We do not just provide a central value when we quote a parameter. We should also *always* provide an estimate of the uncertainty. Our convention in science is to quote uncertainties at 1σ or *one standard deviation*. What we mean when we do this is that 68.3% of the time the true value will lie within 1σ of the central value (frequentist) or that the value is within this 1σ bound with 68.3% probability (Bayesian). Thus when fitting data points with some prediction (*e.g.* fitting a straight line) we only expect the line to pass within the error bars 68.3% of the time (on average).

Later in the course we will discuss goodness-of-fit tests and hypothesis tests. These are often converted, or presented, as p-values. You should be suspicious of fits that give either very low or very high p-values suggesting the data is either very poorly modelled or that the data follows the model suspiciously well. This can be slightly different from over-fitting which in general means the model has too many free parameters and can normally be dealt with in goodness-of-fit tests.

Finally, the inverse c.d.f. of the standard normal distribution, often denoted $\Phi^{-1}(p)$, is also really useful for converting from a probability value (or p-value) back into standard deviations, as the former is much easier to compute when we perform experiments and statistical analysis on them. The inverse c.d.f. is sometimes known as the *percentage point function* or the *quantile* function. You will also sometimes see the standard deviation referred to as the Z-score. However, be slightly wary when you are doing this because it depends whether your statistical test is one-side or two-sided as to how that conversion should be done. I have had some trouble convincing my own field of the proper use here and do not have time to cover the subtleties here. The take home message is that the p-value is normally the more fundamental quantity. We then just chose some convention (how many standard deviations of a standard normal distribution) to convert this into the Z-score or σ or significance.

2.4.4 Multi-variate Normal Distribution

In this case we can extend the normal distribution to many dimensions and include a correlation between them. So instead of depending on one random variable, X, the distribution now depends on a vector of many random variables $\vec{X}$ with the mean given by the vector $\vec{\mu}$ and the covariance given by a symmetric matrix $\boldsymbol{V}$. The p.d.f. for an n-dimensional normal distribution is given by

$$f(\vec{X}) = \frac{1}{(2\pi)^{n/2}\sqrt{|\boldsymbol{V}|}} \exp\left[-\frac{1}{2}\left(\vec{X}-\vec{\mu}\right)^{\mathrm{T}} \boldsymbol{V}^{-1}\left(\vec{X}-\vec{\mu}\right)\right]. \tag{2.81}$$

The expectation values are the vector of means, $E[\vec{X}] = \vec{\mu}$. The variances are the diagonal elements of the covariance matrix, $V(X_i) = V_{ii}$ and the covariances between variables given by the off-diagonal elements of the covariance matrix. The covariance matrix can be written in terms of the standard deviations, σ, and the correlation coefficients

$$\boldsymbol{V} = E\left[(\vec{X}-\vec{\mu})(\vec{X}-\vec{\mu})^{\mathrm{T}}\right] = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \dots & \rho_{1n}\sigma_1\sigma_n \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \dots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{1k}\sigma_1\sigma_n & \dots & \dots & \sigma_n^2 \end{bmatrix}. \tag{2.82}$$

Some interesting properties of the multivariate normal distribution are

- contours of constant probability are given by

$$(\vec{X}-\vec{\mu})^{\mathrm{T}}\boldsymbol{V}^{-1}(\vec{X}-\vec{\mu}) = \text{constant}$$

- any slice through the multivariate normal distribution are a constant in one of the variables gives an $n-1$ normal distribution with a covariance matrix $\boldsymbol{V}_{n-1}$ obtained by removing the relevant column and row from the inverted $^{-1}$ matrix and then re-inverting that
- any projection into a lower dimensionality (*i.e.* the marginal distribution) gives again a normal distribution, with a covariance matrix obtained by removing the relevant row and column of $\boldsymbol{V}$.
- the points about linear sums and the sample mean and variance being independent that were made for the normal distribution above also hold for the multivariate normal (indeed the one dimensional normal distribution is just a special case of the multi-dimensional normal with dimension of one)

Of course it is rather difficult to plot the multi-variate normal distribution in any dimensionality above 2 (we are restricted to the dimensionality of space and time with which we observe our local, macroscale universe). However it's still useful to look at some two dimensional normal distributions for which I make some plots below in Fig. 2.10. I will show

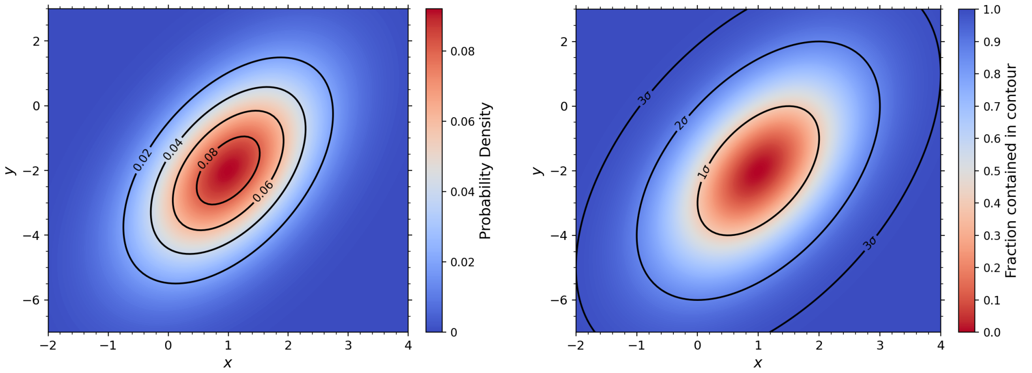

Figure 2.10: A visual demonstration of a two-dimensional Gaussian with $\vec{\mu} = (1, -2)$, $\vec{\sigma} = (1, 2)$ and $\rho = 0.5$. Left: the probability density itself. Right: the shaded contour shows the fraction of the distribution contained within a given contour ellipse, the lines convert this probability into a standard deviation.

the code for these after we have discussed the chi-squared distribution below which is what allows us to compute the probability and Z-score contours.

2.4.5 The exponential decay distribution

Quite a few scientific phenomena are described by an exponential decay. This is obviously particularly relevant in particle and nuclear physics whose decays are exponential with a slope given by the "half-life" or "lifetime" of the state. However, there are many other occurences in nature and sociology mostly relating to the *time until some event.* For example the amount of time until the next earthquake, the number of minutes spent on phone call *etc.*.

The p.d.f. for an exponential decay is given by

$$f(X) = \lambda e^{-\lambda X} \tag{2.83}$$

where λ is the decay constant. The c.d.f. for an exponential decay is

$$F(X) = 1 - e^{-\lambda X}. \tag{2.84}$$

The mean is given by $E[X] = \mu = 1/\lambda$ and the variance is $\sigma^2 = 1/\lambda^2$ and thus the standard deviation is the same as the mean $\sigma = 1/\lambda$.

2.4.6 Polynomial distributions

It is often useful when fitting probability distributions to be able to use some kind of polynomail basis. In other words giving the flexibility to "fit" a shape that we may not *a priori* know. The problem with polynomials is that they can quite easily go negative and this is not good for probability distributions (recall the first Kolmogorov axiom that $p(X) \geq 0$ everywhere).

A solution to this is to use polynomials in the "Bernstein" basis, also called *Bernstein* polynomials. These are defined on the interval $0 \geq X \geq 1$ (which can also be straightforwardly transformed to the interval required) and will always give an output distributions in the interval $0 \geq y \geq$ making them suitable for use in probability distribtutions.

A polynomial of a given degree, n, are built out of **Bernstein basis** polynomials. The $n+1$ Bernstein basis polynomials of degree n are defined as:

$$b_{\nu,n} = \binom{n}{\nu} x^{\nu}(1-x)^{n-\nu} \text{ for } \nu = 0, \ldots, n. \tag{2.85}$$

The coefficient in the bracket is the Binomial coefficient and there is a clear likeness to the Binomial distribution. The first few Bernstein basis polynomails are:

$$b_{0,0} = 1, \tag{2.86}$$

$$b_{0,1} = 1-x, \qquad b_{1,1} = x \tag{2.87}$$

$$b_{0,2} = (1-x)^2, \quad b_{1,2} = 2x(1-x), \quad b_{2,2} = x^2 \tag{2.88}$$

$$b_{0,3} = (1-x)^3, \quad b_{1,3} = 3x(1-x)^2, \quad b_{2,3} = 3x^2(1-x), \quad b_{3,3} = x^3. \tag{2.89}$$

The *Bernstein basis* polynomails for degree n can then be linearly summed with coefficients to form a **Bernstein polynomial** of degree n, with $n+1$ free parameters as the coefficients,

$$B_n(x) = \sum_{\nu=0}^{n} c_\nu b_{\nu,n}. \tag{2.90}$$

The probability distribution (appropriately normalised and prooved in the lectures) is given by

$$p(X;\vec{c}) = \frac{n+1}{\sum_{i=0}^{n} c_i} \sum_{i=0}^{n} c_i b_{i,n}(x). \tag{2.91}$$

A few different Bernstein basis polynomials are given in Fig. 2.11. Notice that Bernstein polynomials have the property of a partition of unity, which means that if the coefficients are all the same then they will sum to unity for all X. Thus with the same value of the Bernstein coefficients they will give a flat line.

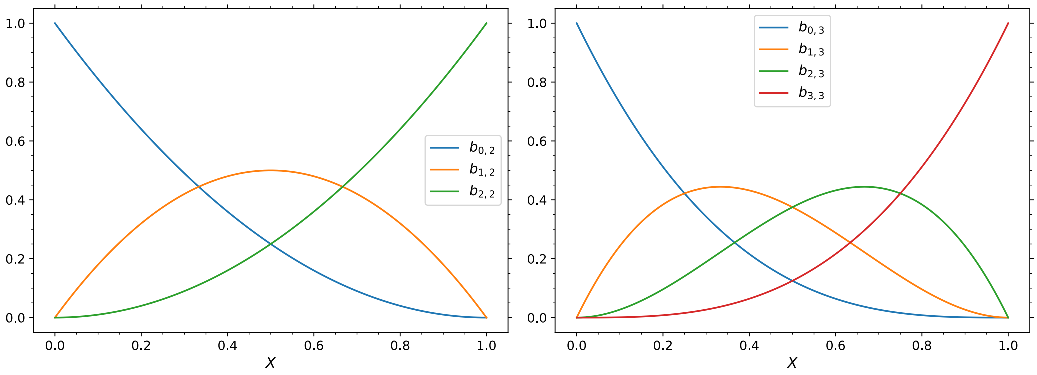

Figure 2.11: Bernstein polynomials up to 3rd order (left) and 4th order (right).

2.4.7 Chi-squared Distribution

The chi-squared, often written χ^2, distribution with k degrees of freedom gives the distribution of the sum of squares of k independent standard normal variables. It is widely used in hypothesis testing and allows us to construct confidence intervals. We have already seen an example of this in Fig. 2.10 above, where I exploited the χ^2 distribution to produce confidence intervals for the multivariate normal plots. The χ^2 distribution is a special case of the "Gamma" or Γ distribution (which I don't have time to cover here but is worth reading up on).

If X_i are a set of k independent standard normal variables then the sum of their squares

$$Z = \sum_{i=1}^{k} X_i^2 \tag{2.92}$$

is distributed according to a χ^2 distribution with k degrees of freedom. Clearly the p.d.f. of a χ^2 is then only non-zero for $x > 0$ and the p.d.f. is described by

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}. \tag{2.93}$$

When we come to discuss fitting later in the course we will see the importance of the χ^2 distribution. It has a mean of k and a variance of $2k$, thus a standard deviation of $\sqrt{2k}$.

Below, Fig 2.12, I plot the p.d.f. and c.d.f. of the χ^2 distribution for different degrees of freedom. One nice property is that the c.d.f. of the χ^2 distribution allows us to read off the fraction contained within a given standard deviation of the multivariate normal distribution. For example, we see that for $k = 1$ degrees of freedom the χ^2 c.d.f. of 1 gives 0.683, the c.d.f. of 2 gives 0.954 *etc.*. An important consequence of this is what it means for the fraction contained within an interval of a given Z-score for different degrees of freedom. Once again reading from the c.d.f. in Fig. 2.12 you can see that the fraction contained within 1σ is 0.683 for $k = 1$, but only 0.393 for $k = 2$ and 0.199 for $k = 3$. Below I put the code (Code Block 2.1) to produce the plots in Fig. 2.12 and now that we have a better understanding of the χ^2 distribution and how it changes for different degrees of freedom I also put below that the code (Code Block 2.2) which generates the 2-d Gaussian plots in Fig. 2.10.

```python
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import chi2

fig, ax = plt.subplots(1,2, figsize=(12,4))
x = np.linspace(0,8,400)
for k in range(1,5):
    ax[0].plot( x, chi2.pdf(x,k), label=f"$k={k}$" )
    ax[1].plot( x, chi2.cdf(x,k), label=f"$k={k}$" )
ax[0].legend()
ax[1].legend()

plt.show()
```

Code Block 2.1: The code snippet to generate the χ^2 distribution plots in Fig. 2.12

```python
import numpy as np
import matplotlib.pyplot as plt
```

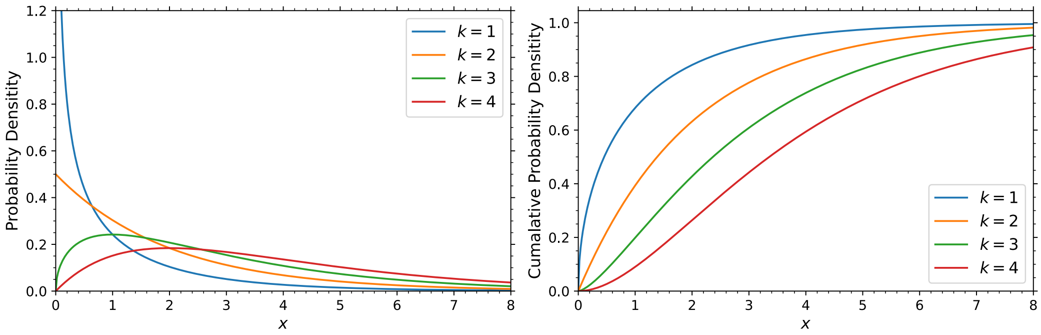

Figure 2.12: The p.d.f. (left) and c.d.f. (right) of the χ^2 distribution for different degrees of freedom

```python
from scipy.stats import multivariate_normal as mvn
from scipy.stats import chi2

# make the pdf
mus = [1,-2]
cov = [[1,1],[1,4]]
pdf = mvn(mus,cov)

# make the grid of points to draw
x = np.linspace(-2,4,100)
y = np.linspace(-7,3,100)
X, Y = np.meshgrid(x,y)
x, y = X.T, Y.T
pos = np.dstack((x,y))
z = pdf.pdf(pos)
dlnz = 2*(pdf.logpdf(mus) - pdf.logpdf(pos))
frz = chi2.cdf(dlnz,2)
sgz = (chi2.ppf( frz, 2 ))**0.5

# plot pdf
fig, ax = plt.subplots()
im = ax.contourf( x, y, z, levels=250, cmap='coolwarm' )
cs = ax.contour( x, y, z, levels=5, colors='k' )
ax.clabel(cs, cs.levels, inline=True)
fig.colorbar(im,ax=ax)

# plot contours for prob contained
fig, ax = plt.subplots()
im = ax.contourf( x, y, frz , levels=250, cmap='coolwarm_r' )
cs = ax.contour( x, y, sgz, levels=3, colors='k')
ax.clabel(cs, cs.levels, inline=True, fmt=lambda x: rf"${x:.0f}\sigma$")
fig.colorbar(im,ax=ax)
```

Code Block 2.2: The code snippet to generate the 2-d Gaussian distributions plotted in Fig. 2.10

2.4.8 Convolutions of Distributions

In scientific experiments we often measure the sum of two random variables. There are many examples of this *e.g.* in particle physics when we measure the decay angle of decay time of a particle our measurement is the sum of the random variable in question (the decay angle according to some distribution given the particles spin, the decay time according to a decaying exponential) and of the measurement error. If we have two independent random variables, X and Y, and define their sum $Z = X + Y$, then the joint probability density is the sum of the individual densities (because they are independent):

$$f(X, Y) = g(X)h(Y). \tag{2.94}$$

We now want to find the probability distribution of their sum Z. The cumalative density can be found by integrating up to the region which encloses $Z = X + Y$:

$$F(Z) = P(X + Y < Z) = \int_{-\infty}^{\infty} g(X)dX \int_{-\infty}^{Z-X} h(Y)dY \tag{2.95}$$

$$= \int_{-\infty}^{\infty} h(Y)dY \int_{-\infty}^{Z-Y} g(X)dX. \tag{2.96}$$

To obtain the p.d.f. we can differentiate the c.d.f. to give

$$f(Z) = \frac{dF(Z)}{dZ} = \int_{-\infty}^{\infty} g(X)h(Z - X)dX, \tag{2.97}$$

$$= \int_{-\infty}^{\infty} h(Y)g(Z - Y)dY. \tag{2.98}$$

This is known as a *convolution* of the two p.d.f.s and is often written as $f = g \otimes h$. To be more precise it is actually the *Fourier convolution* of g and h. There is also the less common *Mellin convolution* of g and h which arises from the product of two random variables, $Z = XY$. The p.d.f. in the Mellin case is given by

$$f(Z) = \int_{-\infty}^{\infty} \frac{g(X)h(Z/X)}{|X|} dX, \tag{2.99}$$

$$= \int_{-\infty}^{\infty} \frac{g(Z/Y)h(Y)}{|Y|} dY. \tag{2.100}$$

$$\tag{2.101}$$

The Fourier convolution is the more common type, but both are identically referred to as convlolutions, so you will often have to infer the type from the context.

Convolutions most commonly occur when we have some resolution function convolved with the distirbution we are measuring. I show some examples of this below in Fig. 2.13. You can see that in the case of an exponential decay this leads to potentially unphysical observations (*i.e.* negative decay times).

In practise it is almost never possible to analytically compute the convolved p.d.f. so instead we have to use numerical methods which always come with some subtleties and pitfalls. There are some basic numerical convolution algorithms as part of `scipy.signal` *e.g.* `scipy.signal.convolve` and `scipy.signal.fftconvolve`.

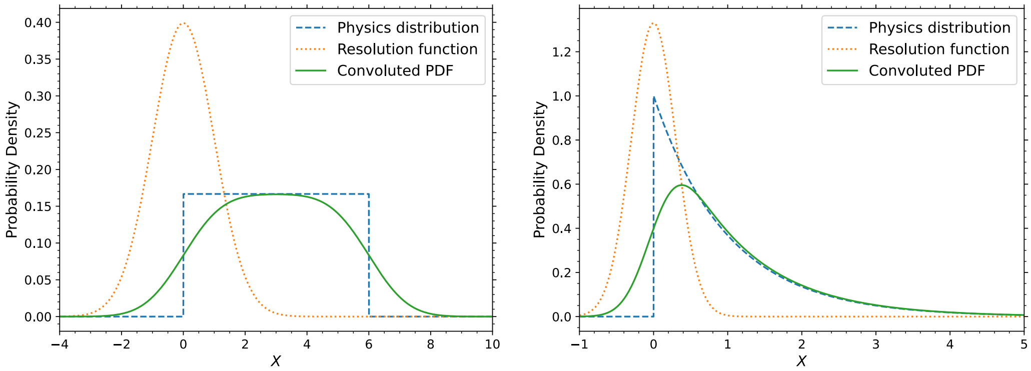

Figure 2.13: Some examples of convoluted p.d.f.s. The true signal distribution is shown as the blue dashed line, a top-hat / uniform distribution on the left and an exponential decay distribution on the right. The resolution function, a Gaussian noise term, is shown as the orange dotted line, and the observed distribution is the Fourier convolution of these shown in green.

2.4.9 Generating samples from distributions and the inverse c.d.f.

As part of our data analysis toolkit we are often interested in simulating different processes. This is done by assuming some probability distribution for the process we want to simulate and then generating events according to that distribution. For example if I want so simulate how a particle detector responds to incoming particles of a given momentum, I could parameterise my momentum resolution in terms of that momentum and would then want to generate from a Gaussian distribution centered at the momentum value with a width of the momentum resolution. If I draw many samples from this I will get a *smeared* momentum distribution which will mimic my particle detector.

Random number generation itself is not covered in this course, for that I suggest the computing oriented courses or the description given in Chapter 4 of Brandt [3]. Suffice to say that we can quite straightforwardly generate random numbers uniformly distributed between 0 and 1.

```python
import numpy as np
x = np.random.uniform()
```

The question is how can we generate random numbers according to some other distribution which is not uniform? There are two approaches to this, one is very fast but relies on knowing the inverse c.d.f. of the distribution, the other is much slower but only requires knowledge of the p.d.f. of the distribution. We will now discuss these two simulation methods.

Simulation using the inverse c.d.f.

Recall our discussion of the cumalative distribution function, c.d.f., in Sec. 2.2.6 which maps from a random variable, X, onto the cumalative probabilty which by defintion lives in the interval between 0 and 1. Consequently, if we can define the inverse c.d.f. this will map us from an interval in $[0,1]$ back onto the random variable X, thus we can simply generate uniform random numbers in $[0,1]$ and then pass them through the inverse c.d.f. to obtain random numbers distributed accordingly. We have already seen one example of the inverse c.d.f. for the Gaussian distribution where we computed the Z-score based on the probability. The inverse c.d.f. is sometimes called the *percentage point function* (p.p.f.) and can be denoted $F^{-1}(p)$ or $\Phi^{-1}(p)$. Below is a little example exploiting the built-in methods of `scipy` with a little graphical representation in Fig. 2.14.

```python
# generate 1000 uniform random numbers
import numpy as np
x = np.random.uniform(size=1000)

# translate them to Gaus(2,1) distributed using
# the inverse cdf, also called the ppf
from scipy.stats import norm, uniform
y = norm.ppf(x, loc=2, scale=1)
```

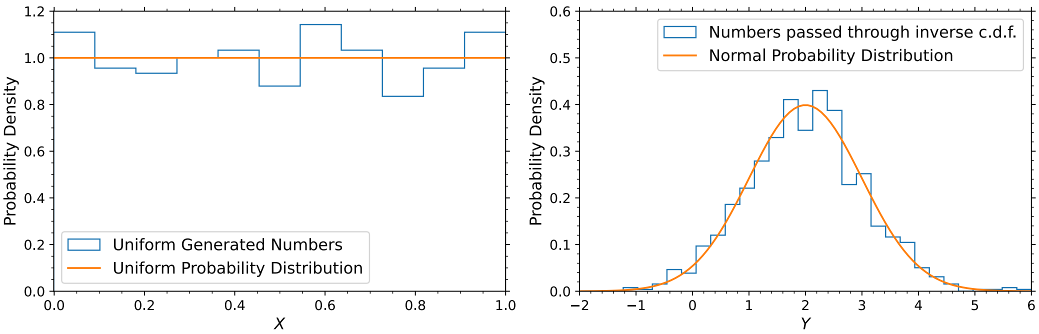

Figure 2.14: A graphical representation of a how uniformly generated random numbers can be translated into a different probability distribution.

In practise most random number generators include built in methods to generate from common distributions. For my little example scripts I always make use of `np.random` which includes the distributions we have discussed here.

```python
import numpy as np
# uniform distributed in interval a to b
x = np.random.uniform(a, b, size=100)
# gaussian distributed centered on mu with width sigma
g = np.random.normal(mu, sigma, size=100)
# exponentially distributed with rate parameter beta
e = np.random.exponential(beta, size=100)
# chisquare distributed with k degrees of freedom
c = np.random.chisquare(k, size=100)
```

The accept-reject method

In some cases you may have a probability distribution for which you do not know the p.p.f. but you still want to generate from it. In this case you can use a more brute-force approach to simulation which is a lot slower than the approach above. It requires generating two random numbers per event and can also be very inefficient because many of the generations are thrown away (improving this efficiency is a topic of it's own in Markov Chain Monte Carlo and other simulation methods which are covered in the other M1 and S2 modules).

The accept-reject method procedure can be defined as follows providing you have a description of the p.d.f.:

- Find the maximum value of the p.d.f., $f_{\max}$. Note that you therefore do not necessarily need the p.d.f. to be normalised for this to work.

- Generate a random position along the random variable axis with a uniform distribution. This gives you an "x_i" position to start from.

- Then generate another uniform random number between 0 and $f_{\max}$.

- Evaluate the p.d.f. at the value x_i to give $f(x_i)$.

- If $f(x_i) > f_{\max}$ the you throw the event away and start the procedure again (you *reject* the event).

- If $f(x_i) <= f_{\max}$ you keep (or *accept*) the event.

In doing so you eventually build up the distribution of the p.d.f..

The inefficiency of this algorithm is very dependent on the range you start generating from. For example if you do it for a Gaussian distribution with mean 0 and width 1 and you are generating in a range of X between $[-5, 5]$ you will *reject* a huge number of the simulated events you throw. *One of the problems gets you to assess how inefficienct this is?*. A graphical representation of accept-reject is shown below in Fig. 2.15.

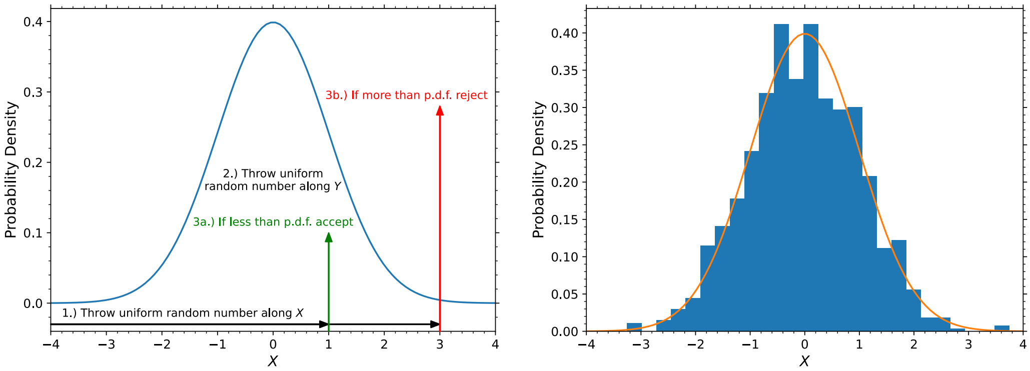

Figure 2.15: Left: Graphical representaion of the accept-reject method. Right: normally distributed data simulated according to accept-reject along with the normal p.d.f..

2.5 Limit Theorems

In this section we'll see a few important theorems about probability and distributions in the infinite statistics limit. These are vital concepts in probability theory where we see that seemingly random behaviour will actually converge to a certain average value or distribution over enough trials. This has particular relevance for the central limit theorem which states how the sum of any random variables from any distribution will approach a normal distribution as the sample size approaches infinity.

2.5.1 Convergence

Convergence in distribution

In words this simply states that if we have a sufficiently large sample of random numbers drawn from a distribution, then as the sample size tends to infinity, the distirbution of the sample will approach the probability distribution. Mathematically, imagine a sequence of random variables $(X_1, X_2, \ldots, X_n)$ with c.d.f.s $(F_1, F_2, \ldots, F_n)$ then the sequence X_n *converges in distribution* when $n \to \infty$ to X distributed by c.d.f. $F(X)$ if,

$$\lim_{n\to\infty} F_n(X) = F(X). \tag{2.102}$$

Convergence in probability

This is a stronger requirement than convergence in distribution and *convergence in probability* actually implies convergence in distribution. It is stronger because it places a requirement on how far X_n is from X. In words it means that the probability for an unusual outcome becomes progessively smaller as the samples size increases. Mathematically, the sequence $(X_1, X_2, \ldots, X_n)$ *converges in probability* to the random variable X if for all $\epsilon > 0$

$$\lim_{n\to\infty} P(|X_n - X|) > \epsilon) = 0. \tag{2.103}$$

The law of large numbers

Often probabilities for different types of events are not *a priori* known, so we have to obtain estimates of them using experiments. We do this by measuring the *frequency* of the given outcome in n experiments

$$h = \frac{1}{n}\sum_i x_i. \tag{2.104}$$

The frequency itself is a random variable as it depends on the outcome of the n experiments we ran but we can compute its expectation (see details on the Binomial distribution in Sec. 2.4.1, where the expectation of x is given by $E[x] = np$) and thus the expectation of the frequency is

$$E[h] = \hat{h} = \frac{1}{n}np = p. \tag{2.105}$$

The nomenclature of a “hat” over a variable, *e.g.* $\hat{h}$, we will use throughout these notes to mean “an estimate of”. The above kind of seems obvious and is a fundamental cornerstone of frequentist (or classical) statistics, but it is important to know that this works because as $n \to \infty$ then $\hat{h} \to p$. This is known as *the law of large numbers.*

2.5.2 Central Limit Theorem

This fundamental theoerem is one of the cornerstones of probability and statistical theory. It is why sums of independent observations tend towards the normal distribution. It is also then why we often quote errors as if they are Gaussian distributed, because we know in the infinite statistics limit that they will be. It is also what allows us to perform error propagation where we treat individual central values and uncertainties as if they are normally distributed.

We have already seen, when discussing the expectation operator in Sec. 2.3.1, that because the expectation operator is linear then the sum of a sequence of independent random variables X_i from any distribution with mean μ_i and variance σ_i^2 will have a mean which is the sum of the individual means and a variance which is the sum of the individual variances. The sum is a random variable, S, where

$$S = \sum X_i \quad \text{which has} \quad \mu = \sum \mu_i \quad \text{and} \quad \sigma^2 = \sum \sigma_i^2. \tag{2.106}$$

The central limit theorem states that the distribution of S will tend to a normal distribution with $\mu = \mu$ and $\sigma = \sigma$ as $N \to \infty$. In particular if I redefine the random variable so that it is shifted by the mean and scaled by the standard deviation so that,

$$S \to S' = \frac{S - \sum_{i=1}^{N} \mu_i}{\sqrt{\sum_{i=1}^{N} \sigma_i^2}}, \tag{2.107}$$

then the distribution of S' in the limit as $N \to \infty$ is the standard normal distribution ($\mu = 0$, $\sigma = 1$).

As the Central Limit Theorem is of such importance I have decided to show you a proof of it, I may not cover this proof in the lectures and it is not examinable. For simplicity we can imagine that the independent random variables X_i are all thrown from the same distribution so that $\mu_i = \mu$ and $\sigma_i = \sigma$. The theorem still holds for all cases but is easier to show under this simplification. First we'll define a variable which is the sample average of the random variable $\bar{X} = \sum X_i / N$. We'll then define a standardised version which is shifted by the mean and scaled by the standard deviation,

$$Y = \frac{\bar{X} - \mu}{\sigma / \sqrt{N}} \tag{2.108}$$

It is then straightforward to show that the expected mean and expected variance of Y are $E[Y] = 0$ and $V(Y) = 1$ (*I leave this for you to show at home*). What we then want to show is that the random variable Y *converges in distribution* to the standard normal distribution, *i.e.* that

$$\lim_{N \to \infty} p(Y < x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{X^2/2} dX. \tag{2.109}$$

If we now write down the first few orders in the expansion of the characteristic function of X, and we also make the assumption that $\mu = 0$ (remembering we can always arbitaily shift a distribution without loss of generality), it looks like

$$\phi_X(t) = E[e^{itX}] = 1 + \sigma^2 \frac{(it)^2}{2} + \mathcal{O}(t^3) + \ldots. \tag{2.110}$$

The characteristic function of Y is then given by

$$\phi_Y(t) = \left[\phi \left(\frac{t}{\sigma \sqrt{N}} \right) \right]^N, \tag{2.111}$$

where the power of N comes from the fact that we are summing over random variables X_i and that equates to a product of the individual characteristic functions. We can now take the logarithm of $\phi_Y(t)$ to find that

$$\begin{aligned}\ln \phi_Y(t) &= N \ln \phi\left(\frac{t}{\sigma\sqrt{N}}\right) && (2.112)\\ &= N \ln\left[1 + \frac{\sigma^2}{2}\frac{(it)^2}{N\sigma^2} + \mathcal{O}\left(\frac{t^3}{N^{3/2}}\right)\right] && (2.113)\\ &= \frac{(it)^2}{2} + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right). && (2.114)\end{aligned}$$

And therefore as $N \to \infty$

$$\lim_{N\to\infty} \phi_Y(t) = e^{-t^2/2}, \tag{2.115}$$

which is the characteristic function of the standard normal distribution.

A rather beautiful demonstration of the Central Limit Theorem can be seen using a Galton Board, of which there is a demonstration in the lectures.

2.5.3 Errors

We have now seen from the central limit theorem why we represent uncertainties as Gaussian distributed, because if we have a large enough sample size then the error will be Gaussian. The convention in science is to quote uncertainties (and draw errors bars) at 1 standard deviation of a normal distribution. Thus we know that 68.3% of the distribution should be contained with the error bar or uncertainty interval. If in reality less than 68.3% is contained within the uncertainty it is said to *under-cover*. If more, then it is said to *over-cover*. When we discuss frequentist interval estimation later, we will talk more about *coverage*.

Propagation of errors

Of course we may measure (or know) the uncertainty on some random variable but want to quote the uncertainty on some other random variable which is a function of it. This means we have to *propagate* the error. Imagine some linear function, $Z = aX + b$, where a and b are constants and X is a random variable with a measured or known variance. We can compute the variance of Z using linear expectation properties:

$$\begin{aligned}V(Z) &= E[Z^2] - E[Z]^2 && (2.116)\\ &= E[(aX+b)^2] - E[aX+b]^2 && (2.117)\\ &= a^2E[X^2] + abE[X] + b^2 - a^2E[X]^2 - 2abE[X] - b^2 && (2.118)\\ &= a^2\left(E[X^2] - E[X]^2\right) && (2.119)\\ &= a^2V(X). && (2.120)\end{aligned}$$

In terms of the standard deviation we can write, $\sigma_Z = |a|\sigma_X$, so that the shift of b does nothing to the spread and the scaling by a increases the spread by a factor of $|a|$.

We can now extend this to any generic function $f(X)$ if we only consider small errors and the linear term in a Taylor expansion of X around some point X_0,

$$f(X) \approx f(X_0) + (X - X_0)\left(\frac{df}{dX}\right)\bigg|_{X=X_0} + \mathcal{O}(X^2). \tag{2.121}$$

Following the logic above in (2.120) we can then write the variance and standard deviation as

$$V(f) \approx \left(\frac{df}{dX}\right)^2 V(X) \tag{2.122}$$

$$\sigma_f \approx \left|\frac{df}{dX}\right| \sigma_X. \tag{2.123}$$

We can now extend this formalism to functions of many random variables. For example if we have two random variables, X and Y, and a generic function $f(X, Y)$ then,

$$V(f) = \left(\frac{df}{dX}\right)^2 V(X) + \left(\frac{df}{dY}\right)^2 V(Y) + 2\left(\frac{df}{dX}\right)\left(\frac{df}{dY}\right) \mathrm{cov}(X, Y) \tag{2.124}$$

$$\sigma_f^2 = \left(\frac{df}{dX}\right)^2 \sigma_X^2 + \left(\frac{df}{dY}\right)^2 \sigma_Y^2 + 2\left(\frac{df}{dX}\right)\left(\frac{df}{dY}\right) \rho\sigma_X\sigma_Y. \tag{2.125}$$

This can then be further extended to many functions of many variables and written in a completely generic notation as

$$\mathrm{cov}(f_k, f_l) = \sum_i \sum_j \left(\frac{\partial f_k}{\partial X_i}\right)\left(\frac{\partial f_l}{\partial X_j}\right) \mathrm{cov}(X_i, X_j) \tag{2.126}$$

or even more compactly in a matrix notation using the Jacobian matrix,

$$\boldsymbol{V}_f = \boldsymbol{J}\boldsymbol{V}_X\boldsymbol{J}^{\mathrm{T}}. \tag{2.127}$$

I assume you will have seen some simple formula for propagation of errors before. As one of the problem sheet exercises I ask you to derive the uncertainties σ_f in terms of σ_x and σ_y for transformations of the type $f = x + y$, $f = xy$, $f = x/y$, $f = sin(x)$, $f = cos(x)$.

Averaging and combining measurements

As scientists we often want to make averages of repeated measurement, these are sometimes referred to as *combinations*. For example, when the Higgs boson was discovered, there were independent measurements from two different detectors, ATLAS and CMS, who each had searched for the Higgs boson decaying in two different ways, $H \to \gamma\gamma$ and $H \to ZZ$. In this case there are 4 independent measurements which we want to statistically combine into one measurement. In particle physics there is an entire worldwide collaboration dedicated to making these combinations of all sorts of parameters in particle physics, it's called the Particle Data Group.

Once again we can exploit the properties of the Central Limit Theorem for this. In the case where each individual result has the same uncertainty, σ, then the average of N measurements is simply given by

$$\boldsymbol{\mu} = \frac{1}{N}\sum_i \mu_i, \tag{2.128}$$

where μ_i is the quoted central value of each result. The variance is

$$V(\boldsymbol{\mu}) = \frac{1}{N^2}\sum V_i = \frac{\sigma^2}{N}. \tag{2.129}$$

So the important thing to realise here is that the error, standard deviation, falls like $1/\sqrt{N}$, which is an important relationship to remember. This means that if you want to double your precision you need four time as much data, and consequently an experiment which only collects data linearly with time will give diminishing returns.

Quite often when we average, or combine, measurements then they do not all have the same uncertainty. In the Higgs discovery example above that is the case, different experiments and different decay channels will have different precision to the parameters of interest. In this case we perform a *weighted average* where the weight is given by the precision (*i.e.* the variance of each input measurement). The average now becomes

$$\boldsymbol{\mu} = \frac{\sum_i \mu_i/\sigma_i^2}{\sum_i 1/\sigma_i^2} \tag{2.130}$$

and the variance becomes

$$V(\boldsymbol{\mu}) = \frac{1}{\sum 1/\sigma_i^2}. \tag{2.131}$$