2

Survey of Distributions

Contents

In this chapter we give a survey of some of the most frequently encountered distributions. In Chapters 4 and 5 we will cover some of these in more detail. Many of the common distributions belong to the *exponential* family of distributions. We present this family and some of its properties in Section 2.4. We conclude the chapter by two appendices covering the gamma and beta functions; both functions that arise in some of the common distributions.

If probability is the language for discussing uncertainties, then Chapter 1 could be viewed as learning the basic grammar and simple verbs, while this chapter is more like learning the important nouns or objects. With these in place, we will be in a position in subsequent chapters to use our language to convey ideas.

https://doi.org/10.1017/9781108635349.003 Published online by Cambridge University Press

2.1 Discrete Distributions

It is easy to be put off by the seemingly endless number of probability distributions, but there are only a handful of distributions that keep on cropping up. The sooner you make friends with them the easier your life is going to be. We start with discrete distributions, which are those that involve random variables which take values that lie in a countable set.

2.1.1 Binomial Distribution

One of the frequently met probability distributions that pops up in a huge number of applications is the *binomial distribution*. It arises when we sample n objects that belong to two classes, A and B say. We assume that the probability of choosing an object of class A is p. This does not change over time. We can think of randomly choosing red and blue balls from a bag where the ratio of red to blue balls is p. Each time we choose a ball we put it back and mix up the balls before drawing the next sample. The probability of choosing m objects of class A in n trials is given by

$$\mathbb{P}\left(X = m|n, p\right) = \text{Bin}(m|n, p) = \binom{n}{m} p^m \left(1 - p\right)^{n-m}, \tag{2.1}$$

where $\binom{n}{m} = \frac{n!}{m!(n-m)!}$ is the binomial coefficient (often referred to as 'n choose m'). Figure 2.1 shows examples of the binomial distribution.

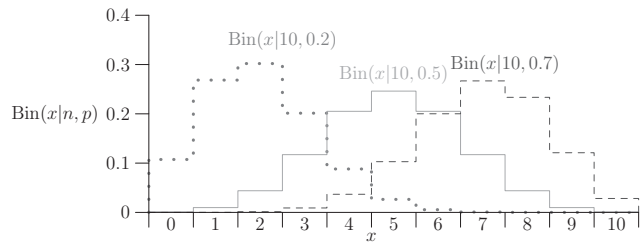


Figure 2.1 Example of binomial mass function for $n = 10$, and $p = 0.2$ (dotted), $p = 0.5$ (solid line) and $p = 0.7$ (dashed).

The mean of a binomial distribution is $n\,p$ and the variance is $n\,p\,(1 - p)$. We return to the binomial distribution in Chapter 4. A large number of distributions are in some way related to the binomial distribution: the hypergeometric distribution describes the situation of sampling without replacement; the Poisson distribution corresponds to a limit of the binomial as $p \to 0$ and $n \to \infty$, but with $p/n \to \mu$, a constant; the multinomial distribution is a generalisation of the binomial distribution to the case where there are more than two classes; finally the Gaussian distribution is a limit of the binomial distribution as $n \to \infty$.

Example 2.1 Rolling Dice

What is the probability of getting three sixes in 10 rolls of an honest dice?

This situation describes a case of repeating independent random binary trials, which gives rise to a binomial probability. In this case,

https://doi.org/10.1017/9781108635349.003 Published online by Cambridge University Press

we have a success probability of $p = 1/6$ and have $n = 10$ trials so the probability of three successes is

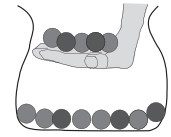
$$\text{Bin}(3|10, 1/6) = \binom{10}{3}\left(\frac{1}{6}\right)^3\left(1-\frac{1}{6}\right)^7 = 0.155.$$

Thus we can expect this to happen around 15% of the times we attempt it.

2.1.2 Hypergeometric Distribution

Although binomial distributions are the most common type of discrete distribution when dealing with two classes, they are not the only one. The hypergeometric distribution describes the probability of choosing k samples of class A out of n attempts, given that there is a total of N objects, m of which are of class A. For example, if you have a bag of N balls of which m are red and the rest are blue, and you sample n balls from the bag without replacement, then the hypergeometric distribution

$$\mathbb{P}\left(K = k|N, m, n\right) = \text{Hyp}(k|N, m, n) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} \tag{2.2}$$



$N = 13,\ m = 8$
$n = 5,\ k = 3$

tells you the probability that k of the drawn balls are red. (N is the total number of balls and is not a random variable – we use a capital to follow a commonly used convention for describing this distribution.) This probability is just the number of different ways of choosing k red balls from the m red balls, times the number of ways of choosing $n - k$ blue balls from the $N - m$ blue balls, divided by the total number of ways of choosing n balls from N balls. There are a number of surprising symmetries, for example, $\text{Hyp}(k|N, m, n) = \text{Hyp}(k|N, n, m)$ (that is, we get the same probability when we exchange the number of red balls and the number of balls that we sample). These arise due to the many identities involving binomial coefficients.

The mean value of K is $n\,m/N$ and its variance is $n(m/N)(1 - m/N)(N - n)/(N - 1)$. Typical probability masses are shown in Figure 2.2. We observe that these figures are not too dissimilar to those for the binomial distribution. Indeed, in the limit $N, m \to \infty$ such that $m/N \to p$ the hypergeometric distribution converges to the binomial distribution (see Exercise 2.4 on page 41).

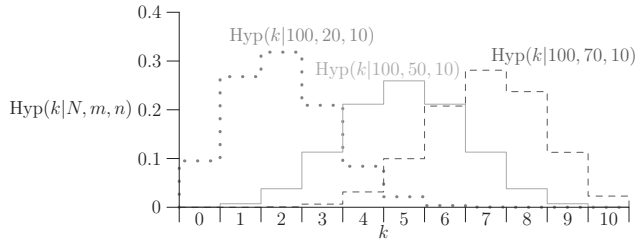


Figure 2.2 Examples of hypergeometric distributions for $N = 100$, $n = 10$, and $m = 30$ (dotted), $m = 50$ (solid line) and $m = 70$ (dashed).

https://doi.org/10.1017/9781108635349.003 Published online by Cambridge University Press

An application of this distribution is when there is a shipment of N objects, m of which are defective. If we sample n of the objects, then the hypergeometric distribution tells us the probability that k of the samples are defective. This is also the distribution you need to use if you want to calculate the probability of winning a prize in the UK National Lottery (see Exercise 2.5 on page 41). The hypergeometric distribution is not so well known as its ubiquity deserves. Possibly its low profile is a consequence of the fact that it is not always that easy to deal with analytically.

Example 2.2 Bridge

In the game of bridge, each player is dealt 13 cards. What is the probability that player 1 has three aces?

Treating a bridge hand as a random sample of 13 cards from a pack of 52 cards where we don't replace the cards, then we see that this is a job for the hypergeometric distribution. The total number of cards is $N = 52$, the number of aces is $m = 4$. A hand is a sample of $n = 13$ cards so that the probability we seek is

$$\mathrm{Hyp}(3|52,4,13) = \frac{\binom{4}{3}\binom{48}{10}}{\binom{52}{13}} = \frac{858}{20\,825} = 0.0412.$$

That is, a player can expect such a bridge hand around 4% of the time.

2.1.3 Poisson Distribution

The Poisson distribution can be regarded as a limiting case of the binomial distribution when $p \to 0$ and $n \to \infty$, but with $pn = \mu$. In this limit, with $m \ll n$, the binomial coefficient simplifies

$$\binom{n}{m} = \frac{n(n-1)\ldots(n-m)}{m!} \approx \frac{n^m}{m!},$$

and

$$(1-p)^{n-m} = \mathrm{e}^{(n-m)\log(1-p)} \approx \mathrm{e}^{-p(n-m)} \approx \mathrm{e}^{-pn} = \mathrm{e}^{-\mu}$$

where we have used the Taylor expansion $\log(1-p) = -p + O(p^2)$. Thus in this limit

$$\lim_{\substack{p\to 0 \\ p\times n=\mu}} \binom{n}{m} p^m (1-p)^{n-m} = \frac{(np)^m \mathrm{e}^{-\mu}}{m!} = \frac{\mu^m \mathrm{e}^{-\mu}}{m!} = \mathrm{Poi}(m|\mu). \tag{2.3}$$

This is the definition of the Poisson distribution. Its somewhat simpler form than the binomial distribution makes it easier to use. It is also important because it describes the distribution of independent point events that occur in space or time (we return to this in Section 12.3). The Poisson distribution arises, for

https://doi.org/10.1017/9781108635349.003 Published online by Cambridge University Press

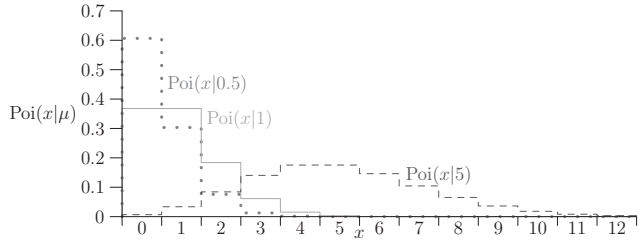


Figure 2.3 Examples of the Poisson distribution for $\mu = 0.5$ (dotted), $\mu = 1$ (solid line), and $\mu = 2$ (dashed).

example, if you want to know the probability of a Geiger counter having 10 counts in a minute given that the background radiation level is six counts per minute (answer 0.0413). Both the mean and variance of the Poisson distribution are equal to μ. Typical examples of the distribution are shown in Figure 2.3.

Example 2.3 Carbon Dating

Carbon dating is traditionally based on counting the number of beta particle emissions associated with the radioactive decay of carbon 14. Carbon 14 is an radioactive isotope of carbon with a relatively short half-life of 5730 years. All carbon 14 that initially existed in the early earth would have decayed a long time ago. However, it is constantly being replenished through neutron capture caused by cosmic rays creating neutrons that react with nitrogen in the atmosphere.

$$^{14}\mathrm{N} + {}^{1}n \rightarrow {}^{14}\mathrm{C} + {}^{1}p.$$

where ${}^{1}n$ is a neutron and ${}^{1}p$ a proton. As a consequence, carbon in the atmosphere (CO_2) has around one part per trillion of carbon 14. This is equivalent to 60 billion atoms per mole of carbon. This carbon is then taken up by plants through photosynthesis. By measuring the ratio of carbon 14 we are then able to deduce its age.

The probability of an atom of carbon decaying in one year is $\lambda = 1.245 \times 10^{-4}$. The number of carbon 14 atoms in a 1 mole sample (i.e. approximately 12 g of carbon) is

$$N = N_0 \, \mathrm{e}^{-\lambda t}$$

where $N_0 = 6 \times 10^{10}$ is the estimated number of atoms absorbed from the atmosphere through photosynthesis and t is the age of the sample. The expected number of decays in a time Δt is $\mu = \lambda N \Delta t$. Now suppose we observe $n = 100$ decays of carbon 14 in one hour. As radioactive decays are well approximated by a Poisson distribution, the probability of the decay is

$$\mathbb{P}\left(n = 100\right) = \frac{\mu^{100}}{100!} \mathrm{e}^{-\mu}$$

where $\mu = \lambda \, \Delta t \, N_0 \, \mathrm{e}^{-\lambda t} = 852.7 \, \mathrm{e}^{-\lambda t}$ (recall that we know λ, Δt, and N_0, but we don't know the age of the sample t). In Figure 2.4 we show

https://doi.org/10.1017/9781108635349.003 Published online by Cambridge University Press

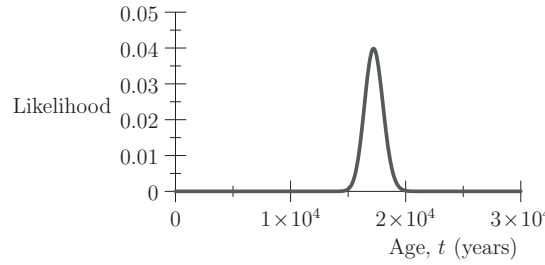


Figure 2.4 Probability of observing 100 decays in a sample with 1 mole of carbon atoms an hour versus the age of the sample.

the probability of observing 100 decays in the sample versus the age t in years. We see that with high likelihood the age of the sample would be between 15,000 and 20,000 years.

We assumed that the proportion of carbon 14 in the atmosphere is constant over time. This is not true (it is not even true over location), thus to obtain a more accurate estimate, the concentration of carbon 14 is calibrated against tree samples that can be aged by counting rings. However, as our probabilistic model shows, there is also a natural uncertainty caused by the underlying Poisson nature of radioactive decay. To obtain precise dates for a small sample requires that measurements over a very long time interval be made. Modern carbon dating tends to measure the proportion of carbon 14 directly using a mass spectrometer to reduce the uncertainty caused by that randomness of beta decays.

2.2 Continuous Distributions

These distributions describe random variables that take on continuous values. By far the most important distribution in this class is the Gaussian or normal distribution. However, there are a number of other continuous distributions that are common enough that they are worth getting to know.

2.2.1 Normal Distribution

The *normal distribution* – also called the *Gaussian distribution* – is by far the most frequently encountered continuous distribution. There are a number of reasons for this. The central limit theorem (see Section 5.3 on page 81) tells us that the distribution of the sum of many random variables (under mild conditions) will converge to a normal distribution as the number of elements in the sum increase. Many of the other distributions converge to the Gaussian distribution as their parameters increase. This means that in practical situations many quantities will be approximately normally distributed. If all you know about a random variable is its mean and variance then there is a line of reasoning (the so-called maximum entropy argument, see Section 9.2.2 on page 267) that says that of all possible distributions with the observed mean and variance the normal distribution is

https://doi.org/10.1017/9781108635349.003 Published online by Cambridge University Press

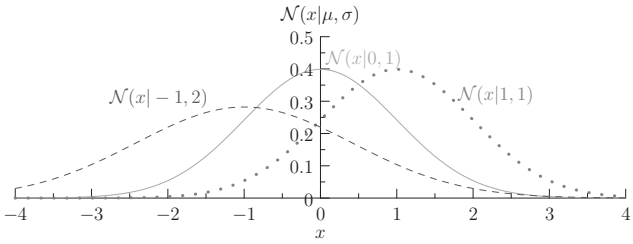


Figure 2.5 Examples of the normal distribution for $(\mu, \sigma) = (0, 1)$ (solid line), $(\mu, \sigma) = (1, 1)$ (dotted), and $(\mu, \sigma) = (-1, 2)$ (dashed).

overwhelmingly the most likely. Thus, assuming it is normally distributed is, in some sense, the optimal decision. However, before you use this argument you need to understand the small print (i.e. you've made a strong assumption about your variables that *ain't necessarily so*). A further reason why normal distributions are so commonly used is simply because they are easy to manipulate mathematically – this is a less contemptible motivation than it may at first appear. All models are abstractions from reality, and an approximate, but an easily solvable model is often much more useful than a more accurate but complex or intractable model.

The probability density for the normal distribution is defined as

$$\mathcal{N}(x|\mu.\sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma}\,\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{2.4}$$

which has mean μ and variance σ^2. Examples of the normal probability density functions are shown in Figure 2.5. We will have much more to say about the normal distribution in Chapter 5.

2.2.2 Gamma Distribution

When considering problems where a continuous random variable only takes positive values, the normal distribution can provide a poor model. Often a more appropriate model is the *gamma distribution* defined for $X > 0$ through the probability density

$$\mathrm{Gam}(x|a,b) = \frac{b^a x^{a-1}\mathrm{e}^{-b\,x}}{\Gamma(a)}, \tag{2.5}$$

where $\Gamma(a)$ is the *gamma function* defined (for real $a > 0$) by

The gamma function is actually defined throughout the complex plane except where a is equal to 0 or a negative integer.

$$\Gamma(a) = \int_0^\infty x^{a-1}\mathrm{e}^{-x}\mathrm{d}x$$

(see Appendix 2.A). It can easily be verified using integration by parts that $\Gamma(a+1) = a\,\Gamma(a)$. For positive integers, $n > 0$, the gamma function is given by $\Gamma(n) = (n-1)!$ (factorial). In some texts the gamma distribution is defined with parameters $\alpha = a$ and $\beta = 1/b$. The mean of the gamma distribution is given by a/b (or $\alpha\,\beta$) while the variance is given by a/b^2 (or $\alpha\,\beta^2$). Examples of the distribution are shown in Figure 2.6.

https://doi.org/10.1017/9781108635349.003 Published online by Cambridge University Press

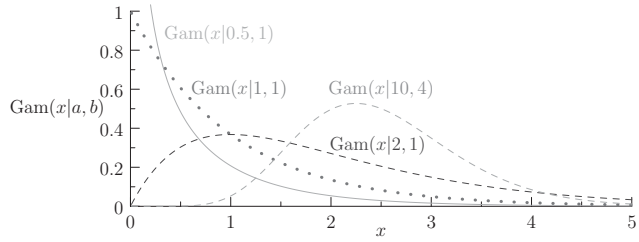


Figure 2.6 Examples of the gamma distribution for $b = 1$ and $a = 0.5$ (solid line), $a = 1$ (dotted) and $a = 2$ (dashed).

The gamma distribution is often used to empirically fit data that are known to always take positive values. For example, if you wish to model the intensity of light from different stars or the sizes of different countries then the gamma distribution is often a reasonable fit. Given an empirically measured mean $\hat{\mu}$ and variance $\hat{\sigma}^2$ a simple fit is to choose $a = \hat{\mu}^2/\hat{\sigma}^2$ and $b = \hat{\mu}/\hat{\sigma}^2$. (Although this gives a reasonably good fit, it is not the maximum likelihood estimator for a and b.) Gamma distributions also arise naturally in many problems. We discuss a few examples here.

The *chi-squared* (or χ^2) distribution is a particular form of the gamma distribution. The distribution arises in sums such as

$$S_k = \sum_{i=1}^{k} X_i^2,$$

where X_i are normally distributed variables with mean 0 and variance 1. Then S_k is distributed according to

$$f_{S_k}(s) = \chi_k(s) = \mathrm{Gam}\left(s|\tfrac{k}{2}, \tfrac{1}{2}\right).$$

The χ^2-distribution arises when evaluating the expected errors in curve fitting.

In the special case of $a = 1$, the gamma distribution reduces to the exponential distribution

$$\mathrm{Exp}(x|b) = \mathrm{Gam}(x|1, b) = b\,\mathrm{e}^{-b\,x}. \tag{2.6}$$

The exponential distribution describes the waiting times between events in a Poisson process (see Section 12.3.1 on page 380).

The velocity of particles in an ideal gas (a model for a real gas which is often very accurate) are normally distributed, such that the components of the velocity V_x, V_y, and V_z have distribution $\mathcal{N}(V_i|0, m/(2\,k\,T))$, where k is the Boltzmann constant and T the temperature. The speed $V = \|\boldsymbol{V}\|$ is consequently distributed according to the Maxwell–Boltzmann distribution, which is related to the gamma distribution

$$\begin{aligned}\mathbb{P}\left(v \leq V < v + \mathrm{d}v\right) = f_V(v)\,\mathrm{d}v &= 4\,\pi\left(\frac{m}{2\,\pi\,k\,T}\right)^{3/2} v^2 \mathrm{e}^{-\dfrac{m\,v^2}{2\,k\,T}}\,\mathrm{d}v \\ &= \mathrm{Gam}\left(v^2\middle|\frac{3}{2}, \frac{m}{2\,k\,T}\right)\mathrm{d}v^2.\end{aligned}$$

https://doi.org/10.1017/9781108635349.003 Published online by Cambridge University Press

Example 2.4 Escaping Helium

Molecules can escape the atmosphere if their velocity exceeds the escape velocity and they are pointing in the right direction. This is known as Jeans' escape. The gravitational escape velocity of an object from a mass M at a radius r from the centre of the mass is given by

$$v_e = \sqrt{\frac{2GM}{r}}.$$

For a molecule 500 km above earth this is around 10.75 km/s. The upper level of the atmosphere is known as the exosphere, which starts at the exobase at a height of around 500 km. In the exosphere the mean free path of a gas molecule is sufficiently large that a molecule can easily escape the gravitational pull of earth if it has sufficient velocity. The temperature of atmospheric gas is surprising large at around 1600 K. The velocity for hydrogen and helium is given by the Maxwell–Boltzmann distribution. Figure 2.7 shows the distribution of velocities for both molecular hydrogen and helium. Although small, there is a sufficiently high probability of reaching the escape velocity for the escape of hydrogen and helium to be important. Although hydrogen is lost to space, most of it is retained as it forms molecules with heavy atoms (e.g. water, H_2O), however, helium does not form any molecules and so will, over time, become lost into outer space. The presence of helium in the atmosphere is the result of radioactive alpha decays. The concentration of helium in the atmosphere (5.2 parts per billion) is determined by an equilibrium between its production through alpha decays and its loss from the atmosphere through Jeans' escape.

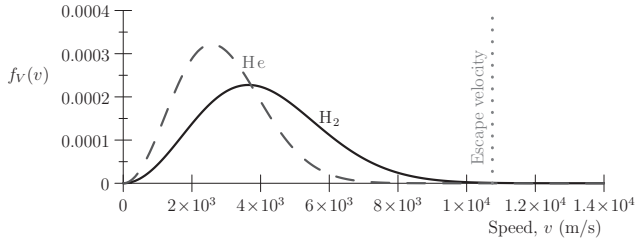


Figure 2.7 Distribution of velocity of hydrogen and helium molecules at 1600 K.

Yet another distribution obtained by making a suitable change of variable is the Weibull distribution. If $Y \sim \text{Exp}(1) = \text{Gam}(1, 1)$ then the random variable $X = \sqrt[k]{\lambda Y}$ is distributed according to a Weibull probability density:

$$\text{Wei}(x|\lambda, k) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} \mathrm{e}^{-(x/\lambda)^k}.$$

The mean of the Weibull distribution is $\Gamma(1 + 1/k)$ and the variance is $\lambda^2(\Gamma(1 + 2/k) - \Gamma^2(1 + 1/k))$. Weibull distributions can also be used to fit data involving

https://doi.org/10.1017/9781108635349.003 Published online by Cambridge University Press

positive real random variables. It is slightly less convenient than a gamma distribution, although easier enough to fit numerically. It provides slightly differently shaped density profiles to the gamma distribution. This is explored in Exercise 2.6.

Although gamma distributions are not so well known, as these many examples illustrate they deserve to be widely appreciated.

2.2.3 Beta Distribution

The beta distribution is a two-parameter continuous distribution that is defined in the interval $[0, 1]$. It is therefore useful for modelling situations where the random variable lies in a range. It is defined by

$$\text{Bet}(x|a, b) = \frac{x^{a-1}(1-x)^{b-1}}{\text{B}(a, b)} \tag{2.7}$$

where $\text{B}(a, b)$ is the beta function (see Appendix 2.B) given by

$$\text{B}(a, b) = \int_0^1 x^{a-1}(1-x)^{b-1}\text{d}x = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

The mean and variance of the beta distribution are $a/(a+b)$ and $a\,b/((a+b)^2(a+b+1))$, respectively. Examples of the beta probability density functions are shown in Figure 2.8.

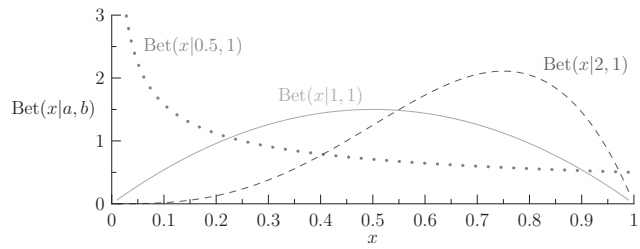


Figure 2.8 Examples of the beta distribution for $(a, b) = (0.5, 1)$ (solid line), $(a.b) = (2, 2)$ (dotted), and $(a, b) = (4, 2)$ (dashed).

A typical application of the beta distribution is to model an unknown probability. The uncertainty might be because you don't know what the value of probability is. For example, you might want to model the probability of a cell dividing in the next hour. In this case, there is some fixed probability p, but you don't know it. To model your uncertainty you can treat p as a random variable that takes some value in the interval $[0, 1]$. Alternatively, you might have different types of cells with different probabilities of dividing. Here, the uncertainty arises because you don't know which type of cell you are looking at. In this case, you are modelling the distribution of p in a population of cells. The beta distribution has a limited parametric form, nevertheless it is sufficiently flexible that it can fit many observed distributions for quantities bounded in an interval quite well, provided the distributions are single peaked (unimodal).

https://doi.org/10.1017/9781108635349.003 Published online by Cambridge University Press

2.2.4 Cauchy Distribution

There are a large number of other continuous distributions, many of which are rather esoteric. However, one type of distribution which you need to be aware of are those with *long tails* – that is, with a significant probability of drawing a sample which is many standard deviations away from the mean. (These are also, perhaps more accurately called *thick-tailed distributions*, since they are usually characterised by a power-law fall-off rather than an exponential fall-off in probability.) A classic example of a distribution with very long (thick) tails is the *Cauchy distribution* (aka Cauchy–Lorentz, Lorentzian, Breit–Wigner), defined through the probability density

$$\mathrm{Cau}(x) = \frac{1}{\pi(1+x^2)}. \tag{2.8}$$

The median and mode of the Cauchy distribution is zero, but rather shockingly the distribution has no mean or variance. That is, if X is drawn from Cau then the improper integrals $\mathbb{E}\left[X\right]$ and $\mathbb{E}\left[X^2\right]$ diverge. We will see that distributions like this behave rather differently to the other distributions we have looked at so far. The Probability Distribution Function (PDF) for the Cauchy distribution is shown in Figure 2.9.

It is tempting to assume the mean must be zero by symmetry. Don't be tempted!

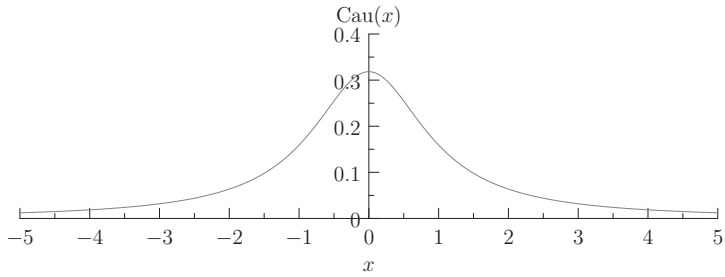


Figure 2.9 The Cauchy distribution.

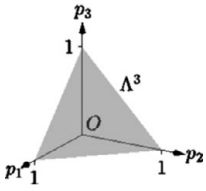
2.3 Multivariate Distributions

So far we have considered distributions involving a single random variable. Often we have situations where there are many correlated random variables. Distributions that describe more than one random variable are known as *multivariate distributions*, in contrast to distributions of a single variable, which are known as *univariate distributions*. There are multivariate extensions for most univariate distributions, although they often become rather too complex to work with. However, there are three well-known and useful multivariate distributions that are relatively easy to work with: the multinomial distribution, the multivariate normal distribution, and the Dirichlet distribution.

2.3.1 Multinomial Distribution

The *multinomial distribution* is the generalisation of the binomial distribution to more than two classes. We assume that we have k classes. The probability

https://doi.org/10.1017/9781108635349.003 Published online by Cambridge University Press

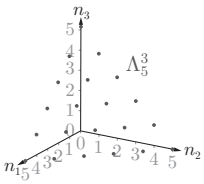


of drawing a sample from class i is given by p_i. Thus the model is described by a vector of probabilities $\boldsymbol{p} = (p_1, p_2, \ldots, p_k)$, with $p_i \geq 0$ for all i and $\sum_i p_i = 1$. The vector of probabilities satisfying these constraints live in the $(k-1)$-*dimensional unit simplex*

$$\Lambda^k = \left\{ \boldsymbol{p} = (p_1, p_2, \ldots, p_k) \,\middle|\, \forall i,\ p_i \geq 0 \text{ and } \sum_{i=1}^{k} p_i = 1 \right\}. \tag{2.9}$$

Note that the simplex is a $(k-1)$-dimensional surface that lives in a k-dimensional space. Suppose we draw a sample of n objects without replacement and we wish to know what is the probability that we have drawn n_1 objects from class 1, n_2 objects from class 2, etc. This probability, $\mathbb{P}\left(\boldsymbol{N} = \boldsymbol{n}\right)$, is given by the multinomial distribution

$$\text{Mult}(\boldsymbol{n}|n, \boldsymbol{p}) = n! \prod_{i=1}^{k} \frac{p_i^{n_i}}{n_i!} \left[\!\left[\boldsymbol{n} \in \Lambda_n^k \right]\!\right] \tag{2.10}$$



where we use the notation $\left[\!\left[\textit{predicate} \right]\!\right]$ to be the *indicator function* as defined in Section 1.5.1 (note that the bold $\boldsymbol{n}$ signifies a vector with components n_i equal to the number of samples in class i, while the italic n denotes the total number of samples). The set Λ_n^k is the *discrete simplex* defined by

$$\Lambda_n^k = \left\{ \boldsymbol{n} = (n_1, n_2, \ldots n_k) \,\middle|\, \forall i,\ n_i \in \{0, 1, 2, \ldots\} \text{ and } \sum_{i=1}^{k} n_i = n \right\}. \tag{2.11}$$

The mean of the multivariate distribution is $\mathbb{E}\left[\boldsymbol{N}\right] = n\,\boldsymbol{p}$. For multivariate distributions you not only have a variance for each variable $\mathbb{V}\text{ar}\left[N_i\right] = \mathbb{E}\left[N_i^2\right] - \mathbb{E}\left[N_i\right]^2$, you also have a covariance between variables $C_{ij} = \mathbb{E}\left[N_i\, N_j\right] - \mathbb{E}\left[N_i\right]\mathbb{E}\left[N_j\right]$. In general, the *second-order statistics* for a multivariate distribution are described by a *covariance matrix*, $\mathbf{C}$, defined as

$$\mathbf{C} = \mathbb{C}\text{ov}\left[\boldsymbol{N}\right] = \mathbb{E}\left[\boldsymbol{N}\,\boldsymbol{N}^{\mathsf{T}}\right] - \mathbb{E}\left[\boldsymbol{N}\right]\mathbb{E}\left[\boldsymbol{N}^{\mathsf{T}}\right].$$

A positive semi-definite matrix has the property that for any vector $\boldsymbol{x}$

$$\boldsymbol{x}^{\mathsf{T}}\mathbf{C}\,\boldsymbol{x} \geq 0.$$

The covariance matrix is both symmetric and positive semi-definite. For the multinomial distribution the covariance between N_i and N_j is given by

$$\mathbb{C}\text{ov}\left[N_i, N_j\right] = C_{ij} = \mathbb{E}\left[N_i\, N_j\right] - \mathbb{E}\left[N_i\right]\mathbb{E}\left[N_j\right] = n\left[\!\left[i = j \right]\!\right] p_i - n\, p_i\, p_j$$

or in matrix form

$$\mathbf{C} = n\,\left(\text{diag}(\boldsymbol{p}) - \boldsymbol{p}\,\boldsymbol{p}^{\mathsf{T}}\right),$$

where $\text{diag}(\boldsymbol{p})$ is a diagonal matrix with elements p_i.

The random variables N_i are not independent since their sum adds up to n. A consequence is that each row (or column) of the covariance matrix sums to zero. The multinomial distribution for just two variables only has one degree of freedom (i.e. given p_1 then $p_2 = 1 - p_1$) and in this case the multinomial reduces to the binomial distribution. With three variables, the multinomial is sometimes referred to as the trinomial distribution.

https://doi.org/10.1017/9781108635349.003 Published online by Cambridge University Press

Multinomial distributions are fairly common. Suppose, for example, you had a (possibly biased) dice which you rolled n times. Letting N_i for $i = 1, 2, \ldots, 6$ denote the number of times the dice lands on i, then the probability of the outcome $\boldsymbol{N} = \boldsymbol{n} = (n_1, n_2, \ldots, n_6)$ is given by the multinomial $\mathbb{P}\left(\boldsymbol{N} = \boldsymbol{n}\right) = \text{Mult}(\boldsymbol{n}|n, \boldsymbol{p})$, where the components of the vector $\boldsymbol{p} = (p_1, p_2, \ldots, p_6)$ describe the probability of each possible outcome of a dice roll.

2.3.2 Multivariate Normal Distribution

The most commonly used multivariate distribution for continuous variables is the multivariate normal distribution defined as

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \mathrm{e}^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}, \tag{2.12}$$

which has mean vector $\mathbb{E}\left[\boldsymbol{X}\right] = \boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. A two-dimensional normal distribution is shown in Figure 2.10. Like its univariate counterpart, the multivariate normal (or Gaussian) distribution can be manipulated analytically. This can be a somewhat complicated or awkward business requiring some practice, but it pays off handsomely. A large number of state-of-the-art algorithms from Gaussian processes to Kalman filters rely on being able to manipulate normal distributions analytically. We discuss the multivariate normal distribution in more detail in Section 5.6 on page 96. Applications of multivariate normal distribution reoccur throughout this book.

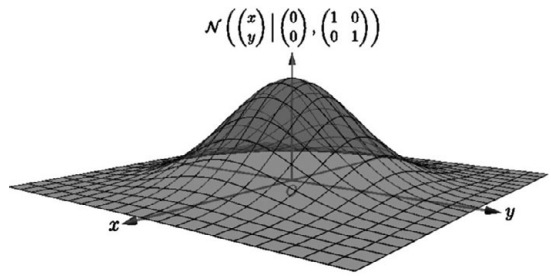


Figure 2.10 A two-dimensional normal distribution.

2.3.3 Dirichlet Distribution

Although the most common multivariate distributions by far are the multinomial and multivariate normal distributions, there exist many others. A particularly convenient distribution for describing a vector of random variables defined on the unit simplex, Λ^k, is the Dirichlet distribution, defined as

$$\text{Dir}(\boldsymbol{x}|\boldsymbol{\alpha}) = \Gamma(\alpha_0) \prod_{i=1}^{k} \frac{x_i^{\alpha_i - 1}}{\Gamma(\alpha_i)}, \tag{2.13}$$

https://doi.org/10.1017/9781108635349.003 Published online by Cambridge University Press

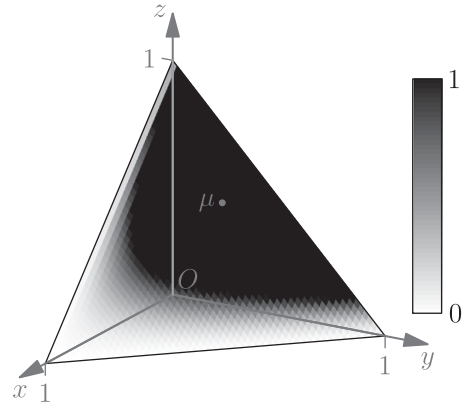


Figure 2.11 A three-dimensional Dirichlet distribution, $\mathrm{Dir}(\boldsymbol{X} = (x, y, z) | \boldsymbol{\alpha} = (1, 2, 3))$. Note that this distribution is defined on the simplex.

where $\alpha_0 = \sum_{i=1}^{n} \alpha_k$. The means are equal to $\mathbb{E}\left[X_i\right] = \alpha_i / \alpha_0$ and the covariance

$$C_{i,j} = \mathbb{E}\left[X_i\, X_j\right] - \mathbb{E}\left[X_i\right] \mathbb{E}\left[X_j\right] = \frac{\alpha_i(\alpha_0 [\![i = j]\!] - \alpha_j)}{\alpha_0^2(\alpha_0 + 1)}.$$

An example of a Dirichlet distribution with three variables is shown in Figure 2.11.

Suppose you have a dice and you are not sure whether it is biased. You could model your uncertainty about the probability of rolling any number using a Dirichlet distribution. The random variables X_i, drawn from $\mathrm{Dir}(\boldsymbol{X}, \boldsymbol{\alpha})$, are not all independent as their sum adds up to one. In the two component case there is only one independent variable and the Dirichlet distribution reduces to a beta distribution.

2.4 Exponential Family⋆

⋆ *Warning, this is more advanced material than the rest of this chapter. It can be skipped.*

Although distributions vary considerably, they also share many properties, sometimes more so than is immediately obvious. One very important family of distributions which share similar properties is the exponential family. These are distributions which can be written in the form

$$f_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\eta}) = g(\boldsymbol{\eta})\; h(\boldsymbol{x})\; \mathrm{e}^{\boldsymbol{\eta}^{\mathsf{T}} \boldsymbol{u}(\boldsymbol{x})}, \tag{2.14}$$

where $\boldsymbol{\eta}$ are *natural parameters* of the distribution, $g(\boldsymbol{\eta})$ and $h(\boldsymbol{x})$ are scalar functions, and $\boldsymbol{u}(\boldsymbol{x})$ is a vector function (i.e. a function for each natural parameter). The distribution can be either for a single random variable or for a random vector, e.g. in the case of a multinomial distribution. The importance of the exponential family is that many properties are known to hold true for distributions belonging to this family. Thus, once we know a distribution belongs to this family we know many of its properties.

It is not immediately obvious, though, which distributions are in the exponential family, because they are often written in ways that don't appear to fit the standard form. Examples of distributions that belong to the exponential family

https://doi.org/10.1017/9781108635349.003 Published online by Cambridge University Press

Distribution	$\boldsymbol{\eta}$	$\boldsymbol{u}(\boldsymbol{x})$	$h(\boldsymbol{x})$	$g(\boldsymbol{\eta})$
$\text{Bern}(x\|\mu) = \mu^x(1-\mu)^{1-x}$	$\log\left(\frac{\mu}{1-\mu}\right)$	x	1	$\frac{1}{1+\mathrm{e}^{\eta}}$
$\mathcal{N}(x\|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{-\frac{(x-\mu^2)}{2\sigma^2}}$	$\frac{1}{\sigma^2}\begin{pmatrix}\mu\\-1/2\end{pmatrix}$	$\begin{pmatrix}x\\x^2\end{pmatrix}$	$\frac{1}{\sqrt{2\pi}}$	$\sqrt{-2\eta_2}\,\mathrm{e}^{\eta_1^2/(4\eta_2)}$
$\text{Gam}(x\|a,b) = \frac{b^a x^{a-1}\mathrm{e}^{-b\,x}}{\Gamma(a)}$	$\begin{pmatrix}-b\\a-1\end{pmatrix}$	$\begin{pmatrix}x\\\log(x)\end{pmatrix}$	1	$\frac{(-\eta_1)^{\eta_2+1}}{\Gamma(\eta_2+1)}$
$\text{Bet}(x\|a,b) = \frac{x^{a-1}(1-x)^b}{B(a,b)}$	$\begin{pmatrix}a-1\\b-1\end{pmatrix}$	$\begin{pmatrix}\log(x)\\\log(1-x)\end{pmatrix}$	1	$\frac{1}{B(\eta_1+1,\eta_2+1)}$
$\text{Dir}(\boldsymbol{x}\|\boldsymbol{\alpha}) = \Gamma\left(\sum_{i=1}^n \alpha_i\right)\prod_{i=1}^n \frac{x_i^{\alpha_i}}{\Gamma(\alpha_i)}$	$\begin{pmatrix}\alpha_1-1\\\vdots\\\alpha_n-1\end{pmatrix}$	$\begin{pmatrix}\log(x_1)\\\vdots\\\log(x_n)\end{pmatrix}$	1	$\frac{\Gamma(n+\sum_{i=1}^n \eta_i)}{\prod_{i=1}^n \Gamma(\eta_i+1)}$

Table 2.1 Examples of distributions belonging to the exponential family. $\text{Bern}(x|\mu)$ is a Bernoulli distribution, which we discuss in Section 4.1.

are shown in Table 2.1. A fuller discussion of the exponential family can be found in Duda et al. (2001).

An important property of members of the exponential family is a relation between the natural parameters and an expectation. To see this we start from the normalisation condition

$$g(\boldsymbol{\eta}) \int h(\boldsymbol{x})\,\mathrm{e}^{\boldsymbol{\eta}^\mathsf{T}\boldsymbol{u}(\boldsymbol{x})}\,\mathrm{d}\,\boldsymbol{x} = 1.$$

Differentiating both sides with respect to the natural parameters, $\boldsymbol{\eta}$,

$$\boldsymbol{\nabla} g(\boldsymbol{\eta}) \int h(\boldsymbol{x})\,\mathrm{e}^{\boldsymbol{\eta}^\mathsf{T}\boldsymbol{u}(\boldsymbol{x})}\,\mathrm{d}\,\boldsymbol{x} + g(\boldsymbol{\eta}) \int h(\boldsymbol{x})\,\mathrm{e}^{\boldsymbol{\eta}^\mathsf{T}\boldsymbol{u}(\boldsymbol{x})}\,\boldsymbol{u}(\boldsymbol{x})\,\mathrm{d}\,\boldsymbol{x} = 0.$$

Rearranging and making use of the normalisation condition we find

$$\frac{-1}{g(\boldsymbol{\eta})}\boldsymbol{\nabla} g(\boldsymbol{\eta}) = g(\boldsymbol{\eta}) \int h(\boldsymbol{x})\,\mathrm{e}^{\boldsymbol{\eta}^\mathsf{T}\boldsymbol{u}(\boldsymbol{x})}\,\boldsymbol{u}(\boldsymbol{x})\,\mathrm{d}\,\boldsymbol{x}$$

or, equivalently,

$$-\boldsymbol{\nabla}\log\big(g(\boldsymbol{\eta})\big) = \mathbb{E}\left[\boldsymbol{u}(\boldsymbol{x})\right].$$

Thus, the expectation of $\boldsymbol{u}(\boldsymbol{x})$ can be found by taking a derivative of $\log\big(g(\boldsymbol{\eta})\big)$. Furthermore, the covariance and higher order moments can be obtained by taking higher order derivatives of $g(\boldsymbol{\eta})$.

A related property is to do with the maximum likelihood estimate of the natural parameters. Given a collection of independent data points, $\mathcal{D} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$, the likelihood of the data is equal to

$$f(\mathcal{D}|\boldsymbol{\eta}) = \left(\prod_{i=1}^n h(\boldsymbol{x}_i)\right) g^n(\boldsymbol{\eta})\,\mathrm{e}^{\boldsymbol{\eta}^\mathsf{T}\sum_{j=1}^n \boldsymbol{u}(\boldsymbol{x}_j)}.$$

The maximum likelihood estimator for the natural parameters, $\hat{\boldsymbol{\eta}}$, satisfies

$$\boldsymbol{\nabla} f(\mathcal{D}|\hat{\boldsymbol{\eta}}) = 0$$

https://doi.org/10.1017/9781108635349.003 Published online by Cambridge University Press

or

$$-\nabla \log\bigl(g(\hat{\boldsymbol{\eta}})\bigr) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{u}(\boldsymbol{x}_n).$$

From this we can solve for the maximum likelihood estimate $\hat{\boldsymbol{\eta}}$. Thus, the only statistics needed for the maximum likelihood estimate of the natural parameters are the components of the vector $\sum_{i=1}^{n} \boldsymbol{u}(\boldsymbol{x}_n)$. These are therefore *sufficient statistics* for any member of the exponential family. We discuss maximum likelihood estimators and sufficient statistics in more detail in Chapter 4.

Although there are a plethora of probability distributions, a few of them are so common and important that they simply can't be ignored. Of these the binomial and normal (or Gaussian) distributions stand out as particularly important. In the second rank sits the hypergeometric, Poisson, gamma, beta, multinomial, and Dirichlet distributions. You also need to be aware that some distributions can have very long tails and nasty properties. Although not very frequently met in practice, the Cauchy distribution is a particularly pretty example of a long-tailed distribution. We'll meet other long-tailed distributions along the way. Appendix B on page 445 provides tables showing the properties of some of the more commonly encountered distributions.

Additional Reading

A useful table of results for different distributions can be found in the compendium of mathematical formula by Abramowitz and Stegun (1964). If you know which distribution you are interested in, then performing a Google or Wikipedia search on the distribution is a very quick way to find most of the common relationships that you might be interested in.

Exercise for Chapter 2

Exercise 2.1 (answer on page 396)

What distribution might you use to model the following situations:

i. the proportion of the gross national product (GDP) from different sectors of the economy;
ii. the probability of three buses arriving in the next five minutes;
iii. the length of people's stride;
iv. the salary of people;
v. the outcome of a roulette wheel spun many times;
vi. the number of sixes rolled in a fixed number of trials; or
vii. the odds of a particular horse winning the Grand National.

(Note that there is not necessarily a single correct answer.)

https://doi.org/10.1017/9781108635349.003 Published online by Cambridge University Press

Exercise 2.2 (answer on page 397)

Assume that one card is chosen from an ordinary pack of cards. The card is then replaced and the pack shuffled. This is repeated 10 times. What is the chance that an ace is drawn exactly three times?

Exercise 2.3 (answer on page 397)

Assume that a pack of cards is shuffled and 10 cards are dealt. What is the probability that exactly three of the cards are aces?

Exercise 2.4 (answer on page 397)

Show that the hypergeometric distribution, $\mathrm{Hyp}(k|N,m,n)$, converges to a binomial distribution, $\mathrm{Bin}(k|n,p)$ in the limit where N and m go to infinity in such a way that $m/N = p$. Explain why this will happen in words.

Exercise 2.5 (answer on page 398)

In the UK National Lottery players choose six numbers between 1 to 59. On draw day six numbers are chosen and the players who correctly guess two or more of the drawn numbers win a prize. The prizes increase substantially as you guess more of the chosen numbers. Write down the probability of guessing k balls correctly using the hypergeometric distribution and compute the probabilities for k equal 2 to 6.

Exercise 2.6 (answer on page 398)

Show that if $Y \sim \mathrm{Exp}(1)$ then the random variable $X = \lambda \sqrt[k]{Y}$ (or $Y = (X/\lambda)^k$) is distributed according to the Weibull density

$$\mathrm{Wei}(x|\lambda,k) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} \mathrm{e}^{-(x/\lambda)^k}.$$

Plot the Weibull densities $\mathrm{Wei}(x|1,k)$ and the gamma distribution with the same mean and variance

$$\mathrm{Gam}\left(\frac{\Gamma^2(1+1/k)}{\Gamma(1+2/k)-\Gamma^2(1+1/k)}, \frac{\Gamma(1+2/k)}{\Gamma(1+2/k)-\Gamma^2(1+1/k)}\right)$$

for $k = 1/2, 1, 2$, and 5.

Appendix 2.A The Gamma Function

The gamma function, $\Gamma(z)$, occurs frequently in probability. For $\Re(z) > 0$ (i.e. the real part of z is positive), the gamma function is defined by the integral

$$\Gamma(z) = \int_0^\infty x^{z-1}\mathrm{e}^{-x}\mathrm{d}x. \tag{2.15}$$

Using integration by parts (assuming $z > 1$) we obtain the relationship

$$\begin{aligned}\Gamma(z) &= \left[-x^{z-1}\mathrm{e}^{-x}\right]_0^\infty + (z-1)\int_0^\infty x^{z-2}\mathrm{e}^{-x}\mathrm{d}x \\ &= (z-1)\Gamma(z-1).\end{aligned}$$

$$\int_a^b u\,\frac{dv}{dx}\,dx = [u\,v]_a^b - \int_a^b v\,\frac{du}{dx}\,dx$$

Since

$$\Gamma(1) = \int_0^\infty \mathrm{e}^{-x}\mathrm{d}x = 1$$

https://doi.org/10.1017/9781108635349.003 Published online by Cambridge University Press

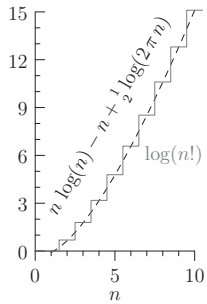
Gauss much more sensibly defined the Pi-function

$$\Pi(z) = \int_0^\infty x^z \mathrm{e}^{-x} dx$$

so that $\Pi(n) = n!$. *Alas, history left us with the gamma function.*

we find for integer, n, that $\Gamma(n) = (n-1)\Gamma(n-1) = (n-1)(n-2)\cdots 2 \cdot 1 = (n-1)!$. Thus, the gamma function is intimately related to factorials (although annoyingly with an offset of 1). As the gamma function increases so fast, it will tend to cause overflows or underflows in numerical calculations if used in its raw form. To overcome this it is usual to work with $\log(\Gamma(z))$. In C-based programming languages this function is called `lgamma`. It is also useful for computing factorials. For example, to compute the binomial coefficient $\binom{n}{k}$ we can use

$$\texttt{exp(lgamma(n + 1) - lgamma(k + 1) - lgamma(n - k + 1))}.$$



The gamma distribution is very well approximated by Stirling's approximation

$$\Gamma(z) = \sqrt{\frac{2\pi}{z}} \left(\frac{z}{e}\right)^z \left(1 + O\left(\frac{1}{z}\right)\right).$$

For factorials this is equivalent to

$$n! \approx \left(\frac{n}{e}\right)^n \sqrt{2\pi n}.$$

In proving theorems involving factorials it is occasionally useful to use a bound provided by Stirling's approximation

$$\sqrt{2\pi}\, n^{n+1/2} e^{-n} \leq n! \leq e\, n^{n+1/2} e^{-n}.$$

Although the integral in Equation (2.15) is only defined for $\Re(z) > 0$, the gamma function can be defined everywhere in the complex plane except at $a = 0, -1, -2, \cdots$, where the function diverges. There are a number of relationships between the gamma function at different values that often help to simplify formulae. We have already seen that $\Gamma(a) = (a-1)\Gamma(a-1)$. Another important relationship is *Euler's reflection formula*

$$\Gamma(1-z)\,\Gamma(z) = \frac{\pi}{\sin(\pi z)}$$

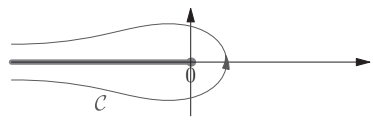
and the *duplication formula*

$$\Gamma(z)\,\Gamma\left(z + \frac{1}{2}\right) = 2^{1-2z}\,\sqrt{\pi}\,\Gamma(2z).$$

For those readers with a more mathematical background, a formula, due to Hermann Hankel, which is occasionally useful is an integral form for the reciprocal of the gamma function in terms of a contour integration

$$\frac{1}{\Gamma(z)} = \frac{1}{2\pi \mathrm{i}} \int_{\mathcal{C}} x^{-z}\, \mathrm{e}^x\, \mathrm{d}x,$$

where $\mathcal{C}$ is a path that starts at $-\infty$ below the branch cut, goes around 0, and returns to $-\infty$ above the branch cut.



https://doi.org/10.1017/9781108635349.003 Published online by Cambridge University Press

Derivatives of the gamma function arise when computing maximum likelihood estimates for distributions such as the gamma distribution. Rather than using the derivative of the gamma function it is more usual (and often more convenient) to consider the derivative of the (natural) logarithm of the gamma function (throughout this book we use $\log(x)$ to denote the natural logarithm of x). This is known as the *digamma function* which is usually written as

$$\psi(z) = \frac{\mathrm{d}\log(\Gamma(z))}{\mathrm{d}z} = \frac{1}{\Gamma(z)}\frac{\mathrm{d}\Gamma(z)}{\mathrm{d}z}.$$

Although the digamma function is not part of the standard C library, it exists in many numerical packages. The derivative of the digamma function is known as the trigamma function $\psi'(z)$, while higher order derivatives are known as polygamma functions. Like the gamma function the polygamma functions, and particularly the digamma function, have interesting properties that are well documented (e.g. in most tables of mathematical functions as well as in Wikipedia, etc.).

The incomplete gamma functions are defined for $\Re(a) > 0$ by

$$\gamma(a, z) = \int_0^z x^{a-1}\,\mathrm{e}^{-x}\,\mathrm{d}x, \qquad \Gamma(a, z) = \int_z^\infty x^{a-1}\,\mathrm{e}^{-x}\,\mathrm{d}x,$$

with $\gamma(a, z) + \Gamma(a, z) = \Gamma(a)$. The normalised incomplete gamma functions are defined as

$$P(a, z) = \frac{\gamma(a, z)}{\Gamma(a)}, \qquad Q(a, z) = \frac{\Gamma(a, z)}{\Gamma(a)},$$

with $P(a, z) + Q(a, z) = 1$.

Appendix 2.B The Beta Function

The beta function is defined (for $\Re(a), \Re(b) > 0$) through the integral

$$\mathrm{B}(a, b) = \int_0^1 x^{a-1}\,(1-x)^{b-1}\mathrm{d}x.$$

Remarkably, it is related to the gamma function through

$$\mathrm{B}(a, b) = \frac{\Gamma(a)\,\Gamma(b)}{\Gamma(a+b)}.$$

To prove this we start from

$$\Gamma(a)\,\Gamma(b) = \int_0^\infty u^{a-1}\,\mathrm{e}^{-u}\mathrm{d}u \int_0^\infty v^{b-1}\,\mathrm{e}^{-v}\mathrm{d}v = \int_0^\infty \int_0^\infty u^{a-1}\,v^{b-1}\,\mathrm{e}^{-u-v}\mathrm{d}u\,\mathrm{d}v.$$

We make the change of variables $u = z\,t$, $v = z\,(1-t)$, with $t \in [0, 1]$ and $z \in [0, \infty]$ (in the u-v plane t determines the angle and z the magnitude). The Jacobian is given by

$$\frac{\partial(u, v)}{\partial(t, z)} = \begin{vmatrix} z & t \\ -z & (1-t) \end{vmatrix} = z\,(1-t) + z\,t = z.$$

https://doi.org/10.1017/9781108635349.003 Published online by Cambridge University Press

With this change of variables we find

$$\begin{aligned}\Gamma(a)\,\Gamma(b) &= \int_0^1 \int_0^\infty (z\,t)^{a-1}\,(z\,(1-t))^{b-1}\,\mathrm{e}^{z\,t-z\,(1-t)}\,z\,\mathrm{d}t\,\mathrm{d}z \\ &= \int_0^1 t^{a-1}\,(1-t)^{b-1}\mathrm{d}t\,\int_0^\infty z^{a+b-1}\mathrm{e}^{-z}\mathrm{d}z = \mathrm{B}(a,b)\,\Gamma(a+b).\end{aligned}$$

The incomplete beta function is defined as

$$\mathrm{B}_z(a,b) = \int_0^z x^{a-1}\,(1-x)^{b-1}\,\mathrm{d}x.$$

The normalised incomplete beta function is defined as $I_z(a,b) = \mathrm{B}_z(a,b)/\mathrm{B}(a,b)$.

https://doi.org/10.1017/9781108635349.003 Published online by Cambridge University Press