1 Probability basics

This chapter reviews the basic ideas of probability and statistics that we will use in the rest of this book. I will outline some fundamental concepts, introduce some of the most common discrete and continuous probability distributions, and work through a few examples. But we'll start with an example that illustrates the role of information in solving problems.

1.1 The three doors problem

You are a contestant in a game show, presented with three closed doors. Behind one of the doors – chosen at random without your knowledge – is a prize car. Behind each of the other doors is a goat. Your objective is to reveal and thus win the car (the assumption is you're less keen on goats). You first select a door at random, which you do not open. The game show host, who knows where the car is, then opens one of the *other* doors to reveal a goat (she would never show you the car). She then gives you the opportunity to change your choice of door to the other closed one. Do you change?

This is a classic inference problem, also known as the Monty Hall problem. If you've not encountered it before – and even if you have – do think about it before reading on.

Your initial probability of winning (before any door is opened) is $1/3$. The question is whether this probability is changed by the actions of the game show host. It appears that her actions don't change anything. After all, she will always show you a goat. In that case you may think that your probability of winning remains at $1/3$ whether or not you change doors. Or you may think that once the host has revealed a goat, then because there are only two doors left, your chance of winning has changed to $1/2$ whether you change or not. In either of these cases changing doors does not improve your chances of winning.

In fact, if you change to the other closed door your chance of winning increases to $2/3$. This may seem counter-intuitive, but it is explained by how the host behaves. If you initially selected the door with the car behind it, then her choice of which door to open is random. But if you initially selected a door with a goat behind it, then her choice is not random. She is in fact forced to open the door with the other goat behind it. Let's call the door you initially select "door 1". The three possible arrangements are as follows.

door 1	door 2	door 3	door opened by host	result if staying	result if changing
car	goat	goat	2 or 3	car	goat
goat	car	goat	3	goat	car
goat	goat	car	2	goat	car

In two of the three initial possible choices you would select a goat, so if you change you will win the car in both cases. In the third case you select the car, so will lose if you change doors. Thus your chance of success is $2/3$. If you don't change then you will only win the car if your initial choice was indeed the car, a probability of $1/3$.

The outcome of this inference depends on how the game show host behaves: she always opens one of the other doors to reveal a goat, and does so at random if she can.

1.1.1 Earthquake variation

The fundamental point of the three doors problem is that your understanding of the problem, and therefore your decision of what to do, depends crucially on your model for the door opening (i.e. the game show host's behaviour). A variant on the problem should make this clear.

Just after making your initial choice, and before the game show host has had a chance to respond, there is an earthquake that causes one of the doors you did not select to open, revealing a goat. After recovering from the shock and seeing the goat behind the door, the game show host decides to continue, and offers you the opportunity to change. Do you?

If we assume that the earthquake opened a door at random, then it was just chance that it opened a door with a goat behind it. (Had it revealed the car, presumably the host would have declared the game invalid.) The difference now is that the choice of door to open could never have been forced by your initial selection, as it was in the original game. So the probability that the car lies behind one of the two remaining doors is $1/2$, and your chance of winning is $1/2$ whether you change or not.

If you are not convinced that you should change doors in the original problem, consider the case of one hundred doors, again with only one prize. After making your initial choice (let's define that as door 1 again), the game show host will now open 98 other doors (to reveal a lot of goats), leaving closed doors 1 and, let's say, 76. Would you stay with door 1 or switch to door 76? I shall present a formal analysis at the end of the chapter (section 1.10).

The take-home message of this problem is that not just data, but also background information, are important for drawing a conclusion. The background information here is our understanding of how the game show host operates. This information is in fact vital because the data – being shown a goat – will always be the same. Changing the background information will generally change how we interpret data and therefore what we conclude.

1.2 Probability

1.2.1 Fundamental properties

What is probability?

Probability is a way of quantifying our state of knowledge of a proposition. Examples of propositions are "this is the spectrum of a star", "it will rain tomorrow", "model M is correct", and "the period of oscillation lies between 1 and 1.1 seconds". Our state of knowledge is inseparable from our belief: the probability reflects what *we* think about the proposition, based on information, data, etc., that *we* have. Other people with other information may well assign a different probability.

Let A and B be two propositions which have probabilities $P(A)$ and $P(B)$ respectively. Without being too formal, the following are the essential axioms of probability.[1] Probabilities are real numbers that satisfy

$$0 \leq P(A) \leq 1. \tag{1.1}$$

Denote as A' the complement of A (i.e. A doesn't happen or isn't true). Then

$$P(A') \;=\; 1 - P(A). \tag{1.2}$$

If two propositions A and B are *mutually exclusive* then

$$P(A \text{ or } B) \;=\; P(A) + P(B) \qquad \text{(exclusive propositions)}\,. \tag{1.3}$$

If $\{E_i\}$ is the set of *mutually exclusive* and *exhaustive* propositions (no others are possible), then

$$\sum_i P(E_i) \;=\; 1. \tag{1.4}$$

The *conditional probability* of A occurring given that B is true (e.g. A = "rain", B = "cloudy") is written $P(A\,|\,B)$. The *joint probability* that both are true is written $P(A, B)$. It follows logically that we can factorize this and write it as

$$P(A \text{ and } B) \;\equiv\; P(A, B) \;=\; P(A\,|\,B)P(B) \;=\; P(B\,|\,A)P(A). \tag{1.5}$$

If A and B are mutually exclusive, $P(A, B) = 0$. If A and B are independent, $P(B\,|\,A) = P(B)$ and $P(A\,|\,B) = P(A)$, in which case

$$P(A, B) \;=\; P(A)P(B) \qquad \text{(independent propositions)}. \tag{1.6}$$

Equation 1.5 is simple, yet it is one of the most important for inference. It can be rearranged and written as

$$P(A\,|\,B) \;=\; \frac{P(A, B)}{P(B)}. \tag{1.7}$$

Note that $P(A\,|\,B) \;\neq\; P(B\,|\,A)$. If you're not convinced, take an example: think of

[1] Much has been written on the foundations of probability theory. Two important contributions are by Kolmogorov (1933) and Cox (1946).

$A =$ "being female", $B =$ "being pregnant". Suppose that the average human female gives birth to two children in her life (gestation period $9/12$ of a year) and the life expectancy is 80 years, then for a female chosen at random, $P(B\,|\,A) \simeq 2 \times (9/12)/80 \simeq 0.02$. Different assumptions will give a slightly different result, but unless you work in a maternity ward it will be small. Yet I think all will agree that $P(A\,|\,B) = 1$, for all common definitions of "female" and "pregnant".

In general it must hold that

$$P(A) \;=\; P(A\,|\,B)P(B) + P(A\,|\,B')P(B') \tag{1.8}$$

because B is either true or is not true, so all possibilities are covered. More generally, if $\{B_i\}$ is the set of all possible propositions, i.e. $\sum_i P(B_i) = 1$, then it follows that

$$\begin{aligned} P(A) \;&=\; \sum_i P(A, B_i) \\ &=\; \sum_i P(A\,|\,B_i)P(B_i). \end{aligned} \tag{1.9}$$

This is called the *marginal probability* of A, where "marginal" just means "average", here over the $\{B_i\}$.

1.2.2 Some problems

Drawing cards

We draw two cards from a well-shuffled deck of 52 playing cards. What is the probability they are both aces given that (a) we replace the first card, (b) we don't?

$$\text{(a)}\;\frac{4}{52}\frac{4}{52} \;=\; \frac{1}{169} \;\simeq\; 0.0059. \tag{1.10}$$

This is also the probability of drawing an ace then a king, or a three then a nine, etc. It's only we who attach meaning to it being an ace both times.

$$\text{(b)}\;\frac{4}{52}\frac{3}{51} \;=\; \frac{1}{221} \;\simeq\; 0.0045. \tag{1.11}$$

Firing missiles

A missile has a probability $p = 0.1$ of destroying its target. What is the probability that the target is destroyed when we fire two missiles?

It is tempting to say 0.2. But this is obviously wrong, because if we instead had $p = 0.7$ we would not say the probability is 1.4. Logically this is because we cannot destroy the target twice. Mathematically it's because we do not add probabilities for non-exclusive events. The question we are really asking is "what is the probability that the target is destroyed by either missile or by both missiles?". In the general case of n missiles it would take a while to work out all combinations of some missiles hitting the target and others not. In such problems it is easier to work out the opposite probability, namely the probability that the target is not destroyed at all. This is $(1-p)^n$. The probability that *at least one*

missile hits the target is then $1-(1-p)^n$. For $n=2$ this is 0.19 with $p=0.1$, and 0.91 with $p=0.7$.

When writing the probability that proposition A or B is true, we need to be careful whether the "or" is exclusive or not. Non-exclusive means A or B or both could be true, so it follows that

$$P(A \text{ or } B) = 1 - P(A', B') \qquad \text{(non-exclusive propositions)}. \tag{1.12}$$

This is the missile example, in which the events are also independent so $P(A', B') = P(A')P(B')$ and $P(A') = P(B') = 1-p$. If the events are exclusive, equation 1.3 applies instead.

1.2.3 Frequency

Probabilities are sometimes equated with frequencies. For example, with a deck of 52 cards we could say that the probability of drawing the ace of spades is $1/52$, perhaps because we imagine repeating this process many times. This frequency interpretation is often useful, but it doesn't always work, as we can see with unique events. For example, if we ask what is the probability that a helicopter will make an emergency landing outside my house next 21 May, we cannot resort to frequencies of actual past or even imagined events. We must instead resort to similar events. But what counts as similar is a matter of personal judgement and experience, so has no unique answer. Likewise, when it comes to inferring something from data, such as the probability that the mass of a particle lies in a particular range, then the frequency interpretation is of little help. It is often more useful to think of probability as a degree of belief.

1.2.4 Bayes' theorem and the principle of inference

If we rearrange the equation for joint probabilities (equation 1.5), we get

$$P(B\,|\,A) = \frac{P(A\,|\,B)P(B)}{P(A)}. \tag{1.13}$$

This is called *Bayes' theorem*. It was once referred to as *inverse probability*, because it relates the probability of B given A to the probability of A given B. It is fundamental in inference for the following reason. Taking M as a model (or hypothesis) and D as some data that we have obtained to test the model, then we can write Bayes' theorem as

$$P(M\,|\,D) = \frac{P(D\,|\,M)P(M)}{P(D)}. \tag{1.14}$$

When we set up an experiment we can normally determine $P(D\,|\,M)$, the probability of observing the data (or perhaps a quantity derived from the data) under the assumption that the model is true. For example, we might be looking for the presence of spectral line (D = line is present) under the assumption that it is produced by a particular trace gas in an exoplanet's atmosphere. We then set up a model M for the atmosphere and our measurements, and compute the probability $P(D\,|\,M)$ that this would allow us to observe

the line. Note that this is not the probability that the model is true, given the data. That quantity is $P(M\,|\,D)$ (recall the example above about pregnancy and females). Observing the line does not necessarily indicate the presence of the gas, because there may be other origins of the line (i.e. other models).

To infer $P(M\,|\,D)$ we use Bayes' theorem. Equation 1.14 tells us that we need two other quantities in addition to $P(D\,|\,M)$. The first is $P(M)$, the unconditional probability that the model M is true. This is called the *prior* probability, prior here meaning "prior to using the data". The other quantity we need is $P(D)$, the unconditional probability of the data, meaning its probability independent of any particular model. Let $\{M_i\}$ be the complete set of all models (mutually exclusive and exhaustive). Using equation 1.9 we can write

$$P(D) = \sum_i P(D\,|\,M_i)P(M_i). \tag{1.15}$$

We now have all terms needed to compute $P(M\,|\,D)$, the *posterior probability* that M is true. This differs from the prior probability in that it is conditional on the data. Bayes' theorem underpins much of inference, so I will have a lot more to say about what it is and how we use it. We will return to it again in chapter 3.

1.2.5 Discrete and continuous probability distributions

So far we have considered discrete variables, whereby $P(B_i)$ gives the actual probability of proposition B_i being true or of event B_i occurring. We sometimes refer to the set of probabilities over all events as the *probability mass function* (PMF).

For a continuous variable x we instead deal with the *probability density function* (PDF) $P(x)$. This is a density – probability per unit x – not a probability. $P(x)dx$ is the infinitesimal probability of x in the range x to $x + dx$, so a finite probability is obtained by integrating over a region. Specifically, the probability of x lying between x_1 and x_2 is

$$\int_{x_1}^{x_2} P(x)\,dx. \tag{1.16}$$

Probability is dimensionless: it has no units. Therefore the density $P(x)$ has units $1/x$. Note that $P(x\,|\,y)$ is a PDF in x, not in y. The thing after the bar just gives conditional information. The variables x and y need not have the same units, which is another reason why $P(x\,|\,y) \neq P(y\,|\,x)$ in general. I will use the same symbol – upper case P – to refer to both PMFs and PDFs. It will be clear from the context which I mean.

The range of the variable x over which the PDF (or PMF) is non-zero is called the *support* of the distribution. If it is non-zero over all real values, the distribution is said to have *infinite support*. If it has a bound, e.g. it is non-zero only for positive values, it is referred to as having *semi-infinite support*.

1.2.6 Normalization

A bona fide PDF is always normalized, i.e.

$$\int_{-\infty}^{+\infty} P(x)\,dx \;=\; 1 \tag{1.17}$$

and likewise

$$\sum_i P(x_i) \;=\; 1 \tag{1.18}$$

for discrete distributions. Distributions that cannot be normalized are called *improper*. An example is the uniform distribution with infinite support (section 1.4.4). It is not normalizable because its integral is infinite. We will nonetheless see that improper distributions can be useful in certain contexts.

We will sometimes work with unnormalized distributions, which I will indicate using an asterisk P^*. For an unnormalized PDF $P^*(x)$ the normalization constant is $\int P^*(x)\,dx$, which is independent of x (because we've integrated over it). The normalized PDF is therefore

$$P(x) \;=\; \frac{P^*(x)}{\int P^*(x)\,dx} \tag{1.19}$$

and the normalized discrete distribution (probability mass function) is

$$P(x_i) \;=\; \frac{P^*(x_i)}{\sum_i P^*(x_i)}. \tag{1.20}$$

1.3 Expectation, variance, and moments

We often want to summarize probability distributions with just a few numbers. Two summary statistics of particular interest are the mean, which indicates the location of the distribution, and the standard deviation, which is a measure of its width. Many of the most commonly occurring distributions are fully defined by these two parameters.

The *expectation value* or *mean* of a continuous variable x with a PDF $P(x)$ is defined as

$$E[x] \;=\; \int x\,P(x)\,dx. \tag{1.21}$$

For a discrete variable with PMF $P(x_i)$ it is defined as

$$E[x] \;=\; \sum_i x_i\,P(x_i). \tag{1.22}$$

The expectation value is sometimes written as μ. The expectation value of a function $f(x)$ is defined as

$$E[f(x)] \;=\; \int f(x)\,P(x)\,dx. \tag{1.23}$$

If we have a set of N data points drawn from a continuous distribution, then we can use equation 1.22 as an approximation for equation 1.21. Each data point occurs once, so implicitly $P(x_i) = 1/N$, and

$$E[x] \simeq \frac{1}{N}\sum_{i=1}^{N} x_i = \overline{x}, \tag{1.24}$$

which is called the *sample mean*. The *law of large numbers* says that in the limit $N \to \infty$ the sample mean tends towards the (true) mean.

Some useful properties of the expectation operator follow from its definition. Let x and y be random variables and a a fixed constant.

$$E[a] = a \tag{1.25}$$
$$E[ax] = aE[x] \tag{1.26}$$
$$E[x+y] = E[x] + E[y] \tag{1.27}$$
$$E[E[x]] = E[x] \tag{1.28}$$
$$E[xE[y]] = E[x]E[y] \tag{1.29}$$

but in general $E[xy] \neq E[x]E[y]$.

The *variance* of a random variable x with expectation value $E[x]$ is defined as

$$\begin{aligned}\mathrm{Var}(x) &= \int (x - E[x])^2 P(x)\, dx \\ &= E[(x - E[x])^2] \\ &= E[x^2 - 2xE[x] + E[x]^2] \\ &= E[x^2] - 2E[x]E[x] + E[x]^2 \\ &= E[x^2] - E[x]^2. \end{aligned} \tag{1.30}$$

We can remember this as "the variance is the expectation of the square minus the square of the expectation" (or just "mean square minus square mean"). If we know the true mean μ, then for a set of N data points $\{x_i\}$, the variance can be approximated by

$$\mathrm{Var}(x) \simeq \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2. \tag{1.31}$$

This is an approximation because of the finite sample size. If we don't know μ then we might want to estimate the variance as

$$V_x = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2 = \overline{x^2} - \overline{x}^2. \tag{1.32}$$

However, if we calculate the expected value of this we find

$$E[V_x] = \frac{N-1}{N}\,\mathrm{Var}(x) \tag{1.33}$$

which is not equal to the true variance for finite N. It is therefore a biased estimate; specifically it is an underestimate (see section 2.1 for a discussion of estimator bias). The origin

of this bias is that V_x uses the estimated (sample) mean of the data (equation 1.24), whereas the definition of variance uses the true mean. As we have computed this bias we can correct for it, to give the *sample variance*, which is defined as[2]

$$\widehat{\mathrm{Var}}(x) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2 \tag{1.34}$$

which is an unbiased estimate of the true variance.

The *standard deviation* is defined as

$$\sigma = \sqrt{\mathrm{Var}(x)} \tag{1.35}$$

which for a set of data of known mean is

$$\sigma \simeq \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \tag{1.36}$$

where the approximation symbol has again been used because we have a finite-sized sample. The *sample standard deviation* is the square root of the sample variance (equation 1.34)

$$\hat{\sigma} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}. \tag{1.37}$$

The variance and standard deviation of a set of data are computed in R by the functions `var` and `sd` respectively. They both use the $N-1$ term.

As variance is a squared quantity it tells you nothing about the asymmetry of a distribution. This we can measure with the *skew*

$$\gamma = \frac{1}{N\sigma^3}\sum_{i=1}^{N}(x_i - \mu)^3 \tag{1.38}$$

which is a dimensionless number. A positive value for the skew means the data have an asymmetry about the mean with a tail to positive values. Here μ and σ are the true mean and true standard deviation respectively. In practice we would estimate these from the data, so we should use a slightly different expression for the skew that involves a small correction analogous to the one applied for the sample variance. The *kurtosis* is the next higher power and measures how centrally concentrated a distribution is

$$\kappa = \frac{1}{N\sigma^4}\sum_{i=1}^{N}(x_i - \mu)^4 - 3. \tag{1.39}$$

It is also dimensionless. The -3 is in the definition so that a Gaussian distribution has zero kurtosis. More negative values are more centrally concentrated.

[2] Here I use the hat symbol to distinguish the sample variance from the variance. I also do this in equations 1.37 and 1.66 to indicate sample standard deviation and sample covariance, respectively. More generally the hat symbol is often used to indicate an estimate.

Mean, variance, skew, and kurtosis are related to the first four *moments* of a distribution. The kth moment of $P(x)$ is defined as

$$\int x^k P(x)\,dx \tag{1.40}$$

for integer $k \geq 0$ (the zeroth moment is the normalization). This is sometimes called the kth raw moment to distinguish it from the kth *central moment*, which is defined as

$$\int (x-\mu)^k P(x)\,dx. \tag{1.41}$$

The kth *sample moment* of a set of data $\{x_i\}$ drawn from the distribution is

$$\frac{1}{N}\sum_{i=1}^{N} x_i^k. \tag{1.42}$$

There is no $N-1$ correction term in the definition of these (non-central) sample moments. Although we can always calculate the raw moments of a set of data, not all distributions have defined moments. We shall see an example of this in section 2.3.2.

1.4 Univariate probability distributions

We turn now to some of the most common univariate probability distributions. The first two of these, the binomial and Poisson distributions, describe discrete variables. The others describe continuous variables. There are of course many other distributions, and several more will be defined in later chapters as we encounter them (see the index under "distributions" for a full list). The derivation of some of these distributions relies on combinations and permutations, so you may want to read section 1.7 before proceeding.

1.4.1 Binomial

This probability distribution describes processes in which an event can have only one of two possible outcomes. Examples include tossing a coin, detecting something at a security check, or winning the lottery. Let p be the probability of one event, call it "success"; $1-p$ is the probability of the other event ("failure"). If the trial is repeated independently n times, then we are interested in the probability of getting exactly r successes, which we can label $P(r\,|\,p,n)$. Suppose that the first r trials are successes and the remainder, $n-r$, are all failures. As the trials are independent, the probability of this sequence is just the product of the probabilities, $p^r(1-p)^{n-r}$. This is the probability of just one particular sequence of trials. The number of unique sequences with this probability is the number of ways of selecting r from n (without replacement), which is ${}_nC_r$ (see section 1.7). As these sequences are mutually exclusive we just sum the probability ${}_nC_r$ times. This gives us the

binomial distribution

$$P(r\,|\,p,n) \;=\; {}_nC_r\, p^r(1-p)^{n-r} \;=\; \binom{n}{r} p^r(1-p)^{n-r} \;=\; \frac{n!}{r!(n-r)!}\, p^r(1-p)^{n-r}. \tag{1.43}$$

This is the probability of getting r successes from n trials if the probability of one success is p, where of course $0 \le r \le n$.

Here is an example: $P(r\,|\,p,n)$ is the probability of drawing r red balls blindly from a set of n red and white balls, in which the fraction of red balls is p. The draws are done one by one with replacement (so p remains constant).[3]

Using equations 1.22 and 1.30 we can show that the expected number of successes is

$$E[r] \;=\; np \tag{1.45}$$

(which agrees with our intuition) and the variance is

$$\mathrm{Var}(r) \;=\; np(1-p). \tag{1.46}$$

Examples of the binomial distribution are shown in the left panel of figure 1.1. The right panel shows instead how the probability density varies as a function of p for a given r; this is not a density function in p. The plots were produced using the following R script, which uses the function `dbinom` to compute the densities.

R file: `binomial_distribution.R`

```
##### Plot the binomial distribution

# Plot P vs. r for fixed n for a range of p
n <- 10
r <- 0:n
pseq <- c(0.1, 0.2, 0.5, 0.8, 0.9)
pdf("dbinom1.pdf", 4, 4)
par(mfrow=c(1,1), mgp=c(2.0,0.8,0), mar=c(3.5,3.5,1,1), oma=0.1*c(1,1,1,1))
plot(r, r, type="n", xlim=c(0,max(r)), ylim=c(0,0.4), xlab="r",
     ylab="P(r | p,n)")
for (p in pseq) {
  points(r, dbinom(x=r, size=n, prob=p), pch=20)
  lines(r,  dbinom(x=r, size=n, prob=p), lty=2)
```

[3] If we did the selection without replacement then the probability that there are r red balls in a sample of n balls, which we draw without replacement from an original set of N balls of which m were red, is

$$P(r\,|\,n,m,N) = \frac{\binom{m}{r}\binom{N-m}{n-r}}{\binom{N}{n}} \tag{1.44}$$

which is called the *hypergeometric distribution*. Here $\binom{m}{r}$ is the number of ways of selecting r red balls from the total m red balls, $\binom{N-m}{n-r}$ is the number of ways in which the remaining $n-r$ white balls can be chosen from the $N-m$ white balls, and $\binom{N}{n}$ is the number of ways in which we could select any sample of n balls (of any colour). All of these selections are without replacement.

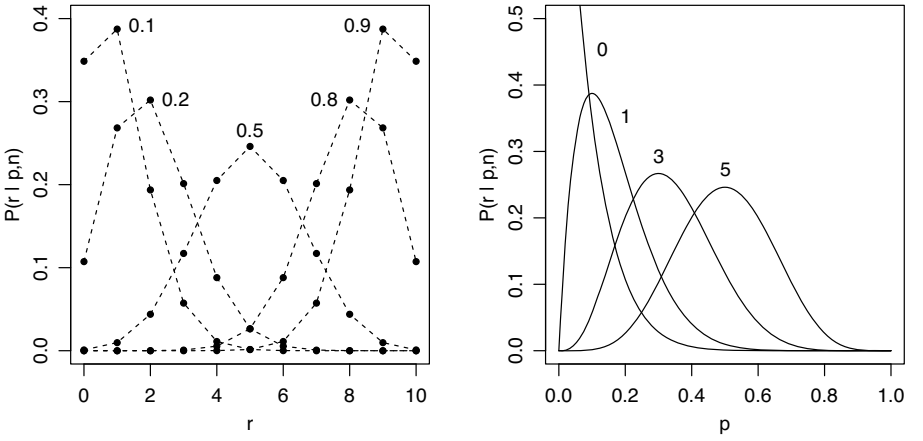

Fig. 1.1 The binomial distribution for $n = 10$. Left: $P(r\,|\,p,n)$ vs r for the five values of p indicated. The distribution is discrete: points are joined with dashed lines just to help identification of the points with common p. Right: $P(r\,|\,p,n)$ vs p for the four values of r indicated. Note that this is still a probability distribution over r, not p.

```
}
text(c(1,2,5,8,9), c(0.39,0.30,0.285,0.30,0.39), pseq, pos=c(4,4,1,2,2))
dev.off()

# Plot P vs. p for fixed n for a range of r
p    <- seq(from=0, to=1, by=0.001)
rseq <- c(0,1,3,5)
n    <- 10
pdf("dbinom2.pdf", 4, 4)
par(mfrow=c(1,1), mgp=c(2.0,0.8,0), mar=c(3.5,3.5,1,1), oma=0.1*c(1,1,1,1))
plot(p, p, type="n", xlim=range(p), ylim=c(0,0.5), xlab="p",
     ylab="P(r | p,n)")
for (r in rseq) {
  lines(p,  dbinom(x=r, size=n, prob=p))
}
text(c(0.08,0.15,0.25,0.45), c(0.45,0.35,0.29,0.27), rseq, pos=4)
dev.off()
```

The binomial distribution is a discrete distribution, so $P(r\,|\,p,n)$ (i.e. `dbinom`) is a probability mass function; it gives actual probabilities.

Example: coin tossing

What is the probability of getting exactly two heads in six tosses of a fair coin?

$$P(r=2|p=1/2,n=6) = \binom{6}{2}\left(\frac{1}{2}\right)^2\left(1-\frac{1}{2}\right)^4 = \frac{15}{64} \simeq 0.23. \qquad (1.47)$$

1.4.2 Poisson

The binomial distribution describes events in which there is a definite event taking place that has a two-way result: it's either a "success" or a "failure"; something happens or it doesn't. Yet a lot of natural processes are only one-way, by which I mean they are only identifiable by having happened. Examples are lightning strikes and α particle emission from a radioactive source. We cannot count non-events because we cannot identify a sequence of events where something is supposed to happen or not. Suppose that on average these events occur at a rate λ, so that λ is the expected number of events in some unit time interval. We would like to find the probability that we get r events in this interval.

We can describe this as the limit of a binomial process. If we divide the time series into n divisions, then in the limit as the divisions become very small such that n gets large, we can write $p = \lambda/n$. The binomial distribution is

$$P(r\,|\,\lambda/n, n) \;=\; \frac{n!}{r!(n-r)!}\left(\frac{\lambda}{n}\right)^r\left(1-\frac{\lambda}{n}\right)^{n-r}. \tag{1.48}$$

We now take n to infinity to produce a continuum of events. As $n \to \infty$ with r finite,

$$\frac{n!}{(n-r)!} \;=\; n(n-1)(n-2)\ldots(n-r+1) \to n^r \tag{1.49}$$

(because each term tends towards n) and

$$\left(1-\frac{\lambda}{n}\right)^{n-r} \;\to\; \left(1-\frac{\lambda}{n}\right)^{n} \;\to\; e^{-\lambda} \tag{1.50}$$

which is a definition of e (Euler's constant). Inserting these two terms into equation 1.48, n^r cancels and we end up with the *Poisson distribution*

$$P(r\,|\,\lambda) \;=\; \frac{e^{-\lambda}\lambda^r}{r!} \quad \text{where} \quad \lambda > 0,\; r \geq 0. \tag{1.51}$$

This is the probability of getting r events if the mean expected number is λ, i.e. $E[r] = \lambda$. Note that although r is an integer, λ does not have to be. An important property of this distribution is that its variance is $\mathrm{Var}(r) = \lambda$, i.e. equal to the mean. It is a discrete distribution so $P(r\,|\,\lambda)$ is a probability mass function.

The following R code plots the distribution (see figure 1.2). The larger λ the closer the Poisson distribution becomes to the Gaussian distribution (notwithstanding the fact that the Poisson distribution is only defined for integer r).

```
pdf("dpois1.pdf", 4, 4)
par(mfrow=c(1,1), mgp=c(2.0,0.8,0), mar=c(3.5,3.5,1,1), oma=0.1*c(1,1,1,1))
r <- 0:20
plot(r, r, type="n", xlim=c(0,max(r)), ylim=c(0,0.4),
  xlab="r", ylab=expression(paste(, "P(r | ", lambda, ")")))
for (lambda in c(1,2,3.5,10)) {
  points(r, dpois(x=r, lambda=lambda))
  lines(r,  dpois(x=r, lambda=lambda), lty=2)
}
dev.off()
```

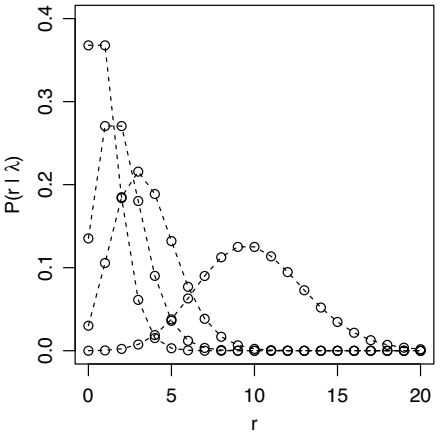

Fig. 1.2 The Poisson distribution for $\lambda = 1, 2, 3.5, 10$ (maxima from left to right). The distribution is discrete: points are joined with dashed lines just to help identify the points with common λ.

A useful property of the Poisson distribution is

$$\begin{aligned} P(r+1\,|\,\lambda) &= \frac{e^{-\lambda}\lambda^{r+1}}{(r+1)!} \\ &= \frac{e^{-\lambda}\lambda^{r}}{r!}\frac{\lambda}{r+1} \\ &= P(r\,|\,\lambda)\frac{\lambda}{r+1} \end{aligned} \tag{1.52}$$

and

$$P(0;\lambda) = e^{-\lambda}. \tag{1.53}$$

It can also be proven that if two Poisson processes have means λ_a and λ_b, then the probability of getting a total of r events from the two processes (without distinguishing which came from where – consider counting radioactive decays from two isotopes) is described by a Poisson distribution with mean $\lambda_a + \lambda_b$.

Example: radioactive decay

Consider a radioactive source with half-life $t_{1/2}$. If n_0 is the initial number of radioactive atoms, then the number of radioactive atoms left after time t is

$$n \;=\; n_0\, e^{-t/\tau} \;=\; n_0\, e^{-\lambda t} \tag{1.54}$$

where $\tau = t_{1/2}/\ln 2$. The mean (expected) number of decays per unit time is $\lambda = 1/\tau$. The theoretical distribution of the number of decays per time is a Poisson distribution. To see this from a set of data, we record the times at which decays occur, then divide up the observed time span into equal-sized time intervals. We then count how many of these

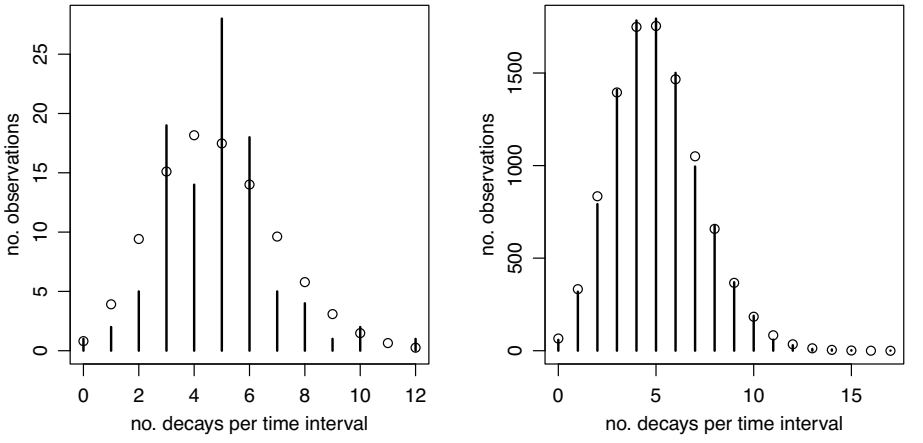

Fig. 1.3 The radioactive decay example showing the frequency (scaled probability mass function) of decays per unit time interval for the simulated data (vertical bars), and for the Poisson distribution (open circles) with λ determined from the data. The left panel is the result for 100 time intervals; the right panel is to a much longer sequence of data with 10 000 time intervals. In both cases the true value of λ (used in the simulation) is 5.

intervals have 0, 1, 2, 3, etc., decays. When normalized, this is approximately a Poisson distribution with mean $1/\tau$.

The R code below demonstrates this using a simulation of radioactive decay. Let's assume a source has a decay rate of $\lambda = 5$ per unit time interval. The number of decays in a unit time interval can be simulated by drawing once from a Poisson distribution with this mean. I repeat this `nint` $= 100$ times to simulate a sequence of observation intervals. Based just on these data I then count how many of these intervals have 0, 1, 2, 3, etc., decays. I do this using the function `table`. The resulting frequency distribution is shown in the left panel of figure 1.3. I then overplot a Poisson density distribution with its mean derived from the data (which is `mean(ndecay)` $= 4.81$), and which I have scaled to give the expected number of counts per time interval. If I increase the amount of data – the number of time intervals – the data follow the Poisson distribution more closely. This is shown in the right panel of the figure, which uses `nint` $= 10\,000$ (the empirical mean is 5.01). Note that doubling the length of the time interval would double both λ and the typical number of events per time interval, so the Poisson distribution would look more like a Gaussian.

R file: `poisson.R`

```
##### Compare data drawn from a Poisson with its theoretical distribution

pdf("dpois2.pdf", 4, 4)
par(mfrow=c(1,1), mgp=c(2.0,0.8,0), mar=c(3.5,3.5,1,1), oma=0.1*c(1,1,1,1))
truelambda <- 5 # = 1/tau
nint <- 100     # number of time intervals
```

```
set.seed(200)
ndecay <- rpois(n=nint, lambda=truelambda) # no.decays in each time interval
nobs   <- table(ndecay) # frequency distribution of ndecay
x <- 0:max(ndecay)
# multiply Poisson density by nint to get expected counts
plot(x, dpois(lambda=mean(ndecay), x=x)*nint, xlim=range(x),
     ylim=c(0,max(nobs)), xlab="no. decays per time interval",
     ylab="no. observations")
points(nobs)
dev.off()
```

Example: why do we build large telescopes?

The main reason for building telescopes with large mirrors is to be able to detect faint sources. The larger the telescope mirror the more photons we gather from the source (for a given exposure time). But why do we need more photons? The reason is that the emission of photons is governed by the Poisson distribution: the number of photons emitted will differ from the expected number due to the finite variance of the distribution. This is noise. How reliably we can detect something is determined by the signal-to-noise ratio (SNR), which is proportional to the signal r divided by its standard deviation σ_r. For the Poisson process $\sigma_r = \sqrt{r}$, so the SNR is proportional to $\sqrt{r}$, i.e. it increases with the number of photons collected.[4] By building a large telescope, we get a larger SNR in a given exposure time.

1.4.3 Beta

A useful PDF for a quantity p bound to lie between 0 and 1 is the beta distribution. This is described by two shape parameters, α and β. Its PDF is

$$P(p) = \frac{1}{B(\alpha,\beta)}\, p^{\alpha-1}(1-p)^{\beta-1} \quad \text{where} \quad \alpha > 0,\ \beta > 0,\ 0 \le p \le 1 \tag{1.55}$$

where

$$B(\alpha,\beta) = \int_0^1 p^{\alpha-1}(1-p)^{\beta-1} dp, \tag{1.56}$$

the *beta function*, is the normalization constant. Figure 1.4 shows some examples of the distribution plotted using the `dbeta` function in R (`shape1` is α and `shape2` is β). The mean, mode (maximum), and variance of the beta distribution are

$$\text{mean} = \frac{\alpha}{\alpha+\beta} \tag{1.57}$$

$$\text{mode} = \frac{\alpha-1}{\alpha+\beta-2} \quad \text{for} \quad \alpha > 1,\ \beta > 1 \tag{1.58}$$

$$\text{variance} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{1.59}$$

[4] The standard deviation is actually equally to the square root of the *expected* number of photons, but we don't know that, so we use r as our best estimate thereof.

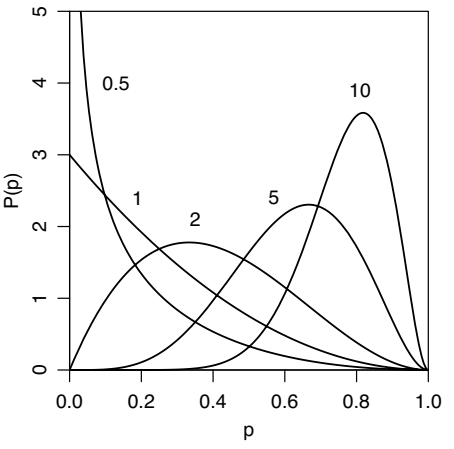

Fig. 1.4 The beta distribution for $\alpha = 0.5, 1, 2, 5, 10$ with $\beta = 3$.

Notice how the dependence on p is similar in form to that of the binomial distribution. They are in fact the same when $\alpha = r + 1$ and $\beta = n - r + 1$. We shall see the relevance of this in section 5.1.

1.4.4 Uniform

A uniform distribution $\mathcal{U}(a, b)$ is constant between a and b and zero outside. In order for it to be normalized the value of the distribution must be $1/(b - a)$. Its mean and median are obviously $(a + b)/2$, but its mode is not uniquely defined (it could have any value between a and b). Using equation 1.30 we find the variance to be $(b - a)^2/12$.

Sometimes we make use of a uniform distribution which has infinite support (no bounds) or which has semi-infinite support (e.g. a lower bound at zero). Such a distribution is constant over an infinite range. It cannot be normalized and so is an improper distribution (section 1.2.6).

1.4.5 Gaussian

The Gaussian or normal distribution is probably the best known and most commonly used distribution in the physical sciences. Its domain is the set of all real numbers and its density function is

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \quad \text{where} \quad \sigma > 0. \tag{1.60}$$

This is sometimes abbreviated with the notation $\mathcal{N}(\mu, \sigma)$. The notation $\epsilon \sim \mathcal{N}(\mu, \sigma)$ means that ϵ is a random number drawn from this distribution. The mean μ and standard deviation σ fully characterize the Gaussian. This PDF is symmetric and the mean is equal to the mode and the median. Approximately 68% of the probability mass lies between $\mu - \sigma$ and $\mu + \sigma$, the "1σ range". The full-width at half-maximum (FWHM) is $2\sqrt{2 \ln 2}\,\sigma \simeq 2.35\sigma$.

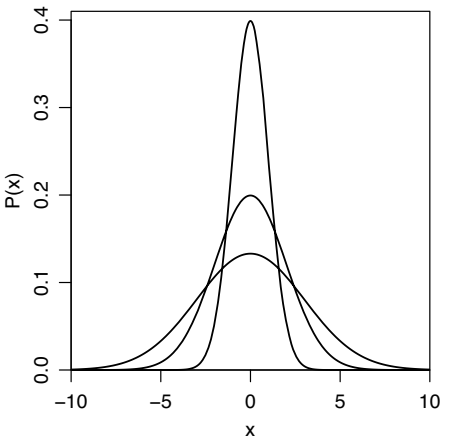

Fig. 1.5 The Gaussian (normal) distribution with zero mean and standard deviation of 1, 2, and 3 (larger is wider/lower).

Examples of the distribution are shown in figure 1.5. As the distribution is normalized, a larger σ means a lower maximum. The *standardized Gaussian* is the distribution $\mathcal{N}(0, 1)$.

There are at least three reasons why the Gaussian is ubiquitous: (1) it has some convenient properties, making it easy to use; (2) the central limit theorem (to be discussed in section 2.3); (3) the principle of maximum entropy says that if we have a continuous variable for which we know only its mean and variance, then the Gaussian is the distribution that makes the fewest assumptions; it is the most conservative choice.

1.4.6 Gamma

The gamma distribution has semi-infinite support: it is non-zero only for $x \geq 0$. It is characterized by two parameters, the shape k and the scale θ. Its PDF is

$$P(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-x/\theta} \quad \text{where} \quad k > 0,\ \theta > 0,\ x \geq 0 \tag{1.61}$$

and $\Gamma(k)$ is the *gamma function*, defined as

$$\Gamma(k) = \int_0^\infty x^{k-1} e^{-x}\, dx \tag{1.62}$$

which can be seen as a generalization of the factorial function for non-integer k, because $\Gamma(k) = (k-1)!$ for integer k. (The gamma function is `gamma` in R.) The mean and variance of the gamma distribution are $k\theta$ and $k\theta^2$ respectively. It has a mode only when $k \geq 1$, and then at $(k-1)\theta$. When $k < 1$ the function tends to infinity as x goes to zero, but the area under the curve of course remains equal to one. Examples of the distribution are shown in figure 1.6. The gamma distribution may be convenient for defining distributions over quantities that cannot be negative.

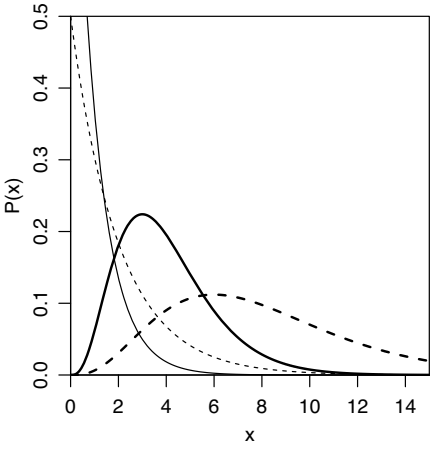

Fig. 1.6 The gamma distribution for two values of its shape k (thin = 1, thick = 4) and scale θ (solid = 1, dashed = 2) parameters.

1.4.7 Cauchy

The Cauchy distribution (also called the Lorentz distribution) is characterized by a location parameter a and a width (or scale) parameter b. Its PDF is

$$P(x) = \frac{1}{\pi b[1 + (\frac{x-a}{b})^2]} = \frac{b}{\pi[b^2 + (x-a)^2]} \quad \text{where} \quad b > 0. \tag{1.63}$$

If you plot it (figure 1.7), it looks quite harmless. It has a mode (and median) at $x = a$ and is symmetric about this. But it is the canonical example of a pathological distribution, because it has no mean and no standard deviation. Why? The distribution has a finite zeroth moment (equation 1.40) so it is normalizable. But it turns out that all its higher moments are either undefined or infinite (we'll investigate this further in section 2.3.2). We can nonetheless characterize the width of the distribution using the FWHM, for example. It turns out this is $2b$, so the parameter b is the half-width at half-maximum (HWHM) (and it is also the interquartile range, defined in section 1.5).

Figure 1.7 plots two different Cauchy distributions. Owing to the normalization, a larger width b means a lower maximum. The plot compares the Cauchy distribution with Gaussians with the same FWHM. For values nearer to the centre the Gaussian is broader, but as we move into the lower probability density regions the Cauchy is much wider. The Cauchy has "heavier tails" than the Gaussian. Other than being nicely pathological, the Cauchy distribution is also the distribution of the variable which is the ratio of two independent Gaussian variables. It is sometimes used to describe the broadening of spectral lines.

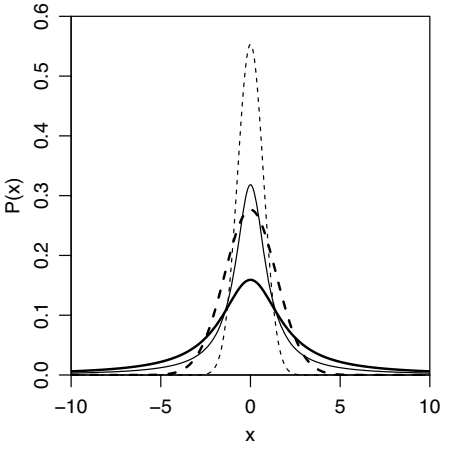

Fig. 1.7 The Cauchy distribution (solid lines) with zero mode and b equal to 1 (thin) and 2 (thick). The dashed lines shows a Gaussian with zero mode and the same FWHM as the two Cauchy distributions.

1.5 Cumulative distribution function and quantile function

Given a continuous, one-dimensional PDF $P(x)$, the *cumulative distribution function* (CDF) is the probability that x is less than some value q. For a PDF with infinite support this is

$$C(q) \;=\; P(x < q) \;=\; \int_{-\infty}^{q} P(x)\,dx. \tag{1.64}$$

If $P(x)$ is only defined over a limited range $(x_{\min}, x_{\max})$, the lower limit of the above integral is $x_{\min}$. $C(q)$ is a monotonically non-decreasing[5] function of q, and as $P(x)$ is normalized, $C(q)$ lies between 0 and 1. The derivative of the CDF is the PDF. The CDF is only defined for univariate probability distributions.

The inverse function of the CDF is the *quantile function* $Q(p)$. It gives the value of x below which the integral of $P(x)$ is equal to p. These two functions are shown in figure 1.8 for Gaussian and Cauchy distributions. Of particular interest as a measure of the location of a distribution is the 50% quantile ($p = 0.5$), the *median*: this is the value of x for which half the integrated probability lies below that value and half above. Together with the 25% and 75% quantiles these are the *quartiles*. The difference between the 75% and 25% quantiles, $Q(p = 0.75) - Q(p = 0.25)$, is the *interquartile range* (IQR), and is sometimes used as a measure of the width of a distribution.

Given a set of data, $Q(p)$ can be estimated by sorting the data and finding the point

[5] A function is *monotonic* if its gradient is always either non-negative ("monotonically non-decreasing") or non-positive ("monotonically non-increasing"), both of which allow the gradient to be zero. A function is *strictly monotonic* if its gradient is either always positive or always negative, i.e. it also cannot be zero.

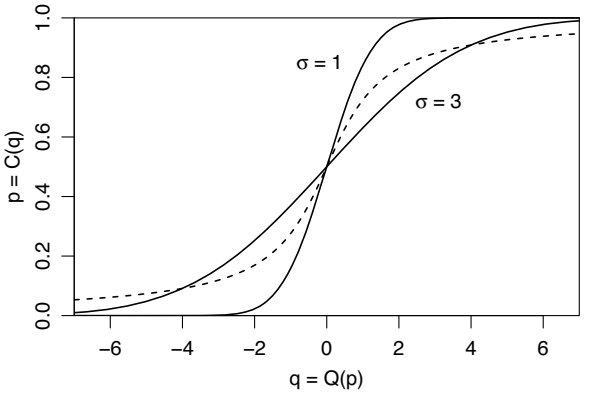

Fig. 1.8 The cumulative distribution function $C(q)$ and the quantile function $Q(p)$ for two zero-mean Gaussian distributions (solid lines) with $\sigma = 1$ (steeper function) and $\sigma = 3$, and for a Cauchy distribution (dashed line) with mode zero and the same FWHM as the Gaussian with $\sigma = 1$.

which puts a fraction p of the data below that point. For a finite set of data, there will not usually be a value of x that corresponds to the required p. We must therefore smooth or interpolate the data in some way. There are various different ways of doing this, so not all algorithms will return the same results.

For many standard distributions R has a set of built-in functions for calculating quantities from the distribution. For example, for the normal distribution they are

- `dnorm(x)` $= P(x)$, the probability density function (PDF)
- `pnorm(q)` $= p = \int_{-\infty}^{q} P(x)\,dx$, the cumulative distribution function (CDF)
- `qnorm(p)` the quantile function, which is the inverse function of `pnorm(q)`
- `rnorm` draws a random number from the distribution.

The CDF of a discrete set of data can be calculated in R using `cumsum` and the quantiles can be found using `quantile`. The parameter called `type` in the latter determines how the quantile is defined for a discrete data set.

1.6 Multiple variables

1.6.1 Covariance

Often we have problems involving multiple variables, and want to measure how closely two variables vary with one another. The *covariance* between two variables x and y is

defined as

$$\begin{aligned}\mathrm{Cov}(x,y) &= \iint P(x,y)(x-E[x])(y-E[y])\,dx\,dy \\ &= E[(x-E[x])(y-E[y])] \\ &= E[xy - yE[x] - xE[y]] + E[x]E[y] \\ &= E[xy] - E[x]E[y]. \end{aligned} \tag{1.65}$$

We can remember this as "expectation of the product minus product of expectations". It generalizes the definition of variance in equation 1.30. In analogy to the sample variance, the *sample covariance* is what we should compute from a set of data when we estimate the means from the data. This is

$$\widehat{\mathrm{Cov}}(x,y) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\overline{x})(y_i-\overline{y}). \tag{1.66}$$

It is often helpful to normalize the covariance according to the standard deviation in each variable, σ_x and σ_y, which leads to the definition of the *correlation coefficient*

$$\rho(x,y) = \frac{\mathrm{Cov}(x,y)}{\sigma_x\sigma_y} \tag{1.67}$$

$$= \frac{\overline{xy} - \overline{x}\,\overline{y}}{\sigma_x\sigma_y}. \tag{1.68}$$

The *sample correlation coefficient* is computed using the sample covariance and the sample standard deviations (equation 1.37). The correlation coefficient lies in the range $-1 \le \rho \le +1$. A value of $\rho = +1$ corresponds to perfect correlation: the two variables are equal apart from some scale factor (which cancels in the ratio) and an offset (which is removed by the mean). Similarly $\rho = -1$ corresponds to perfect anticorrelation. If $\rho = 0$ the variables are uncorrelated.

The R functions that calculate the covariance and correlation between two vectors of variables are `cov` and `cor` respectively.

If we have two or more variables then we can form a *covariance matrix*, in which element c_{ij} is the covariance between variable i and variable j. Thus the diagonal elements of the matrix are the variances and the off-diagonal elements are the covariances (we will see an example in the next section). This is a symmetric matrix.

1.6.2 Multivariate probability distributions

When we have multiple variables we are often interested in their *joint probability distribution*. Let $\mathbf{x}$ be a J-dimensional vector of the variables, the mean of which is $\boldsymbol{\mu}$ (also a J-dimensional vector). The J-dimensional Gaussian PDF is defined as

$$P(x_1,\ldots,x_J) = \frac{1}{(2\pi)^{J/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \tag{1.69}$$

where Σ is the $J \times J$ covariance matrix of the data, $|\Sigma|$ is its determinant, and $(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}$ denotes the transpose of the $J \times 1$ column vector $(\mathbf{x}-\boldsymbol{\mu})$. (My vectors are column vectors

by default.) The argument of the exponential is of course a scalar. The dimension of this probability density is $(\prod_i x_i)^{-1}$, as can also be seen by the fact that the determinant of the covariance enters as the square root in the denominator. If the variables are independent of one another then the covariance matrix is diagonal. In two dimensions, writing $\mathbf{x} = (x, y)$, the covariance matrix is

$$\Sigma = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix} \tag{1.70}$$

and the distribution is

$$P(x, y) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right) \tag{1.71}$$

where $|\Sigma| = \sigma_x^2\sigma_y^2(1 - \rho^2)$. The locus of points with constant probability density is given when the argument of the exponential is constant. Consider for convenience the zero mean Gaussian with zero covariance, in which case

$$\Sigma^{-1} = \begin{bmatrix} \frac{1}{\sigma_x^2} & 0 \\ 0 & \frac{1}{\sigma_y^2} \end{bmatrix}. \tag{1.72}$$

Performing the matrix multiplications we see that the locus of points of constant probability is

$$\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} = \text{constant}. \tag{1.73}$$

This is the equation for an ellipse. We can show that if the covariance is non-zero then this just tilts the ellipse. So we can always transform a two-dimensional Gaussian distribution with non-zero covariance into a two-dimensional Gaussian with zero covariance by rotating the axes (and likewise for higher dimensional Gaussians). The following code plots a bivariate Gaussian with non-zero covariance both as a three-dimensional perspective mesh and as a contour plot. This is shown in figure 1.9.

R file: `2D_gaussian.R`

```
##### Plot bivariate Gaussian as a 3D mesh plot
##### and as contours of constant probability density

library(mvtnorm) # for dmvnorm
sigma.x <- 1
sigma.y <- 1
rho <- 0.5 # correlation coefficient
Cov <- matrix(data=c(sigma.x^2, rho*sigma.x*sigma.y, rho*sigma.x*sigma.y,
                     sigma.y^2), nrow=2, ncol=2)
Nsig  <- 3.5
Nsamp <- 100
x <- seq(from=-Nsig*sigma.x, to=Nsig*sigma.x, length.out=Nsamp)
y <- seq(from=-Nsig*sigma.y, to=Nsig*sigma.y, length.out=Nsamp)
z <- matrix(dmvnorm(x=expand.grid(x,y), mean=c(0,0), sigma=Cov),
            nrow=length(x), ncol=length(y))
z <- z/max(z)

pdf("2D_gaussian_3Dmesh.pdf", 4, 4)
```

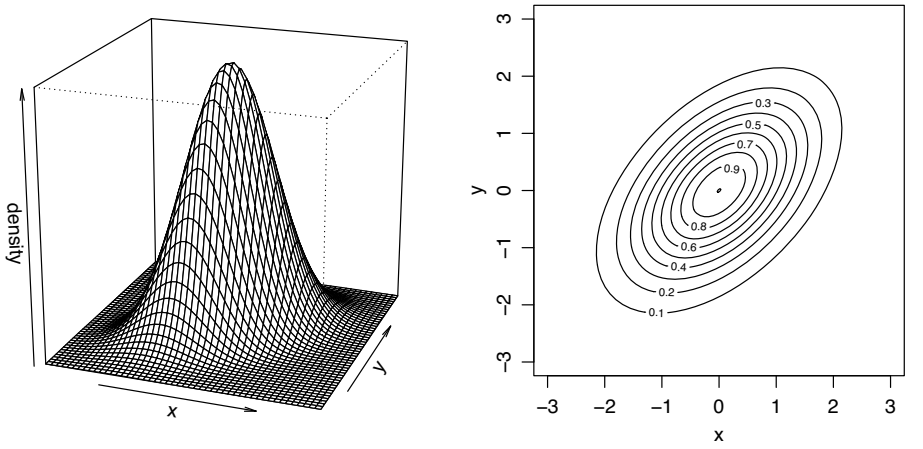

Fig. 1.9 A bivariate Gaussian with $\rho = 0.5$ and unit standard deviation in x and y, shown as a three-dimensional perspective mesh (left) and as contours of constant probability density (right), labelled with the density relative to the peak.

```
par(mfrow=c(1,1), mar=c(1,1,1,1), oma=c(0,0,0,0), mgp=c(2.2,0.8,0), cex=1.0)
persp(x=x, y=y, z=z, phi=20, theta=20, d=5, zlab="density")
dev.off()

pdf("2D_gaussian_contours.pdf", 4, 4)
par(mfrow=c(1,1), mgp=c(2.0,0.8,0), mar=c(3,3,1,1), oma=0.1*c(1,1,1,1))
contour(x, y, z, asp=1, xlim=c(-3,3), ylim=c(-3,3), xlab="x", ylab="y")
dev.off()
```

Given a two-dimensional PDF we might want to know $P(x\,|\,y=y_0)$. This is the *conditional* PDF on x given $y = y_0$ and is a one-dimensional distribution. Equation 1.7 gives this for discrete variables. For continuous variables it is

$$P(x\,|\,y=y_0) \;=\; \frac{P(x, y=y_0)}{P(y_0)}. \tag{1.74}$$

This is a slice through the two-dimensional PDF at constant y, with normalization constant $P(y_0)$. If necessary, we can work out this normalization constant numerically (see chapter 5).

We may also want to know $P(x)$ regardless of the value of y. This is the *marginal* PDF of x over y. We can think of it as viewing the two-dimensional PDF along the y-axis. From the laws of probability this is

$$\begin{aligned} P(x) \;&=\; \int_{-\infty}^{+\infty} P(x, y)\, dy \\ &=\; \int_{-\infty}^{+\infty} P(x\,|\,y) P(y)\, dy \end{aligned} \tag{1.75}$$

which is known as marginalizing over y. The corresponding marginalization for discrete variables was given by equation 1.9.

1.7 Combinations and permutations

Expressions for the number of ways of combining and arranging objects come up in many places, including – as we have seen – in the derivation of some probability distributions.

Unique pairings of objects

You have n objects. How many ways are there of selecting unique pairs (without caring about the ordering)? Think of an $n \times n$ array. Every point in the array is a pairing, except for the leading diagonal, where an object is paired with itself. The two halves on each side of the diagonal are identical (A paired with B is indistinguishable from B paired with A). So the number of ways must be $(n^2 - n)/2 = n(n-1)/2$.

Unique orderings

How many ways can you order r objects (e.g. the three letters abc)? There are r options for the first object, $r-1$ for the second, $r-2$ for the third, etc. Therefore it is $r!$.

Selection with replacement, order relevant

Imagine a word written in an alphabet with n different letters. How many unique words can you make with r characters? The first character has n different possibilities, so does the second, third, etc. Therefore it is n^r. This is the number of ways of selecting r objects from a set of n *with replacement*, where the order is relevant.

Selection without replacement, order relevant

You select r unique objects from a set of n ($r \leq n$) and place them in a line. How many unique arrangements are there of this line? There are n ways of selecting the first, $n-1$ for the second, ..., and $n-r+1$ for selecting the rth. This is *selection without replacement*. So the total number of ways of selecting is $n(n-1)(n-2)\ldots(n-r+1)$. The number of arrangements is called the *permutations* ${}_nP_r$, and can be written

$$ {}_nP_r = \frac{n!}{(n-r)!} \quad \text{where} \quad n \geq 0,\, r \leq n. \tag{1.76} $$

Note that ${}_nP_n = n!/0! = n!$ (as $0! = 1$ by definition). We are concerned here with the order of the objects, so an arrangement abc is distinct from bca.

Selection without replacement, order irrelevant

You now select the r objects in the previous case but you are no longer concerned about the order. The number of different unique samples you can select must be smaller than ${}_nP_r$ by a factor of $r!$, because this is the number of unique orderings of r objects. The number

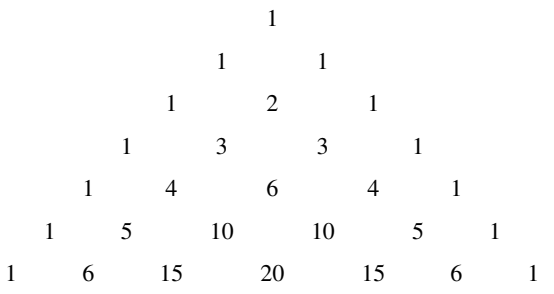

Fig. 1.10 The first few rows of Pascal's triangle. The nth row (counting the top one as $n = 0$) gives the values of "n choose r" for all r between 0 and n inclusive (equation 1.77). The nth diagonal (counting the first one as $n = 1$) gives the values of "n multichoose r" for all $r \geq 0$ (equation 1.78).

of ways of selecting r objects from a set of n without regard to the order of selection is called the number of *combinations* ${}_nC_r$, and is

$$ {}_nC_r = \frac{{}_nP_r}{r!} = \binom{n}{r} = \frac{n!}{r!(n-r)!} \quad \text{where} \quad n \geq 0,\, r \leq n \tag{1.77} $$

which we often call "n choose r". It is also called the binomial coefficient, because it occurred in the derivation of the binomial distribution (see equation 1.43). This is of course also selection without replacement. The set of values for all $r \leq n$ for a given n are the values in the rows of Pascal's triangle (figure 1.10).

Selection with replacement, order irrelevant

The number of ways of choosing r objects from n with replacement, in which we don't care about the order, is

$$ \binom{n+r-1}{r} = \frac{(n+r-1)!}{r!(n-1)!} \quad \text{where} \quad n > 0,\, r \geq 0 \tag{1.78} $$

which is sometimes called "n multichoose r". For example, the number of ways of choosing two letters from the set abc is six: aa, ab, ac, bb, bc, cc. The order doesn't matter: ba is identical to ab so is not counted separately, for example. If we choose three letters, there are ten ways. The set of values for all r for a given n are given in the diagonals of Pascal's triangle (figure 1.10), and are also known as the figurate numbers.[6]

[6] There are various ways to derive this. One is to think of an n-dimensional space with integer steps along each axis. The number sought is the number of points on the flat (hyper)surface that have an L^1 distance of r from the origin (counting just the positive quadrant). The L^1 distance between two points is the sum of projected distances along each axis, $\sum_i |\Delta x_i|$. That is, we can take any number of (positive) steps along each axis we like (we can select from each n as often as we like) subject to the constraint that the total number of steps (choices) is r. We don't care about the path (the order of selection). For $n = 2$ the "surface" is a straight line, and with $r = 2$ there are just three points (choices) on this line: $(2, 0)$, $(1, 1)$, $(0, 2)$. For $n = 3$ the surface is a triangle, and the number of points with L^1 distance from the origin of $r = 2$ is six. For $n = 4$ the surface is the surface of a four-dimensional triangle, etc.

The birthday problem

In a large room full of people, how many people do you have to ask before there is a 50% chance that any two or more of them share a common birthday?

As with the missiles problem it is easier to work out the probability of the complementary event. Assume that $n = 365$ birthdays are equally probable (not actually true, but it's fine for this problem). You ask r people. Let

$$A = n(n-1)\dots(n-r+1) = \frac{n!}{(n-r)!} = \text{number of ways for } r \text{ people to share no common birthday} \quad (1.79)$$

and

$$B = n^r = \text{number of ways of assigning } n \text{ birthdays to } r \text{ people.} \quad (1.80)$$

The probability of there being no common birthday among the r people is A/B, so the probability of at least one common birthday is $1 - A/B$. Setting this equal to 0.5 and solving for r is a bit complicated. But as $1 - A/B$ increases monotonically with r, and r is an integer, it's easier instead to write a program to compute the probability for a range of values of r. We find that the probability of no common birthday for $r = 22$ is 0.476 and for $r = 23$ it is 0.507. So we need to ask 23 people. This might seem low, but that is because we are interested in *any* pair of birthdays.

How many people do you instead have to ask before there is a 50% chance that at least one of them shares *your* birthday?

This will be larger because we are interested in only one specific pair of birthdays. Intuition may suggest $n/2 \simeq 183$. But this is wrong because when we ask people we sample birthdays with replacement: the tenth person we ask might have the same birthday as the third person we asked. Again we work out the complementary probability. The probability that each person we ask doesn't share your birthday is $(n-1)/n$, so the probability that none of the r people asked share your birthday is $[(n-1)/n]^r$. Thus the probability that at least one of these people shares your birthday is $1 - [(n-1)/n]^r$. We can solve by taking logs

$$\log 0.5 = r \log\left(\frac{n-1}{n}\right) \quad (1.81)$$

the solution to which is $r = 253$ (rounded up to the nearest integer). This is more than the naive $n/2$ "without replacement" estimate.

1.7.1 Using R

The factorial is provided by `factorial`. It even works on non-integers because it is implemented in R using the gamma function (equation 1.62), as $n! = \Gamma(n+1)$. In some analytic calculations you might want to use Stirling's approximation to $n!$ for large n

$$n! \simeq \sqrt{2\pi n}\, n^n e^{-n}. \quad (1.82)$$

The number of combinations ${}_nC_r$ is given by the function `choose(n,r)`. You can use `choose(n,r)*factorial(r)` to compute ${}_nP_r$. If you want to list the actual permutations and combinations, use `combinations` and `permutations` in the `gtools` package.

1.8 Random number generation

We shall see in later chapters that the practice of inference often requires us to draw random numbers from a distribution. We can do this with computer algorithms despite the fact that we use computers as deterministic machines. These algorithms create extremely long sequences of numbers which are apparently random, but which are in fact determined by a random number seed: use the same seed and you get the same sequence of numbers. Although not random, the sequence has many properties of randomness, such as having no shortscale correlations, having a distribution and a Fourier spectrum consistent with random, etc. Such sequences of numbers are *pseudo random*, meaning that they are random enough for our purposes (but often not for other applications such as cryptography).

A simple algorithm for generating pseudo random numbers is the *linear congruential generator*, which generates a sequence of integers $\{x_i\}$. Given large positive integers a, b, and m, the sequence is defined by the recurrence relation

$$x_{i+1} = \mathrm{Mod}(ax_i + b, m) \tag{1.83}$$

where Mod is the modulus operator. This gives the remainder after integer division of the first argument by m, i.e.

$$\mathrm{Mod}(x, m) = x - \mathrm{IntegerPart}(x/m) \times m. \tag{1.84}$$

For example $\mathrm{Mod}(17, 5) = 2$. This can be achieved on a digital computer by truncating the number of bits used in the storage. The algorithm creates a sequence of integers between 0 and $m - 1$ which will repeat after no more than m steps. If the values of a, b, and m are chosen sensibly, then this maximum can be achieved and each integer will appear once. The seed just determines the starting point. Of course, in any practical application, m should be much larger than the number of samples required (a typical value is 2^{64}).

How do we use such a sequence of numbers to draw from (sample) a distribution? For a uniform distribution we simply need to scale the sequence range to the range of our distribution. Provided m is large enough this will produce a set of real numbers with high enough numerical precision such that they are practically indistinguishable from numbers which really have been drawn from the distribution. When we scale to the distribution $\mathcal{U}(0, 1)$, for example, the smallest difference between any two pseudo random numbers obtained in this way is $1/m$.

For other univariate distributions we can in principle use the cumulative distribution function (CDF), which was defined in section 1.5, and ranges from 0 to 1. Suppose we want to draw from the PDF $P(x)$. We first draw a random number from $\mathcal{U}(0, 1)$. Transforming this via the inverse of the CDF gives a sample which has been drawn from $P(x)$. This can be understood in reference to figure 1.8: a uniform distribution along the vertical axis (the

CDF) will be transformed to a non-uniform one along the horizontal axis. There will be relatively few values of q far from zero – the tails of these distributions – because they are spread out over a larger range by the transformation. This method is convenient if the inverse of the CDF (the quantile function) is fast to compute, which is the case for some standard distributions. The R functions like `rnorm`, `rpois`, and `rbinom` select random numbers from standard distributions for you. You can test them by using a histogram to plot the distribution of the sample, for example as follows.

```
x <- rnorm(1e5)
hist(x, breaks=100, prob=TRUE)
xp <- seq(from=-4, to=+4, by=0.01)
lines(xp, dnorm(xp))
```

To draw random variables from a multivariate Gaussian distribution you can use `mvrnorm` in the `MASS` package, for which you need to specify the mean vector and the covariance matrix.

In practice, more sophisticated algorithms than the linear congruential generator are used in order to produce much longer sequences of pseudo random numbers. The default algorithm in R is the Mersenne-Twister algorithm, with a repeatability length of about 10^{6000}.

To sample a given set of numbers you can use the function `sample`, which will draw samples at random from a defined set, either with or without replacement. The function `set.seed` is used to define the seed. For a given seed you will always get the same sequence of numbers.

```
sample(10) # samples ten times from 1:10 without replacement
sample(x=10, size=3) # draws three samples
sample(x=10, size=3) # gives a different sequence
set.seed(100)
sample(x=10, size=3)
set.seed(100)
sample(x=10, size=3) # now gives the same sequence
sample(x=10, size=15, replace=TRUE) # samples with replacement
sample(c(-7,0,4,56,-76,128,17), size=3)
```

I advise that you set the seed whenever your data analysis procedure involves random number generation. You should modify the seed to ensure that there is no relevant sensitivity to this randomness, but recording the seed ensures repeatability and thus easier bug detection. Pay particular attention to how seeds and random number sequences are dealt with if you are doing parallel processing.

Drawing random numbers from an arbitrary PDF is difficult, because the CDF may not be easy to find or the PDF may not be easy to normalize. Drawing from multivariate PDFs is generally non-trivial. For this we can use Monte Carlo methods, which we will start to explore in chapter 8.

1.9 Change of variables

1.9.1 One-dimensional

Suppose we have a univariate PDF $P(x)$ over x, and want to express it as a PDF over y, where $y = f(x)$. Consider a small interval δx around the point x that corresponds to the small interval δy around the point y. In making the transformation we need to conserve probability,

$$P(x)\,\delta x \;=\; P(y)\,\delta y. \tag{1.85}$$

In the limit of infinitesimally small intervals this becomes

$$P(x) \;=\; P(y)\left|\frac{dy}{dx}\right|. \tag{1.86}$$

The term dy/dx is called the *Jacobian*. We take the modulus because we are interested only in the ratio of the lengths, not their signs: probabilities and probability densities are always positive (or zero).

Example

If $y = \ln x^a$ then $dy/dx = a/x$. Thus a distribution uniform in $\ln x$ has $P(x) \propto 1/x$. Indeed, for any $a \neq 0$ we see that

$$P(\ln x^a) = \text{constant} \quad\Leftrightarrow\quad P(x) \propto \frac{1}{x}. \tag{1.87}$$

Using a different base logarithm only changes the constant of proportionality.

We also see that if $P(x) \propto x^b$, then $P(\ln x) \propto x^{b+1}$. Setting $b = -1$ we see that the only power law distribution for $P(x)$ that produces a distribution uniform in $P(\ln x)$ is $P(x) \propto x^{-1}$.

1.9.2 Multi-dimensional

The change of variables generalizes to multiple dimensions (here J). We have $P(\mathbf{x})$ where $\mathbf{x} = (x_1, x_2, \ldots, x_J)^{\mathsf{T}}$ and want to transform this to $P(\mathbf{y})$ where $\mathbf{y} = (y_1, y_2, \ldots, y_J)^{\mathsf{T}}$. The J-dimensional volume element in the $\mathbf{x}$ space is $\delta x_1 \delta x_1 \ldots \delta x_J$ which I write as $\delta^J V_{\mathbf{x}}$, and the corresponding volume element in the $\mathbf{y}$ space is $\delta^J V_{\mathbf{y}}$. To conserve probability we have

$$P(\mathbf{x})\delta^J V_{\mathbf{x}} \;=\; P(\mathbf{y})\delta^J V_{\mathbf{y}} \tag{1.88}$$

and it can be shown that

$$\delta^J V_{\mathbf{y}} \;=\; \left|\frac{\partial(y_1, y_2, \ldots, y_J)}{\partial(x_1, x_2, \ldots, x_J)}\right| \delta^J V_{\mathbf{x}} \tag{1.89}$$

which involves the determinant of the Jacobian matrix, which is the $J \times J$ matrix of first partial derivatives with elements (row i, column j) $\partial y_i / \partial x_j$. Thus

$$P(\mathbf{x}) \;=\; P(\mathbf{y}) \left| \frac{\partial(y_1, y_2, \ldots, y_J)}{\partial(x_1, x_2, \ldots, x_J)} \right|. \tag{1.90}$$

One can only take the determinant of a square matrix, so the above only shows how we can transform from one J-dimensional space to another J-dimensional space. Sometimes we will want to reduce the dimensionality in a transformation (we cannot increase it). In that case we can marginalize over the unwanted variables, as is shown in the following example.

Example

Consider the transformation from two-dimensional Cartesian coordinates (x, y) to two-dimensional radial coordinates (r, θ),

$$\begin{aligned} x &= r\cos\theta \\ y &= r\sin\theta. \end{aligned} \tag{1.91}$$

We get

$$\begin{aligned} \left| \frac{\partial(x, y)}{\partial(r, \theta)} \right| &= \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} \\ &= r(\cos^2\theta + \sin^2\theta) \;=\; r. \end{aligned} \tag{1.92}$$

We could have taken instead a geometric approach to show that $\delta x \delta y = r \delta r \delta\theta$, in which case we see from equation 1.89 that the Jacobian determinant is r. Suppose that $P(x, y)$ is a bivariate, zero mean, isotropic Gaussian. This is

$$P(x, y) \;=\; \frac{1}{2\pi\sigma^2} \exp\left[-\frac{(x^2 + y^2)}{2\sigma^2} \right] \tag{1.93}$$

because it is just the product of two independent Gaussians. Using equation 1.90 we can write the PDF in the radial coordinates as

$$P(r, \theta) \;=\; \frac{r}{2\pi\sigma^2} \exp\left[-\frac{r^2}{2\sigma^2} \right]. \tag{1.94}$$

To get the marginal PDF of r we integrate over θ

$$P(r) \;=\; \int_0^{2\pi} P(r, \theta)\, d\theta \;=\; \frac{r}{\sigma^2} \exp\left[-\frac{r^2}{2\sigma^2} \right]. \tag{1.95}$$

The marginal PDF of θ is of course just $1/2\pi$ (the distribution is isotropic so must be uniform in θ).

In this example the maximum of $P(x, y)$ is at $(x, y) = (0, 0)$, which corresponds to $r = 0$. But the maximum of $P(r, \theta)$ is at $r = \sigma$ for all θ, as you can verify by differentiation or just by plotting. This difference arises because we are looking at *density* functions, which here are probability per unit area. The density function in Cartesian coordinates (equation 1.93) varies as $\exp(-r^2)$. When we express this in radial coordinates

the Jacobian stretches the area elements in proportion to r. The combination of these two dependencies is to give a maximum of the density function in radial coordinates at $r > 0$.

It follows from equation 1.86 that probability density functions are generally not invariant under transformations of their variables. This is because the Jacobian is usually a function of the variable too. Thus statistics of the density function, such as the mean or maximum, are in general also non-invariant. An exception arises when the transformation is strictly monotonic, which ensures that the transformation is invertible (one-to-one). It then follows from the definition of the cumulative density function that the quantiles (of univariate density functions), such as the median, are invariant.

1.10 The three doors problem revisited

I opened this chapter with the classic three doors problem. While a more formal analysis is not necessary for understanding the problem, writing down formulae for the relevant probabilities helps us to solve more general problems, such as one with multiple doors, multiple cars, or a different behaviour of the game show host.

The problem can be formulated probabilistically in a number of ways (and some are less useful than others). I define door 1 as the one we choose, and door 2 as the one which is opened by the game show host (to always show a goat). Let W be the proposition *the car is behind door 1*, and C be the proposition *we select the car when changing doors (from 1 to 3)*. From the rules of probability we can write

$$P(C) = P(C|W)P(W) + P(C|W')P(W') \tag{1.96}$$

where W' is the complement of W, i.e. the car is not behind door 1. For the original problem as described, $P(W) = 1/3$, $P(W') = 2/3$, $P(C|W) = 0$, $P(C|W') = 1$. Thus $P(C) = 2/3$.

This approach generalizes easily to the case of n doors and the game show host opening k doors to reveal as many goats ($0 \leq k \leq n-2$). There is still one car. What is the probability of winning if we switch to another closed door at random? C is now the proposition *we select the car when changing doors (from 1 to one specific closed door)*. In equation 1.96 we now have $P(W) = 1/n$, $P(W') = (n-1)/n$, $P(C|W) = 0$, $P(C|W') = 1/(n-k-1)$. Thus

$$P(C) = \frac{1}{n-k-1}\frac{n-1}{n}. \tag{1.97}$$

If $k = n-2$ (all doors but no. 1 and another are opened), $P(C) = (n-1)/n$. This gives $2/3$ for the original problem ($n = 3$). With $k \geq 1$ then $(n-1)/(n-k-1) > 1$, so $P(C) > 1/n$. The probability of wining is increased over the initial probability (which is $1/n$) whenever one or more doors are opened. Hence we should always switch doors.