Useful Probability Distributions 5

We will use probability as a tool to resolve practical questions about data. Here are important example questions. We could ask **what process produced the data?** For example, I observe a set of independent coin flips. I would now like to know the probability of observing a head when the coin is flipped. We could ask **what sort of data can we expect in the future?** For example, what will be the outcome of the next election? Answering this requires collecting information about voters, preferences, and the like, then using it to build a model that predicts the outcome. We could ask **what labels should we attach to unlabelled data?** For example, we might see a large number of credit card transactions, some known to be legitimate and others known to be fraudulent. We now see a new transaction: is it legitimate? We could ask **is an effect easily explained by chance variations, or is it real?** For example, a medicine appears to help patients with a disease. Is there a real effect, or is it possible that by chance the patients we tested the medicine on felt better?

These questions do not lend themselves to "right" answers. Instead, we will need to produce estimates and perhaps some measure of our confidence in those estimates. Sensible answers to questions like these have great practical value. Producing sensible answers to these questions requires some form of probability model. In this chapter, I describe the properties of some probability distributions that are used again and again in model building.

5.1 Discrete Distributions

5.1.1 The Discrete Uniform Distribution

Assume we have a random variable that can take one of k different values. We can relabel these values $1, \ldots, k$ without losing anything significant. If each of these values has the same probability (and all others have probability zero), then the probability distribution is the discrete uniform distribution. We have seen this distribution before, numerous times. For example, I define a random variable by the number that shows face-up on the throw of a fair die. This has a uniform distribution. As another example, write the numbers 1–52 on the face of each card of a standard deck of playing cards. The number on the face of the first card drawn from a well-shuffled deck is a random variable with a uniform distribution.

Definition 5.1 (Uniform Random Variable, Discrete) A random variable has the discrete uniform distribution if it takes each of k values with the same probability $1/k$, and all other values with probability zero.

One can construct expressions for the mean and variance of a discrete uniform distribution, but they're not usually much use (too many terms, not often used). Keep in mind that if two random variables have a uniform distribution, their sum and difference will not (recall Example 4.3).

D. Forsyth, *Probability and Statistics for Computer Science*,
https://doi.org/10.1007/978-3-319-64410-3_5

5.1.2 Bernoulli Random Variables

A Bernoulli random variable models a biased coin with probability p of coming up heads in any one flip.

Definition 5.2 (Bernoulli Random Variable) A Bernoulli random variable takes the value 1 with probability p and 0 with probability $1 - p$. This is a model for a coin toss, among other things.

Useful Facts 5.1 (Mean and Variance of a Bernoulli Random Variable)

A Bernoulli random variable that takes the value 1 with probability p has:

1. mean p;
2. variance $p(1 - p)$.

5.1.3 The Geometric Distribution

We have a biased coin. The probability it will land heads up, $P(\{H\})$ is given by p. We flip this coin until the first head appears. The number of flips required is a discrete random variable which takes integer values greater than or equal to one, which we shall call X. To get n flips, we must have $n - 1$ tails followed by 1 head. This event has probability $(1 - p)^{(n-1)}p$. We can now write out the probability distribution that n flips are required.

Definition 5.3 (Geometric Distribution) The geometric distribution is a probability distribution on positive integers n (i.e. $n > 0$). It has the form

$$P(\{X = n\}) = (1 - p)^{(n-1)}p.$$

for $0 \leq p \leq 1$ and $n \geq 1$ (for other n the distribution is zero). p is called the **parameter** of the distribution.

Notice that the geometric distribution is non-negative everywhere. It is straightforward to show that it sums to one, and so is a probability distribution (exercises).

Useful Facts 5.2 (Mean and Variance of a Geometric Distribution)

A geometric distribution with parameter p has

1. mean $\frac{1}{p}$;
2. variance $\frac{1-p}{p^2}$.

It should be clear that this model isn't really about coins, but about repeated trials. The trial could be anything that has some probability of failing. Each trial is independent, and the rule for repeating is that you keep trying until the first success. Textbooks often set exercises involving missiles and aircraft; I'll omit these on grounds of taste.

5.1.4 The Binomial Probability Distribution

Assume we have a biased coin with probability p of coming up heads in any one flip. The binomial probability distribution gives the probability that it comes up heads h times in N flips. Recall there are

$$\binom{N}{h} = \frac{N!}{h!(N - h)!}$$

outcomes of N coin flips that have h heads. These outcomes are disjoint, and each has probability $p^h(1-p)^{(N-h)}$. As a result, we must have the probability distribution below.

Definition 5.4 (Binomial Distribution) In N independent repetitions of an experiment with a binary outcome (ie heads or tails; 0 or 1; and so on) with $P(H) = p$ and $P(T) = 1 - p$, the probability of observing a total of h H's and $(N-h)T$'s is

$$P_b(h; N, p) = \binom{N}{h} p^h (1-p)^{(N-h)}$$

as long as $0 \leq h \leq N$; in any other case, the probability is zero.

The binomial distribution really is a probability distribution. For $0 \leq p \leq 1$, it is clearly non-negative for any i. It also sums to one. Write $P_b(i; N, p)$ for the binomial distribution that one observes i H's in N trials. Then, by pattern matching to the binomial theorem, we have

$$(p + (1-p))^N = \sum_{i=0}^{N} P_b(i; N, p) = 1.$$

The binomial distribution satisfies a recurrence relation. You can get h heads in N flips either by having $h-1$ heads in $N-1$ flips, then flipping another head, or by having h heads in N flips then flipping a tail. This means that

$$\begin{aligned} P_b(h; N, p) &= pP_b(h-1; N-1, p) \\ &+ (1-p)P_b(h; N-1, p) \end{aligned}$$

(exercises).

Useful Facts 5.3 (Mean and Variance of the Binomial Distribution)

The binomial distribution

$$P_b(h; N, p) = \binom{N}{h} p^h (1-p)^{(N-h)}$$

has:

1. mean Np;
2. variance $Np(1-p)$.

The proofs are informative, and so are not banished to the exercises.

Property 5.1 Mean and variance of binomial distribution.

Proposition *The mean of the binomial distribution $P_b(h; N, p)$ is Np. The variance is $Np(1-p)$.*

Proof Write X for a random variable with distribution $P_b(h; N, p)$. Notice that the number of heads in N coin tosses can be obtained by adding the number of heads in each toss. Write Y_i for the Bernoulli random variable representing the i'th toss. If the coin comes up heads, $Y_i = 1$, otherwise $Y_i = 0$. The Y_i are independent. Now

(continued)

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}\left[\sum_{j=1}^{N} Y_i\right] \\ &= \sum_{j=1}^{N} \mathbb{E}[Y_i] \\ &= N\mathbb{E}[Y_1] \text{ because the } Y_i \text{ are independent} \\ &= Np.\end{aligned}$$

The variance is easy, too. Each coin toss is independent, so the variance of the sum of coin tosses is the sum of the variances. This gives

$$\begin{aligned}\mathsf{var}[X] &= \mathsf{var}\left[\sum_{j=1}^{N} Y_i\right] \\ &= N\mathsf{var}[Y_1] \\ &= Np(1-p)\end{aligned}$$

5.1.5 Multinomial Probabilities

The binomial distribution describes what happens when a coin is flipped multiple times. But we could toss a die multiple times too. Assume this die has k sides, and we toss it N times. The distribution of outcomes is known as the **multinomial distribution**.

We can guess the form of the multinomial distribution in rather a straightforward way. The die has k sides. We toss the die N times. This gives us a sequence of N numbers. Each toss of the die is independent. Assume that side 1 appears n_1 times, side 2 appears n_2 times, ... side k appears n_k times. Any single sequence with this property will appear with probability $p_1^{n_1} p_2^{n_2} \dots p_k^{n_k}$, because the tosses are independent. However, there are

$$\frac{N!}{n_1! n_2! \dots n_k!}$$

such sequences. Using this reasoning, we arrive at the distribution below

Definition 5.5 (Multinomial Distribution) Perform N independent repetitions of an experiment with k possible outcomes. The i'th such outcome has probability p_i. The probability of observing outcome 1 n_1 times, outcome 2 n_2 times, etc. (where $n_1 + n_2 + n_3 + \dots + n_k = N$) is

$$P_m(n_1, \dots, n_k; N, p_1, \dots, p_k) = \frac{N!}{n_1! n_2! \dots n_k!} p_1^{n_1} p_2^{n_2} \dots p_k^{n_k}.$$

I don't recall ever using the mean and variance of a multinomial distribution, so they're not in a box. If you happen to need this information, you can derive it with using the reasoning of proof 5.1.

5.1.6 The Poisson Distribution

Assume we are interested in counts that occur in an interval of time (e.g. within a particular hour). Because they are counts, they are non-negative and integer valued. We know these counts have two important properties. First, they occur with some fixed average rate. Second, an observation occurs independent of the interval since the last observation. Then the Poisson distribution is an appropriate model.

There are numerous such cases. For example, the marketing phone calls you receive during the day time are likely to be well modelled by a Poisson distribution. They come at some average rate—perhaps 5 a day as I write, during the last phases of an election year—and the probability of getting one clearly doesn't depend on the time since the last one arrived. Classic examples include the number of Prussian soldiers killed by horse-kicks each year; the number of calls arriving at a call center each minute; the number of insurance claims occurring in a given time interval (outside of a special event like a hurricane, etc.).

Definition 5.6 (Poisson Distribution) A non-negative, integer valued random variable X has a Poisson distribution when its probability distribution takes the form

$$P(\{X = k\}) = \frac{\lambda^k e^{-\lambda}}{k!},$$

where $\lambda > 0$ is a parameter often known as the **intensity** of the distribution.

Notice that the Poisson distribution is a probability distribution, because it is non-negative and because

$$\sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = e^{\lambda}$$

so that

$$\sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} = 1$$

Useful Facts 5.4 (Mean and Variance of the Poisson Distribution)

A Poisson distribution with intensity λ has:

1. mean λ;
2. variance λ (no, that's not an accidentally repeated line or typo).

I described the Poisson distribution as a natural model for counts of randomly distributed points along a time axis. But it doesn't really matter that this is a time axis—it could be a space axis instead. For example, you could take a length of road, divide it into even intervals, then count the number of road-killed animals is in each interval. If the location of each animal is independent of the location of any other animal, then you could expect a Poisson model to apply to the count data. Assume that the Poisson model that best describes the data has parameter λ. One property of such models is that if you doubled the length of the intervals, then the resulting dataset would be described by a Poisson model with parameter 2λ; similarly, if you halved the length of the intervals, the best model would have parameter $\lambda/2$. This corresponds to our intuition about such data; roughly, the number of road-killed animals in two miles of road should be twice the number in one mile of road. This property means that no pieces of the road are "special"—each behaves the same as the other.

We can build a really useful model of spatial randomness by observing this fact and generalizing very slightly. A **Poisson point process** with intensity λ is a set of random points with the property that the number of points in an interval of length s is a Poisson random variable with parameter λs. Notice how this captures our intuition that if points are "very randomly" distributed, there should be twice as many of them in an interval that is twice as long.

This model is easily, and very usefully, extended to points on the plane, on surfaces, and in 3D. In each case, the process is defined on a domain D (which has to meet some very minor conditions that are of no interest to us). The number of points in any subset s of D is a Poisson random variable, with intensity $\lambda m(s)$, where $m(s)$ is the area (resp. volume) of s. These models are useful, because they capture the property that (a) the points are random and (b) the probability you find a point doesn't depend on where you are. You could reasonably believe models like this apply to, say, dead flies on windscreens; the places where you find acorns at the foot of an oak tree; the distribution of cowpats in a field; the distribution of cherries in a fruitcake; and so on.

5.2 Continuous Distributions

5.2.1 The Continuous Uniform Distribution

Some continuous random variables have a natural upper bound and a natural lower bound but otherwise we know nothing about them. For example, imagine we are given a coin of unknown properties by someone who is known to be a skillful maker of unfair coins. The manufacturer makes no representations as to the behavior of the coin. The probability that this coin will come up heads is a random variable, about which we know nothing except that it has a lower bound of zero and an upper bound of one. If we know nothing about a random variable apart from the fact that it has a lower and an upper bound, then a **uniform distribution** is a natural model. A continuous random variable whose probability distribution is the uniform distribution is often called a **uniform random variable**.

Definition 5.7 (Uniform Distribution, Continuous) Write l for the lower bound and u for the upper bound. The probability density function for the uniform distribution is

$$p(x) = \begin{cases} 0 & x < l \\ 1/(u-l) & l \leq x \leq u \\ 0 & x > u \end{cases}$$

5.2.2 The Beta Distribution

It's hard to explain now why the Beta (or β) distribution is useful, but it will come in useful later (Sect. 9.2.1). The Beta distribution is a probability distribution for a continuous random variable x in the range $0 \leq x \leq 1$. There are two parameters, $\alpha > 0$ and $\beta > 0$. Recall the definition of the Γ function from Sect. 15.2.

Definition 5.8 (Beta Distribution) A continuous random variable x in the range $0 \leq x \leq 1$ has a Beta distribution if its probability density function has the form

$$P_\beta(x|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{(\alpha-1)}(1-x)^{(\beta-1)}.$$

where $\alpha > 0$ and $\beta > 0$.

From the expression for the Beta distribution, you can see that:

- $P_\beta(x|1, 1)$ is a uniform distribution on the unit interval.
- For $\alpha > 1, \beta > 1)$, $P_\beta(x|\alpha, \beta)$ has a single maximum at $x = (\alpha - 1)/(\alpha + \beta - 2)$ (differentiate and set to zero).
- Generally, as α and β get larger, this peak gets narrower.
- For $\alpha = 1, \beta > 1$ the largest value of $P_\beta(x|\alpha, \beta)$ is at $x = 0$.
- For $\alpha > 1, \beta = 1$ the largest value of $P_\beta(x|\alpha, \beta)$ is at $x = 1$.

Figure 5.1 shows plots of the probability density function of the Beta distribution for a variety of different values of α and β.

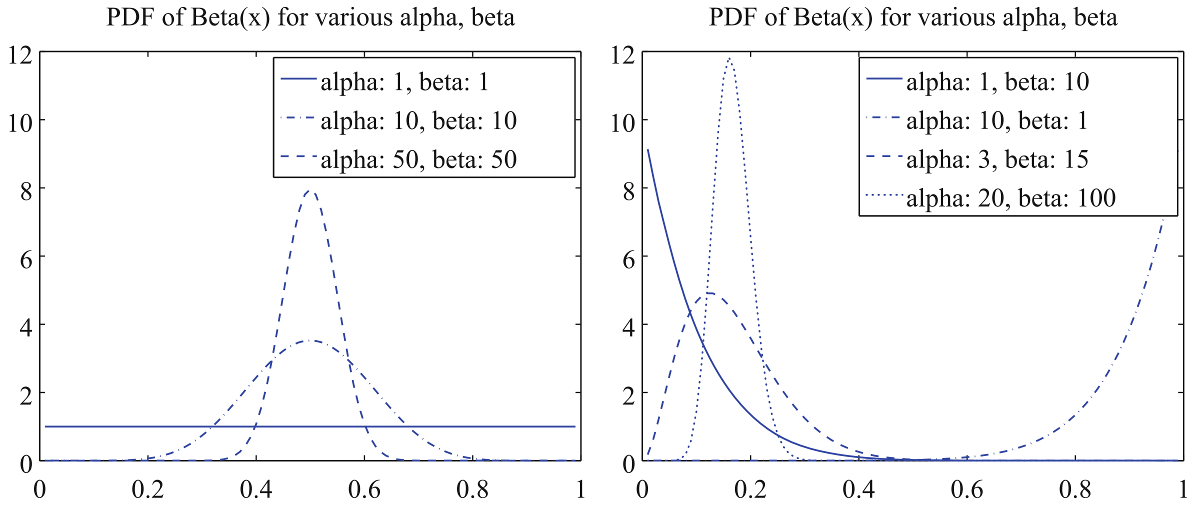

Fig. 5.1 Probability density functions for the Beta distribution with a variety of different choices of α and β

Useful Facts 5.5 (Mean and Variance of a Beta Distribution)

A Beta distribution with parameters α, β has:

1. mean $\frac{\alpha}{\alpha+\beta}$;
2. variance $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

5.2.3 The Gamma Distribution

The Gamma (or γ) distribution will also come in useful later on (Sect. 9.2.1). The Gamma distribution is a probability distribution for a non-negative continuous random variable $x \geq 0$. There are two parameters, $\alpha > 0$ and $\beta > 0$.

Definition 5.9 (Gamma Distribution) A non-negative continuous random variable x has a Gamma distribution if its probability density function is

$$P_\gamma(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{(\alpha-1)} e^{-\beta x}.$$

where $\alpha > 0$ and $\beta > 0$.

Figure 5.2 shows plots of the probability density function of the Gamma distribution for a variety of different values of α and β.

Useful Facts 5.6 (Mean and Variance of the Gamma Distribution)

A Gamma distribution with parameters α, β has:

1. mean $\frac{\alpha}{\beta}$;
2. variance $\frac{\alpha}{\beta^2}$.

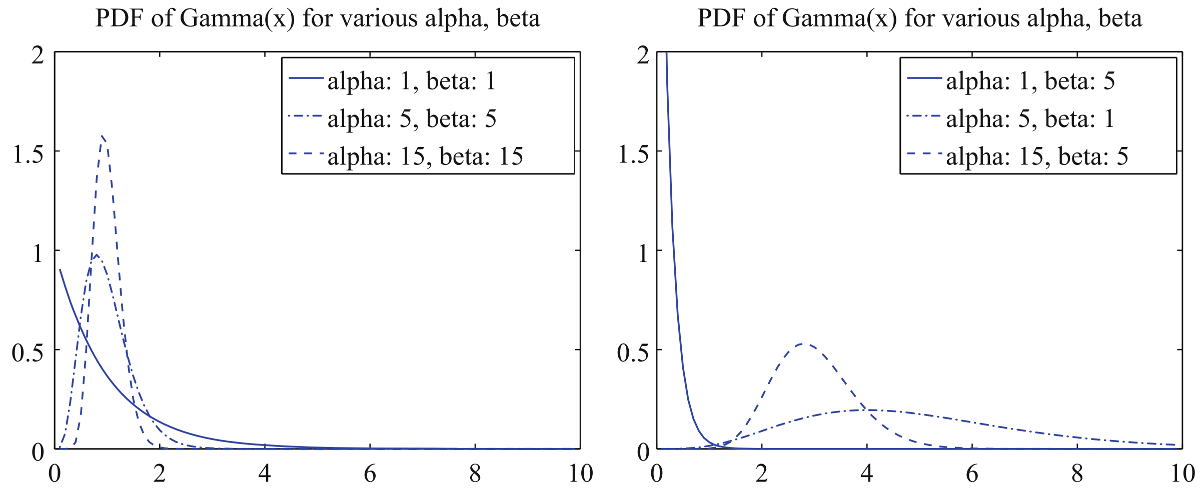

Fig. 5.2 Probability density functions for the Gamma distribution with a variety of different choices of α and β

5.2.4 The Exponential Distribution

Assume we have an infinite interval of time or space, with points distributed on it. Assume these points form a Poisson point process, as above. For example, we might consider the times at which email arrives; or the times at which phone calls arrive at a large telephone exchange; or the locations of roadkill on a road. The distance (or span of time) between two consecutive points is a random variable X. This random variable takes an exponential distribution, defined below. There is a single parameter, $\lambda > 0$. This distribution is often useful in modelling the failure of objects. We assume that failures form a Poisson process in time; then the time to the next failure is exponentially distributed.

Definition 5.10 (Exponential Distribution) A continuous random variable x has an exponential distribution when its probability density function takes the form

$$P_{\exp}(x|\lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{otherwise} \end{cases}.$$

where $\lambda > 0$ is a parameter.

Useful Facts 5.7 (Mean and Variance of the Exponential Distribution)

An exponential distribution with parameter λ has

1. mean $\frac{1}{\lambda}$;
2. variance $\frac{1}{\lambda^2}$.

Notice the relationship between this parameter and the parameter of the Poisson distribution. If (say) the phone calls are distributed with Poisson distribution with intensity λ (per hour), then your expected number of calls per hour is λ. The time between calls will be exponentially distributed with parameter λ, and the expected time to the next call is $1/\lambda$ (in hours).

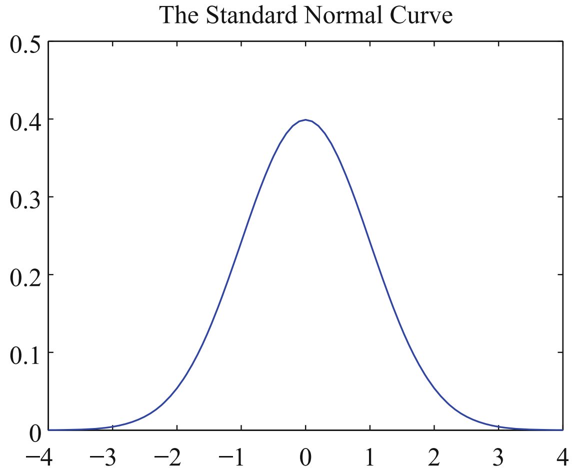

Fig. 5.3 A plot of the probability density function of the standard normal distribution. Notice how probability is concentrated around zero, and how there is relatively little probability density for numbers with large absolute values

5.3 The Normal Distribution

Many real datasets have histograms that look like a "bump", and the probability density function for a normal distribution looks like a "bump", too. Some of this is just an experimental fact of life. But there are important mathematical reasons that normal distributions should be common. Imagine your data is a sum of random variables (say, you are measuring the weight of a net full of fishes). Then pretty much however the original random variables are distributed, your data will be normally distributed.

5.3.1 The Standard Normal Distribution

Definition 5.11 (Standard Normal Distribution) The probability density function

$$p(x) = \left(\frac{1}{\sqrt{2\pi}}\right) \exp\left(\frac{-x^2}{2}\right).$$

is known as the **standard normal distribution**

The first step is to plot this probability density function (Fig. 5.3). You should notice it is quite familiar from work on histograms, etc. in chapter Worked example 14.13. It has the shape of the histogram of standard normal data, or at least the shape that the histogram of standard normal data aspires to.

Useful Facts 5.8 (Mean and Variance of the Standard Normal Distribution)

The standard normal distribution has:

1. mean 0;
2. variance 1.

These results are easily established by looking up (or doing!) the relevant integrals; they are relegated to the exercises.

A continuous random variable is a **standard normal random variable** if its probability density function is a standard normal distribution.

5.3.2 The Normal Distribution

Any probability density function that is a standard normal distribution *in standard coordinates* is a **normal distribution**. Now write μ for the mean of a random variable and σ for its standard deviation; we are saying that, if

$$\frac{x-\mu}{\sigma}$$

has a standard normal distribution, then $p(x)$ is a normal distribution. We can work out the form of the probability density function of a general normal distribution in two steps: first, we notice that for any normal distribution, we must have

$$p(x) \propto \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right].$$

But, for this to be a probability density function, we must have $\int_{-\infty}^{\infty} p(x)dx = 1$. This yields the constant of proportionality, and we get

Definition 5.12 (Normal Distribution) The probability density function

$$p(x) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right).$$

is a normal distribution.

Useful Facts 5.9 (Mean and Variance of the Normal Distribution)

The probability density function

$$p(x) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right).$$

has:

1. mean μ;
2. and variance σ^2.

These results are easily established by looking up (or doing!) the relevant integrals; they are relegated to the exercises.

A continuous random variable is a **normal random variable** if its probability density function is a **normal distribution**. Notice that it is quite usual to call normal distributions **gaussian distributions**.

5.3.3 Properties of the Normal Distribution

Normal distributions are important, because one often runs into data that is well described by a normal distribution. It turns out that anything that behaves like a binomial distribution with a lot of trials—for example, the number of heads in many coin tosses; as another example, the percentage of times you get the outcome of interest in a simulation in many runs—should produce a normal distribution (Sect. 5.4). For this reason, pretty much any experiment where you perform a simulation, then count to estimate a probability or an expectation, should give you an answer that has a normal distribution.

It is a remarkable and deep fact that adding many independent random variables produces a normal distribution pretty much *whatever* the distributions of those random variables. Because it's important, exciting and non-obvious, this has been proved in various forms by many major mathematicians. It was the subject of Alan Turing's Fellowship Thesis in 1934, where the story goes that examiners didn't quite know how to react: enthusiasm for a novel and brilliant form of proof, or irritation because he didn't already know the theorem.

I've not done this in detail because it's a nuisance to state in detail and to prove. However, you should remember that, if you add together many random variables, each of pretty much any distribution, then the answer has a distribution close to the normal distribution. It turns out that many of the processes we observe add up subsidiary random variables. This means that you will see normal distributions very often in practice.

Remember this: *The central limit theorem means that, under some not very worrying technical conditions, the sum of a large number of independent random variables will be very close to normal. The details are beyond our reach technically; the fact is extremely important.*

A normal random variable tends to take values that are quite close to the mean, measured in standard deviation units. We can demonstrate this important fact by computing the probability that a standard normal random variable lies between u and v. We form

$$\int_u^v \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du.$$

It turns out that this integral can be evaluated relatively easily using a special function. The **error function** is defined by

$$\mathsf{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp\left(-t^2\right) dt$$

so that

$$\frac{1}{2}\mathsf{erf}\left((\frac{x}{\sqrt{2}})\right) = \int_0^x \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du.$$

Notice that $\mathsf{erf}(x)$ is an odd function (i.e. $\mathsf{erf}(-x) = \mathsf{erf}(x)$). From this (and tables for the error function, or your favorite math package) we get that, for a standard normal random variable

$$\frac{1}{\sqrt{2\pi}} \int_{-1}^1 \exp\left(-\frac{x^2}{2}\right) dx \approx 0.68$$

and

$$\frac{1}{\sqrt{2\pi}} \int_{-2}^2 \exp\left(-\frac{x^2}{2}\right) dx \approx 0.95$$

and

$$\frac{1}{\sqrt{2\pi}} \int_{-3}^3 \exp\left(-\frac{x^2}{2}\right) dx \approx 0.99.$$

These are very strong statements. They measure how often a standard normal random variable has values that are in the range $-1, 1$, $-2, 2$, and $-3, 3$ respectively. But these measurements apply to normal random variables if we recognize that they now measure how often the normal random variable is some number of standard deviations away from the mean. In particular, it is worth remembering that:

Useful Facts 5.10 (How Often a Normal Random Variable is How Far from the Mean)

- About 68% of the time, a normal random variable takes a value within one standard deviation of the mean.
- About 95% of the time, a normal random variable takes a value within two standard deviations of the mean.
- About 99% of the time, a normal random variable takes a value within three standard deviations of the mean.

5.4 Approximating Binomials with Large N

The Binomial distribution appears to be a straightforward thing. We assume we flip a coin N times, where N is a very large number. The coin has probability p of coming up heads, and so probability $q = 1 - p$ of coming up tails. The number of heads h follows the binomial distribution, so

$$P(h) = \frac{N!}{h!(N-h)!} p^h q^{(N-h)}$$

The mean of this distribution is Np, the variance is Npq, and the standard deviation is $\sqrt{Npq}$.

Evaluating this probability distribution for large N is very difficult, because factorials grow fast. We will construct an approximation to the binomial distribution for large N that allows us to evaluate the probability that h lies in some range.

Notice that h/N is particularly interesting, because this is the fraction of flips that comes up heads. We are dividing by a constant, so the expected value of h/N is p and the standard deviation is $pq/\sqrt{N}$. Our approximation will show that the probability that h/N is within one standard deviation of the mean is approximately 68%. Note the standard deviation of the mean falls as N grows. This is important, because it shows that our model of probability as frequency is consistent. As $N \to \infty$,

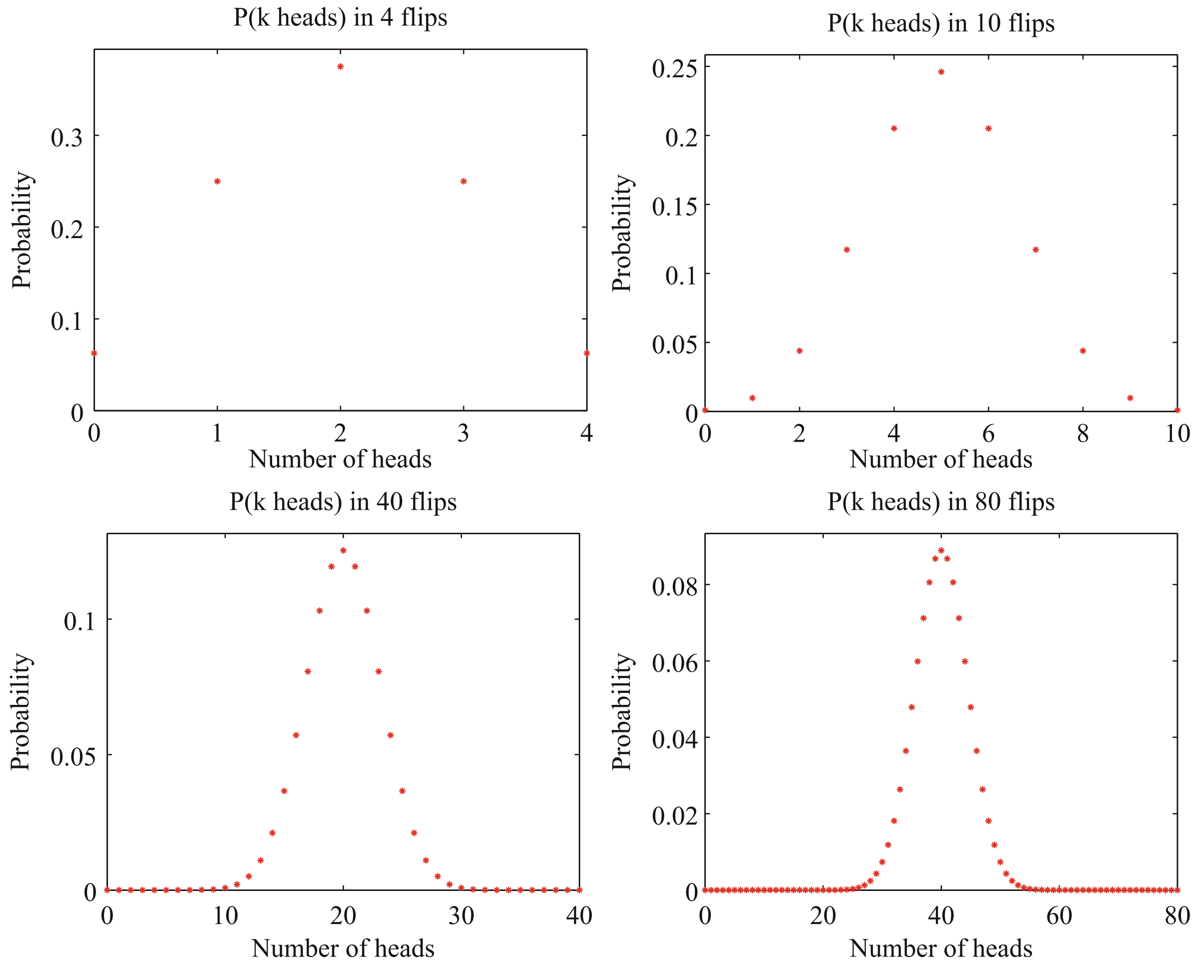

Fig. 5.4 Plots of the binomial distribution for $p = q = 0.5$ for different values of N. You should notice that the set of values of h (the number of heads) that have substantial probability is quite narrow compared to the range of possible values. This set gets narrower as the number of flips increases. This is because the mean is pN and the standard deviation is $\sqrt{Npq}$—so the fraction of values that is within one standard deviation of the mean is $O(1/\sqrt{N})$

$$\frac{h}{N} \to p$$

because h/N will tend to land in an interval around p that gets narrower as N gets larger.

The main difficulty with Fig. 5.4 (and with the argument above) is that the mean and standard deviation of the binomial distribution tends to infinity as the number of coin flips tends to infinity. This can confuse issues. For example, the plots of Fig. 5.4 show narrowing probability distributions—but is this because the scale is compacted, or is there a real effect? It turns out there is a real effect, and a good way to see it is to consider the normalized number of heads.

5.4.1 Large N

Recall that to normalize a dataset, you subtract the mean and divide the result by the standard deviation. We can do the same for a random variable. We now consider

$$x = \frac{h - Np}{\sqrt{Npq}}.$$

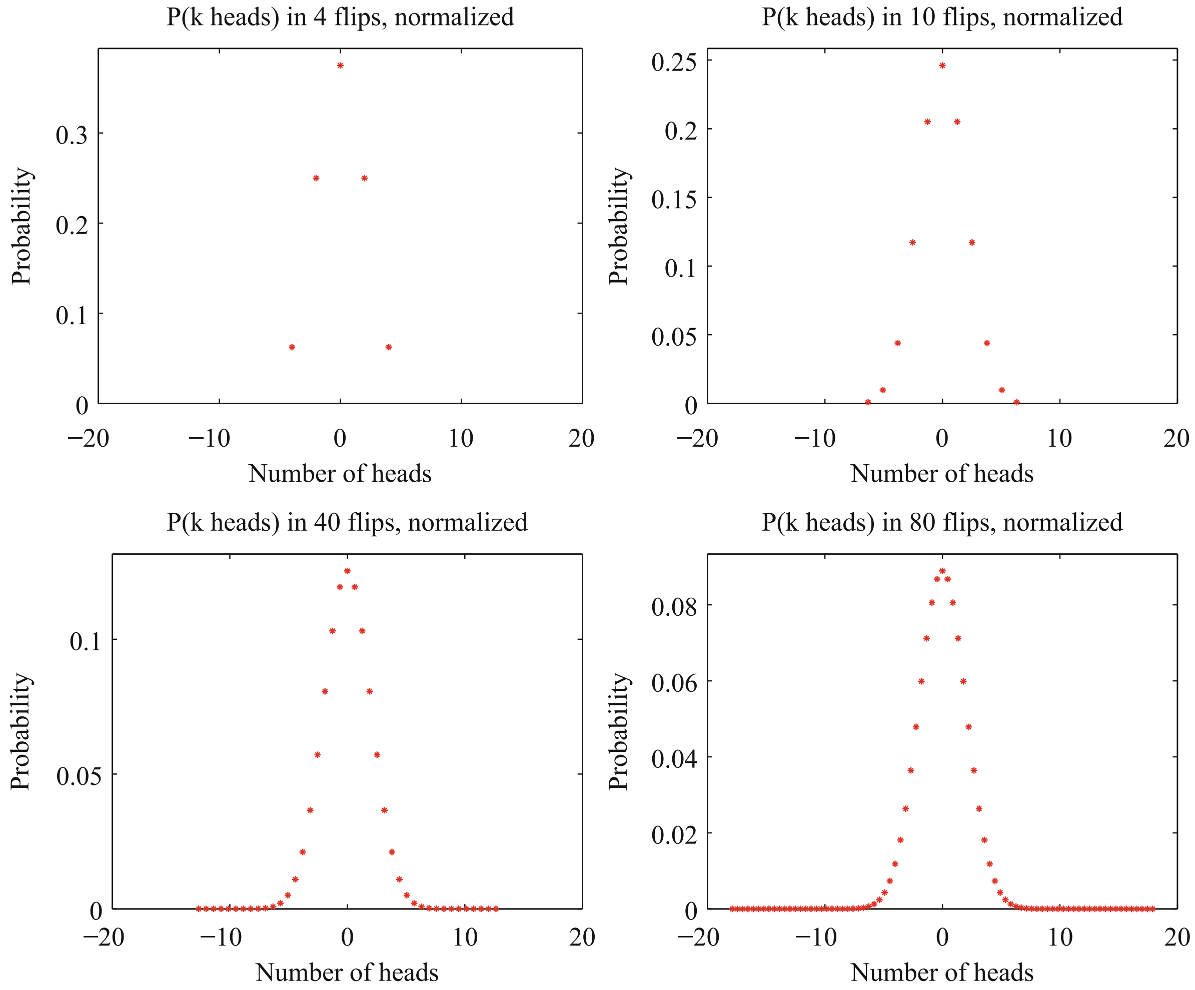

Fig. 5.5 Plots of the distribution for the normalized variable x, with $P(x)$ given in the text, obtained from the binomial distribution with $p = q = 0.5$ for different values of N. These distributions are normalized (mean 0, variance 1. They look increasingly like a standard normal distribution EXCEPT that the value at their mode gets smaller as N gets bigger (look at the vertical axis; this occurs because there are more possible outcomes). In the text, we will establish that the standard normal distribution is a limit, in a useful sense

The probability distribution of x can be obtained from the probability distribution for h, because $h = Np + x\sqrt{Npq}$, so

$$P(x) = \left(\frac{N!}{(Np + x\sqrt{Npq})!(Nq - x\sqrt{Npq})!} \right) p^{(Np+x\sqrt{Npq})} q^{(Nq-x\sqrt{Npq})}.$$

I have plotted this probability distribution for various values of N in Fig. 5.5.

But it is hard to work with this distribution for very large N. The factorials become very difficult to evaluate. Second, it is a discrete distribution on N points, spaced $1/\sqrt{Npq}$ apart. As N becomes very large, the number of points that have non-zero probability becomes very large, and x can be very large, or very small. For example, there is some probability, though there may be very little indeed, on the point where $h = N$, or, equivalently, $x = N(p + \sqrt{Npq})$. For sufficiently large N, we think of this probability distribution as a probability density function. We can do so, for example, by spreading the probability for x_i (the i'th value of x) evenly over the interval between x_i and x_{i+1}. We then have a probability density function that looks like a histogram, with bars that become narrower as N increases. But what is the limit?

5.4.2 Getting Normal

To proceed, we need Stirling's approximation, which says that, for large N,

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

This yields

$$P(h) \approx \left(\frac{Np}{h}\right)^h \left(\frac{Nq}{N-h}\right)^{(N-h)} \sqrt{\frac{N}{2\pi h(N-h)}}$$

Recall we used the normalized variable

$$x = \frac{h - Np}{\sqrt{Npq}}.$$

We will encounter the term $\sqrt{Npq}$ often, and we use $\sigma = \sqrt{Npq}$ as a shorthand. We can compute h and $N - h$ from x by the equalities

$$h = Np + \sigma x \qquad N - h = Nq - \sigma x.$$

So the probability distribution written in this new variable x is

$$P(x) \approx \left(\frac{Np}{(Np+\sigma x)}\right)^{(Np+\sigma x)} \left(\frac{Nq}{(Nq-\sigma x)}\right)^{(Nq-\sigma x)} \sqrt{\frac{N}{2\pi(Np + \sigma x)(Nq - \sigma x)}}$$

There are three terms to deal with here. It is easiest to work with $\log P$. Now

$$\log(1 + x) = x - \frac{1}{2}x^2 + O(x^3)$$

so we have

$$\begin{aligned}\log\left(\frac{Np}{(Np + \sigma x)}\right) &= -\log\left(1 + \frac{\sigma x}{Np}\right) \\ &\approx -\frac{\sigma x}{Np} + (\frac{1}{2})(\frac{\sigma x}{Np})^2\end{aligned}$$

and

$$\log\left(\frac{Nq}{(Nq-\sigma x)}\right)\approx\frac{\sigma x}{Nq}+(\frac{1}{2})(\frac{\sigma x}{Nq})^2.$$

From this, we have that

$$\log\left[\left(\frac{Np}{Np+\sigma x}\right)^{(Np+\sigma x)}\left(\frac{Nq}{Nq-\sigma x}\right)^{(Nq-\sigma x)}\right]$$

is approximately

$$[Np+\sigma x]\left[-\frac{\sigma x}{Np}+\left(\frac{1}{2}\right)\left(\frac{\sigma x}{Np}\right)^2\right]+[Nq-\sigma x]\left[\frac{\sigma x}{Nq}+\left(\frac{1}{2}\right)\left(\frac{\sigma x}{Nq}\right)^2\right]$$

which is

$$-\left(\frac{1}{2}\right)x^2+O((\sigma x)^3)$$

(recall $\sigma=\sqrt{Npq}$ if you're having trouble with the last step). Now we look at the square-root term. We have

$$\begin{aligned}\log\sqrt{\frac{N}{2\pi(Np+\sigma x)(Nq-\sigma x)}}&=-\frac{1}{2}\begin{pmatrix}\log\left[Np+\sigma x\right]\\+\log\left[Nq-\sigma x\right]\\-\log N\\+\log 2\pi\end{pmatrix}\\&=-\frac{1}{2}\begin{pmatrix}\log Np+O\left(\left(\frac{\sigma x}{Np}\right)\right)\\+\log Nq-O\left(\left(\frac{\sigma x}{Nq}\right)\right)\\-\log N\\+\log 2\pi\end{pmatrix}\end{aligned}$$

but, since N is very large compared to σx, we can ignore the $O(\left(\frac{\sigma x}{Np}\right))$ terms. Then this term is not a function of x. So we have

$$\log P(x)\approx\frac{-x^2}{2}+\text{constant}.$$

Now because N is very large, our probability distribution P limits to a probability density function p, with

$$p(x)\propto\exp\left(\frac{-x^2}{2}\right).$$

We can get the constant of proportionality from integrating, to

$$p(x)=\left(\frac{1}{\sqrt{2\pi}}\right)\exp\left(\frac{-x^2}{2}\right).$$

This constant of proportionality deals with the effect in Fig. 5.5, where the mode of the distribution gets smaller as N gets bigger. It does so because there are more points with non-zero probability to be accounted for. But we are interested in the limit where N tends to infinity. This must be a probability density function, so it must integrate to one.

Review this blizzard of terms. We started with a binomial distribution, but standardized the variables so that the mean was zero and the standard deviation was one. We then assumed there was a very large number of coin tosses, so large that the distribution started to look like a continuous function. The function we get is the standard normal distribution.

5.4.3 Using a Normal Approximation to the Binomial Distribution

I have proven an extremely useful fact, which I shall now put in a box.

Useful Facts 5.11 (Binomial Distribution for Large N)
Assume h follows the binomial distribution with parameters p and q. Write

$$x = \frac{h - Np}{\sqrt{Npq}}.$$

Then, for sufficiently large N, the probability distribution $P(x)$ can be approximated by the probability density function

$$\left(\frac{1}{\sqrt{2\pi}}\right) \exp\left(\frac{-x^2}{2}\right)$$

in the sense that

$$P(\{x \in [a, b]\}) \approx \int_a^b \left(\frac{1}{\sqrt{2\pi}}\right) \exp\left(\frac{-u^2}{2}\right) du$$

This justifies our model of probability as frequency. I interpreted an event having probability p to mean that, if I had a large number N of independent repetitions of the experiment, the number that produced the event would be close to Np, and would get closer as N got larger. We know that, for example, 68% of the time a standard normal random variable takes a value between 1 and -1. In this case, the standard normal random variable is

$$\frac{h - (Np)}{\sqrt{Npq}}$$

so that 68% of the time, h must take a value in the range $[Np - \sqrt{Npq}, Np + \sqrt{Npq}]$. Equivalently, the relative frequency h/N must take a value in the range

$$[p - \frac{pq}{\sqrt{N}}, p + \frac{pq}{\sqrt{N}}]$$

but as $N \to \infty$ this range gets smaller and smaller, and h/N limits to p. So our view of probability as a frequency is consistent.

5.5 You Should

5.5.1 Remember These Definitions

Uniform random variable, discrete	115
Bernoulli random variable	116
Geometric distribution	116
Binomial distribution	117
Multinomial distribution	118
Poisson distribution	118
Uniform distribution, continuous	120
Beta distribution	120
Gamma distribution	121
Exponential distribution	122
Standard Normal distribution	123
Normal distribution	124

5.5.2 Remember These Terms

intensity	119
Poisson point process	119

5.5.3 Remember These Facts

5.5.4 Remember These Points

Problems

Sums and Differences of Discrete Random Variables

5.1 Assume X and Y are discrete random variables which take integer values in the range $1 \ldots 100$ (inclusive). Write $S = X + Y$.

(a) Show that

$$P(S = k) = \sum_{u=1}^{u=100} P(\{\{X = k - u\} \cap \{Y = u\}\}).$$

(b) Now assume that both X and Y are uniform random variables. Show that S is not uniform by considering $P(S = 2)$, $P(S = 3)$, and $P(S = 100)$.

5.2 Assume X and Y are discrete random variables which take integer values in the range $1 \ldots 100$ (inclusive). Write $D = X - Y$.

(a) Show that

$$P(D = k) = \sum_{u=1}^{u=100} P(\{X = k + u\})P(\{Y = u\}).$$

(b) Now assume that both X and Y are uniform random variables. Show that D is not uniform by considering $P(D = -99)$, $P(D = 99)$, and $P(D = 0)$.

The Geometric Distribution

5.3 Write $S_\infty = \sum_{i=0}^{\infty} r^i$. Show that $(1-r)S_\infty = 1$, so that

$$S_\infty = \frac{1}{1-r}$$

5.4 Write $P(\{X = n\})$ for the probability that an experiment requires n repeats for success under the geometric distribution model with probability of success in one experiment p. Use the result of the previous exercise to show that

$$\begin{aligned}\sum_{n=1}^{\infty} P(\{X = n\}) &= p\sum_{n=1}^{\infty}(1-p)^{(n-1)} \\ &= 1\end{aligned}$$

5.5 Show that

$$\sum_{i=0}^{\infty} ir^i = (\sum_{i=1}^{\infty} r^i) + r(\sum_{i=1}^{\infty} r^i) + r^2(\sum_{i=1}^{\infty} r^i) + \ldots$$

(look carefully at the limits of the sums!) and so show that

$$\sum_{i=0}^{\infty} ir^i = \frac{r}{(1-r)^2}.$$

5.6 Write $S_\infty = \sum_{i=0}^{\infty} r^i$. Show that

$$\sum_{i=0}^{\infty} i^2 r^i = (S_\infty - 1) + 3r(S_\infty - 1) + 5r^2(S_\infty - 1) + 7r^3(\S_\infty - 1) + \ldots$$

and so that

$$\sum_{i=0}^{\infty} i^2 r^i = \frac{r(1+r)}{(1-r)^3}$$

5.7 Show that, for a geometric distribution with parameter p, the mean is

$$\sum_{i=1}^{\infty} i(1-p)^{(i-1)}p = \sum_{u=0}^{\infty}(u+1)(1-p)^u p.$$

Now by rearranging and using the previous results, show that the mean is

$$\sum_{i=1}^{\infty} i(1-p)^{(i-1)}p = \frac{1}{p}$$

5.8 Show that a geometric distribution with parameter p has variance $(1-p)/p^2$. To do this, note the variance is $\mathbb{E}[X^2] - \mathbb{E}[X]^2$. Now use the results of the previous exercises to show that

$$\mathbb{E}[X^2] = \sum_{i=1}^{\infty} i^2(1-p)^{(i-1)}p = \frac{p}{1-p}\frac{(1-p)(2-p)}{p^3},$$

then rearrange to get the expression for variance.

5.9 You have a coin with unknown probability p of coming up heads. You wish to generate a random variable which takes the values 0 and 1, each with probability $1/2$. Assume $0 < p < 1$. You adopt the following procedure. You start by flipping the coin twice. If both flips produce the same side of the coin, you start again. If the result of the first flip is different from the result of the second flip, you report the result of the first flip and you are finished (this is a trick originally due to John von Neumann).

(a) Show that, in this case, the probability of reporting heads is $1/2$.
(b) What is the expected number of flips you must make before you report a result?

Bernoulli Random Variables

5.10 Write X for a Bernoulli random variable which takes the value 1 with probability p (and 0 with probability $(1-p)$).

(a) Show that $\mathbb{E}[X] = p$.
(b) Show that the variance of X is $p(1-p)$.

5.11 Write $X^{(N)}$ for

$$\frac{1}{N}(X_1 + X_2 + \ldots X_N)$$

where the X_i are independent Bernoulli random variables. Each of these takes the value 1 with probability p (and 0 with probability $(1-p)$).

(a) Show that $\mathbb{E}\left[X^{(N)}\right] = p$.
(b) Show that the variance of $X^{(N)}$ is $p(1-p)$

5.12 Write $S^{(N)}$ for

$$(X_1 + X_2 + \ldots X_N)$$

where the X_i are independent Bernoulli random variables. Each of these takes the value 1 with probability p (and 0 with probability $(1-p)$).

(a) Show that, for $0 \leq k \leq N$, $P(\{X = k\})$ is

$$\binom{N}{k} p^k (1-p)^{(N-k)}$$

(b) Show that $\mathbb{E}\left[S^{(N)}\right] = Np$.
(c) Show that the variance of $S^{(N)}$ is $Np(1-p)$.

The Binomial Distribution

5.13 Show that $P_b(N-i; N, p) = P_b(i; N, (1-p))$ for all i.

5.14 Show that

$$\begin{aligned} &P_b(i; N, p) \\ &\quad = pP_b(i{-}1; N{-}1, p){+}(1{-}p)P_b(i; N{-}1, p). \end{aligned}$$

5.15 Write h_r for the number of heads obtained in r flips of a coin which has probability p of coming up heads. Compare the following two ways to compute the probability of getting i heads in five coin flips:

- Flip the coin three times, count h_3, then flip the coin twice, count h_2, then form $w = h_3 + h_2$.
- Flip the coin five times, and count h_5.

Show that the probability distribution for w is the same as the probability distribution for h_5. Do this by showing that

$$P(\{w = i\}) = \left[\sum_{j=0}^{5} P(\{h_3 = j\} \cap \{h_2 = i - j\})\right] = P(\{h_5 = i\}).$$

5.16 Now we will do the previous exercise in a more general form. Again, write h_r for the number of heads obtained in r flips of a coin which has probability p of coming up heads. Compare the following two ways to compute the probability of getting i heads in N coin flips:

- Flip the coin t times, count h_t, then flip the coin $N - t$ times, count h_{N-t}, then form $w = h_t + h_{N-t}$.
- Flip the coin N times, and count h_N.

Show that the probability distribution for w is the same as the probability distribution for h_N. Do this by showing that

$$P(\{w = i\}) = \left[\sum_{j=0}^{N} P(\{h_t = j\} \cap \{h_{N-t} = i - j\})\right] = P(\{h_N = i\}).$$

5.17 An airline runs a regular flight with six seats on it. The airline sells six tickets. The gender of the passengers is unknown at time of sale, but women are as common as men in the population. All passengers always turn up for the flight. The pilot is eccentric, and will not fly a plane unless at least one passenger is female. What is the probability that the pilot flies?

5.18 An airline runs a regular flight with s seats on it. The airline always sells t tickets for this flight. The probability a passenger turns up for departure is p, and passengers do this independently. What is the probability that the plane travels with exactly three empty seats?

5.19 An airline runs a regular flight with s seats on it. The airline always sells t tickets for this flight. The probability a passenger turns up for departure is p, and passengers do this independently. What is the probability that the plane travels with 1 or more empty seats?

The Multinomial Distribution

5.20 Show that the multinomial distribution

$$P_m(n_1, \dots, n_k; N, p_1, \dots, n_k) = \frac{N!}{n_1!n_2!\dots n_k!}$$

$$p_1^{n_1} p_2^{n_2} \dots p_k^{n_k}$$

must satisfy the recurrence relation

$$P_m(n_1, \dots, n_k; N, p_1, \dots, p_k)$$

$$= p_1 P_m(n_1 - 1, \dots, n_k; N - 1, p_1, \dots, p_k) +$$

$$p_2 P_m(n_1, n_2 - 1, \dots, n_k; N - 1, p_1, \dots, p_k) + \dots$$

$$p_k P_m(n_1, n_2, \dots, n_k - 1; N - 1, p_1, \dots, p_k)$$

The Poisson Distribution

5.21 The exponential function e^x can be represented by the series

$$\sum_{i=0}^{\infty} \frac{x^i}{i!}$$

(which converges absolutely; try the ratio test). Use this information to show that the Poisson distribution sums to one.

5.22 You will show that the mean of the Poisson distribution with intensity parameter λ is λ.

(a) Show that Taylor series for xe^x around $x = 0$ is given by

$$\sum_{i=0}^{\infty} \frac{ix^i}{i!}$$

and use the ratio test to show that this series converges absolutely.

(b) Now use this series and pattern matching to show the mean of the Poisson distribution with intensity parameter λ is λ.

5.23 Compute the Taylor series for $(x^2 + x)e^x$ around $x = 0$. Show that this series converges absolutely, using the ratio test. Use this and pattern matching to show that the variance of the Poisson distribution with intensity parameter λ is λ.

Sums of Continuous Random Variables

5.24 Write p_x for the probability density function of a continuous random variable X and p_y for the probability density function of a continuous random variable Y. Show that the probability density function of $S = X + Y$ is

$$p(s) = \int_{-\infty}^{\infty} p_x(s-u)p_y(u)du = \int_{-\infty}^{\infty} p_x(u)p_y(s-u)du$$

The Normal Distribution

5.25 Write

$$f(x) = \left(\frac{1}{\sqrt{2\pi}}\right)\exp\left(\frac{-x^2}{2}\right).$$

(a) Show that $f(x)$ is non-negative for all x.

(b) By integration, show that

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

so that $f(x)$ is a probability density function (you can look up the integral; few people remember how to do this integral these days).

(c) Show that

$$\int_{-\infty}^{\infty} xf(x)dx = 0.$$

The easiest way to do this is to notice that $f(x) = f(-x)$

(d) Show that

$$\int_{-\infty}^{\infty} xf(x-\mu)dx = \mu.$$

The easiest way to do this is to change variables, and use the previous two exercises.

(e) Show that

$$\int_{-\infty}^{\infty} x^2 f(x) dx = 1.$$

You'll need to either do, or look up, the integral to do this exercise.

5.26 Write

$$g(x) = \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

Show that

$$\int_{-\infty}^{\infty} g(x) dx = \sqrt{2\pi}\sigma.$$

You can do this by a change of variable, and the results of the previous exercises.

5.27 Write

$$p(x) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right) \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right).$$

(a) Show that

$$\int_{-\infty}^{\infty} x p(x) dx = \mu$$

using the results of the previous exercises.

(b) Show that

$$\int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx = \sigma^2$$

using the results of the previous exercises.

The Binomial Distribution for Large *N*

5.28 I flip a fair coin *N* times and count heads. We consider the probability that *h*, the fraction of heads, is in some range of numbers. For each of these questions, you should just write an expression, rather than evaluate the integral. *Hint:* If you know the range of numbers for *h*, you know the range for h/N.

(a) For $N = 1e6$, use the normal approximation to estimate

$$P(\{h \in [49{,}500, 50{,}500]\}).$$

(b) For $N = 1e4$, use the normal approximation to estimate

$$P(\{h > 9000\}).$$

(c) For $N = 1e2$, use the normal approximation to estimate

$$P(\{h > 60\} \cup \{h < 40\}).$$

Programming Exercises

5.29 An airline runs a regular flight with 10 seats on it. The probability that a passenger turns up for the flight is 0.95. What is the smallest number of seats the airline should sell to ensure that the probability the flight is full (i.e. 10 or more passengers turn up) is bigger than 0.99? You'll need to write a simple simulation; estimate the probability by counting.

5.30 You will plot a series of figures showing how the binomial distribution for large N increasingly "looks like" the normal distribution. We will consider the number of heads h in N flips of an unbiased coin (so $P(H) = P(T) = 1/2 = p$, and in this case $q = 1 - p = 1/2$). Write $x = \frac{h-Np}{\sqrt{Npq}}$.

(a) Prepare plots of the probability distribution of x for $N = 10$, $N = 30$, $N = 60$, and $N = 100$. These should be superimposed on the same set of axes. On this set of axes, you should also plot the normal probability distribution.

(b) Evaluate $P(\{x \geq 2\})$ for each case by summing over the appropriate terms in the binomial distribution. Now compare this to the prediction that the approximation would make, which is

$$\int_2^{\infty} \frac{1}{\sqrt{2\pi}} e^{[-u^2/2]} du.$$

You can obtain this number by appropriate evaluation of error functions.

(c) Now you will write a program to simulate coin flips and evaluate the variance of the simulated value of x for different numbers of flips. Again, the coin should be fair. For each N from $10, 40, 90, 160, 250, 490, 640, 810, 1000$, estimate the value of x by simulating that number of flips. You should run each simulation 100 times, and use the set of estimates to evaluate the variance of your estimate of x. Plot this variance against N and against $1/\sqrt{N}$—what do you see?