Unit 5

Events occurring at random and population quantiles

Introduction

The term Bernoulli trial was introduced in Unit 3 to describe an experiment for which there are just two outcomes – success or failure, 1 or 0, win or lose, and so on. In Unit 3, you met two probability distributions directly associated with a collection or sequence of independent Bernoulli trials in which the probability of success remains constant from trial to trial. The first, the binomial distribution, was developed as a model for the total number of successes in n trials. The second, the geometric distribution, is the distribution of the number of trials up to and including the first success.

In Section 1 of this unit, we look again, but from a more practical viewpoint, at the distribution of the number of successes in a sequence of n Bernoulli trials when the number of trials is large, but the chance of success in each trial is small. In this case, where something rather unlikely has many opportunities to happen, the distribution of the number of successes has a very useful approximate form, as you saw in Section 4 of Unit 3: the probability model which gives the appropriate approximation to the binomial distribution in this sort of situation is the Poisson distribution. The result is sometimes known as *Poisson's approximation for the occurrence of rare events*. It will be central to developments later in the unit. (Of course, the Poisson distribution is also an extremely important discrete probability distribution in its own right; some of its properties were given in Units 3 and 4.)

A rare event: weather conditions and the angle of the sun have to be just right for this waterfall in Yosemite National Park, USA, to look like a lava flow

A sequence of independent Bernoulli trials in which the probability of success remains constant from trial to trial is called a *Bernoulli process*. This is just the same notion as a collection or set of Bernoulli trials that you met earlier, but the word *process* is used to stress the sequential nature of trials in the contexts with which we are concerned in this unit. So a Bernoulli process is a useful model when an event either occurs (a success) or does not (a failure) at each of a clearly defined sequence of opportunities (trials) which take place one after another. The process is characterised by the assumptions that the probability that an event occurs is constant and is independent of the outcomes of previous trials. In this sense, occurrences of events are unpredictable: they occur at random points in the sequence of trials.

However, many events occur in a random, unpredictable way in continuous time, where there is no notion of a trial. For example, domestic accidents, air crashes, floods and earthquakes may occur at *any* time, not just at a discrete set of opportunities. In Section 2, we look first at the distribution of the waiting time between events in a Bernoulli process, and then at an analogous continuous model for the waiting times between events that may occur at any time. The latter model is known as the *exponential distribution*.

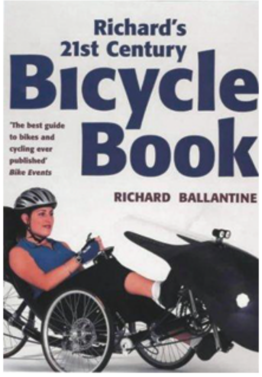
In Section 3, another model is introduced, analogous to the Bernoulli process, but for events occurring at random in continuous time. This model is the *Poisson process*. It transpires that the number of events that occur in a Poisson process in any given interval of time has a Poisson distribution, and the waiting time between successive events in a Poisson process has an exponential distribution. The work of Subsection 3.2 then considers how we can decide, given data on the times at which a sequence of events occur, whether or not a Poisson process is a good model for the occurrences of these events.

The final section of the unit, Section 4, is about something a bit different from the preceding sections of the unit: the population analogues of sample statistics such as the median and the quartiles. The results from this section will be applied to the models introduced in this unit, and will assume even greater importance in a number of other contexts later in the module.

1 Bernoulli trials and the Poisson distribution

In this section, we first investigate numerically the distribution of the number of successes in n Bernoulli trials when the number of trials is large but the chance of success, p, at each trial is small. This investigation gives a practical context to a result obtained mathematically in Section 4 of Unit 3: in the situation of many trials and small probability of success, the Poisson distribution provides a good approximation to the binomial distribution. The approximation is useful partly because Poisson probabilities are simpler than binomial probabilities when n is large and p is small, but also because the approximating Poisson distribution does not depend directly on the value of n. The latter is a feature that will assume central importance in Section 3. Finally, some further examples are given of situations where the Poisson approximation provides a useful model.

The number of successes in n independent Bernoulli trials, each with the same probability of success, p, has the binomial distribution, $B(n, p)$. We will now investigate this distribution for situations where n is large and p is small. The first situation we look at concerns bicycle usage and getting caught in a downpour.


Richard's
21st Century
Bicycle
Book
'The best guide to bikes and cycling ever published' Bike Events
RICHARD BALLANTINE

Ballantine, R. (2000) *Richard's 21st Century Bicycle Book*, Pan Books

Example 1 *Bicycle usage*

One disadvantage of cycling is the possibility of arriving at one's destination drenched with rain, having been caught in a downpour. This is a very unpleasant experience, but actually its frequency has been estimated at only 15 times a year for a regular cyclist.

It is not entirely clear how the estimate of 15 times a year has been calculated. So if we wish to use it for calculating the probability, say, of escaping a drenching over the next month, then we need to make some further assumptions. To simplify the situation, we will assume that the probability of getting drenched does not change from one journey to the next, and that whether or not a drenching occurs on a journey is independent of experiences on previous journeys. That is, we will assume that cycle rides form a sequence of independent Bernoulli trials with constant probability of a drenching (a 'success'!).

We are told that a 'regular cyclist' gets drenched approximately 15 times a year, on average, or $15/12 = 1.25$ times a month. However, we do not know n, how many cycle journeys a month a regular cyclist makes, or the value of p, the proportion of rides that result in a drenching. In Activities 1 and 2 you are asked to investigate the effect of assuming different numbers of journeys a month on the distribution of the number of drenchings in a month.

The simplifications we have made imply the rather unrealistic assumption that the average number of drenchings is the same for each month of the year.

Activity 1 *Drenchings for Chris*

Chris uses her bicycle regularly for travelling to and from work and, occasionally, for shopping and social visits. Suppose that she gets drenched 15 times a year on average. Her own assessment of cycle usage (which is fairly rough) works out at 50 journeys a month of a reasonable distance.

(a) Estimate the proportion of Chris's journeys that result in a drenching.

(b) Suggest a probability model for X, the number of times a month that Chris gets drenched.

(c) According to the model you suggested in part (b), what is the mean number of times that Chris gets drenched in a month?

(d) Use your model to calculate the probability that over a month:

(i) Chris does not get drenched at all

(ii) Chris gets drenched twice.

Chris's guess at a monthly total of 50 rides is just that – a guess. Suppose that a better estimate is 60 rides a month. In this case, using the given average of 15 drenchings a year, the estimated probability of a drenching during a single ride is

$$p = \frac{15}{12 \times 60} = \frac{15}{720} = \frac{1}{48}.$$

In this case, the assumed probability distribution for the number of drenchings a month is $B(60, 1/48)$.

Activity 2 *Probabilities of drenchings*

You calculated two of these probabilities in Activity 1.

Table 1 contains certain probabilities for drenchings assuming the binomial distribution $B(50, 1/40)$. It also contains some of the corresponding probabilities of drenchings assuming the binomial distribution $B(60, 1/48)$. Complete the table by computing the missing probabilities of drenchings assuming the binomial distribution $B(60, 1/48)$.

Table 1 More probabilities of drenchings while cycling

	$P(X = 0)$	$P(X = 1)$	$P(X = 2)$	$P(X = 3)$	$P(X \geq 4)$
$B(50, 1/40)$	0.2820	0.3615	0.2271	0.0932	0.0362
$B(60, 1/48)$		0.3610		0.0932	0.0365

What do you notice about these probabilities?

In Activity 2, you saw that the probabilities of drenchings that were calculated assuming the higher estimate of 60 rides a month are not very different from those calculated assuming 50 rides a month. Also, the mean number of times that Chris gets drenched according to this second model is $n \times p = 60 \times (1/48) = 1.25$, the same as the mean of the model assuming 50 rides a month (Activity 1(c)).

Now suppose that Chris actually averages 40 cycle rides a month. In this case, the estimated probability of a drenching during a single ride is $p = 1/32$; and the assumed probability distribution of the number of drenchings a month is $B(40, 1/32)$, which also has mean 1.25. Some probabilities of drenchings for 40 rides a month are given in Table 2, together with those previously calculated for 50 and 60 rides a month.

Table 2 Probabilities of drenchings while cycling

	$P(X = 0)$	$P(X = 1)$	$P(X = 2)$	$P(X = 3)$	$P(X \geq 4)$
$B(50, 1/40)$	0.2820	0.3615	0.2271	0.0932	0.0362
$B(60, 1/48)$	0.2827	0.3610	0.2266	0.0932	0.0365
$B(40, 1/32)$	0.2808	0.3624	0.2280	0.0931	0.0357

Notice that the values for 40 rides a month are close to those previously calculated for 50 and 60 rides a month.

These calculations have all been based on the assumption that the mean number of drenchings a month for a regular cyclist is 1.25. If the number of cycle rides a month is n, then the corresponding estimate for the probability p of a drenching in a single ride is

$$p = \frac{15}{12 \times n} = \frac{1.25}{n};$$

and the distribution used to model the number of drenchings a month is the binomial distribution $B(n, 1.25/n)$.

The results in Table 2 suggest that, for values of n such as those used (50, 60 and 40), probability calculations for the binomial distribution $B(n, 1.25/n)$ lead to very similar results. Notice that, whatever the value of n, the mean of this binomial distribution is $\mu = n \times p = 1.25$, the average number of drenchings a month. So the question arises: can we satisfactorily model the number of monthly drenchings for a regular cyclist using a random variable whose distribution depends only on a single parameter, the expected number μ, rather than two parameters n and p? The answer is that we can. As you already know from Section 4 of Unit 3, the approximating distribution will be the Poisson distribution. The relevant result is given in the following box.

Poisson's approximation for rare events

For large values of n and small values of p, a random variable following the binomial distribution, $B(n, p)$, has approximately the same distribution as a random variable following the Poisson distribution with parameter np:

$$B(n, p) \approx \text{Poisson}(np).$$

Equivalently, if $\mu = np$, then

$$B(n, \mu/n) \approx \text{Poisson}(\mu).$$

The approximation improves as n increases.

The symbol '$\approx$' is read as 'is approximately the same distribution as'.

In Units 3 and 4, the Poisson parameter was denoted λ, and in Unit 4 you saw that λ is also the mean of the Poisson distribution. In this section, because we're equating the means of the binomial and Poisson distributions and μ denotes the mean, it is convenient to use μ to denote the Poisson parameter.

In the following example, we compare the probabilities associated with the Poisson approximation with those already calculated using binomial models for the example of cyclist Chris getting drenched.

Example 2 *Poisson and binomial probabilities of drenchings*

Several probabilities calculated for the binomial distributions $B(50, 1/40)$, $B(60, 1/48)$ and $B(40, 1/32)$ were given in Table 2. Each of these distributions has mean $\mu = np = 1.25$, so the Poisson approximation is Poisson(1.25). This gives the following probabilities:

$$p(0) = e^{-1.25} \simeq 0.2865,$$

$$p(1) = \frac{e^{-1.25}1.25}{1!} \simeq 0.3581,$$

$$p(2) = \frac{e^{-1.25}(1.25)^2}{2!} \simeq 0.2238,$$

and so on. Table 3 (overleaf) shows the binomial probabilities from Table 2 and the corresponding Poisson probabilities.

The p.d.f. of the Poisson(μ) distribution is $p(x) = e^{-\mu}\mu^x/x!$, $x = 0, 1, 2, \ldots$.

Table 3 Yet more probabilities of drenchings while cycling

	$P(X=0)$	$P(X=1)$	$P(X=2)$	$P(X=3)$	$P(X \geq 4)$
$B(50, 1/40)$	0.2820	0.3615	0.2271	0.0932	0.0362
$B(60, 1/48)$	0.2827	0.3610	0.2266	0.0932	0.0365
$B(40, 1/32)$	0.2808	0.3624	0.2280	0.0931	0.0357
Poisson(1.25)	0.2865	0.3581	0.2238	0.0933	0.0383

As you can see, the probabilities agree in the first decimal place and differ by at most 1 in the second decimal place, so the approximation is a reasonable one.

So far we have looked at only one example involving a situation where the parameter n of the binomial distribution is large and the corresponding value of p is small. We have calculated probabilities only for binomial distributions with mean $\mu = 1.25$. Do similar results hold for binomial distributions with other means? It seems from the mathematics that they do, but whether or not the approximation is a good one depends on the values of n and $p = \mu/n$.

To investigate the circumstances in which the approximation is a good one, we must consider a variety of values of n and $p = \mu/n$. However, you doubtless found the calculations involved in Activities 1 and 2, where n is large, to be time-consuming using a calculator. This is a situation where the computer comes into its own: it will do these sorts of calculations quickly and allow you to concentrate on the question being investigated. You are asked to carry out such an investigation in Chapter 9 of Computer Book A.

Refer to Chapter 9 of Computer Book A for another example of the approximation of binomial probabilities by Poisson probabilities.

A very rare but predictable event

So you have seen that when the parameter p of a binomial distribution is small and the parameter n is fairly large, the actual values of the binomial parameters are not critical to the values of the binomial probabilities (see, for example, Table 3). All that is required to find approximate values for these probabilities is the product np, which is the mean of the binomial distribution. You also know that the model which provides a good approximation to the binomial distribution when p is small and n is large is the Poisson distribution. (The Poisson distribution depends on only one parameter, np, rather than on the two separate parameters, n and p, of the binomial distribution.) A rough rule is now given for deciding when the Poisson distribution provides a good approximation to the binomial distribution.

A rough rule for using Poisson's approximation

- If n is large and p is small ($n \geq 50$ and $p \leq 0.05$, say), the approximation is good.
- When p is small enough, the approximation is good even for quite small values of n. The smallest value of n for which the approximation is good decreases as the value of p decreases.

In other texts, you may see different rules. One such rule requires $n \geq 100$ and $np \leq 10$.

It is unlikely that the rough rule given above agrees exactly with the rule you formulated when investigating binomial distributions in Chapter 9 of Computer Book A. Whether you regard the values of two probabilities as 'close' is a subjective matter; and how close you require them to be depends on what they are to be used for. There are a few (rare) situations where the third or fourth decimal place is critical. On the other hand, it is unlikely that you would feel the assessment of the chances of several drenchings a month for a regular cyclist to be one of these!

There are many other practical situations in which a binomial model with a small value of p would apply, and hence for which the Poisson distribution could also reasonably be used. A few more examples are described and investigated in the rest of this section.

Example 3 *Birthdays*

Suppose we have a group of 100 unrelated people. Let X be the random variable denoting the number of people in a group of 100 unrelated people whose birthday falls on 1 April. Then, making the assumption that births are independent and distributed evenly over the year, and ignoring Leap Day, the number of people in the group with a birthday on 1 April has a binomial distribution: $X \sim B(100, 1/365)$. So the probability that, say, precisely one person in the group was born on 1 April is given by

$$P(X = 1) = \binom{100}{1} \left(\tfrac{1}{365}\right) \left(\tfrac{364}{365}\right)^{99} \simeq 0.2088.$$

Actually, the proportion of births on any day varies from day to day and from month to month. For example, during a holiday period, fewer births than usual are induced. The birth rate falls on Sundays, too.

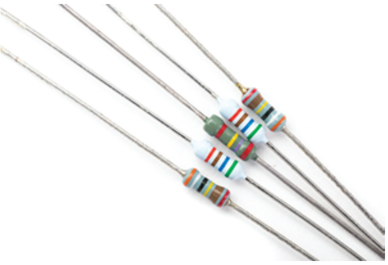
The binomial distribution has mean $\mu = 100 \times 1/365 \simeq 0.2740$. So, using Poisson's approximation, we have

$$P(X = 1) \simeq 0.274e^{-0.274} \simeq 0.2083.$$

As you can see, the two probabilities are very close.

Activity 3 *Faulty resistors*

Suppose that a process manufacturing electrical resistors results in a small proportion, approximately 1 in 20, not working when they leave the factory. This is a tolerable risk accepted by purchasers: these resistors are not quality tested before they are packaged, because to do so would add considerably to their cost. The resistors are boxed in packages of 50.



Suppose also that whether or not any particular resistor is faulty is independent of whether or not any other resistor is faulty, and that the probability that a resistor is faulty is the same for all resistors.

(a) State an 'exact' model for the number of faulty resistors in a box. Use this model to calculate the probabilities that there is one faulty resistor in a box and that there are three faulty resistors in a box.

(b) Write down an approximate distribution for the number of faulty resistors in a box, and use this to calculate probabilities corresponding to those you found in part (a).

(c) Comment on any differences between the values you obtained in parts (a) and (b).

Activity 4 *Getting a seat*

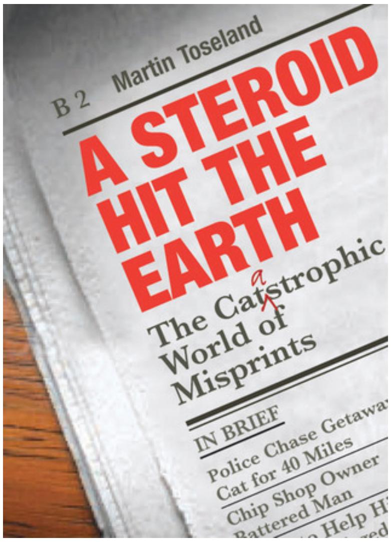
On average, 1% of passengers who book seats on a particular flight do not turn up for the departure of the flight. The airline sells 200 tickets for this flight; only 198 seats are available. Find an approximate value for the probability that everyone who turns up for the flight will have a seat. You may assume that passengers who have booked a seat either turn up or do not turn up for the flight independently of each other.

What happens if you don't get a seat?

Exercises on Section 1

Exercise 1 *Typographical errors*

One of the necessary stages in getting a book published is the careful checking of the proofs (the first version of the printed pages). Usually there are typographical errors that need correcting. With a good publisher and printer, such errors, even in the first proofs, are not too numerous. Experience with one publisher of children's books suggests an average of 3.6 errors per page.



For example: 'And now, the Superstore – unequalled in size, unmatched in variety, unrivalled inconvenience.' Taken from Toseland, M. (2009) *A Steroid Hit the Earth*, Anova Books.

Taking a children's paperback and counting the number of words on a randomly selected page gave a word count of 320. If we assume there are 320 words on a page, then an estimate of the probability that a word is mistyped is

$$p = \frac{3.6}{320} = 0.011\,25.$$

If we assume independence from word to word, then X, the number of errors per page, has a binomial distribution with parameters $n = 320$ and $p = 0.011\,25$.

Of course, the number of words per page is not constant from page to page. Turning to another page, the number of words would almost certainly have been different – 360, perhaps. In this case, we would have assumed a binomial model for the number of errors per page with parameters $n = 360$ and $p = 3.6/360 = 0.01$.

Table 4 contains some probabilities for the number of errors on a page using binomial models with $n = 320$ and $n = 360$.

Table 4 Probabilities of errors

	$P(X=0)$	$P(X=1)$	$P(X=2)$	$P(X=3)$	$P(X \geq 4)$
$B(320, 0.011\,25)$	0.0268	0.0975	0.1769	0.2134	0.4854
$B(360, 0.01)$	0.0268	0.0976	0.1769	0.2133	0.4854

(a) Suggest an approximate model for the number of errors on a page.

(b) Use this approximate model to calculate probabilities corresponding to those given in Table 4.

(c) Comment on any differences between the results given in Table 4 and those you obtained in part (b).

Exercise 2 *Blue–yellow colour vision deficiency*

Blue–yellow colour vision deficiency, or tritanopia, is very rare, and it occurs equally in men and women. The probability that a person has this type of colour vision deficiency is about 1 in 10 000, or 0.0001.

(a) Use an exact model to find the probability that there will be exactly one person with tritanopia in a group of 100 people.

(b) Use a Poisson model to find an approximate value for the probability that there will be exactly one person with tritanopia in a group of 100 people. Comment on the comparison between the probabilities calculated in parts (a) and (b).

2 Waiting times between events

Many events occur repeatedly, but ‘at random’. For example, if a fair six-sided die is rolled repeatedly, sometimes a six is obtained (a success) and sometimes it is not. Similarly, if an arrow is aimed at a target on many occasions, some shots will hit the centre of the target (successes) and some will miss. In either case, an event (a success) may occur at each of a discrete set of opportunities (trials) taking place one after the other. It is not possible to predict at which trials the event will occur and so, in this sense, the events occur ‘at random’.

However, there are many situations in which events seem to occur unpredictably, that is ‘at random’, but where there is no notion of a specific trial or opportunity when an event may occur. For instance, events such as floods, earthquakes and accidents may occur at *any* time.

Two models for the times between events that occur ‘at random’ are discussed in this section: one discrete probability distribution and one continuous probability distribution. The discrete distribution – which is one of the ones that you have met already – is a model for the times between successive events in situations where events may occur only at a discrete set of opportunities or trials; this is discussed in Subsection 2.1. The continuous distribution, which is new to you and is discussed in Subsection 2.2, is a model for the times between events that may occur at any time.

2.1 Bernoulli processes

The notion of a collection or sequence of independent Bernoulli trials for which the probability of success remains constant from trial to trial was introduced in Unit 3. Such collections of Bernoulli trials were fundamental to the development of the binomial and geometric distributions in that unit. Thought of as sequences – that is, when trials take place sequentially through time – these collections of Bernoulli trials are called Bernoulli processes. The formal definition of a Bernoulli process is given in the box below.

Bernoulli processes

A **Bernoulli process** is a sequence of Bernoulli trials in which

- trials are independent
- the probability of success remains the same from trial to trial.

In a Bernoulli process it is assumed that an event either occurs, or does not occur, at each of a clearly defined sequence of opportunities (called trials). Three typical realisations of a Bernoulli process – that is, three sets of observed outcomes of a Bernoulli process – are shown in Table 5. In each of these three realisations, the parameter p, the probability of a 1 (a success) at each trial, was equal to 0.3. In each case, the outcomes of the first 25 trials in the realisation have been recorded.

Table 5 The first 25 trials in three realisations of a Bernoulli process

(a)	0	1	0	0	0	0	0	1	1	0	1	1	0	0	0	1	0	0	0	1	0	1	1	0	0
(b)	0	0	0	1	0	0	1	1	1	0	0	0	0	0	1	0	1	1	0	0	1	1	0	0	0
(c)	0	1	0	1	1	0	1	0	0	0	0	1	0	0	0	0	1	1	0	1	0	0	0	0	0

In Unit 3, you saw that X, the number of trials up to and including the first success in a Bernoulli process, has a geometric distribution with parameter p (and mean $1/p$). In the three realisations in Table 5, the observed values of X are 2, 4 and 2, respectively. But what is the distribution of the number of trials from one success to the next, that is from the first success to the second, or from the second success to the third, and so on?

One of the properties of the Bernoulli process is that the probability of a success at any trial remains the same and is equal to the probability of

success at all preceding trials, whatever the history of the process up to that trial. So even a long run of failures (that is, of 0s) does not make the probability of a success (a 1) at the next trial any more likely. There is no question of 'luck being bound to change'. Similarly, a long run of 1s (a 'run of good luck') does not alter the probability of a 1 at subsequent trials. The incidence of 1s may be regarded in a special sense as 'random': knowledge of the whole history of the process so far – that is, the outcome of all the trials so far – does not alter the future development of the process, or the probability laws describing its future development.
A statistical experimenter, who knows the probability of success in each trial and is about to embark on trial number 1001, is not helped at all in her assessment of what might or might not happen at that trial by knowing what has happened at any or all of the preceding one thousand trials. This is known as the *memoryless property* of the Bernoulli process.

Since a Bernoulli process has this memoryless property, it follows that at any point in a Bernoulli process, the number of trials from that point until the next success has the same probability distribution as X, the number of trials from the start of the process until the first success. So, in particular, the number of trials from one success to the next has a *geometric distribution* with parameter p, written $G(p)$. Hence we can use $X \sim G(p)$ to represent the number of trials from one success up to and including the next – the 'waiting time' to the next success – not just for the number of trials from the start up to and including the first success.

In each of the realisations in Table 5, several independent values of X can be observed. In realisation (a), they are 2, 6, 1, 2, 1, 4, 4, 2, 1; and the next value of X must be 3 or more. Notice that, to match what we did when considering the number of trials from the start to the first success, after each success we count the number of failures, 0s, plus one for the next success, 1; that is, we include in our waiting time the trial on which the next success happens.

Activity 5 *Observed values of X*

Write down the observed values of X in each of realisations (b) and (c) in Table 5. In each realisation, what can you say about the next value of X?

Recall from Unit 3 that the geometric distribution has p.m.f.

$$p(x) = P(X = x) = (1 - p)^{x-1} p$$

and c.d.f.

$$F(x) = P(X \leq x) = 1 - (1 - p)^{x},$$

respectively, on range $x = 1, 2, \ldots$.

Activity 6 *Times between successes*

Suppose that the probability of a success in each trial in a Bernoulli process is equal to 0.3.

(a) Find the probability that the number of trials from one success to the next is equal to (i) 2, (ii) 6.

(b) Find the probability that the number of trials from one success to the next is 3 or more.

Activity 7 *Obtaining your next six*

Suppose that in a board game you roll a fair six-sided die repeatedly.

(a) Find the probability that after obtaining a six you will have to roll the die exactly four times to obtain another six.

(b) Find the probability that after obtaining a six you will have to roll the die at least five times to obtain your next six.

Example 4 *Intervals between serious earthquakes*

Table 6 contains some data on serious earthquakes worldwide. An earthquake has been included if over 1000 people were killed. The data cover the period from 15 August 1950, when 1530 people were killed in the Assam–Tibet earthquake, to 11 March 2011, when 1479 people were killed in the Tōhoku earthquake and tsunami, in Japan. Altogether 68 serious earthquakes occurred in this period. Table 6 contains the 67 'waiting times' in days between successive earthquakes. The numbers should be read across the rows.

Table 6 Time intervals between serious earthquakes (days)

186	77	682	449	91	1027	164	808	83	832
328	1120	743	328	163	83	64	356	324	257
503	232	252	151	142	32	120	100	561	755
44	200	47	503	320	690	386	147	533	109
560	26	460	420	291	474	131	643	71	270
115	240	204	34	493	423	422	219	366	92
194	230	717	506	104	91	332			

(Source: adapted from the *Global Significant Earthquake Database* of the US National Geophysical Data Center)

There are no instances of two or more serious earthquakes occurring on the same day, though such a coincidence is not impossible. So it might be reasonable, at least as a first approximation, to regard each day as an opportunity or trial during which a serious earthquake might or might not happen.

Activity 8 *Modelling the times between serious earthquakes*

Suppose that the occurrences of serious earthquakes are modelled as a Bernoulli process.

(a) What assumptions about the occurrences of serious earthquakes are made?

(b) What distribution is used to model the time between serious earthquakes?

(c) A histogram of the data in Table 6 is shown in Figure 1. By considering the shape of this histogram, say whether or not you think the model you proposed in part (b) might be a reasonable one.

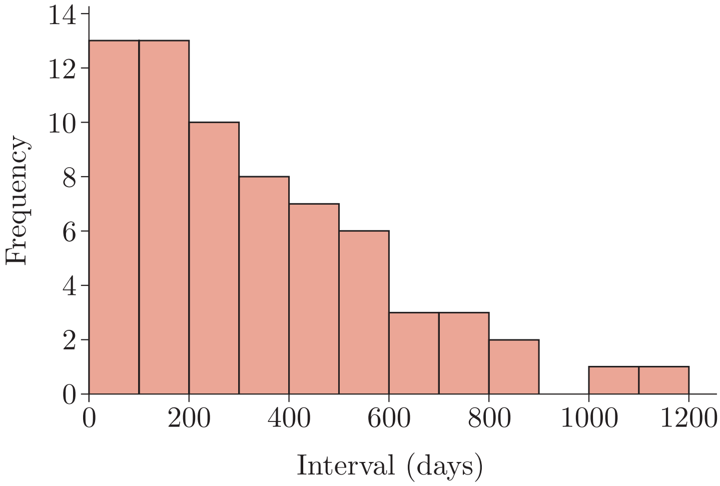


Figure 1 Time intervals between serious earthquakes (days)

(d) If a Bernoulli process is a good model, how could you use the data to estimate the probability that on any individual day a serious earthquake will occur?

Serious damage caused by a major earthquake

This is another instance of using a histogram to portray discrete data with a large range.

Example 5 *Intervals between serious earthquakes, continued*

For the data in Table 6, the mean time in days between serious earthquakes is

$$\frac{186+77+\cdots+332}{67} = \frac{22\,120}{67} \simeq 330$$

(which is just under 11 months). So, as suggested in the solution to part (d) of Activity 8, an estimate of the probability that a serious earthquake occurs somewhere in the world on any individual day is given by the reciprocal of this sample mean; that is, an estimate of p is

$$p = \frac{67}{22\,120} \simeq 0.0030.$$

Reminder: you will notice that the module authors are happy to allow the number of decimal places in the results of calculations to vary, as you may, in ways that seem appropriate to the context at hand.

The estimated probability of a gap between serious earthquakes exceeding three years (say) can be found using the c.d.f. of X, the waiting time between successive serious earthquakes. Ignoring the possibility of a leap year, there are 1095 days in three years, so the probability of a gap longer than three years is given by

$$\begin{aligned} P(X > 1095) &= 1 - F(1095) = 1 - \left(1 - (1-p)^{1095}\right) \\ &= (1-p)^{1095} = \left(1 - \frac{67}{22\,120}\right)^{1095} \simeq 0.0361, \end{aligned}$$

or about 1 in 28. Notice that, for the period covered by the data in Table 6, one interval out of the 67 exceeded three years: there was an interval of 1120 days between the twelfth and thirteenth earthquakes recorded. (The twelfth was on 26 July 1963 in Skopje, Macedonia, and the thirteenth, in Varto, Turkey, was on 19 August 1966.) There were fewer gaps of more than three years than are predicted by the model.

Activity 9 *Gaps longer than two years*

Following on from Examples 4 and 5, according to the geometric model, what is the probability that the interval between serious earthquakes will exceed two years? (You may ignore the possibility of a leap year.) How many cases of gaps longer than two years between serious earthquakes are recorded in Table 6?

Both as a summary and for comparison with results further on in this unit, the main properties of a Bernoulli process are highlighted below.

The Bernoulli process: two main results

For a Bernoulli process, that is, a sequence of independent Bernoulli trials with constant probability of success, p:

- the random variable which represents the number of successes in n trials has a binomial distribution with parameters n and p
- the waiting time from after one success up to and including the next success has a geometric distribution with parameter p.

An underlying assumption implicit in using the geometric distribution to model the intervals between serious earthquakes is that at most one serious earthquake can occur on any day: each day was regarded as a trial at which an earthquake might or might not happen. However, although it is unlikely that two or more serious earthquakes will occur on the same day, it is not impossible. In reality, an earthquake may occur at any time on any day. Also, the exact length of the interval from one serious earthquake to the next is unlikely to be a whole number of days, but is measured in days, hours and minutes. A continuous model for the times between earthquakes is needed. In Subsection 2.2, such a continuous model is discussed; this model is the continuous-time analogue of the geometric distribution.

2.2 The exponential distribution

The histogram in Figure 1 suggests that a suitable continuous probability model for the intervals between serious earthquakes will be one with a highly skew probability density function, starting high for values near 0 and tailing off for longer waiting times (rather as the geometric probability mass function does). Many probability density functions could be constructed with this sort of shape. However, one of these probability density functions turns out to belong to the distribution with properties most like those of the geometric distribution, and therefore forms a possible model for the distribution of the waiting time from one serious earthquake to the next. This is the *exponential distribution*. It is defined as follows.

The exponential distribution

A random variable X is said to have an **exponential distribution with parameter $\boldsymbol{\lambda}$**, where $\lambda > 0$, if it has probability density function

$$f(x) = \lambda e^{-\lambda x}, \quad x > 0. \tag{1}$$

This is written $X \sim M(\lambda)$. Here, e is the exponential function which also appears in the Poisson probability mass function.

The exponential distribution is the continuous analogue of the geometric distribution. The probability density function of a typical exponential distribution is shown in Figure 2(a). Notice the similarity between its general shape and that of the probability mass function of a geometric random variable (see Figure 2(b)). The exponential p.d.f. is in fact a decreasing function of x for any value of its (positive) parameter λ.

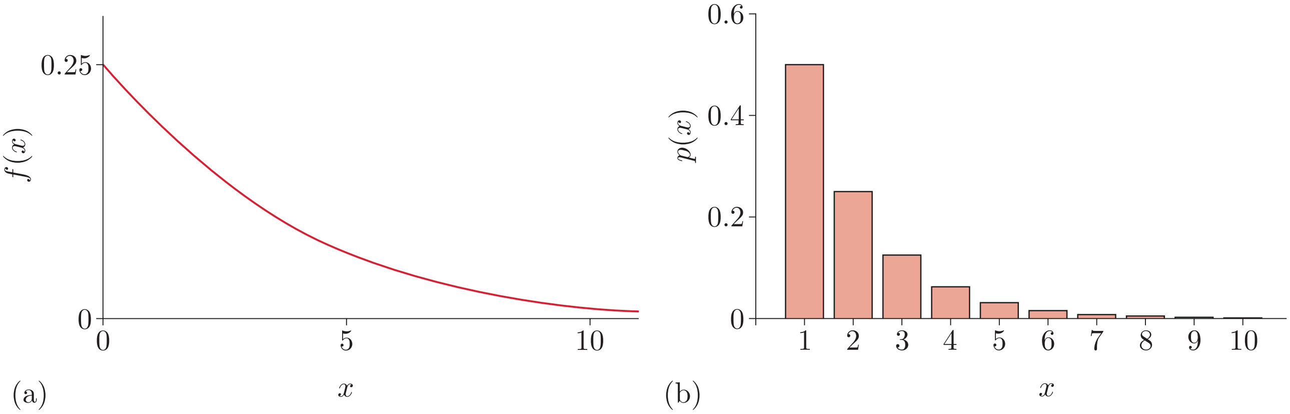


Figure 2 (a) The p.d.f. of a typical exponential distribution; (b) the p.m.f. of a typical geometric distribution

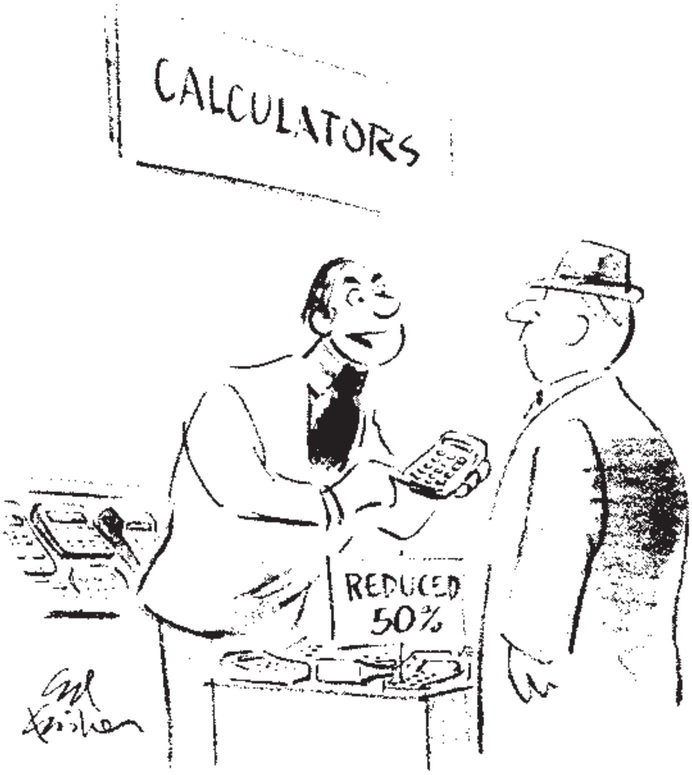
You will recall that for a continuous random variable, the cumulative distribution function is very useful for calculating probabilities. The c.d.f. of an exponential distribution is given in the box below.

The c.d.f. of an exponential random variable

If the random variable X has an exponential distribution with parameter λ, then its c.d.f. is given by

$$F(x) = 1 - e^{-\lambda x}, \quad x > 0. \tag{2}$$

As usual, to derive the c.d.f. of a continuous distribution from its p.d.f., integration is used. You will do this for yourself in Activity 11. First, though, in the case of the exponential distribution, it is clear from Equation (1) that you need to be able to integrate an exponential function. Integrating an exponential function is a fairly straightforward thing to do, the results you need being given next.



"It's got more special-function keys than you'll find on many of the larger models: square root, cosine, logarithmic, integral and exponential keys. I'd say that more than makes up for the fact that it doesn't have the number nine!"

Integrating the exponential function

Suppose we want to integrate e^{ax} where a is a non-zero constant. The formula for its indefinite integral is

$$\int e^{ax}\,dx = \frac{e^{ax}}{a} + c.$$

In words, the integral of the exponential function e^{ax} is the same exponential function, e^{ax}, divided by the constant a. Definite integrals of e^{ax} therefore have the form

$$\int_{x_1}^{x_2} e^{ax}\,dx = \left[\frac{e^{ax}}{a}\right]_{x_1}^{x_2} = \frac{1}{a}\left(e^{ax_2} - e^{ax_1}\right).$$

Example 6 *Integrating exponentials*

To illustrate applying these rules, we have

$$\int e^{2x} = \frac{e^{2x}}{2} + c$$

and

$$\int e^{-x} = \frac{e^{-x}}{-1} + c = -e^{-x} + c.$$

Also,

$$\int_0^1 e^{3x} = \left[\frac{e^{3x}}{3}\right]_0^1 = \frac{1}{3}\left(e^3 - e^0\right) = \frac{1}{3}\left(e^3 - 1\right) \simeq 6.362.$$

Activity 10 *Integrating exponentials*

Find each of the following integrals.

(a) $\displaystyle\int e^x\,dx$

(b) $\displaystyle\int_1^5 e^{-x}\,dx$

(c) $\displaystyle\int_0^t e^{-2x}\,dx$

Activity 11 *From exponential p.d.f. to exponential c.d.f.*

The exponential p.d.f. is given by Equation (1). Confirm that the c.d.f. of the exponential distribution is given by Equation (2).

The exponential c.d.f. corresponding to the exponential p.d.f. in Figure 2(a) is shown in Figure 3.

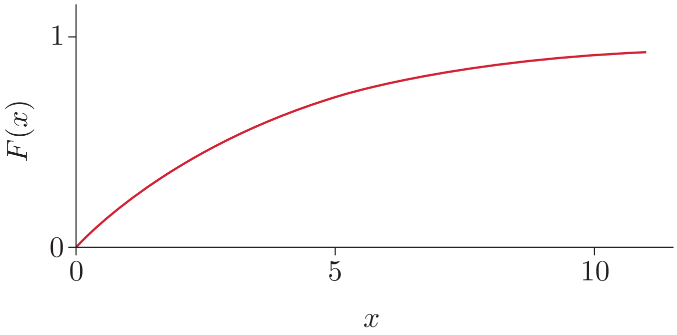


Figure 3 The c.d.f. of the exponential distribution whose p.d.f. is shown in Figure 2(a)

As you can see, the family of exponential distributions is indexed by a single parameter $\lambda > 0$. But what does this parameter represent?

For the geometric distribution, the mean is given by $1/p$, where p is the parameter of the distribution. The parameter λ of an exponential distribution has a similar role: the mean μ of an exponential distribution is equal to $1/\lambda$.

This result may be obtained using integration but you need not see the details.

The mean of an exponential distribution

If the random variable X has an exponential distribution with parameter λ, then

$$E(X) = \frac{1}{\lambda}. \qquad (3)$$

So λ is equal to the reciprocal of the mean μ. Hence, returning to the earthquake example of Subsection 2.1, if μ represents the mean waiting time between serious earthquakes, then λ represents the average *rate* at which serious earthquakes occur. Indeed, λ is often referred to as the *rate parameter*. Given data, this result can be used to estimate the parameter λ.

Example 7 *More on modelling the times between serious earthquakes*

The mean of the waiting times in Table 6 is $22120/67 \simeq 330$ days. If an exponential distribution is used to model the variation in the waiting times between serious earthquakes, then an estimate of the parameter λ of the distribution is $67/22120 \simeq 0.0030$.

In Subsection 2.1, a geometric distribution was used to model the waiting times between serious earthquakes. Using this model we found that, for instance, the probability of a gap between serious earthquakes exceeding three years (i.e. 1095 days) was approximately 0.0361. Using an exponential model with parameter $\lambda = 67/22120$, the corresponding probability is

$$\begin{aligned} P(X > 1095) &= 1 - F(1095) \\ &= 1 - \left(1 - e^{-(67/22120)\times 1095}\right) \quad \text{using Equation (2)} \\ &\simeq e^{-3.31668} \simeq 0.0363. \end{aligned}$$

So, to three decimal places, the exponential model gives the same result as the geometric model.

Activity 12 *Understanding and using the exponential model*

(a) The p.d.f. of the exponential model with parameter $\lambda = 67/22120$ is shown in Figure 4.

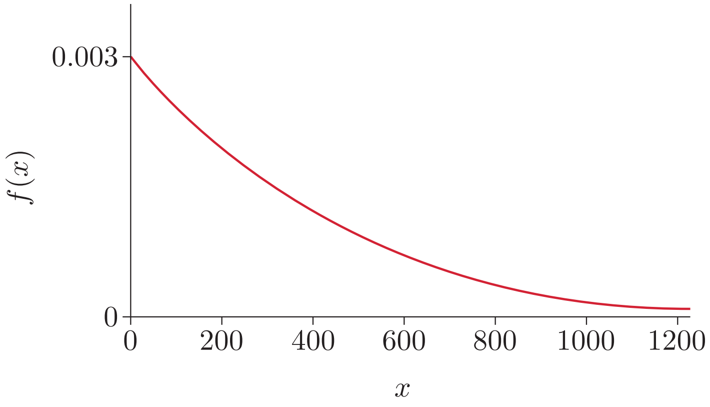


Figure 4 The p.d.f. of the exponential distribution with $\lambda = 67/22120$

(i) Mark on a copy of Figure 4 the area corresponding to the probability that the interval between serious earthquakes will exceed two years (that is, 730 days).

(ii) Use the formula to find what is, according to this model, the probability that the interval between serious earthquakes will exceed two years. Comment on the comparison between this probability and that provided by the geometric distribution in Activity 9.

(b) According to this exponential model, what is the probability that the gap between successive earthquakes will be a week or less?

(c) Use the exponential model to find the proportion of waiting times between serious earthquakes that are longer than average, that is, the proportion that exceed $22120/67 \simeq 330$ days. Compare this proportion with the proportion of waiting times in Table 6 that exceed 330 days.

As you know from Subsection 4.3 of Unit 2, the probability that a random variable X takes values between, say, x_1 and x_2, $x_1 < x_2$, is also available from the c.d.f. of X:

$$P(x_1 \leq X \leq x_2) = F(x_2) - F(x_1).$$

Applying this to the exponential distribution,

$$P(x_1 \leq X \leq x_2) = 1 - e^{-\lambda x_2} - (1 - e^{-\lambda x_1}),$$

which reduces to the following formula for probabilities of lying within intervals under this model.

Probabilities associated with an exponential random variable

If the random variable X has an exponential distribution with parameter λ, then

$$P(x_1 \leq X \leq x_2) = e^{-\lambda x_1} - e^{-\lambda x_2}. \qquad (4)$$

Activity 13 *More on understanding and using the exponential model*

(a) Figure 4 showed the p.d.f. of the exponential model, with $\lambda = 67/22120$, that we are using for the data on time intervals between serious earthquakes. Mark on a copy of that figure the area corresponding to the probability that the interval between serious earthquakes will be between one and two years (that is, between 365 and 730 days).

(b) Use Equation (4) to find what is, according to the model, the probability that the interval between serious earthquakes will be between one and two years.

(c) According to the exponential model, what is the probability that the gap between successive serious earthquakes will be between a week and a month (taken to be 30 days)?

Presumably, if you can catch one of these number 55 buses you'd have been waiting longer than average for one of them to come along!

Activity 14 *Longer than average waiting times*

Suppose that the waiting times between events occurring at random may be modelled by an exponential distribution with parameter λ. According to the model, what proportion of waiting times will be longer than the mean waiting time?

In this subsection, the exponential distribution has been defined. A formula for the c.d.f. of an exponential distribution has been given; and the result that the mean of an exponential distribution with parameter λ is $1/\lambda$ has been stated without proof. A formula for the variance is also stated, without proof, in the box below, where several properties of exponential distributions are summarised.

Like the mean, the variance may also be found using integration.

Properties of exponential distributions

If the random variable X has an exponential distribution with parameter λ, $X \sim M(\lambda)$, then

- the p.d.f. of X is
$$f(x) = \lambda e^{-\lambda x}, \quad x > 0$$
- the c.d.f. of X is
$$F(x) = 1 - e^{-\lambda x}, \quad x > 0$$
- the mean of X is
$$\mu = E(X) = \frac{1}{\lambda}$$
- the variance of X is
$$\sigma^2 = V(X) = \frac{1}{\lambda^2}$$
- and hence the standard deviation of X is
$$\sigma = \sqrt{V(X)} = \frac{1}{\lambda}. \tag{5}$$

Notice that the mean and standard deviation of an exponential distribution are equal. This provides a method for checking quickly, given data, whether an exponential model is worth considering: if the sample mean and sample standard deviation differ greatly, then an exponential model is not appropriate. Of course, even if the sample mean and sample standard deviation are roughly equal, an exponential distribution may not be a suitable model. For instance, an exponential model would not be appropriate if the shape of a histogram of the data was very different from that of an exponential p.d.f.

Exercises on Section 2

Exercise 3 *Explosive volcanic eruptions*

The histogram in Figure 5 (overleaf) represents the 55 intervals in months between 56 successive major explosive volcanic eruptions in the northern hemisphere. The dataset includes all such eruptions that took place between the beginning of 1851 and the beginning of 1985. (The first occurred in March 1853 and the last in July 1983.)

The sample mean is 28.4 months and the sample standard deviation is 31.7 months.

(a) Explain briefly whether or not you think an exponential distribution might be a reasonable model for the intervals between major explosive eruptions in the northern hemisphere.

(b) Which model do you think is more appropriate in this case – a geometric distribution or an exponential distribution? (You will investigate using both models in Exercises 4 and 5.)

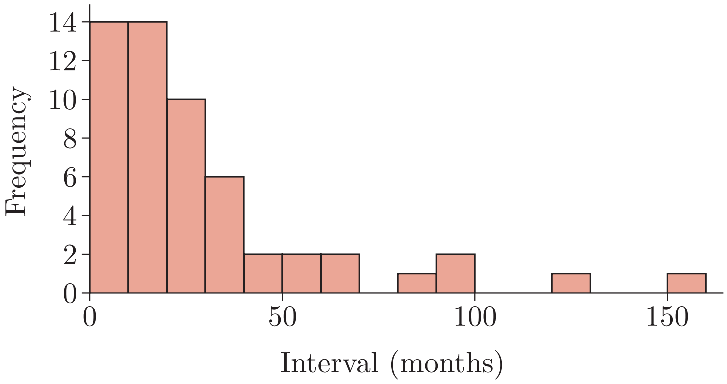


Figure 5 Intervals between eruptions

(Source: adapted from Solow, A.F. (1991) 'An exploratory analysis of occurrence of explosive volcanism in the northern hemisphere', *Journal of the American Statistical Association*, vol. 86, no. 413, pp. 49–54)

Exercise 4 *Using a geometric model*

Despite the solution to Exercise 3(b), this exercise concerns using a geometric model to model the intervals between successive major explosive volcanic eruptions in the northern hemisphere.

(a) What assumptions about the occurrences of major explosive volcanic eruptions in the northern hemisphere are you making by using a geometric model?

(b) Use the information given in Exercise 3 to find an estimate for the parameter p of the geometric model.

(c) According to the geometric model, what is the probability that the interval between successive major explosive volcanic eruptions in the northern hemisphere will exceed ten years? How many cases of intervals exceeding ten years were there?

Exercise 5 *Using an exponential model*

This exercise concerns using an exponential model for the intervals between successive major explosive volcanic eruptions in the northern hemisphere.

(a) Use the information given in Exercise 3 to find an estimate for the parameter λ of the exponential model.

(b) According to the exponential model, what is the probability that the interval between successive major explosive volcanic eruptions in the northern hemisphere will exceed ten years? Compare this result with the probability you obtained using a geometric model in Exercise 4.

(c) According to the exponential model, what is the probability that the interval between successive major explosive volcanic eruptions in the northern hemisphere will be between one and five years?

3 The Poisson process

In this section a process rather similar to the Bernoulli process is introduced. However, the process considered here develops in continuous time, rather than as a sequence of 0s and 1s occurring at a discrete set of trials (as in the Bernoulli process). This process is called the *Poisson process.*

In Subsection 3.1 the Poisson process and some of its properties are discussed. But given data on the occurrences of events, how can we decide whether a Poisson process is a reasonable model? In Subsection 3.2 we take a look at this problem.

3.1 A continuous-time analogue of the Bernoulli process

A Bernoulli process consists of a sequence of trials which result in a sequence of 0s and 1s occurring at a discrete set of points. In the sorts of situation considered in this section, events occur after intervals, or *waiting times*, that are random in the sense that knowing how many events have occurred so far, or how long it is since an event last occurred, is of no use in forecasting when the next event will occur. These waiting times correspond roughly to sequences, or runs, of 0s in a realisation of a Bernoulli process: when an event happens, it is rather as though a 1 has been recorded. But the waiting times are not integers, and successive events are not restricted to a discrete set of opportunities.

Here are some further examples of the sorts of phenomenon that could be modelled in this way, in addition to the occurrences of earthquakes and volcanic eruptions.

Example 8 *Random events occurring in continuous time*

(a) Even in a machine shop where a regular maintenance programme is carried out, there are occasional machine breakdowns. However, their occurrence cannot be forecast. Sometimes the repair staff are overworked, and sometimes there is nothing for them to do. All that is known from repair records is that the rate at which breakdowns occur remains constant.

In the context of events occurring at random in time, the rate is the average number of occurrences per unit time.

(b) Roughly twice a year, extreme wind conditions sufficient to ground aircraft are recorded at a small airport on the south coast of England, but their time of occurrence does not seem to be linked to the time of the year.

(c) Spam emails arrive in mailboxes seemingly at random and, over long stretches of time, at an apparently constant rate. (Emails and other electronic messages which are of interest to the receiver tend to arrive in a less random pattern as they are interacted with.)

(d) In England, the Highways England Traffic Officer Service monitors motorways, and has breakdowns reported to it. There are many more during the day than there are at night, and noticeably more on Friday and Sunday evenings than on Saturday evenings. However, within shorter periods, such as, for example, the middle of the day or the early hours of the morning, the rate at which breakdowns are reported may be regarded as constant. And within any such period, there does not seem to be any pattern to the times at which the reported breakdowns occur.

(e) Arrivals at a hospital accident and emergency department occur without warning but at rates that, depending broadly on the time of day and the day of the week, are reasonably stable.

Note, however, that for many situations in which there is some uncertainty about when events will occur, those events do not occur completely at random in the kind of way exemplified in Example 8. Hence such situations will not be amenable to the same kind of statistical modelling. Here are some examples that illustrate this.

Example 9 *Events showing a non-random pattern*

(a) The owner of a battery-operated garden thermometer has noticed that, over the past six years, batteries advertised to last six months have, to within two or three weeks, lasted six months.

(b) Manufacturers of car tyres build into their tyres a lifetime with as small a variability as possible. It is possible to forecast roughly when a particular set of tyres will have to be replaced.

(c) Patients at a dental surgery arrive by appointment and are seen at roughly the time stated on their appointment card or email.

Let us return to processes, like those in Example 8, that can be modelled as events occurring at random. A continuous-time analogue of the Bernoulli process should possess properties similar to those of a Bernoulli process. For instance, only one event can occur at each trial in a Bernoulli process; that is, events should occur singly. Also, the probability that an event occurs at each trial remains constant, so the rate at which events should occur remains constant. In addition, a Bernoulli process has the memoryless property: the occurrence of events in the future does not depend in any way on how many events have occurred in the past or on when they occurred.

In fact, these properties can be used to *define* a continuous-time analogue of the Bernoulli process. This process is called the *Poisson process*; its definition is given in the following box.

The Poisson process

The **Poisson process** is a model for the occurrence of events in continuous time in which the following assumptions are made.

- Events occur *singly*.
- The rate of occurrence of events remains *constant*.
- The incidence of future events is *independent* of the past.

Events in a Poisson process can be visualised as occurring along a line representing time (or distance) as shown in Figure 6.

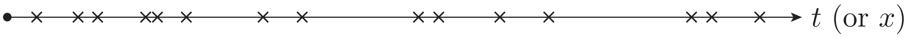


Figure 6 Events occurring in continuous time

For a Bernoulli process, you have seen that the number of successes (or events) in n trials has a binomial distribution, and the waiting time from one success to the next has a geometric distribution. What are the corresponding results for a Poisson process? That is, what is the distribution of the number of events that occur in an interval of length t? And what is the distribution of the waiting time from one event to the next?

The number of events in an interval of length t

It can be shown that the number of events that occur in an interval of length t has a Poisson distribution. To prove this requires mathematical techniques beyond the scope of this module. However, an informal argument leading to the required result is given in Example 10.

Example 10 *Calls at a switchboard*

Suppose that the arrivals of calls at a telephone switchboard may be modelled as a Poisson process. This means that calls are received at random at a constant rate. Let this rate be λ per unit time. Then the average number of calls received in an interval of length t is $\lambda \times t$.

Now imagine that this time interval is split into a large number n of very short time intervals, each of length t/n.

Provided that n is large enough, the probability that more than one call is received in any interval of length t/n may be regarded as negligible (since events occur singly in a Poisson process).

Since the rate at which calls are received remains constant, the probability p that a call is received in a short interval remains constant from one interval to the next. We can, therefore, think of the arrivals of calls as a sequence of n independent Bernoulli trials with probability p of a success (a call) at each trial.

Telephone switchboards, old and new

It follows that the total number of calls received in n intervals of length t/n (the trials) has a binomial distribution with parameters n and p, $B(n,p)$, which has mean np. However, the mean number of calls received in an interval of length t is λt, so $p = \lambda t/n$. Thus, provided that n is large, the distribution of the number of calls received in an interval of length t is binomial, $B(n, \lambda t/n)$. Using Poisson's approximation for rare events, this distribution may be approximated by a Poisson distribution with parameter λt.

From here on, it is useful to revert back to using λ (instead of μ) for the Poisson parameter.

In fact, this result is exact, not approximate. If the arrivals of calls may be modelled as a Poisson process with rate λ and X is a random variable representing the number of calls received in an interval of length t, then X has a Poisson distribution with parameter λt: $X \sim \text{Poisson}(\lambda t)$.

Notice that the parameter of the probability distribution of the number of calls received in an interval depends on both the length of the interval, t, and the rate at which the calls are arriving per unit time, λ, but only through the single quantity that is their product, λt.

The result described in Example 10 for calls arriving at a switchboard generalises as follows to events in any Poisson process.

- If the occurrences of events may be modelled as a Poisson process with rate λ, and X is a random variable representing the number of events that occur in an interval of length t, then X has a Poisson distribution with parameter λt: $X \sim \text{Poisson}(\lambda t)$.

This result states that the number of events in an interval has a Poisson distribution and that the parameter of this distribution is equal to the rate at which events occur per unit time multiplied by the length of the interval. That is, the value of the parameter is the expected number of events that will occur in the interval. The calculation of the parameter of the Poisson distribution is illustrated in Example 11. The use of this distribution to calculate probabilities associated with events in a Poisson process will be illustrated shortly.

Example 11 *Telephone calls*

Suppose that the arrivals of telephone calls at a solicitor's office may be modelled as a Poisson process with rate 9 per hour, that is, with $\lambda = 9$. Then, for instance, the number of calls received in $t = 1\frac{1}{2}$ hours has a Poisson distribution with parameter

$$\lambda t = (9 \text{ per hour}) \times \left(1\tfrac{1}{2} \text{ hours}\right) = 13.5;$$

and the number of calls received in a ten-minute period has a Poisson distribution with parameter

$$\lambda t = (9 \text{ per hour}) \times \left(\tfrac{1}{6} \text{ hour}\right) = 1.5.$$

Activity 15 *The number of events in an interval*

(a) Suppose that the arrivals of telephone calls at a call centre on a weekday morning may be modelled as a Poisson process with rate 3 per minute. Find the distribution of each of the following.

(i) The number of calls that arrive in a five-minute period.

(ii) The number of calls that arrive in a quarter of an hour.

(b) Suppose that arrivals at a hospital accident and emergency department on a weekday afternoon may be modelled as a Poisson process with rate 5 per hour. Find the distribution of each of the following.

(i) The number of arrivals in an afternoon ($3\frac{1}{2}$ hours).

(ii) The number of arrivals in a 15-minute period.

The waiting times between successive events

Now consider the waiting times between successive events in a Poisson process. Since the waiting time may take any positive value, a continuous model is required. In addition, we require a probability distribution with the memoryless property, that is, with the property that the occurrence of events in the future does not depend on how many events have occurred in the past or on when they occurred. It can be shown, though this will not be done here, that the only continuous distribution with the memoryless property is the exponential distribution. As mentioned in Section 2, the exponential distribution is the continuous-time analogue of the geometric distribution.

In fact, assuming the result that the number of events that occur in an interval of length t in a Poisson process with rate λ has a Poisson distribution with parameter λt, it can be shown that the waiting time T between successive events has an exponential distribution with parameter λ. The method involves finding the c.d.f. of the waiting time. The argument is rather neat and goes as follows.

We now denote the waiting time random variable by T to avoid confusion with the number of events random variable X.

The c.d.f. of T is defined to be

$$F(t) = P(T \leq t) = 1 - P(T > t).$$

The waiting time T from one event to the next is greater than t if and only if no event occurs in the interval of length t following the first of the two events. But the number of events that occur in an interval of length t is a random variable X, where $X \sim \text{Poisson}(\lambda t)$. So

$$P(T > t) = P(X = 0).$$

But $P(X = 0) = e^{-\lambda t}$, so

$$P(T > t) = e^{-\lambda t},$$

and hence

$$F(t) = 1 - P(T > t) = 1 - e^{-\lambda t}.$$

This is the c.d.f. of an exponential distribution with parameter λ, so $T \sim M(\lambda)$.

The two results just discussed are stated formally in the following box.

The Poisson process: two main results

Suppose that events occur at random at rate λ per unit time in such a way that their occurrence may be modelled as a Poisson process.

- The random variable X, which represents the number of events that occur during a time interval of length t, has a Poisson distribution with parameter λt:

$$X \sim \text{Poisson}(\lambda t). \tag{6}$$

- The waiting time T between successive events has an exponential distribution with parameter λ:

$$T \sim M(\lambda). \tag{7}$$

The parallels between the Bernoulli process and the Poisson process are therefore complete. These parallels are emphasised in the following screencast.

Screencast 5.1 The Bernoulli and Poisson processes

If a Poisson process is thought to be a reasonable model for the occurrence of events, then the two results in the box above may be used to calculate probabilities associated with the events. Their use is illustrated in Example 12.

Example 12 *Modelling occurrences of serious earthquakes*

See Examples 4, 5 and 7.

In Section 2 an exponential distribution was used to model the waiting times between serious earthquakes. For the data in Table 6, the mean time between earthquakes was 22120/67 days, or approximately 330 days.

If we assume that the occurrences of serious earthquakes may be modelled as a Poisson process, then an estimate for the rate λ of the process per day is

$$\lambda = \frac{67}{22\,120} \simeq 0.0030 \text{ per day}.$$

Distributional Result (6) may be used to answer questions concerning the number of serious earthquakes in any period of fixed length. For example, in a typical year of 365 days, X, the number of serious earthquakes that occur, has a Poisson distribution with parameter

$$\lambda t = \left(\frac{67}{22\,120} \text{ per day}\right) \times (365 \text{ days}) \simeq 1.10556.$$

This is consistent with serious earthquakes happening around every 11 months on average.

Since the mean of a Poisson distribution is equal to its parameter, this is also, according to the Poisson process model, the mean number of serious earthquakes per year. The probability that exactly three serious earthquakes occur in a typical year is

$$P(X=3) = \frac{e^{-1.10556}(1.10556)^3}{3!} \simeq 0.0746;$$

and the probability that two or more serious earthquakes occur in a typical year is

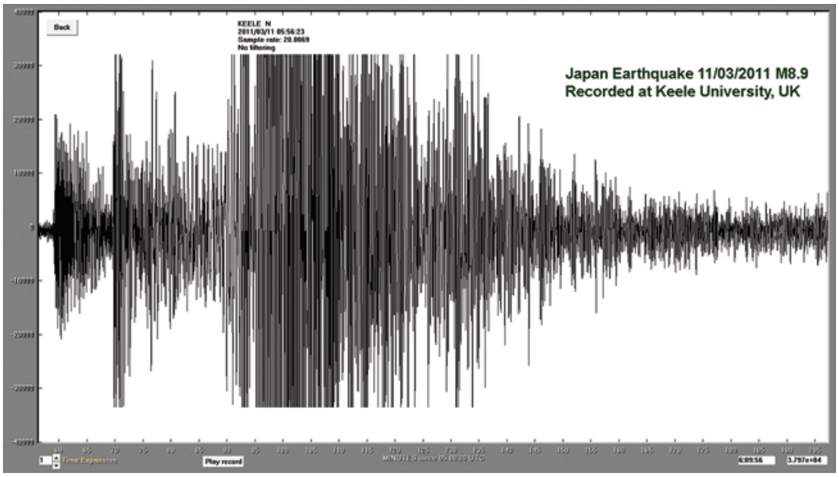
$$\begin{aligned} P(X \geq 2) &= 1 - (P(X=0) + P(X=1)) \\ &= 1 - e^{-1.10556}(1 + 1.10556) \simeq 0.3030. \end{aligned}$$

Distributional Result (7) may be used to answer questions about the waiting times between serious earthquakes. The waiting time T in days between serious earthquakes has an exponential distribution with parameter $\lambda = 67/22120$, so the c.d.f. of T is

$$F(t) = 1 - e^{-\lambda t} = 1 - e^{-(67/22120)\times t}, \quad t > 0.$$

So, for instance, the probability of a gap of more than one year between serious earthquakes is

$$\begin{aligned} P(T > 365) = 1 - F(365) &= 1 - \left(1 - e^{-(67/22120)\times 365}\right) \\ &= e^{-1.10556} \simeq 0.3310. \end{aligned}$$



A seismometer records the moment the Japanese earthquake (and then tsunami) of 2011 occurred

Activity 16 *Earthquakes in a decade*

Suppose that the occurrences of serious earthquakes may be reasonably modelled as a Poisson process with rate $\lambda = 67/22120$ per day.

(a) Write down the distribution of the number of serious earthquakes worldwide in a typical decade. (Allow for two leap years.)

(b) Find the probability that there will be exactly five serious earthquakes in a typical decade.

(c) Find the probability that there will be at least five and at most seven serious earthquakes in a typical decade.

(d) What is the probability of a gap of more than five years between serious earthquakes?

Activity 17 *Grounded aircraft*

Suppose that the occurrences of wind conditions severe enough to ground aircraft at an airport may be modelled as a Poisson process with rate $\lambda = 3$ per year.

(a) Find the probability that aircraft are grounded because of severe wind conditions twice or more in a typical year.

(b) Find the probability that aircraft are grounded exactly once in a two-year period.

(c) Find the probability of a gap of more than two years between occasions when aircraft are grounded because of severe wind conditions.

Some more examples of calculations of probabilities of numbers of events and of waiting times between events are contained in Screencast 5.2.

Screencast 5.2 Calculations associated with the Poisson process

3.2 Is a Poisson process a good model?

Many events appear to occur ‘at random’ in continuous time in such a way that it is impossible to forecast when the next event will happen. For example, lightning strikes, road accidents, pleasant surprises, air crashes and earthquakes are unpredictable. But does this mean that a Poisson process is a good model for such events?

Given appropriate data, Distributional Results (6) and (7) suggest a possible approach to the problem of deciding whether a Poisson process is a reasonable model for the occurrences of unpredictable events. For a Poisson process, the number of events that occur in an interval of length t has a Poisson distribution. So if data are available on the number of events that occur in intervals of some fixed length, then these data can be compared with a Poisson distribution. Or, since the waiting times between successive events in a Poisson process have an exponential distribution, given data on waiting times, we could investigate whether an exponential model is a good fit for the data. Another possibility, given suitable data, is to check whether it is reasonable to assume that the rate at which events occur remains constant: this is a defining property of a Poisson process.

These possible approaches are illustrated in the examples that follow and in Chapter 10 of Computer Book A.

Example 13 *Particle counting*

The physical nature of radioactive particle emission is such that emissions occur entirely at random, independently of one another, but at a rate which remains virtually constant with passing time. This suggests that a Poisson process ought to be a good model for emissions.

In 1910, the scientists Ernest Rutherford and Hans Geiger reported an experiment in which they counted the number of alpha particles emitted from a radioactive source during intervals of length $7\frac{1}{2}$ seconds. Counts were obtained for 2612 intervals, 10 126 particles being recorded in total. The numbers of particles emitted and the frequencies with which the different counts occurred are shown in Table 7. (Source: Rutherford, E. and Geiger, H. (1910) 'The probability variations in the distribution of alpha particles', *Philosophical Magazine*, Sixth Series, vol. 20, pp. 698–704.)

Rutherford (right) and Geiger (left) in the physics laboratory of the Victoria University of Manchester circa 1912

If the Poisson process is a good model for the emission of alpha particles, then the data in Table 7 should be very like observations from a Poisson distribution. Recall that the mean and variance of a Poisson distribution with parameter μ are both equal to μ. So if the Poisson process is a good model, then the sample mean and sample variance should be similar. For these data, the sample mean is 3.877 and the sample variance is 3.696, so the values of the two statistics are indeed close. Moreover, if a Poisson distribution is fitted to these data, then an appropriate estimate of the parameter μ is the sample mean, 3.877.

In Figure 7 the relative frequencies of the counts and the probability mass function of a Poisson distribution with parameter 3.877 are drawn side by side. As you can see, the Poisson distribution appears to fit the data very well.

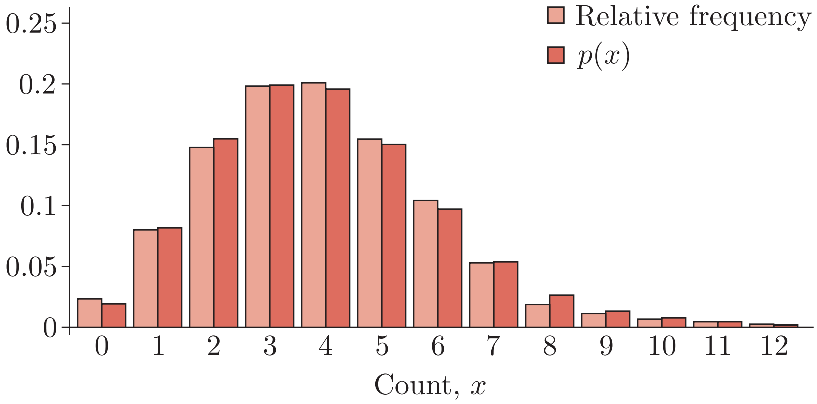


Figure 7 Relative frequencies of emissions and the p.m.f. of the Poisson(3.877) distribution

Table 7 Emissions of alpha particles

Count	Frequency
0	57
1	203
2	383
3	525
4	532
5	408
6	273
7	139
8	49
9	27
10	10
11	4
12	2
> 12	0

Table 8 Emissions and expected frequencies

Count	Frequency	Expected frequency
0	57	54
1	203	210
2	383	407
3	525	525
4	532	509
5	408	395
6	273	255
7	139	141
8	49	68
9	27	30
10	10	11
11	4	4
12	2	1
> 12	0	1

Table 8 contains the data from Table 7 and, in the third column, the frequencies that we would expect to obtain for a Poisson(3.877) distribution. These expected frequencies were calculated in the following way. For a Poisson(3.877) distribution, $p(0) = e^{-3.877} \simeq 0.0207$, so of the 2612 $7\frac{1}{2}$-second intervals, we would expect $0.0207 \times 2612 \simeq 54$ intervals to contain no emissions. Similarly, $p(1) = 3.877e^{-3.877} \simeq 0.0803$, so we would expect $0.0803 \times 2612 \simeq 210$ intervals to contain one emission. And so on. (The sum of the expected frequencies is 2611 rather than 2612; this discrepancy is due to rounding error.) You do not need to be able to calculate expected frequencies like these at this stage in the module.

Because the observed frequencies and the expected frequencies are close, a Poisson(3.877) distribution seems to fit the data very well. In this respect, the data are consistent with a Poisson process being a good model for the emission of alpha particles.

A word of caution is appropriate here. Note that if a Poisson distribution is a good model for the counts, this does not necessarily mean that a Poisson process is a good model for the occurrences of the events. Suppose, for instance, that all the low counts occurred early on during the period of observation, and all the high counts towards the end. This would suggest that the rate of occurrence of events was not constant but was increasing; and in that case, a Poisson process would not be a good model for the occurrences of the events. Given the data only in the form of Table 7, we cannot check this out further.

Example 14 *Serious earthquakes*

More earthquake damage

If a Poisson process is a good model for the occurrences of serious earthquakes around the world, then the waiting times between successive serious earthquakes will be observations from an exponential distribution with parameter λ, where λ is the rate of occurrence of serious earthquakes.

In Section 2, data on waiting times between serious earthquakes were introduced. These data are repeated in Table 9 for convenience.

Table 9 Time intervals between serious earthquakes (days)

186	77	682	449	91	1027	164	808	83	832
328	1120	743	328	163	83	64	356	324	257
503	232	252	151	142	32	120	100	561	755
44	200	47	503	320	690	386	147	533	109
560	26	460	420	291	474	131	643	71	270
115	240	204	34	493	423	422	219	366	92
194	230	717	506	104	91	332			

An exponential model was proposed in Subsection 2.2. But is the model a good fit to the data?

See Equations (3) and (5). The mean and standard deviation are both equal to $1/\lambda$.

For an exponential distribution, the mean and standard deviation are equal. The sample mean and the sample standard deviation for these data are approximately 330 days and 255 days, respectively. These values are not *too* different, but there may be a little doubt that the data come from an exponential distribution.

A unit-area histogram of the data is shown in Figure 8. As already noted in Subsection 2.2, these data are highly skew: they peak on the left of the histogram and the frequencies tail off to the right. This shape is reasonably consistent with the data coming from an exponential distribution. The p.d.f. of the exponential distribution with parameter $\lambda = 67/22120$ is also shown in Figure 8, superimposed on the histogram.

Figure 8 is essentially Figures 1 and 4 combined.

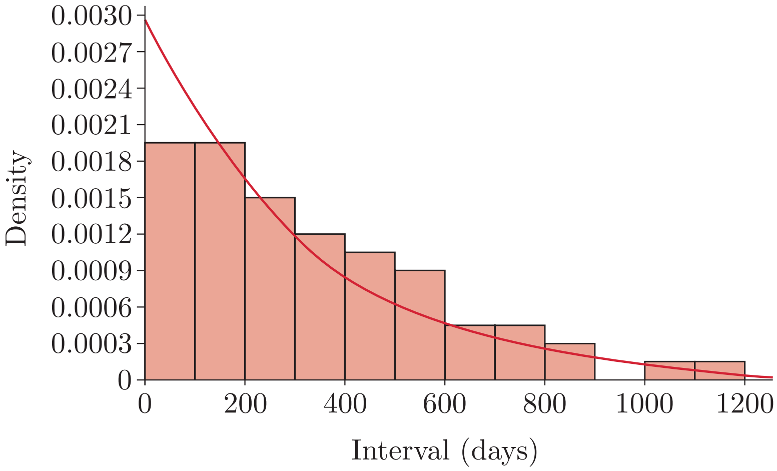


Figure 8 A unit-area histogram of the earthquake data with the p.d.f. of an exponential model superimposed

With slight misgivings, it looks as though the exponential model might be a reasonable one for these data.

The data in Table 9 may also be used to obtain a visual check on whether or not the rate of occurrence of earthquakes remains constant over time. Reading across the rows, the time intervals are given in order of occurrence. So the first earthquake after the one on 15 August 1950 occurred after 186 days; the second, after 263 days $(186 + 77)$; the third, after 945 days $(186 + 77 + 682)$; and so on. That is, 1 earthquake had occurred after 186 days, 2 had occurred after 263 days, 3 after 945 days, etc. In Figure 9(a) (overleaf), the time of each earthquake (that is, the cumulative time) is plotted along a line in the manner of Figure 6. The same information is relayed in possibly more useful form in Figure 9(b). There the number of earthquakes that had occurred since the start of the period of observation is plotted against the times at which the earthquakes occurred. If the earthquakes are occurring at a constant rate, the points in Figure 9(b) should approximately follow a straight line through the origin.

The points in Figure 9(b) seem to lie quite close to a straight line through the origin, suggesting that over the period of observation the rate of occurrence of serious earthquakes remained roughly constant. So it would seem that a Poisson process might be a reasonable model for the occurrences of serious earthquakes.

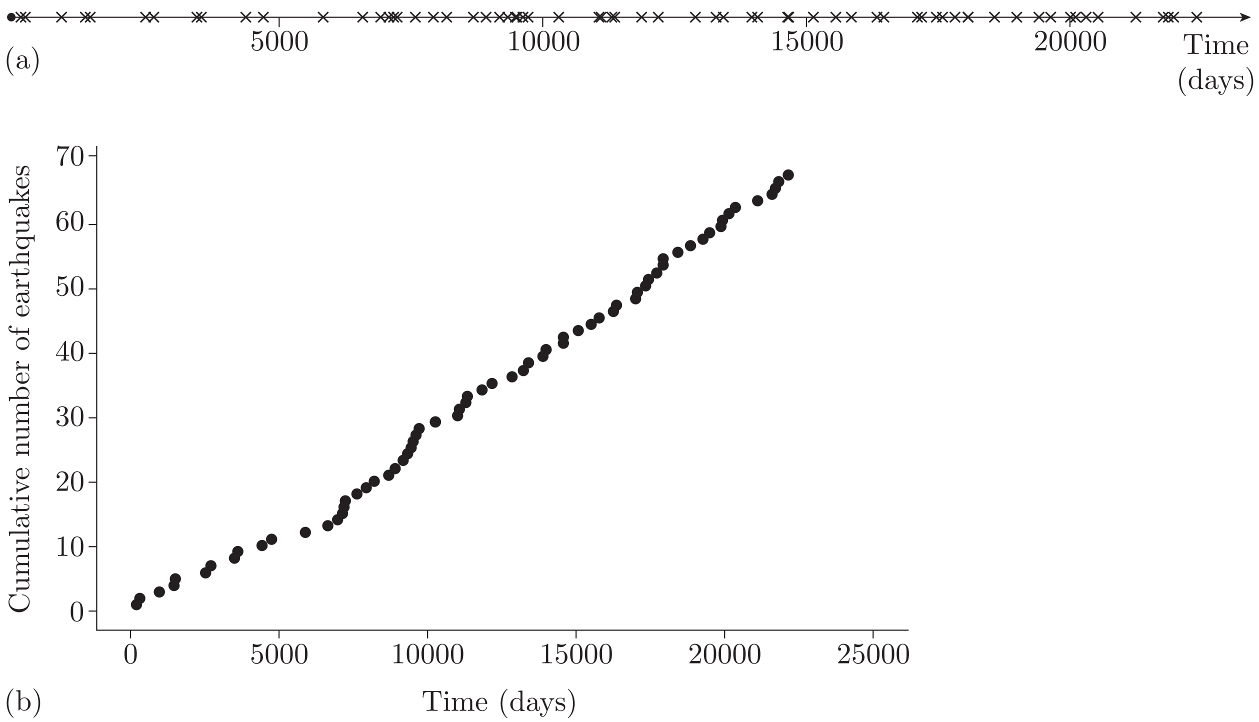


Figure 9 Times of occurrence of serious earthquakes plotted in two ways: (a) as points along a timeline, (b) as a scatterplot of cumulative number of earthquakes against their times of occurrence

The rest of the work in this subsection is contained in Computer Book A. For each of a number of datasets, you will investigate informally whether a Poisson process might be a good model for the occurrences of the events. In each case, data are given either on counts or on waiting times. The following box is intended as a reminder of the different equalities between means and dispersion parameters that we are looking for if Poisson and exponential distributions are good models for our data.

Equalities between means and dispersion parameters

For a Poisson(λ) distribution,

$$\text{mean} = \text{variance } = \lambda;$$

for an exponential distribution with parameter λ,

$$\text{mean} = \text{standard deviation } = 1/\lambda.$$

You will also use Minitab to calculate probabilities associated with events occurring in a Poisson process.

Refer to Chapter 10 of Computer Book A for the rest of the work in this subsection.

Exercises on Section 3

Exercise 6 *Explosive volcanic eruptions*

Data on the intervals between successive major explosive volcanic eruptions in the northern hemisphere were introduced in Exercise 3. Suppose that occurrences of such eruptions may be modelled as a Poisson process with rate $\lambda = 0.0352$ per month.

(a) Write down the distribution of the number of eruptions that occur in a year.

(b) (i) Find the probability that exactly one eruption occurs in a year.

(ii) Find the probability that there is more than one eruption in a year.

(c) Write down the distribution of the waiting times in months between successive eruptions.

(d) Find the probability that the gap between successive eruptions will exceed five years.

Exercise 7 *Telephone orders at a restaurant*

Suppose that the arrivals of telephone orders at a popular pizza outlet on a Saturday evening may be modelled as a Poisson process with rate 15 per hour.

(a) Write down the distribution of the number of telephone orders received in ten minutes.

(b) (i) Find the probability that exactly three orders are received in ten minutes.

(ii) Find the probability that at most two orders are received in ten minutes.

(c) Write down the distribution of the waiting times in hours between the arrivals of successive telephone orders.

(d) Find the probability that the gap between two successive orders will exceed five minutes.

4 Population quantiles

To conclude this unit, we consider something a bit different. In Unit 4, the population analogues of the sample mean, sample variance and sample standard deviation were defined. In this section, the population analogues of the sample median and sample quartiles are defined and discussed.

Quantiles are also pretty much synonymous with 'percentiles'

More generally, the idea of *population quantiles* is introduced: the population median and population quartiles are examples of population quantiles. This idea is important for many of the statistical techniques discussed later in the module.

Quantiles of continuous distributions are discussed in Subsection 4.1 and quantiles of discrete distributions in Subsection 4.2.

4.1 Quantiles of continuous distributions

Before the idea of population quantiles in general is introduced, the population analogues of the sample median and sample quartiles will be discussed briefly.

The population median

In Unit 1, the sample median was defined to be the middle value of a set of values (or halfway between the two middle values when there is an even number of values). Roughly speaking, the sample median divides the data into two halves, one with values lower than the median and the other with values greater than the median.

For a continuous random variable X, the population analogue of the sample median is the value m that splits the probability distribution of X into two halves in such a way that

$$P(X \le m) = \tfrac{1}{2} \quad \text{and} \quad P(X \ge m) = \tfrac{1}{2}.$$

Recall that for continuous distributions, $P(X = m)$ is effectively 0, as it is for any other value in the range of X.

Equivalently, if $F(x)$ is the c.d.f. of X, then

$$F(m) = 1 - F(m) = \tfrac{1}{2}.$$

Thus we have the following definition.

The population median

For a continuous random variable X with c.d.f. $F(x)$, the **population median**, or simply the **median**, is the value, m, which satisfies

$$F(m) = \tfrac{1}{2}.$$

Notice that, for continuous population distributions, there is no ambiguity in the definition of the (population) median: there is no analogue of the need to treat samples of odd and even sizes separately.

Example 15 *Finding the median*

Suppose that a random variable X has p.d.f.

$$f(x) = 3x^2, \quad 0 < x < 1,$$

and c.d.f.

$$F(x) = P(X \le x) = x^3, \quad 0 < x < 1.$$

$F(x)$ derives from $f(x)$ in the usual way via $\int_0^x f(y)\,dy$.

To find the median of this distribution, we need to find the value m such that

$$F(m) = \tfrac{1}{2}.$$

That is, we need to solve the equation

$$m^3 = \tfrac{1}{2},$$

which gives

$$m = \left(\tfrac{1}{2}\right)^{1/3} \simeq 0.794.$$

The median is shown on a plot of the p.d.f. of this distribution in Figure 10.

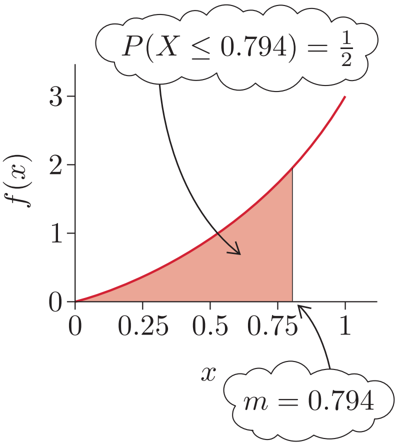


Figure 10 The median of X

Activity 18 *The median of a continuous uniform distribution*

The continuous random variable X is uniformly distributed over the interval $a < x < b$: $X \sim U(a,b)$. This distribution has c.d.f.

$$F(x) = \frac{x-a}{b-a}, \quad a < x < b.$$

(a) What is the median of X?

(b) How does the median compare with the mean $\mu = (a+b)/2$?

The result in part (b) of Activity 18 is a consequence of the symmetry of the uniform distribution. The 'centre of symmetry' is then both the median and the mean of the distribution. For distributions that are not symmetric, means and medians differ. The exponential distribution is a good example, as you will investigate next.

Activity 19 *The median of an exponential distribution*

A random variable X is exponentially distributed with parameter λ: $X \sim M(\lambda)$. This distribution has c.d.f.

$$F(x) = 1 - e^{-\lambda x}, \quad x > 0.$$

(a) Find the median of X.

(b) Express the median of X as a multiple of the mean of X.

Activity 20 *Waiting times between earthquakes*

The sample median waiting time between successive earthquakes for the data given in Table 6 is approximately 257 days. In Example 7 an exponential distribution with mean 330 days was suggested as a model for the waiting time between serious earthquakes. According to this model, what is the (population) median waiting time between serious earthquakes? How does this compare with the sample median?

Population quartiles

The idea behind the definition of sample quartiles is to choose values which split the data into proportions of one-quarter and three-quarters. There are, of course, two ways to do this. Consequently, as you saw in Unit 1, there are two quartiles: the values of approximately a quarter of the data are smaller than the lower quartile, and three-quarters are larger; and the values of approximately three-quarters of the data are smaller than the upper quartile, and a quarter are larger. The difference between the upper and lower quartiles gives the sample interquartile range.

To define population quartiles for continuous distributions, ‘proportions’ are replaced by ‘probabilities’ in the following way. The lower population quartile of a continuous random variable X, denoted by q_L, satisfies

$$P(X \leq q_L) = F(q_L) = \tfrac{1}{4},$$

and the upper population quartile, denoted by q_U, satisfies

$$P(X \leq q_U) = F(q_U) = \tfrac{3}{4}.$$

So population quartiles of continuous distributions are defined as follows.

Population quartiles

For a continuous random variable X with c.d.f. $F(x)$, the **lower quartile** is the value, q_L, which satisfies

$$F(q_L) = \tfrac{1}{4}.$$

The **upper quartile** is the value, q_U, which satisfies

$$F(q_U) = \tfrac{3}{4}.$$

The **interquartile range** is $q_U - q_L$.

Notice that the population interquartile range has also been defined in the above box, in terms of the population quartiles in the obvious way. Just as the sample interquartile range is a measure of the spread or dispersion in a sample of data, so the population interquartile range is a measure of the spread or dispersion in a population model.

Example 16 *Finding the lower quartile*

The c.d.f. of the random variable X of Example 15 is

$$F(x) = x^3, \quad 0 < x < 1.$$

The lower quartile, q_L, is the solution of $F(x) = \frac{1}{4}$. That is,

$$F(q_L) = (q_L)^3 = \tfrac{1}{4},$$

and hence

$$q_L = \left(\tfrac{1}{4}\right)^{1/3} \simeq 0.630.$$

Activity 21 *Finding the upper quartile and the interquartile range*

(a) Find the upper quartile q_U for the random variable X of Example 16.

(b) Hence find the interquartile range of X.

Activity 22 *Quartiles of an exponential distribution*

A random variable X is exponentially distributed with parameter λ: $X \sim M(\lambda)$.

(a) Find the lower quartile and the upper quartile of X.

(b) Hence find the interquartile range of X.

(c) The (sample) lower quartile of the waiting times between successive serious earthquakes for the data given in Table 6 is 115 days and the (sample) upper quartile is 493 days, so the (sample) interquartile range is $493 - 115 = 378$ days. In Example 7, an exponential distribution with mean 330 days was suggested as a model for the waiting time between serious earthquakes. According to this model, what are the lower quartile, the upper quartile and the interquartile range of the waiting time between serious earthquakes? How do these compare with the values calculated from the data?

Quantiles

Finding a population median involves solving the equation $F(x) = \frac{1}{2}$. The quartiles, q_L and q_U, are found by solving $F(x) = \frac{1}{4}$ and $F(x) = \frac{3}{4}$, respectively. So each of the median, the lower quartile and the upper quartile is the solution of an equation of the form $F(x) = \alpha$, where $0 < \alpha < 1$. More generally, the solution of an equation of this form is called a *population quantile*. The definition is as follows.

Quantiles of continuous distributions

For a continuous random variable X with c.d.f. $F(x)$, the **α-quantile** is the value x which is the solution of the equation

$$F(x) = \alpha, \quad 0 < \alpha < 1.$$

This value is denoted q_α. Put another way, q_α is the value such that

$$F(q_\alpha) = P(X \leq q_\alpha) = \alpha.$$

Notice that α is a value between 0 and 1 that labels which quantile you require, while the α-quantile itself, q_α, is a value in the range of the distribution.

The idea of a population quantile is illustrated in Figure 11. Each diagram shows the graph of a typical continuous c.d.f. $F(x)$. For a given value a, say, on the horizontal axis, you could evaluate (or read off the graph) the corresponding value $F(a)$ on the vertical axis. ($F(a) = P(X \leq a)$ is a probability, so $0 \leq F(a) \leq 1$.) This is shown in Figure 11(a).

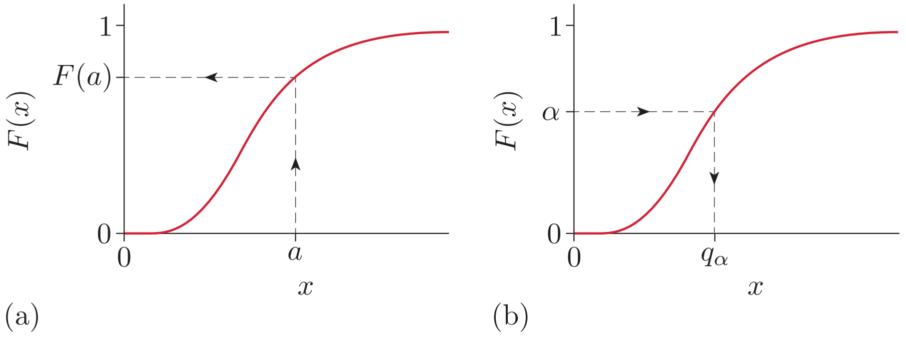


Figure 11 Illustrating population quantiles

In Figure 11(b), the starting point is a number α $(0 < \alpha < 1)$ on the vertical axis. The corresponding value on the horizontal axis is the α-quantile of X, the value q_α. That is, the α-quantile q_α is the value q_α which satisfies $F(q_\alpha) = \alpha$. (Mathematically, what we are doing is *inverting* the cumulative distribution function. We could therefore write $q_\alpha = F^{-1}(\alpha)$, where F^{-1} is the inverse function of F. But we won't usually do so explicitly in this module.)

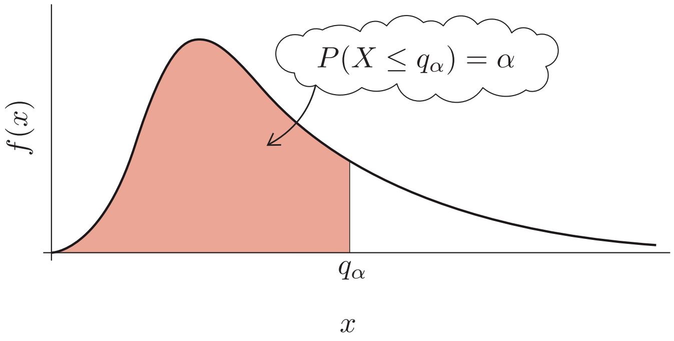


Figure 12 Illustrating a population quantile on the p.d.f.

By definition, $F(x) = P(X \leq x)$, so

$$F(q_\alpha) = P(X \leq q_\alpha) = \alpha.$$

So the definition of the α-quantile, q_α, says that the values of a proportion α of the population are less than or equal to the α-quantile. This can also be illustrated on the p.d.f. (as opposed to the c.d.f.) as shown in Figure 12. Note that Figures 11(a), 11(b) and 12 are different ways of showing the same thing. In particular, as you already know, half the values are below or equal to the median m, so $m = q_{0.5}$; and a quarter of the values are below or equal to the lower quartile q_L, so $q_L = q_{0.25}$. Similarly, $F(q_U) = 0.75$, so $q_U = q_{0.75}$. The median and quartiles are illustrated on a c.d.f. in Figure 13.

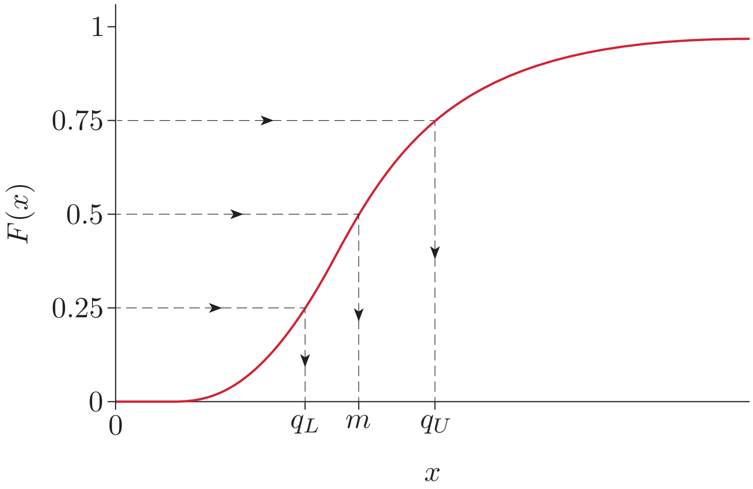


Figure 13 The median and quartiles

The number α is often expressed as a percentage. When this is the case, the quantile is often called a **percentile** or **percentage point**; these terms are synonymous with quantile. So, for example, the 0.3-quantile is also the 30th percentile, and the 95% percentage point is also the 0.95-quantile. Also, if α is an integer multiple of 0.1, the corresponding quantiles are sometimes called **deciles**: that is, the deciles are $q_{0.1}, q_{0.2}, \ldots, q_{0.9}$.

‘Almost half of Britain’s private wealth owned by top 10% of households’, headline in *The Guardian*, 18 December 2015. The ‘top 10%’ of people are those whose wealth exceeds the $q_{0.9}$ decile, or equivalently the 0.9-quantile or 90th percentile.

The median and quartiles have been generalised to population α-quantiles because many statistical techniques depend on the idea of quantiles. In particular, for some of the methods introduced later in the module, the ‘extremes’ or ‘tails’ of a distribution are important. So quantiles associated with values such as 0.9, 0.95, 0.99, at one extreme, and 0.1, 0.05, 0.01, at the other, are often used.

You can get more of a feel for the definition and meaning of quantiles of continuous distributions using the computer animation provided in Chapter 11 of Computer Book A.

Refer to Chapter 11 of Computer Book A to explore quantiles further.

Example 17 *Finding quantiles mathematically*

The c.d.f. of the random variable X of Example 15 is

$$F(x) = x^3, \quad 0 < x < 1.$$

So the 0.1-quantile, $q_{0.1}$, is the solution of $F(x) = 0.1$. That is,

$$F(q_{0.1}) = (q_{0.1})^3 = 0.1,$$

and hence

$$q_{0.1} = (0.1)^{1/3} \simeq 0.464.$$

Similarly, $q_{0.9}$ is the solution of $F(x) = 0.9$. So

$$(q_{0.9})^3 = 0.9,$$

and hence

$$q_{0.9} = (0.9)^{1/3} \simeq 0.965.$$

Activity 23 *Finding more quantiles*

The c.d.f. of the random variable Z is

$$F(z) = z^4, \quad 0 < z < 1.$$

Find the quantiles $q_{0.05}$ and $q_{0.99}$ for Z.

Activity 24 *Quantiles for an exponential distribution*

A random variable X has an exponential distribution with parameter λ: $X \sim M(\lambda)$.

(a) Find the quantiles $q_{0.01}$ and $q_{0.8}$ of X.

(b) Following Figure 3, make a very rough sketch of the c.d.f. of the exponential distribution with parameter λ, and on it mark $\alpha = 0.8$ and $q_{0.8}$.

(c) Following Figure 18 in the solution to Activity 19, make a very rough sketch of the p.d.f. of the exponential distribution with parameter λ, and on it mark the mean and $q_{0.8}$.

In this subsection, you have seen how quantiles for a continuous random variable are calculated using the c.d.f. $F(x)$. However, for some continuous distributions, a simple formula for the c.d.f. is not available. In such cases, printed tables or computer software must be used to find quantiles.

There are also cases where the c.d.f. formula is available but it cannot be explicitly inverted.

4.2 Quantiles of discrete distributions

Quantiles of continuous distributions are found by solving the equation $F(x) = \alpha$ for x. In all the examples you have met, and in all the cases you will encounter in this module, this equation has a unique solution.

Unfortunately, when quantiles of discrete distributions are discussed, this simplicity and lack of ambiguity disappears. Quantiles of discrete distributions are not as important as those of continuous distributions, so they will be discussed only very briefly here. In this subsection, we will look at some of the problems and how they might be overcome.

Example 18 *The median score on a fair die*

We begin with the example of a uniform distribution on the integers $1, 2, \ldots, 6$. This is the model adopted for X, the score obtained when a fair six-sided die is rolled. The c.d.f. $F(x)$ of X is shown in Table 10.

Table 10 The c.d.f. of the score on a fair die

x	1	2	3	4	5	6
$F(x)$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{5}{6}$	1

If we solve the equation $F(x) = \frac{1}{2}$ to find the median, then this yields $m = 3$.

However, when the population analogue of a sample median was introduced at the beginning of Subsection 4.1, the underlying idea was to find a value m which splits the distribution of X into two halves in such a way that $P(X \leq m) = \frac{1}{2}$ and $P(X \geq m) = \frac{1}{2}$. For the random variable X in this example, $P(X \leq 3) = \frac{1}{2}$ and $P(X \geq 4) = \frac{1}{2}$. So, which of the values 3 and 4 should we take as the median of X? If the median is regarded as an indication of the 'centre' of the distribution, then neither value is satisfactory. So perhaps we should use neither value: any (non-integer) value x strictly between 3 and 4 does satisfy $P(X \leq x) = \frac{1}{2}$ and $P(X \geq x) = \frac{1}{2}$. Should we take one of these values? Considerations of symmetry suggest 3.5, which is at the centre of the distribution (and is also the mean of the distribution). But, as you will soon see, this is not the choice usually made.

If a sample turned out to take the values 1, 2, 3, 4, 5, 6, then 3.5 would be the sample median.

Example 19 *The median of a uniform distribution*

Another problem is illustrated by considering a random variable Y which is uniformly distributed on the integers $1, 2, \ldots, 5$. The c.d.f. is given in Table 11.

Table 11 The c.d.f. of a uniform distribution

y	1	2	3	4	5
$F(y)$	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{3}{5}$	$\frac{4}{5}$	1

In this case, if we try to solve the equation $F(y) = \frac{1}{2}$, then we find that the equation is not satisfied for any value in the range of Y. (In fact, it is not satisfied for any value y, such as 2.5, between values in the range of Y either.) Considerations of symmetry suggest $m = 3$, which is at the centre of the distribution (and is also the mean of the distribution). But is this satisfactory since $F(3) = \frac{3}{5}$?

These examples indicate that the definition of the median needs to be modified for discrete distributions. One possibility is to define the population median to be the *minimum* value m such that $F(m) \geq \frac{1}{2}$. For the uniform distribution on the integers $1, 2, \ldots, 5$ of Example 19, this yields the solution $m = 3$, which is in the middle of the range. However, this definition also yields the value $m = 3$ for the distribution of Example 18, and 3 is not the 'centre' of that distribution. This is not entirely satisfactory. Nevertheless, this is the definition that is used in this module.

'Perverse' and 'arbitrary' might also come to mind, but this is a fairly standard definition of the population median of a discrete distribution

Problems similar to those discussed above arise when finding other quantiles for discrete distributions. The definition of quantiles is modified in a similar way. The definition used in this module, which includes the definition of the median, is therefore as follows.

Quantiles of discrete distributions

For a discrete random variable X with c.d.f. $F(x)$, the **α-quantile** q_α is defined to be the *smallest* value of x in the range of X satisfying

$$F(x) \geq \alpha.$$

Example 20 *Quartiles of a binomial distribution*

Table 12 contains the cumulative distribution function $F(x)$ of a random variable X having a binomial distribution with parameters $n = 6$ and $p = 0.6$: $X \sim B(6, 0.6)$.

Table 12 The c.d.f. of a binomial distribution, $B(6, 0.6)$

x	0	1	2	3	4	5	6
$F(x)$	0.0041	0.0410	0.1792	0.4557	0.7667	0.9533	1

To find the median of X, we need to find the smallest value of x in the range of X such that $F(x) \geq 0.5$. From the table, $F(3) = 0.4557$, which is less than 0.5, and $F(4) = 0.7667$, which is greater than 0.5, so

$$F(3) < 0.5 \leq F(4).$$

Hence the median m is 4.

The lower quartile q_L is the smallest value of x in the range of X for which $F(x) \geq 0.25$. Since $F(2) < 0.25 \leq F(3)$, the lower quartile q_L is 3. Similarly, since $F(3) < 0.75 \leq F(4)$, the upper quartile q_U is 4.

Notice that in this example, the median and the upper quartile are equal.

Activity 25 *Quantiles of a binomial distribution*

Use the values of the c.d.f. given in Table 12 to find the quantiles $q_{0.01}$, $q_{0.05}$, $q_{0.8}$ and $q_{0.95}$ for the binomial random variable X of Example 20: $X \sim B(6, 0.6)$.

The final work in this subsection consists of another short chapter of Computer Book A. This chapter explains how to use Minitab to find quantiles for both continuous and discrete distributions.

Refer to Chapter 12 of Computer Book A for the rest of the work in this subsection.

Exercises on Section 4

Exercise 8 *Quartiles of a continuous uniform distribution*

The random variable X is uniformly distributed over the interval $a < x < b$: $X \sim U(a,b)$. It has c.d.f.

$$F(x) = \frac{x-a}{b-a}, \quad a < x < b.$$

Find the lower quartile, upper quartile and interquartile range of X.

Exercise 9 *Finding quantiles of another continuous distribution*

The c.d.f. of the continuous random variable Y is given by

$$F(y) = \tfrac{1}{4}y^2, \quad 0 < y < 2.$$

(a) Find the median, the quartiles and the interquartile range of Y.

(b) The c.d.f. of interest in this question is shown in Figure 14.

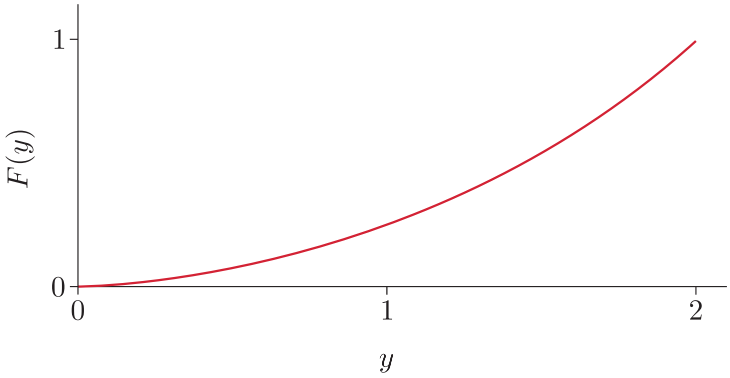


Figure 14 The c.d.f. $F(y) = \frac{1}{4}y^2$

On a copy of this figure, sketch the positions of the median and quartiles, and the corresponding values of α.

Exercise 10 *Finding quantiles of yet another continuous distribution*

The c.d.f. of the continuous random variable Z is given by

$$F(z) = z^\beta, \quad 0 < z < 1,$$

with $\beta > 0$.

(a) Find a general formula for the α-quantile of Z.

(b) Find the median, the lower quartile and the 0.8-quantile of Z.

(c) Hence give the 0.8-quantile for the particular choice $\beta = \frac{1}{2}$.

Exercise 11 *Quartiles of a discrete uniform distribution*

The random variable X is uniformly distributed on $1, 2, \ldots, 10$.

(a) Draw up a table showing values of the c.d.f. $F(x)$ of X for $x = 1, 2, \ldots, 10$.

(b) Find the lower quartile, the median and the upper quartile of X.

Summary

In this unit, you first saw, numerically, how the Poisson distribution with parameter $\mu = np$ provides a useful approximation to the binomial distribution $B(n, p)$ for large values of n and small values of p. You also met the exponential distribution, denoted $M(\lambda)$, a continuous distribution with (rate) parameter $\lambda > 0$. The exponential distribution has mean $1/\lambda$ and variance $1/\lambda^2$.

Two models for events that occur at random as time progresses have been discussed. The Bernoulli process is a model for events that may occur only at a discrete set of opportunities, or trials, in such a way that the probability that an event occurs at each trial is constant and is independent of the outcomes of previous trials. The Poisson process is the continuous-time analogue of the Bernoulli process: events occur singly at a constant rate, and the incidence of future events is independent of the past.

The parallels between the Bernoulli and Poisson processes are summarised in Table 13.

Table 13 The parallels between the Bernoulli and Poisson processes

Process	Type	Distribution of number of events	Distribution of waiting time between events
Bernoulli	discrete	Binomial	Geometric
Poisson	continuous	Poisson	Exponential

The population analogues of the sample median and sample quartiles have also been discussed, and the idea generalised to population quantiles.

Learning outcomes

After you have worked through this unit, you should be able to:

- understand and use the fact that the Poisson distribution with parameter $\mu = np$ provides a useful approximation to the binomial distribution $B(n, p)$ for large values of n and small values of p
- use the fact that the parameter λ of an exponential distribution may be estimated from data by the reciprocal of the sample mean
- recognise that the number of events in a Poisson process with rate λ per unit time that occur in an interval of length t has a Poisson distribution with parameter λt
- recognise that the waiting time between successive events in a Poisson process with rate λ has an exponential distribution with parameter λ
- use the Poisson distribution and the exponential distribution to calculate probabilities associated with events occurring in a Poisson process, both mathematically and by using Minitab
- explore data to investigate whether a Poisson process is a plausible model for the occurrences of events
- appreciate that, in the continuous case, the α-quantile, q_α, solves $F(q_\alpha) = \alpha$, $0 < \alpha < 1$, where F is the cumulative distribution function
- appreciate that, in the discrete case, the α-quantile, q_α, is the smallest value of x in the range of X satisfying $F(x) \geq \alpha$, $0 < \alpha < 1$
- calculate population α-quantiles, both mathematically and by using Minitab
- remember that the median, m, is the 0.5-quartile, the lower quartile, q_L, is the 0.25-quartile, and the upper quartile, q_U, is the 0.75-quartile, and that $q_U - q_L$ is the interquartile range.

Solutions to activities

Solution to Activity 1

(a) Assuming 50 journeys a month and using the given average of 15 drenchings a year, the proportion of journeys that result in a drenching is

$$p = \frac{15}{12 \times 50} = \frac{15}{600} = \frac{1}{40}.$$

(b) The number of times a month that Chris gets drenched, X, has a binomial distribution:

$$X \sim B(50, 1/40).$$

(c) Assuming the binomial model, the mean number of drenchings a month is

$$E(X) = n \times p = 50 \times \frac{1}{40} = 1.25.$$

(d) (i) The probability that Chris does not get drenched at all in a month is

$$P(X = 0) = \binom{50}{0} \left(\tfrac{1}{40}\right)^0 \left(\tfrac{39}{40}\right)^{50} = \left(\tfrac{39}{40}\right)^{50} \simeq 0.2820.$$

(ii) The probability that Chris gets drenched twice is

$$P(X = 2) = \binom{50}{2} \left(\tfrac{1}{40}\right)^2 \left(\tfrac{39}{40}\right)^{48} = 1225 \left(\tfrac{1}{40}\right)^2 \left(\tfrac{39}{40}\right)^{48} \simeq 0.2271.$$

Solution to Activity 2

If $X \sim B(60, 1/48)$, then

$$P(X = 0) = \left(\tfrac{47}{48}\right)^{60} \simeq 0.2827.$$

To obtain this result, full calculator accuracy was retained for 47/48. Rounding 47/48 to 0.979, for instance, would give $(0.979)^{60} \simeq 0.2799$, and so would introduce quite serious rounding errors. In calculations like these, it is important not to round intermediate values.

Also, if $X \sim B(60, 1/48)$, then

$$P(X = 2) = \binom{60}{2} \left(\tfrac{1}{48}\right)^2 \left(\tfrac{47}{48}\right)^{58} = 1770 \left(\tfrac{1}{48}\right)^2 \left(\tfrac{47}{48}\right)^{58} \simeq 0.2266.$$

The completed table is as follows.

Table 14

	$P(X = 0)$	$P(X = 1)$	$P(X = 2)$	$P(X = 3)$	$P(X \geq 4)$
$B(50, 1/40)$	0.2820	0.3615	0.2271	0.0932	0.0362
$B(60, 1/48)$	0.2827	0.3610	0.2266	0.0932	0.0365

Notice that the two sets of probabilities do not differ greatly: all the differences between values are less than 0.001 in absolute value.

Solution to Activity 3

(a) The number of faulty resistors in a box, X, has a binomial distribution: $X \sim B(50, 1/20)$ or $X \sim B(50, 0.05)$. So

$$P(X = 1) = \binom{50}{1} (0.05)\,(0.95)^{49} \simeq 0.2025,$$

$$P(X = 3) = \binom{50}{3} (0.05)^3\,(0.95)^{47} \simeq 0.2199.$$

(b) Since the mean of X is $50 \times 0.05 = 2.5$, the approximating probability distribution is Poisson(2.5). So

$$P(X = 1) \simeq \frac{e^{-2.5}2.5}{1!} \simeq 0.2052,$$

$$P(X = 3) \simeq \frac{e^{-2.5}(2.5)^3}{3!} \simeq 0.2138.$$

(c) The probabilities are ‘similar’, but are not really very close – certainly not as close as in some previous activities and examples. The parameter p is 0.05, which is at the limit of the range of values given in the rough rule for when a Poisson approximation will be a good one. Also, n is not as large as it has been in some previous examples.

Solution to Activity 4

X, the number of passengers who have booked a seat but do not turn up, may be modelled by a binomial distribution: $X \sim B(200, 0.01)$. Since p is small and n is large, this may be approximated by a Poisson distribution with mean $\mu = 200 \times 0.01 = 2$.

Everyone who turns up will get a seat if at least two people fail to turn up. Hence the required probability is

$$P(X \geq 2) = 1 - (P(X = 0) + P(X = 1)) \simeq 1 - (e^{-2} + 2e^{-2}) \simeq 0.594.$$

Solution to Activity 5

In realisation (b), the observed values of X are 4, 3, 1, 1, 6, 2, 1, 3 and 1. The next value of X must be 4 or more.

In realisation (c), the observed values of X are 2, 2, 1, 2, 5, 5, 1 and 2. The next value of X must be 6 or more.

Solution to Activity 6

The number of trials from one success to the next, X, has a geometric distribution with parameter $p = 0.3$.

(a) $P(X = x) = (0.7)^{x-1} \times 0.3, \quad x = 1, 2, \ldots,$

so we have the following.

(i) $P(X = 2) = 0.7 \times 0.3 = 0.21.$

(ii) $P(X = 6) = (0.7)^5 \times 0.3 \simeq 0.050.$

(b) $$P(X \leq x) = 1 - (0.7)^x, \quad x = 1, 2, \ldots,$$

so

$$P(X \geq 3) = 1 - P(X \leq 2) = 1 - \left(1 - (0.7)^2\right) = 0.49.$$

Alternatively,

$$\begin{aligned} P(X \geq 3) &= 1 - (P(X = 1) + P(X = 2)) \\ &= 1 - 0.3 - 0.21 = 0.49. \end{aligned}$$

Solution to Activity 7

(a) For a fair six-sided die, $p = 1/6$. The p.m.f. of X, the number of rolls from one six to the next, is therefore given by

$$p(x) = \left(\tfrac{5}{6}\right)^{x-1} \times \tfrac{1}{6}, \quad x = 1, 2, \ldots.$$

So

$$P(X = 4) = \left(\tfrac{5}{6}\right)^3 \times \tfrac{1}{6} \simeq 0.096.$$

(b) The c.d.f. of X is given by

$$F(x) = P(X \leq x) = 1 - \left(\tfrac{5}{6}\right)^x, \quad x = 1, 2, \ldots.$$

So

$$P(X \geq 5) = 1 - P(X \leq 4) = 1 - \left(1 - \left(\tfrac{5}{6}\right)^4\right) \simeq 0.482.$$

Solution to Activity 8

(a) Two assumptions are made. First, whether or not a serious earthquake occurs on any particular day is independent of whether or not a serious earthquake occurs on any other day. Second, the probability that a serious earthquake occurs on any single day remains constant from day to day.

(b) The time between successive serious earthquakes has a geometric distribution with parameter p, where p is the probability of a serious earthquake on any single day.

(c) The histogram is highly skew, with the largest frequencies for low values and with frequencies tailing off towards the right. So a geometric model might be a reasonable one.

(d) We could calculate the mean time between serious earthquakes for the given data. Then, since the mean μ of a geometric distribution is equal to $1/p$, we could estimate p by the reciprocal of the sample mean. (This is equivalent to dividing the number of serious earthquakes after 15 August 1950, which is 67, by the total number of days, thus calculating the proportion of days on which a serious earthquake occurred.)

Solution to Activity 9

A time interval of two years, with no leap year, corresponds to 730 days altogether. The probability of a gap exceeding 730 days is

$$\begin{aligned} P(X>730) = 1 - F(730) &= 1 - \left(1 - \left(1 - \frac{67}{22\,120}\right)^{730}\right) \\ &= \left(1 - \frac{67}{22\,120}\right)^{730} \simeq 0.1092, \end{aligned}$$

or a little above 1 in 10. Since $0.1092 \times 67 \simeq 7.3$, we might expect around seven intervals in a list of 67 to exceed 730 days. In fact, in Table 6 there are six such gaps; they are of lengths 1027, 808, 832, 1120, 743 and 755 days.

Solution to Activity 10

(a) $\displaystyle\int e^x\,dx = \frac{e^x}{1} + c = e^x + c.$

(b) $\displaystyle\int_1^5 e^{-x}\,dx = \left[\frac{e^{-x}}{-1}\right]_1^5 = -\left(e^{-5} - e^{-1}\right) = \left(e^{-1} - e^{-5}\right) \simeq 0.361.$

(c) $\displaystyle\int_0^t e^{-2x}\,dx = \left[\frac{e^{-2x}}{-2}\right]_0^t = -\frac{1}{2}\left(e^{-2t} - e^{0}\right) = \frac{1}{2}\left(1 - e^{-2t}\right).$

Solution to Activity 11

The range of the exponential distribution has lower limit 0. So, taking f as in Equation (1),

$$\begin{aligned} F(x) &= \int_0^x f(y)\,dy = \int_0^x \lambda e^{-\lambda y}\,dy \\ &= \lambda\left[\frac{e^{-\lambda y}}{-\lambda}\right]_0^x = -\left(e^{-\lambda x} - e^{0}\right) = 1 - e^{-\lambda x}. \end{aligned}$$

This is Equation (2), as required.

Solution to Activity 12

(a) (i)

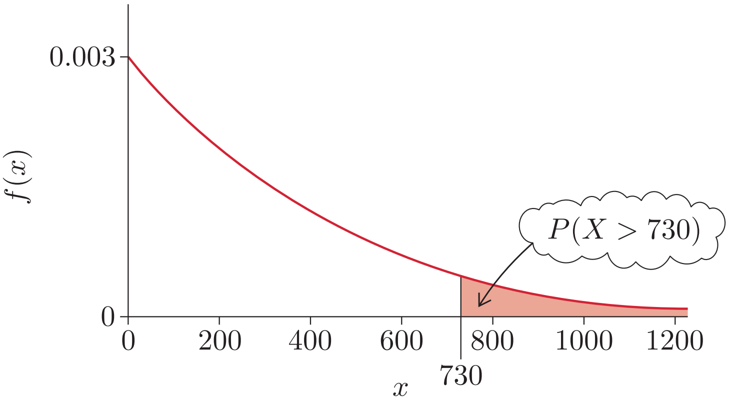


Figure 15 Area corresponding to $P(X > 730)$ on Figure 4

(ii) The probability that the interval between successive serious earthquakes will exceed two years is

$$\begin{aligned}P(X > 730) = 1 - F(730) &= 1 - \left(1 - e^{-(67/22120)\times 730}\right)\\ &\simeq e^{-2.21112} \simeq 0.1096.\end{aligned}$$

This is very close to 0.1092, the value found in Activity 9 using a geometric model.

(b) The probability that the gap between successive earthquakes will be a week (7 days) or less is

$$P(X \leq 7) = F(7) = 1 - e^{-(67/22120)\times 7} \simeq 1 - e^{-0.02120} \simeq 0.0210.$$

(c) The proportion of intervals that exceed the mean is

$$\begin{aligned}P(X > 22120/67) &= 1 - F(22120/67)\\ &= 1 - \left(1 - e^{-(67/22120)\times(22120/67)}\right)\\ &= e^{-1} \simeq 0.3679.\end{aligned}$$

So according to the model, a little over a third of gaps are longer than average.

For the data in Table 6, 27 out of 67, or approximately 0.4030, of intervals between serious earthquakes exceed 330 days. The two proportions are quite close, although the observed proportion is the larger.

Solution to Activity 13

(a)

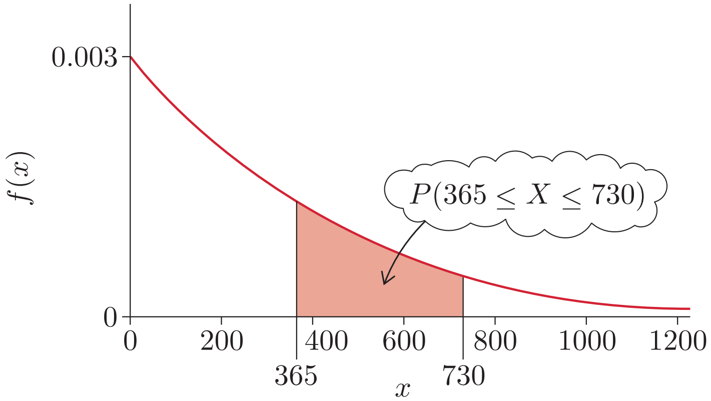


Figure 16 Area corresponding to $P(365 \leq X \leq 730)$ on Figure 4

(b) The probability that the interval between successive serious earthquakes will be between one and two years is

$$\begin{aligned}P(365 \leq X \leq 730) &= e^{-(67/22120)\times 365} - e^{-(67/22120)\times 730}\\ &\simeq e^{-1.10556} - e^{-2.21112} \simeq 0.2214.\end{aligned}$$

(c) The probability that the interval between successive serious earthquakes will be between a week and a month is

$$\begin{aligned}P(7 \leq X \leq 30) &= e^{-(67/22120)\times 7} - e^{-(67/22120)\times 30}\\ &\simeq e^{-0.09087} - e^{-0.02120} \simeq 0.0659.\end{aligned}$$

Solution to Activity 14

If the waiting time between successive events is represented by an exponential random variable X with parameter λ, then the mean waiting time is $1/\lambda$. So the proportion of waiting times that are longer than the mean waiting time is given by

$$P(X > 1/\lambda) = 1 - F(1/\lambda) = 1 - \left(1 - e^{-(1/\lambda)\times\lambda}\right) = e^{-1} \simeq 0.3679.$$

This proportion is independent of the value of the parameter λ. (You saw this value arise in the solution to Activity 12(c) when $\lambda = 67/22120$, but you have now shown that it is the correct proportion whatever the value of λ.)

Solution to Activity 15

In each case X, the number of events in an interval, has a Poisson distribution with parameter λt, where λ is the rate at which events occur per unit time and t is the length of the interval.

(a) In this case, $\lambda = 3$ per minute.

(i) The length of the interval is 5 minutes, so

$$\lambda t = (3 \text{ per minute}) \times (5 \text{ minutes}) = 15.$$

Therefore $X \sim \text{Poisson}(15)$.

(ii) For an interval of a quarter of an hour,

$$\lambda t = (3 \text{ per minute}) \times (15 \text{ minutes}) = 45.$$

So $X \sim \text{Poisson}(45)$.

(b) In this case, $\lambda = 5$ per hour.

(i) For an interval of length $3\frac{1}{2}$ hours,

$$\lambda t = (5 \text{ per hour}) \times \left(3\tfrac{1}{2} \text{ hours}\right) = 17.5.$$

So $X \sim \text{Poisson}(17.5)$.

(ii) For an interval of 15 minutes,

$$\lambda t = (5 \text{ per hour}) \times \left(\tfrac{1}{4} \text{ hour}\right) = 1.25.$$

So $X \sim \text{Poisson}(1.25)$.

Solution to Activity 16

(a) Allowing for two leap years, there are 3652 days in a decade. So X, the number of serious earthquakes in a typical decade, has a Poisson distribution with parameter

$$\lambda t = \left(\frac{67}{22\,120} \text{ per day}\right) \times (3652 \text{ days}) \simeq 11.06167.$$

(b) The probability that there will be exactly five serious earthquakes in a decade is

$$P(X = 5) = \frac{e^{-11.06167}(11.06167)^5}{5!} \simeq 0.0217.$$

(c) The probability that there will be at least five and at most seven serious earthquakes in a decade is

$$\begin{aligned} P(5 \leq X \leq 7) &= P(X=5) + P(X=6) + P(X=7) \\ &= \frac{e^{-11.06167}(11.06167)^5}{5!} \\ &\quad \times \left(1 + \frac{11.06167}{6} + \frac{(11.06167)^2}{7 \times 6}\right) \\ &\simeq 0.1248. \end{aligned}$$

(d) The time T between serious earthquakes has an exponential distribution with parameter $\lambda = 67/22120$, so the probability of a gap of more than five years, or $3652/2 = 1826$ days, is

$$P(T > 1826) = 1 - \left(1 - e^{-(67/22120)\times 1826}\right) \simeq 0.0040.$$

Solution to Activity 17

(a) If X is the number of groundings in a typical year, then X has a Poisson distribution with parameter

$$\lambda t = (3 \text{ per year}) \times (1 \text{ year}) = 3.$$

So

$$\begin{aligned} P(X \geq 2) &= 1 - (P(X=0) + P(X=1)) \\ &= 1 - e^{-3}(1+3) \simeq 0.8009. \end{aligned}$$

(b) If Y is the number of groundings in a two-year period, then Y has a Poisson distribution with parameter

$$\lambda t = (3 \text{ per year}) \times (2 \text{ years}) = 6.$$

So

$$P(Y = 1) = 6e^{-6} \simeq 0.0149.$$

(c) The time T between groundings has an exponential distribution with parameter $\lambda = 3$, so the probability of a gap of more than two years between groundings is

$$P(T > 2) = 1 - F(2) = 1 - \left(1 - e^{-3\times 2}\right) = e^{-6} \simeq 0.0025,$$

or about 1 in 400.

Solution to Activity 18

(a) We need to find m such that

$$F(m) = \frac{m-a}{b-a} = \frac{1}{2}.$$

Multiplying both sides by $b - a$, then adding a to both sides, yields

$$m = a + \tfrac{1}{2}(b-a) = \tfrac{1}{2}(b+a).$$

(b) The median is equal to the mean.

Both the mean and the median are midway between a and b, as shown in Figure 17.

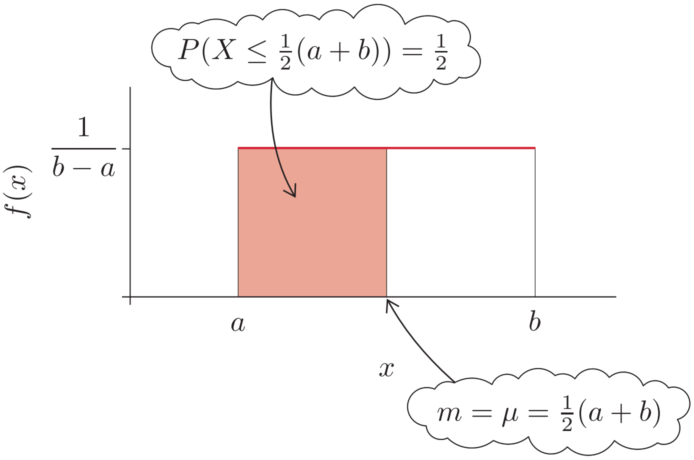


Figure 17 The p.d.f., mean and median of $X \sim U(a,b)$

Solution to Activity 19

(a) The median m satisfies the equation

$$F(m) = 1 - e^{-\lambda m} = \frac{1}{2}.$$

So

$$e^{-\lambda m} = \frac{1}{2}$$

or

$$-\lambda m = \log \frac{1}{2} = -\log 2$$

and hence

$$m = \frac{\log 2}{\lambda} \simeq \frac{0.693}{\lambda}.$$

Here, as throughout this module, the notation log is used to refer to the 'natural' logarithm (the inverse of the exponential function).

(b) The mean of X is $\mu = 1/\lambda$, so

$$m = \log 2 \times \mu \simeq 0.693\mu.$$

So for an exponential random variable, the median is approximately 70% of the mean. This result, which is illustrated in Figure 18 (overleaf), is due to the considerable skewness of the exponential distribution.

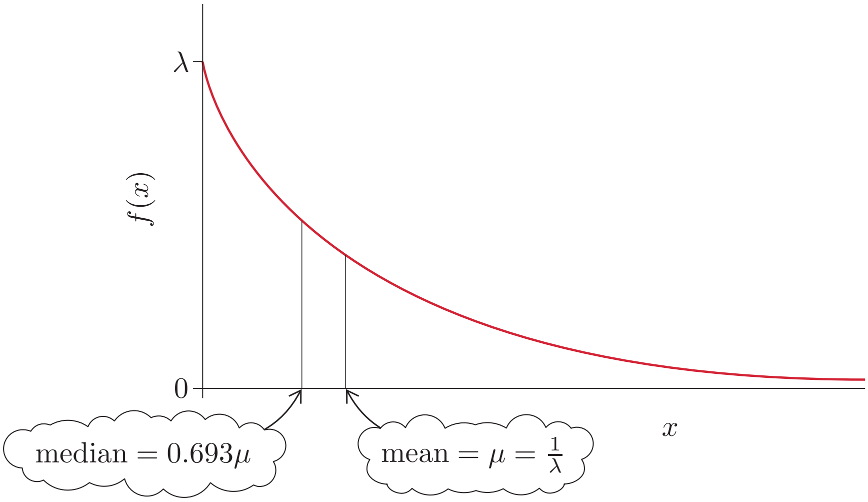


Figure 18 The mean and median of an exponential distribution

Solution to Activity 20

From the solution to Activity 19, the median waiting time m, using the exponential model, is

$$m = \log 2 \times \mu = \log 2 \times 330 \simeq 229 \text{ days.}$$

This is a bit less than the sample median of 257 days.

Solution to Activity 21

(a) The upper quartile, q_U, is the solution of $F(x) = \frac{3}{4}$. That is,

$$F(q_U) = (q_U)^3 = \tfrac{3}{4},$$

and hence

$$q_U = \left(\tfrac{3}{4}\right)^{1/3} \simeq 0.909.$$

(b) The interquartile range of X is

$$q_U - q_L = \left(\tfrac{3}{4}\right)^{1/3} - \left(\tfrac{1}{4}\right)^{1/3} \simeq 0.279.$$

Solution to Activity 22

(a) The c.d.f. of X is

$$F(x) = 1 - e^{-\lambda x}, \quad x > 0.$$

So the lower quartile q_L satisfies

$$F(q_L) = 1 - e^{-\lambda q_L} = \frac{1}{4};$$

that is,

$$e^{-\lambda q_L} = \frac{3}{4},$$

or

$$-\lambda q_L = \log \frac{3}{4},$$

so

$$q_L = -\frac{1}{\lambda}\log\frac{3}{4} \simeq \frac{0.288}{\lambda}.$$

The upper quartile q_U satisfies

$$F(q_U) = 1 - e^{-\lambda q_U} = \frac{3}{4};$$

that is,

$$e^{-\lambda q_U} = \frac{1}{4},$$

so

$$q_U = -\frac{1}{\lambda}\log\frac{1}{4} \simeq \frac{1.386}{\lambda}.$$

(b) The interquartile range of X is

$$\begin{aligned} q_U - q_L &= -\frac{1}{\lambda}\log\frac{1}{4} - \left(-\frac{1}{\lambda}\log\frac{3}{4}\right) \\ &= \frac{1}{\lambda}\left(\log\frac{3}{4} - \log\frac{1}{4}\right) \simeq \frac{1.099}{\lambda}. \end{aligned}$$

(Notice that this is slightly larger than the standard deviation of an exponential distribution, which is equal to $1/\lambda$.)

(c) If we use the results from parts (a) and (b) with $1/\lambda = 330$, then, according to the model, the quartiles (in days) are

$$q_L \simeq 0.288 \times 330 \simeq 95, \quad q_U = 1.386 \times 330 \simeq 457.$$

So the interquartile range is approximately $457 - 95 = 362$ days. The lower quartile given by the model is a bit smaller than the sample lower quartile, and the upper quartile for the model is also a bit smaller than the sample upper quartile. The population and sample interquartile ranges turn out to be similar to one another.

Solution to Activity 23

The 0.05-quantile, $q_{0.05}$, satisfies $F(q_{0.05}) = 0.05$. That is,

$$F(q_{0.05}) = (q_{0.05})^4 = 0.05,$$

and hence

$$q_{0.05} = (0.05)^{1/4} \simeq 0.473.$$

The 0.99-quantile, $q_{0.99}$, satisfies $F(q_{0.99}) = 0.99$. That is,

$$F(q_{0.99}) = (q_{0.99})^4 = 0.99,$$

and hence

$$q_{0.99} = (0.99)^{1/4} \simeq 0.997.$$

Solution to Activity 24

(a) The 0.01-quantile, $q_{0.01}$, of an exponential random variable with parameter λ satisfies

$$F(q_{0.01}) = 1 - e^{-\lambda q_{0.01}} = 0.01.$$

So

$$e^{-\lambda q_{0.01}} = 0.99,$$

leading to

$$-\lambda q_{0.01} = \log 0.99,$$

and hence

$$q_{0.01} = -\frac{1}{\lambda}\log 0.99 \simeq \frac{0.010}{\lambda}.$$

The 0.8-quantile, $q_{0.8}$, satisfies

$$F(q_{0.8}) = 1 - e^{-\lambda q_{0.8}} = 0.8.$$

So

$$e^{-\lambda q_{0.8}} = 0.2,$$

and hence

$$q_{0.8} = -\frac{1}{\lambda}\log 0.2 \simeq \frac{1.609}{\lambda}.$$

(b) The c.d.f. of X, following Figure 3, is shown in Figure 19; $\alpha = 0.8$ is marked, as is the horizontal dashed line connecting it to the graph of F, as well as the vertical dashed line connecting $F(q_{0.8})$ to $q_{0.8} = 1.609/\lambda$.

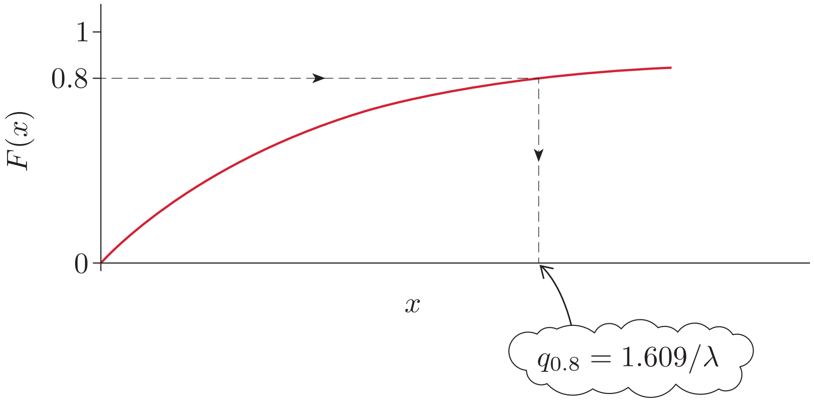


Figure 19 The c.d.f. of $X \sim M(\lambda)$

(c) The p.d.f. of X, following Figure 18, is shown in Figure 20; the 0.8-quantile $q_{0.8} = 1.609/\lambda$ is marked at the correct multiple of the mean $1/\lambda$.

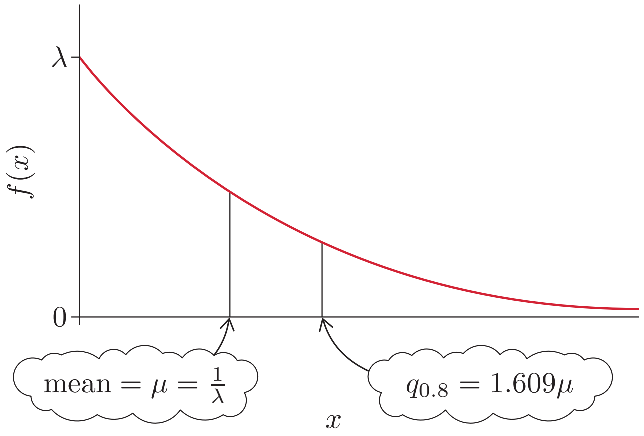


Figure 20 The p.d.f. of $X \sim M(\lambda)$

Solution to Activity 25

From Table 12, we have

$$F(0) = 0.0041 < 0.01 < 0.0410 = F(1),$$

so

$$F(0) < 0.01 \leq F(1),$$

and hence $q_{0.01} = 1$.

Similarly, $F(1) < 0.05 \leq F(2)$, so $q_{0.05} = 2$;

$F(4) < 0.8 \leq F(5)$, so $q_{0.8} = 5$;

and $F(4) < 0.95 \leq F(5)$, so $q_{0.95} = 5$ also.

Solutions to exercises

Solution to Exercise 1

(a) Since the mean number of errors on a page is 3.6, an appropriate approximate model for the number of errors on a page is a Poisson distribution with parameter 3.6.

(b) The Poisson probabilities are given by the formula $P(X = x) = e^{-3.6}(3.6)^x/x!$. They are given alongside the binomial probabilities from Table 4 in the table below.

Table 15 Probabilities of errors

	$P(X = 0)$	$P(X = 1)$	$P(X = 2)$	$P(X = 3)$	$P(X \geq 4)$
$B(320, 0.01125)$	0.0268	0.0975	0.1769	0.2134	0.4854
$B(360, 0.01)$	0.0268	0.0976	0.1769	0.2133	0.4854
Poisson(3.6)	0.0273	0.0984	0.1771	0.2125	0.4847

(c) The probabilities using the Poisson model agree with the binomial probabilities to three decimal places. The approximation is really quite good.

Solution to Exercise 2

(a) The number of people with tritanopia in a group of 100 people has a binomial distribution: $X \sim B(100, 0.0001)$, so

$$P(X = 1) = \binom{100}{1} (0.0001)\,(0.9999)^{99} \simeq 0.0099.$$

(b) The mean number of people with tritanopia in a group of 100 people is $100 \times 0.0001 = 0.01$, so, approximately, $X \sim \text{Poisson}(0.01)$, and

$$P(X = 1) \simeq \frac{e^{-0.01} \times 0.01}{1!} \simeq 0.0099.$$

The Poisson approximation gives the same value as that given by the binomial model to four decimal places.

Solution to Exercise 3

(a) The shape of the histogram is not inconsistent with the data coming from an exponential distribution. In addition, the sample mean and sample standard deviation are fairly close in value. So an exponential model might well be suitable.

(b) The interval between two successive eruptions need not be a whole number of months; it could be any length. So a continuous model is more appropriate here. For this reason, an exponential model is more suitable than a geometric model.

Solution to Exercise 4

(a) Three assumptions are made. First, whether or not an eruption occurs in any particular month is independent of whether an eruption occurs in any other month. Second, the probability that an eruption occurs in any month remains constant from month to month. Finally, only one eruption can occur in any month.

(b) Since the mean of a geometric distribution is equal to $1/p$, an estimate of p is given by the reciprocal of the sample mean. So an estimate of p is

$$p = 1/28.4 \simeq 0.0352.$$

(c) There are 120 months in ten years. So if X is the number of months between successive major explosive eruptions, then the probability required is

$$P(X > 120) = \left(1 - \frac{1}{28.4}\right)^{120} \simeq 0.0136,$$

or approximately 1 in 74. Looking at the histogram in Figure 5, you can see that two out of the 55 gaps exceeded ten years. This is in line with the $0.0136 \times 120 \simeq 1.6$ suggested by the model.

Solution to Exercise 5

(a) Since the mean time between major explosive eruptions is 28.4 months, an estimate of the parameter λ is

$$\lambda = \frac{1}{28.4} \simeq 0.0352.$$

(b) According to the exponential model, the probability that X, the gap between major explosive eruptions, exceeds ten years is given by

$$P(X > 120) = 1 - F(120) = e^{-120/28.4} \simeq 0.0146,$$

or about 1 in 68. This probability is slightly greater than that obtained using the geometric model.

(c) According to the exponential model, the proportion of gaps between major explosive eruptions that are between 1 and 5 years in length, that is, between 12 and 60 months, is (using Equation (4))

$$\begin{aligned} P(12 < X < 60) &= e^{-12/28.4} - e^{-60/28.4} \\ &\simeq 0.5345, \end{aligned}$$

or a little more than a half.

Solution to Exercise 6

(a) If X represents the number of eruptions that occur in a year, then X has a Poisson distribution with parameter $\lambda t = 0.0352 \times 12 = 0.4224$.

(b) (i) The probability that exactly one eruption occurs in a year is given by

$$P(X = 1) = 0.4224e^{-0.4224} \simeq 0.2769.$$

(ii) The probability that there is more than one eruption in a year is given by

$$\begin{aligned} P(X > 1) &= 1 - (P(X = 0) + P(X = 1)) \\ &= 1 - e^{-0.4224}(1 + 0.4224) \simeq 0.0677. \end{aligned}$$

(c) If T is the waiting time in months between eruptions, then T has an exponential distribution with parameter $\lambda = 0.0352$.

(d) The probability that the gap between successive eruptions will exceed five years, which is 60 months, is given by

$$\begin{aligned} P(T > 60) &= 1 - F(60) = 1 - \left(1 - e^{-0.0352\times 60}\right) \\ &= e^{-0.0352\times 60} \simeq 0.1210. \end{aligned}$$

Solution to Exercise 7

(a) If X is the number of telephone orders received in ten minutes, or 1/6 hour, then X has a Poisson distribution with parameter

$$\lambda t = 15 \times \frac{1}{6} = 2.5.$$

(b) (i) The probability that exactly three orders are received in ten minutes is given by

$$P(X = 3) = \frac{e^{-2.5}(2.5)^3}{3!} \simeq 0.2138.$$

(ii) The probability that at most two telephone orders are received in ten minutes is

$$\begin{aligned} P(X \leq 2) &= P(X = 0) + P(X = 1) + P(X = 2) \\ &= e^{-2.5}\left(1 + 2.5 + \frac{(2.5)^2}{2}\right) \simeq 0.5438. \end{aligned}$$

(c) If T is the waiting time in hours between telephone orders, then T has an exponential distribution with parameter $\lambda = 15$.

(d) The probability that the gap between successive telephone orders will exceed five minutes, or 1/12 hour, is given by

$$\begin{aligned} P\left(T > \tfrac{1}{12}\right) &= 1 - F\left(\tfrac{1}{12}\right) = 1 - \left(1 - e^{-15\times 1/12}\right) \\ &= e^{-15\times 1/12} = e^{-1.25} \simeq 0.2865. \end{aligned}$$

Solution to Exercise 8

First, we need to find q_L such that

$$F(q_L) = \frac{q_L - a}{b - a} = \frac{1}{4}.$$

Multiplying both sides by $b - a$, then adding a to both sides, yields

$$q_L = a + \frac{1}{4}(b - a) = \frac{1}{4}(b + 3a).$$

Similarly,

$$F(q_U) = \frac{q_U - a}{b - a} = \frac{3}{4}$$

leads to

$$q_U = a + \frac{3}{4}(b - a) = \frac{1}{4}(3b + a).$$

The interquartile range is therefore

$$\begin{aligned} q_U - q_L &= \frac{1}{4}(3b + a - b - 3a) = \frac{1}{4}(2b - 2a) \\ &= \frac{1}{2}(b - a). \end{aligned}$$

Solution to Exercise 9

(a) The median m is the solution of $F(y) = \frac{1}{2}$. So

$$F(m) = \tfrac{1}{4}m^2 = \tfrac{1}{2},$$

and hence $m^2 = 2$, so $m = \sqrt{2} \simeq 1.414$.

The lower quartile, q_L, is the solution of $F(y) = \frac{1}{4}$, so

$$F(q_L) = \tfrac{1}{4}(q_L)^2 = \tfrac{1}{4},$$

and hence $q_L = 1$.

The upper quartile, q_U, is the solution of $F(y) = \frac{3}{4}$, so

$$F(q_U) = \tfrac{1}{4}(q_U)^2 = \tfrac{3}{4},$$

and hence $(q_U)^2 = 3$ and $q_U = \sqrt{3} \simeq 1.732$.

Therefore the interquartile range is

$$q_U - q_L \simeq 1.732 - 1 = 0.732.$$

(b) The values of $\alpha = \frac{1}{2}, \frac{1}{4}$ and $\frac{3}{4}$ and the corresponding values of m, q_L and q_U are shown in Figure 21.

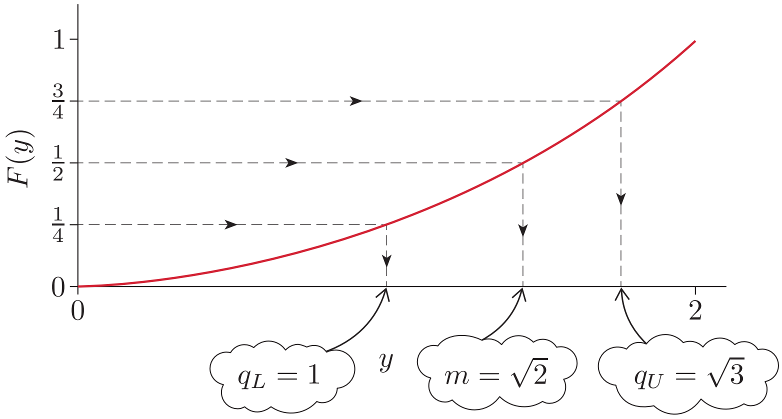


Figure 21 The c.d.f. $F(y) = \frac{1}{4}y^2$ showing m, q_L and q_U

Solution to Exercise 10

(a) The α-quantile q_α is the solution of $F(z) = \alpha$. So

$$F(q_\alpha) = (q_\alpha)^\beta = \alpha,$$

and hence $q_\alpha = \alpha^{1/\beta}$.

(b) The median, lower quartile and 0.8-quantile are the α-quantiles when $\alpha = 0.5$, 0.25 and 0.8, respectively. They are therefore given by $m = (0.5)^{1/\beta}$, $q_L = (0.25)^{1/\beta}$ and $q_{0.8} = (0.8)^{1/\beta}$.

(c) When $\beta = \frac{1}{2}$, $q_{0.8} = (0.8)^2 = 0.64$.

Solution to Exercise 11

(a) The c.d.f. of X is given in the table below.

Table 16

x	1	2	3	4	5	6	7	8	9	10
$F(x)$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1

(b) The lower quartile is the smallest value x in the range of X such that $F(x) \geq 0.25$. From Table 16, $F(2) = 0.2$, which is less than 0.25, and $F(3) = 0.3$, which is greater than 0.25, so $F(2) < 0.25 \leq F(3)$, and hence the lower quartile q_L is 3.

Similarly, $F(4) < 0.5 \leq F(5)$, so the median is 5. And $F(7) < 0.75 \leq F(8)$, so the upper quartile is 8.

Acknowledgements

Grateful acknowledgement is made to the following sources:

Page 267: © Phitha Tanpairoj / www.Shutterstock.com

Page 268: Taken from: http://www.abebooks.co.uk/9780330377171/Richards-21st-Century-Bicycle-Book-0330377175/plp

Page 269: © Photograph Credit: Richard Peyton / thegalwaycow.com

Page 272: © Lutfar Rahman Nirjhar / https://en.wikipedia.org/wiki/File:Solar_eclipse_22_July_2009_taken_by_Lutfar_Rahman_Nirjhar_from_Bangladesh.jpg This file is licensed under the Creative Commons Attribution Licence http://creativecommons.org/licenses/by/3.0/

Page 273: © inus grobler / www.istockphoto.com

Page 274 top: © monkeybusinessimages / www.istockphoto.com

Page 274 bottom: Taken from: http://www.goodreads.com/book/show/7160867-a-steroid-hit-the-earth

Page 278: © Reproduced by permission from Karen Madsen, WT Townshend

Page 279: © somjin klong-ugkara / www.123rf.com

Page 282: © Punch Limited

Page 287: Taken from: https://en.wikipedia.org/wiki/File:Lava_forms.jpg This file is licensed under the Creative Commons Attribution Licence http://creativecommons.org/licenses/by/3.0/

Page 291 top: Taken from: http://www.ack-shop.ru/shop/telefoniya/sistemnye_telefony/ericsson_lg_digital_phone_panel_ldp_48_button_dss_black_color/

Page 291 bottom: © Joseph A. Carr

Page 295: Dr. Ian G. Stimpson / http://hypocentral.com/blog/2011/03/11/japan-earthquake-11032011-recorded-at-keele This file is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike Licence http://creativecommons.org/licenses/by-nc-sa/3.0/

Page 297: This file is licensed under the Creative Commons Attribution-ShareAlike Licence http://creativecommons.org/licenses/by-sa/3.0/

Page 298: © irin717 / www.istockphoto.com

Page 301: © Wavebreakmedia Ltd / www.Dreamstime.com

Page 307: © Paul Wishart / www.123rf.com

Page 310: © iloveotto / www.123rf.com

Page 286: Felix O, Flickr / This file is licensed under the Creative Commons Attribution-ShareAlike Licence
http://creativecommons.org/licenses/by-sa/3.0/

Every effort has been made to contact copyright holders. If any have been inadvertently overlooked, the publishers will be pleased to make the necessary arrangements at the first opportunity.