5.

Random sampling

- Sampling
- How big should our sample be?
- How should we select our sample?
- Random sampling
- Generating random numbers
- Stratified sampling
- Other forms of sampling
- Capture-recapture
- Simulations
- Random number generating from other distributions
- More simulations
- Variability of random samples
- Miscellaneous exercise five

Situation One

Taking a sample

Doctors may at times need an analysis of our blood. Fortunately they can do this by taking a sample, not the whole lot!

Whilst blood samples are commonly taken from the veins around the elbow this is not always the case.

Sometimes they are taken from the wrist,
or perhaps by a pin prick to the thumb or finger,
or maybe a heel prick.

Research each of these and write a few sentences about why each location is chosen and what in particular the blood sample may be used for.

Blood samples are not the only samples members of the medical profession may request of us. Research and write a few sentences about each of the following:

Amniocentesis.
Cerebral spinal fluid sample.
Sweat testing.
Mid-stream urine sample.

Shutterstock.com/Jmiks

Situation Two

The following situation involves a technique sometimes referred to as

Capture – Recapture

This technique has been used for many years to estimate the populations of species of animals.

To estimate the population of a certain species of frog in a particular area a scientist caught and 'tagged' 40 of the frogs and then released them back into the area. Some days later a random catch of 50 frogs of this species from the area found 4 that were tagged from the earlier capture.

Estimate the number of frogs of this species in the area.

imagefolk/Zoonar/K Bain

ISBN 9780170395205

Sampling

The situations on the previous page involved **samples** being taken from a larger **population**.

By *population* we mean the whole amount under consideration – e.g. **all** of your blood, or **all** of the frogs in a particular area.

By *sample* we mean the subgroup of the population that we will use to make inferences about the whole population.

If we take some numerical measure for the sample, perhaps red blood cell count or average length of a frog, we could then use this **sample statistic** to estimate the equivalent measure for the population. Numerical characteristics about an entire population, for example, the average age of all Australians, the mean length of all of the frogs in a region, etc, are called **population parameters**. If the sample statistic is going to give a good indication of the equivalent population parameter it is important that the sample is typical of the population. Two aspects in particular that we need to consider are:

- How big should our sample be?
  The larger our sample the more confident we can be that any information collected about the sample will be indicative of the same information about the population.
- How should we select our sample?
  It is important that the sample is a fair reflection of the makeup of the population and as free from unwanted bias as possible.

How big should our sample be?

Suppose a farmer has 1000 sheep and wants to select a random sample for testing to monitor the likely wool and meat quality of his stock.

How many sheep should he include in his sample?

How many he should choose depends on how confident he wants to feel that the results from his sample fairly reflect the characteristics of the population of 1000. A small sample of just 5 or 6 animals for example is unlikely to give him this confidence.

Getty Images/EyeEm/Maria Greenwood

One quite common 'rule of thumb' is that, when reasonable-sized populations are involved, always choose at least 30. Less than 30 will tend to leave considerable doubt as to whether data obtained from our samples will fairly reflect the whole population.

Should he choose more than 30?

Certainly the more he has in the random sample the more confident he can be that the characteristics of the sample reflect those of the population as a whole. However he must balance this desire to be confident that the results reflect the population with other aspects, two of which are mentioned on the next page.

Note: A 'rule of thumb' is a general rule often based more on experience than precise calculation.

ISBN 9780170395205

- Just how confident does he need to be that the results from his sample will reflect the population as a whole?
  How crucial is it that any data from the sample is a good reflection of the population? Is it just to have a rough idea of the standard of the wool and meat of his sheep or is it perhaps to check for something more serious? If he was worried that some of his sheep were carrying a particularly harmful and transmittable disease he may only gain peace of mind about his animals if he has them all tested – i.e. not a sample at all.
- How much will it cost him to have each animal in the sample tested?
  If the test is expensive he may be more inclined to select a small sample and accept the uncertainty about how typical the sample is. (Though there will come a point where the sample involves so few sheep that the high level of uncertainty will mean that there was little point having any tested at all.)

How should we select our sample?

There are various methods for selecting a sample so as to reduce the likelihood of unwanted bias. Consideration of a number of these methods follows.

Random sampling

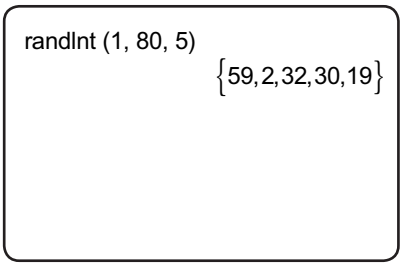
The important aspect of *random* sampling is that each member of the population has an equal chance of being chosen.

Thus to obtain a random sample from a population we could:

- Assign each member of the population a number.
  If people are involved this could simply be a list of names with each name given a number. If an area is involved, e.g. when investigating an area of grass for weed content, divide the area into equal size squares and number each square.
- Select numbers from the desired range of numbers using a random process, e.g. randomly selecting numbered balls from a bag, using random numbers from a random number generator (see below).

Generating random numbers

Some calculators can randomly generate numbers in a particular range. The display on the right for example shows 5 integers from 1 to 80 generated using the random number facility on a calculator.

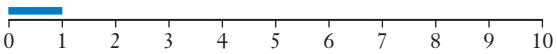


Other calculators may generate random numbers in a particular preset range, e.g. between 0 and 1. However these too can be instructed to output integer values in a desired range, as explained on the next page.

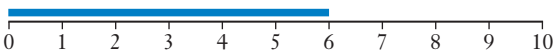
ISBN 9780170395205

To change from a random number with an output in the range 0 to 1 to integers from the set {1, 2, 3, 4, 5, 6}:

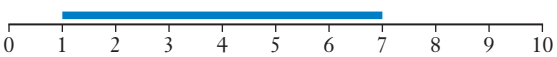
Usual output is between 0 and 1:



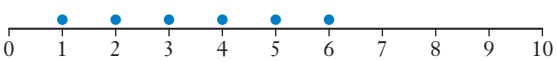
Multiply by 6 to obtain numbers between 0 and 6:



Add 1 to obtain numbers between 1 and 7:



Displaying only the integer part of such numbers will give the integers 1, 2, 3, 4, 5 or 6.
(See the display below.)

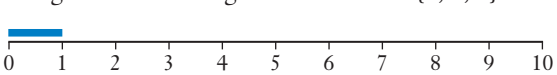


Int(Ran# × 6 + 1)

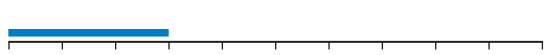
5
3
4
3
6
1

To change from a random number with an output in the range 0 to 1 to integers from the set {7, 8, 9}:

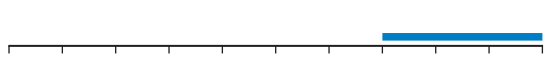
Usual output is between 0 and 1:



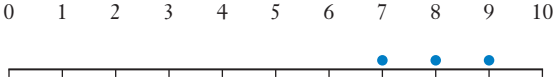
Multiply by 3 to obtain numbers between 0 and 3:



Add 7 to obtain numbers between 7 and 10:



Displaying only the integer part of such numbers will give the integers 7, 8 or 9.
(See the display below.)



Int(Ran# × 3 + 7)

9
7
9
7
8
8

Use your calculator to randomly generate 5 different integers from 1 to 80.
Will you obtain the same five numbers as those shown in the display on the previous page?

ISBN 9780170395205

Stratified sampling

In stratified sampling the population is divided up into layers, or strata, and then samples are randomly selected from each strata. For example, suppose we require a random group of 60 students from a school containing year 7 students to year 12 students. The strata could be the year levels and we then randomly select ten students from each of the 6 years.

This stratified sampling is often selected proportionally to make it more representative of the population. For example, suppose the school just mentioned had 1221 students distributed as follows:

221 in year 7,	240 in year 8,	248 in year 9,
285 in year 10,	124 in year 11,	103 in year 12.

If we want a proportional stratified sample of 60 students we choose as follows:

$\frac{221}{1221} \times 60 \approx 10.9$	$\frac{240}{1221} \times 60 \approx 11.8$	$\frac{248}{1221} \times 60 \approx 12.2$
11 year 7 students.	12 year 8 students.	12 year 9 students.
$\frac{285}{1221} \times 60 \approx 14.0$	$\frac{124}{1221} \times 60 \approx 6.1$	$\frac{103}{1221} \times 60 \approx 5.06$
14 year 10 students.	6 year 11 students.	5 year 12 students.

Other forms of sampling

Another sampling technique involves first listing the population in some order. If the sample size requires, say a 1 in 25 selection, a number between 1 and 25 is randomly selected and then every 25th person is selected after that. Thus if person number 17 is selected first then numbers 42, 67, 92, 117, …, are selected as well. This can be referred to as **systematic sampling** or **array sampling**. This method is unsuitable in cases where there is some periodic feature in the population list. For example if we were sampling components made by a machine and the work load on the machine caused it to make every 20th item faulty, (i.e., the 20th, the 40th the 60th, etc.) the above systematic selection, i.e. 17, 42, 67, … would not feature any of these defective components.

In some sampling, perhaps in an attempt to achieve stratified sampling overall, individual interviewers are given a particular number of people they must interview of various types. For example, they may be required to interview 20 people of which 5 are men aged in their twenties, 8 are married females aged over 50 and 7 are males aged over 60. This is called **quota sampling**.

Convenience sampling, as the name suggests, is when the sample is chosen because it is convenient. For example, if we wanted to collect data about primary school students the local primary school would be a convenient school to choose. Convenience sampling is sometimes used as a preliminary investigation of the situation. It is likely to be an inexpensive option compared to some others but can give some early direction to later, more involved and detailed data collection using more sophisticated sampling methods.

Television 'phone-in' surveys rely on people volunteering their opinion by phoning in. We would not expect the sample to be particularly random and the balance of the opinions expressed may be far from representative of those held by the population as a whole. This is an example of **volunteer sampling**. Those taking part select themselves to be part of the sample. This method is also called **self-selection sampling**.

ISBN 9780170395205

Capture-recapture

Situation Two at the start of this chapter involved *capture-recapture*, a method that uses **sampling** of a population to estimate the size of the population. In both the initial capture, and again in the recapture, a sample of the whole population is taken.

In the given situation 40 frogs from an area were caught, tagged and then released back into the area. Upon their release the proportion of tagged frogs in the area is

$$\frac{40}{\text{Total number of frogs in the area}}.$$

The recapture process caught 50 frogs, of which 4 were found to be ones tagged in the earlier capture. This suggests that the proportion of tagged frogs in the area is

$$\frac{4}{50}.$$

If this second sample reasonably reflects the proportion in the whole population then

$$\frac{4}{50} \approx \frac{40}{\text{Total number of frogs in the area}}.$$

Solving this equation allows the population of frogs in the area to estimated.

Can we use capture-recapture for counting humans?

Your initial reaction to the above question might be

'Of course not! We cannot capture, tag and then release humans!'

Well in fact the capture-recapture process is used to count human populations but we do not really capture and tag people in the same way as we might do with frogs, birds or fish. With humans the first 'capture and tag' involves seeing how many of the group under consideration appear on one list and the 'recapture' is to see how many of those on this first list appear on a second list. For example:

Suppose scientists in a country where only incomplete medical records were kept of the population, wanted to estimate how many people in the capital city had suffered the amputation of a limb. Suppose that a list compiled from a number of the larger hospitals gave the identities of 150 individuals who had experienced such surgery. (These 150 amputees form the initial 'captured and tagged' group.) Hence if in the entire city there were n amputees the proportion of 'tagged' ones (i.e. the proportion appearing on the hospital listing) would be: $\frac{150}{n}$.

Now suppose we check the list of amputees attending a centre that specialises in the fitting of artificial limbs. Let us suppose that this list involved 78 people of whom 23 were also on out hospital listing (i.e. 23 of the 78 were 'tagged' from the first capture). This would indicate that the proportion of 'tagged' individuals in the entire amputee population is: $\frac{23}{78}$.

Thus if the 2nd sample is representative of the amputee population $\frac{23}{78} \approx \frac{150}{n}$,

giving an estimate of the total number in the capital city who have experienced limb amputation as approximately 510.

Note: A survey of this type was carried out for Rio de Janeiro in Brazil by Spichler at al and was published in the Pan American Journal of Public Health, 2001.

ISBN 9780170395205

Exercise 5A

1 For each of the following state whether the sample is 'likely to introduce bias' or 'not likely to introduce bias'.

- **a** People parking their cars at a car park are asked 'Do you think bus travel is good value?'
- **b** People eating at *Speedy Annes* restaurant are asked 'How many times per week do you eat at a restaurant?'
- **c** Every fourth person on a school's alphabetical roll of students is asked 'How many times do you use the school canteen in a week?'
- **d** The colours of 2000 cars on a freeway are noted in an attempt to determine the most common colour of car.
- **e** The heights of all the year eights in an Australian school of 1800 pupils were noted to give an indication of the average heights of Australian year eights.

2 The 497 members of a sports club comprise 100 in the under 20 age range, 154 in their twenties, 175 in their thirties and 68 aged 40 or over.

If the committee is to consist of 10 members and the age balance on the committee is to reflect the age balance of the membership, how many of each of the four age ranges should the committee consist of?

3 **a** Use the random number generator on a calculator or computer to simulate 10 rolls of a normal six-sided die. Tabulate your results and also determine the mean score from the 10 rolls and compare your results to those of others in your class.

b Repeat the above for 20 rolls, 30 rolls, 50 rolls and 100 rolls, each time comparing your results to those of others in your class.

4 A stratified sample of 80 students is to be selected from the year 8 to 12 student population of a school. If the school student population for these years consists of 420 in year eight, 407 in year nine, 389 in year ten, 270 in year eleven and 258 in year twelve how many of each year should be in the sample for it to reflect the proportion in each year group.

5 To estimate the population of a certain species of frog in a particular area a scientist caught and 'tagged' 34 of the frogs and then released them back into the area. Some days later a random catch of 28 frogs of this species from the area found 5 that were tagged from the earlier capture.

Estimate the number of frogs of this species in the area.

6 In an attempt to estimate how many fish were in a particular lake, 64 fish from the lake were netted, tagged and released back into the lake to mix with the rest of the population.

A 'recapture' carried out one week later netted 83 fish and, of these, 7 were found to carry tags indicating they were in the first netting.

According to these figures, estimate the number of fish in the lake at this time.

Auscape/Reg Morrison

ISBN 9780170395205

7 To estimate the numbers of a particular species of bird visiting a favoured breeding ground at breeding time, scientists catch and tag 123 of the birds and then release then back into the population in the breeding ground.

A second capture catches 154 of the birds of which 6 showed the tag that indicated they were in the first catch too.

Use these figures to estimate the number of these birds in the area.

Shutterstock.com/Gerald A. DeBoer

8 Shane used the random number generator on his calculator to simulate the rolling of a normal, fair, six-sided die 12 times.

Similarly, Christine simulated the rolling of a normal, fair, six-sided die 150 times.

The results obtained by one of these students had a mean of 4.08 (maybe rounded).

The results of the other student had a mean of 3.42 (maybe rounded).

Which mean value is likely to belong to which student? Explain your reasoning.

9 Portia used the random number generator on her calculator to simulate the rolling of two normal, fair, six-sided die 12 times, each time noting the sum of the two numbers obtained.

Similarly, Horace simulated the rolling of two normal, fair, six-sided die 150 times, each time noting the sum of the two numbers obtained.

The results obtained by one of these students had a mean of 7.17 (maybe rounded).

The results of the other student had a mean of 6.42 (maybe rounded).

Which mean value is likely to belong to which student? Explain your reasoning.

10 To estimate the number of people missed from the population census carried out in a region, statisticians targeted a particular subsection of the region that they believed to be typical of the region as a whole, interviewed everyone in the subsection, and then checked how many of this targeted group featured on the census.

They found that: 1 235 067 people completed a census form.

1345 people were in the subsection and of these 1338 people had completed a census form.

Based on these figures suggest an approximation for the number in the whole region who did not complete a census form.

11 As part of an attempt to estimate the population of long-necked turtles living in a particular lakeland area, scientists intend to catch and tag a number of the turtles from the area and then release them back into the area. Explain how this can be used to estimate the population, including in your explanation what the process involves, why it works and what possible sources of error might need to be considered and, if possible, avoided?

Getty Images/Ted Mead

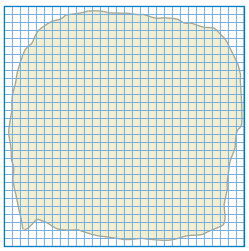
What is a **census**?

The random numbers generated by a calculator or computer spreadsheet are sometimes referred to as **Pseudo-random numbers**. Why is this? Do some research to find out why.

ISBN 9780170395205

Counting seals

In order to estimate the number of seals on a particular island favoured by seals, a number of aerial photographs are to be taken of the island.



A '30 × 30' grid of 900 squares is placed over a map of the entire island with each square covering 10 000 m^2 in real life (i.e. 100 m × 100 m).

220 of these squares were at the edge of the island and showed some sea or all sea, all the others were entirely covering land.

The plan was to randomly select some of the squares showing only land, to photograph these areas and to use the photographs to count the number of seals present in each of the selected squares.

Knowing that the island had a total area of 7.34 km^2 an estimate of the population of seals on the island could be made.

- Explain how we could 'randomly select' the squares to be photographed and suggest how many squares should be selected.
- Will your random selection guarantee that the sample is an accurate representation of the population of seals on the island at the time the photographs were taken? Explain.
- How could the 'seal counts' from the selected photographs be used to estimate the seal population on the island at the time the photographs were taken?
- Suggest any improvements that could be made to the plan.

Dreamstime/Johncarnemolla

ISBN 9780170395205

Simulations

Some questions in the previous exercise mentioned the idea of using a random number generator to **simulate** the rolling of a die.

Just as sampling allows us to 'get a feel for' the characteristics of a population without actually surveying every member of the population, then so running a simulation allows us to 'get a feel for' how a situation might evolve without actually running the real situation – just as a flight simulator attempts to recreate for a trainee pilot the conditions and experiences they will encounter when flying a real aircraft, without actually flying a real plane. Computer games try to make the player feel the excitement of playing a game of soccer, golf, cricket, etc without playing the real game.

A simulation of something attempts to resemble or mimic the real thing without actually being the real thing.

In mathematics we may run a simulation to collect data about some event without actually carrying out the real event. If our simulation is a good imitation of the real thing then the data collected from the simulation may help us predict what might happen in the real thing. If we need our simulation to involve some randomly occurring event we can simulate the outcome with the toss of a coin, the roll of a die or the ability of some calculators and computer spreadsheets to generate random numbers.

Simulation I: Overbooking

An airline company finds that, due to late cancellations, flights often take off with empty seats.

iStock.com/bkindler

The company considers deliberately overbooking on flights so that the cancellations will bring the flight back to capacity and reduce the number of empty seats. Should insufficient cancellations occur on any overbooked flight any passengers who have to miss the flight will be offered an alternative flight plus a refund. One particular route has a plane with a capacity of just 25 passengers. Past records indicate that on average, on this route, one in every 15 customers cancels late. The company wants to investigate the likely consequences of booking 26 passengers for the flight.

We can simulate the situation if we use a calculator or spreadsheet to randomly generate 26 integers taken from the integers 1 to 15, and take the number 15 to indicate a passenger who cancels. Run such a simulation at least ten times and see how many will result in an overbooking problem.

Of course it might be the case that on some flights there would be less than 25 booking anyway. Suppose instead we expect bookings to be in the range 15 to 27 (i.e. up to 2 overbooked) and the cancellation rate is anything from 0% to 10%.

Consider how you might run a simulation to investigate this situation.

ISBN 9780170395205

Simulation II: Spread of illness

If one student in a class of twenty-five students has an infectious illness and comes to school how many others in the class are likely to catch it?

Let us suppose that the class sits in the five-by-five layout shown below:

Let us further suppose that if one student has the illness the probability of one of his or her immediate neighbours catching the illness is one-sixth for each neighbour.

We can simulate the spread of disease in the class using a normal die.

Suppose the student in the middle is the one initially having the illness. This student is the 'active' spreader of the illness to immediate neighbours. (See diagram below left.)

This student has 8 immediate neighbours.

Starting with the neighbour 'below' the initial carrier we roll a normal die and, if a six results, we assume the student catches the illness. We then work clockwise around the 8 neighbours.

The eight rolls: 3, 6, 2, 3, 5, 5, 6, 1 shown in the middle situation leads to the situation below right – three people infected and two 'active' spreaders of the illness whose immediate neighbours now need to be considered.

If this process continues will the entire class become infected? If not how many will? (Assume an infected person does not become active a second time.)

Carry out the simulation a number of times and report on your findings.

You might also like to consider the following:

- Suppose the student initially having the illness is not the middle student?
- Suppose the infection rate is something other that one-sixth? Perhaps one-tenth or one-half or …?

ISBN 9780170395205

Random number generating from other distributions

Suppose we wanted to randomly select five fit adult males aged between 20 and 40, and note their systolic blood pressure. If we were to know that this blood pressure reading for almost all fit adult males aged between 20 and 40 lies in the range 95 mm Hg to 135 mm Hg, we could ask our calculator or computer spreadsheet to generate five random numbers in the range 95 to 135:

RndFix(Ran# × 40 + 95,2)
106.25
113.38
121.37
125.24
99.18

However, such a list would give numbers taken from a uniform distribution. The histogram on the next page shows how the the 150 random numbers listed below tend to demonstrate this uniformity.

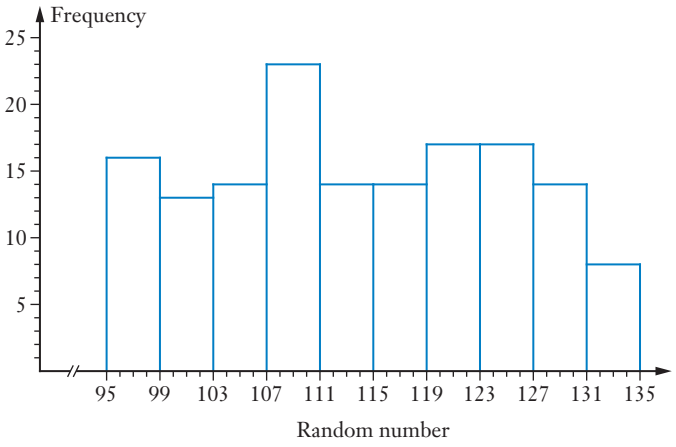
104.51	114.53	109.87	134.79	118.51	132.97	96.35	112.51	108.65	95.95
119.07	113.01	119.7	98.18	122.45	117.73	99.46	132.85	124.45	132.3
102.3	110.43	96.17	132.97	121.19	104.96	98.25	123.73	123.59	101.28
121.27	100.52	124.9	125.47	107.13	97.63	113.06	105.94	130.74	124.39
117.61	122.33	106.05	97.36	115.34	106.3	110.06	122.14	115.86	112.51
97.21	98.65	116.61	109.98	131.75	120.99	128.93	104.34	126.62	104.07
114.07	110.91	129.82	110.57	128.44	110.7	100.98	111.84	110.06	120.2
128.16	114.84	119.55	125.87	95.83	118.18	120.95	117.23	107.93	108.68
130.23	108.66	95.48	108.5	107.91	116.17	105.76	130.74	118.66	105
107.82	95.7	108.34	100.54	133.05	103.72	117.71	118.39	129.04	130.18
120.1	127.89	107.45	128.73	115.64	111.52	111.04	119.93	108.02	128.29
134.13	124.05	98.29	112.2	108.06	113.63	98.21	102.22	115.88	120.42
125.8	110.25	122.02	104.43	110.4	125.95	102.04	103.74	126.44	101.66
112.05	123.47	100.81	125.16	114.17	120.74	100.54	102.31	123.15	95.69
129.28	101.69	103.37	126.59	110.62	106.6	125.1	96.05	127.94	122.17

Number (x)	$95 \le x < 99$	$99 \le x < 103$	$103 \le x < 107$	$107 \le x < 111$	$111 \le x < 115$
Frequency	16	13	14	23	14

Number (x)	$115 \le x < 119$	$119 \le x < 123$	$123 \le x < 127$	$127 \le x < 131$	$131 \le x < 135$
Frequency	14	17	17	14	8

ISBN 9780170395205

However, with blood pressure tending to be normally distributed it would be better if our sample still involved each member of the appropriate population having an equal chance of being selected but, with more people having a blood pressure close to the mean, we need our selection to be more likely to give a random number close to the mean. We can achieve this if our random selection is made from a normal distribution.



Let us suppose that for the population under consideration, systolic blood pressure is normally distributed with mean 115 mm Hg and standard deviation 6 mm Hg.

Some calculators, computer spreadsheet programs and interactive websites can generate random numbers from a normal distribution.

Explore the capability of your calculator or computer spreadsheet in this regard.

randNorm(6,115,5)
116.9482723
113.7535383
109.1955157
112.4588398
119.1266192

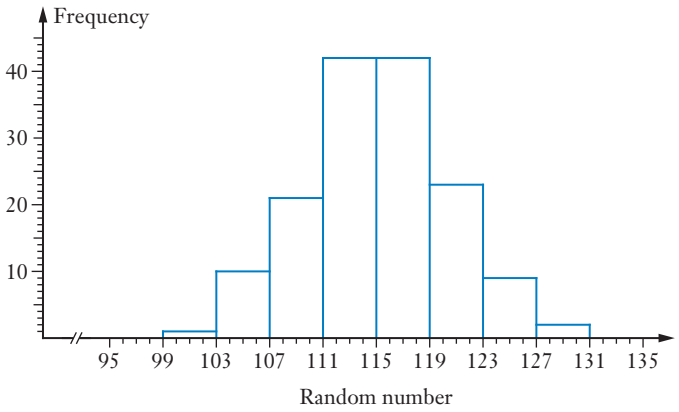
The 150 numbers below have been randomly generated from a normal distribution with a mean of 115 and a standard deviation 6. Notice on the next page how the characteristic bell shaped curve is evident on the histogram.

111.9	115.25	104.81	118.62	120.84	126.37	118.56	111.76	123.68	113.57
109.26	119.48	114.52	118.12	112.29	111.5	112.06	113.73	117.54	115.84
122.34	115.45	118.22	107.86	121.95	110.74	128.94	123.25	115.22	122.9
124.96	120.05	111.7	110.67	116.02	116.02	118.31	120.49	111.63	111.98
106.76	115.09	118.94	108.17	112.49	112.02	113.72	120.02	114.7	117.67
119.92	113.83	122.02	117.85	110.13	115.94	115.64	111.35	106.46	121.06
115.93	118.26	109.34	117.61	106.37	118.61	126.94	107.73	118.42	120.01
106.48	112.76	117.17	107.6	108.96	114.01	119.38	114.76	109.11	118.36
110.85	108.49	119.11	113.69	112.55	120.18	109.01	104.66	111.04	119.8
112.59	117.84	117.26	112.44	111.3	106.78	113.24	120.05	111.23	119.5
116.98	123.68	118.97	116.14	110.89	106.31	113.37	110.4	114.03	112.33
111.87	102.84	124.09	121.17	113.7	115.92	111.83	114.86	112.38	116.82
106.49	122.2	116.04	105.36	110.77	117.33	123.75	113.73	114.37	112.96
110.34	117.25	115.14	111.71	117.57	113.67	119.45	108.63	109.53	118.28
123.21	122.06	115.08	117.1	117.4	120.54	130	113.73	108.78	118.39

ISBN 9780170395205

Number (x)	$95 \le x < 99$	$99 \le x < 103$	$103 \le x < 107$	$107 \le x < 111$	$111 \le x < 115$
Frequency	0	1	10	21	42

Number (x)	$115 \le x < 119$	$119 \le x < 123$	$123 \le x < 127$	$127 \le x < 131$	$131 \le x < 135$
Frequency	42	23	9	2	0



More simulations

Earlier we considered two simulations:

Simulation I: Overbooking.
Simulation II: Spread of illness.

We will now consider two more but now our generation of random numbers can come from non-uniform distributions.

Simulation III: Investing funds

A person wishes to invest \$30 000 into share funds, property trusts and cash.

This person believes that in any one year:

the share funds are likely to do anything from losing about 12% to gaining 18%,
i.e. the multiplication factor will be somewhere between 0.88 and 1.18.

the property trusts are likely to do anything from losing 8% to gaining 16%,
i.e. the multiplication factor will be somewhere between 0.92 and 1.16.

and the cash is likely to do anything from gaining 5% to gaining 8% in a bank account.
i.e. the multiplication factor will be somewhere between 1.05 and 1.08.

Let us suppose that the multiplication factors are each normally distributed as follows:

Share funds multiplication factor ~ $N(1.03, 0.05^2)$.
Property trusts multiplication factor ~ $N(1.04, 0.04^2)$.
Cash multiplication factor ~ $N(1.065, 0.005^2)$.

ISBN 9780170395205

The person decides to simulate a number of years with $10 000 invested in each of the three types of investment, and the multiplication factors randomly generated from appropriate normal distributions.

Ten such simulations are shown below.

	A	B	C	D	E	F	G	
1	Shares		Property		Cash		Total	
2	Invest	× by	Invest	× by	Invest	× by	$	
3	10000	1.026	10000	1.067	10000	1.058	31510	← gain > 5%
4	10000	1.096	10000	1.047	10000	1.064	32070	← gain > 5%
5	10000	0.999	10000	1.028	10000	1.066	30930	
6	10000	0.998	10000	1.018	10000	1.069	30850	
7	10000	1.069	10000	1.053	10000	1.065	31870	← gain > 5%
8	10000	1.001	10000	1.020	10000	1.063	30840	
9	10000	0.993	10000	0.939	10000	1.061	29930	← loss
10	10000	1.092	10000	1.051	10000	1.062	32050	← gain > 5%
11	10000	1.084	10000	1.051	10000	1.063	31980	← gain > 5%
12	10000	1.074	10000	1.050	10000	1.067	31910	← gain > 5%

Based on these ten runs we might suggest that

- the probability that by the end of a year the total value of the investment will have reduced is 0.1.
- the probability that by the end of a year the total value of the investment will have increased by at least 5% is 0.6.

However such statements would not necessarily be regarded as particularly reliable when based on just ten runs of the simulation.

- Perform this simulation yourself for at least 50 runs and use your results to put forward some statements like to the two given earlier. (If you feel that the range of percentage losses or gains used above are not applicable for the financial climate at the time of reading then research and adjust accordingly.)
- Carry out the simulation a number of times but divide the $30 000 between the three types of investment differently to the even split shown above. Consider for example $20 000 to shares, $10 000 to the property trusts and $0 to cash.

Using random numbers to simulate an event involving a number of variables, in the above case the various multiplication factors, and running the simulated event many times, usually with the aid of a computer, is called a **Monte Carlo simulation**.

RESEARCH

Do some research to find out who named such a process a Monte Carlo simulation and why.

Simulation IV: Will the mineral extraction be profitable?

Let us suppose that an Australian company involved with the extraction and processing of minerals wants to investigate the viability of extracting a particular mineral, which we shall call X, from a newly discovered deposit located overseas. The country involved has granted a mining licence to the Australian company.

An ore containing X will have to be mined and then processed to yield the X it contains.

The company wishes to analyse the likely profitability of such a venture.

Whether this new venture will be profitable or not will depend on many variables. For example:

- **The cost of mining and processing each tonne of the ore.**
  This will vary according to the difficulty of extraction and processing, local costs for labour, time lost due to mechanical failure, availability of sufficient skilled labour, etc. If paid in local currency the exchange rate introduces another variable but for simplicity we will assume payments are in American dollars (US$).
- **The amount of X extracted from each tonne of ore.**
  This will vary according to the concentration of X in the ore which could vary across the expanse of the deposit.
- **The amount the company will receive for each tonne of X.**
  The company will sell the X in the 'market place' where the price is quoted in US dollars (US$). This price will vary according to the pressures of supply and demand.
- **The US dollar to Australian dollar exchange rate.**
  The company needs to be able to justify any decision to go ahead with the venture by predicting likely profits in Australian dollars (A$) and the exchange rate will vary with time.

There may well be all sorts of other variables too but for now let us restrict our attention to the ones just listed.

Suppose the company analyses these four aspects using two different models, one which assumes the variables are uniformly distributed across particular ranges and the other that assumes the variables are normally distributed.

Item	Uniform model	Normal model
Mining and processing 1 tonne of the ore (US$)	From 120 to 180	$\sim N(150, 10^2)$
Number of kg of X extracted per tonne of ore	From 135 to 225	$\sim N(180, 15^2)$
US$ received for each tonne of X	From 1200 to 2400	$\sim N(1800, 200^2)$
Exchange rate, i.e. what 1 US$ buys in A$	From 0.96 to 1.44	$\sim N(1.2, 0.08^2)$

ISBN 9780170395205

Thus, for the uniform model, the worst and best case scenarios, for 1 tonne of ore, would be as follows:

Worst case scenario	**Best case scenario**
Costs US$180 to obtain	Costs US$120 to obtain
This produces 135 kg of X	This produces 225 kg of X
Sale of this raises US$162	Sale of this raises US$540
Hence LOSS in US$ is $18	Hence PROFIT in US$ is $420
Loss in A$ is $25.92	Profit in A$ is $604.80

Five simulations of each model are shown below with the profit (or loss) involved from each tonne calculated, in Australian dollars.

	Cost (in US$) of mining and processing 1 tonne of ore	**Amount (kg) of *X* produced per tonne of ore**	**US$ received for each tonne of *X***	**A$ bought by 1 US$**	**Profit (in A$) for each tonne of ore**
Uniformly distributed model	150.93	218.93	1435.63	1.1036	180.30
	121.69	166.49	1308.53	1.0801	103.87
	140.76	143.84	2271.59	1.2089	224.84
	124.26	204.89	2199.99	1.1528	376.38
	177.80	182.53	1549.11	1.3960	146.52
Normally distributed model	156.28	141.60	1781.58	1.1496	110.35
	151.34	198.30	1758.71	1.1958	236.07
	149.62	157.23	1803.02	1.2400	166.00
	139.12	194.80	1422.09	1.1841	163.29
	165.80	194.81	1997.46	1.2687	283.33

Using a computer to simulate thousands of runs the likelihood of the profit falling into a particular range could be investigated for each model.

Use a spreadsheet to run the above simulations and write a report of your findings.

Shutterstock.com/Andriy Solovyov

ISBN 9780170395205

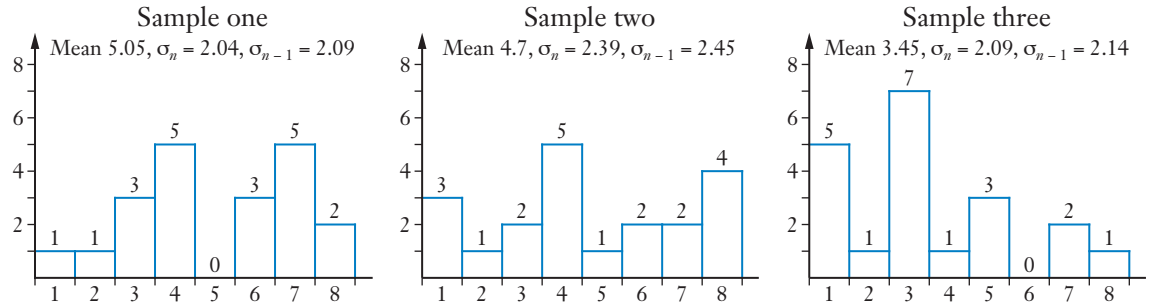
Variability of random samples

Whether we are considering a random sample of numbers taken from a uniform distribution or from a normal distribution, or indeed from any other distribution, we should not expect every such sample of numbers taken from that distribution to have the same characteristics. We would expect a certain amount of variation between samples even when they are taken from the same distributions.

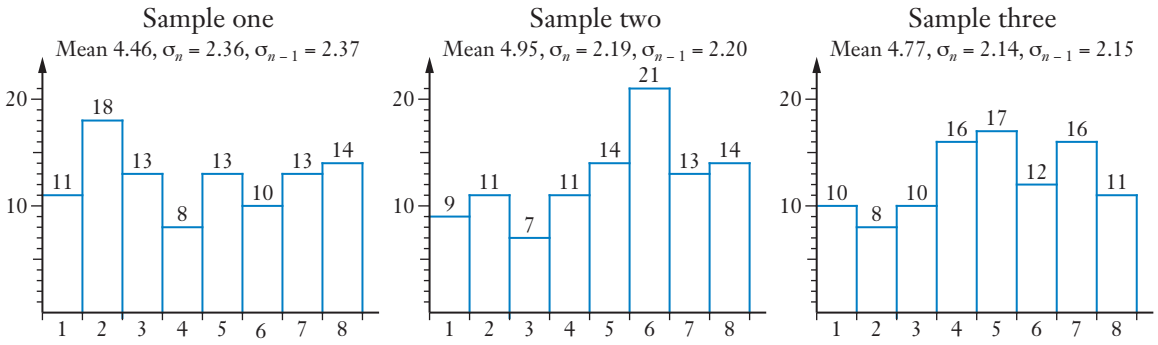
The following graphs each show the distribution of a sample of 20 numbers generated from a random variable, X, with X uniformly distributed across the integers

1, 2, 3, 4, 5, 6, 7, 8.

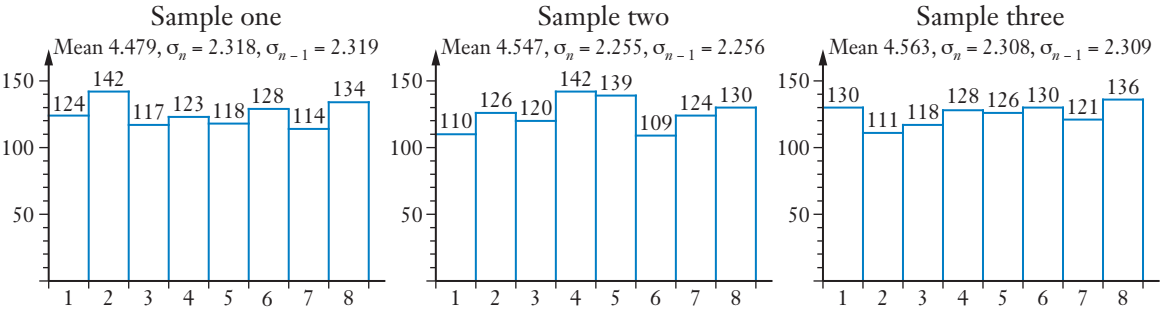
Note: $E(X) = 4.5$, $SD(X) = 2.29$ (2 decimal places).



The following graphs each show the distribution of a sample of 100 numbers generated from a random variable, X, with X uniformly distributed across the integers 1 to 8.



The following graphs each show the distribution of a sample of 1000 numbers generated from a random variable, X, with X uniformly distributed across the integers 1 to 8.

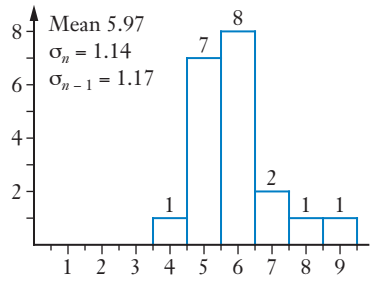


ISBN 9780170395205

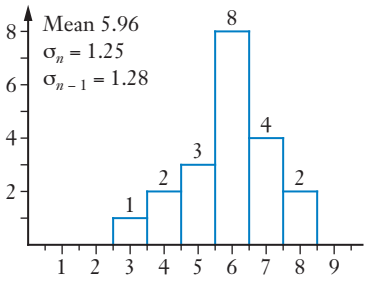
The following graphs each show the distribution of a sample of 20 numbers generated from a random variable, X, with $X \sim N(6, 1.2^2)$

Note: On this page, whilst each graph shows the numbers grouped into columns, each mean and standard deviation have been calculated from the original numbers.

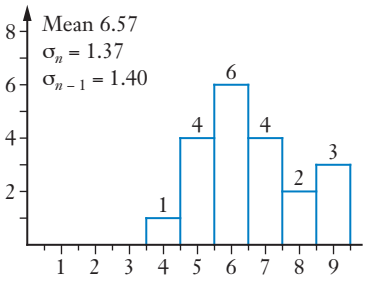
Sample one



Sample two

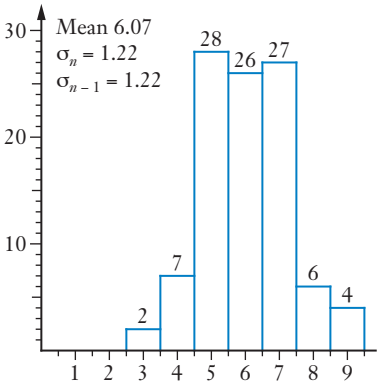


Sample three

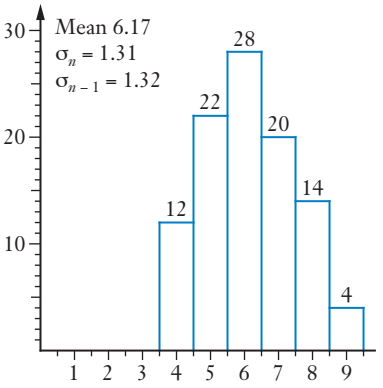


The following graphs each show the distribution of a sample of 100 numbers generated from a random variable, X, with $X \sim N(6, 1.2^2)$.

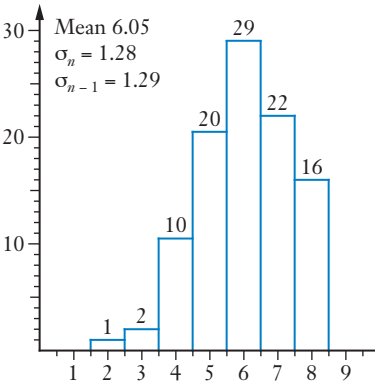
Sample one



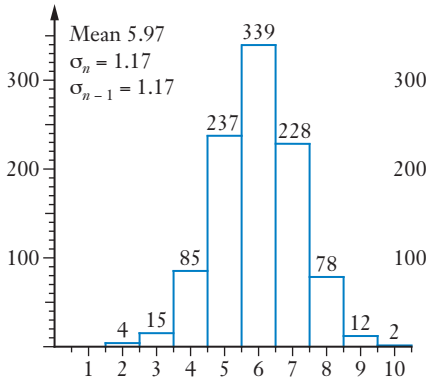
Sample two

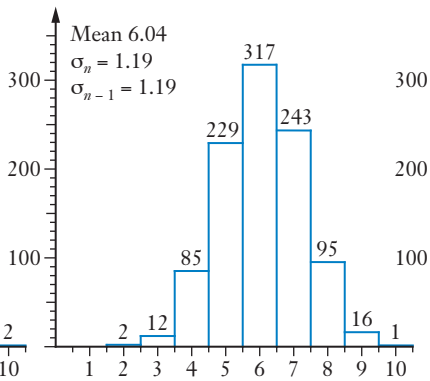


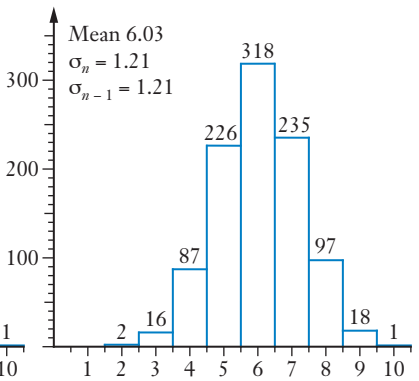
Sample three



The following graphs each show the distribution of a sample of 1000 numbers generated from a random variable, X, with $X \sim N(6, 1.2^2)$.



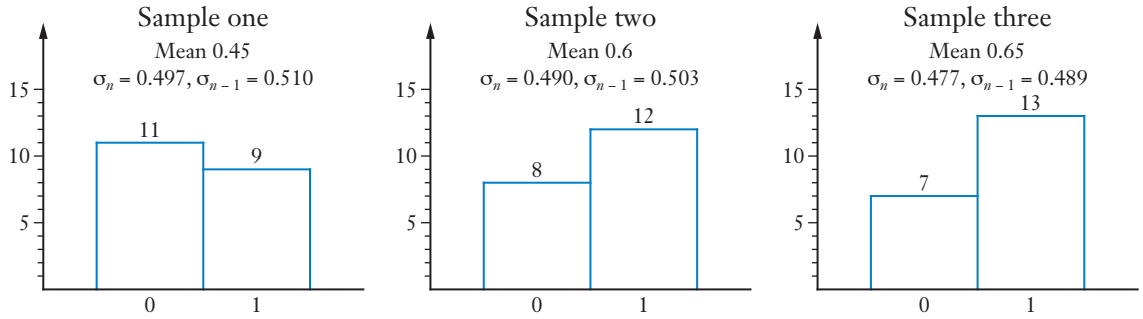




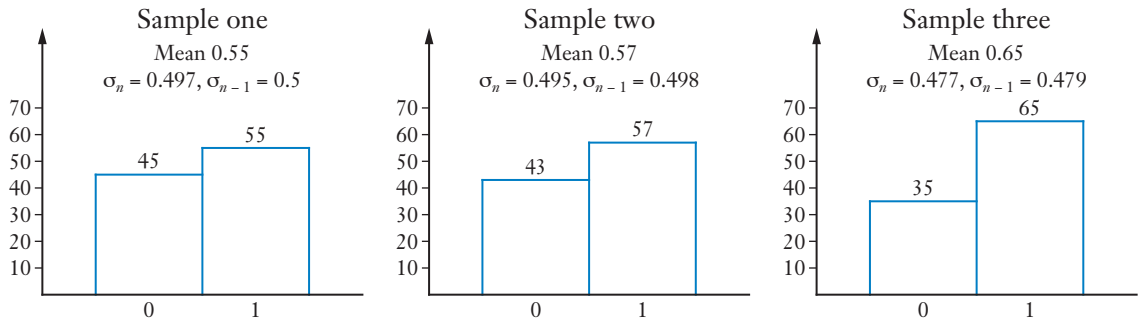
ISBN 9780170395205

The following graphs each show the distribution of a sample of 20 numbers generated from a Bernoulli distribution, X, with P(0) = 0.4 and P(1) = 0.6.

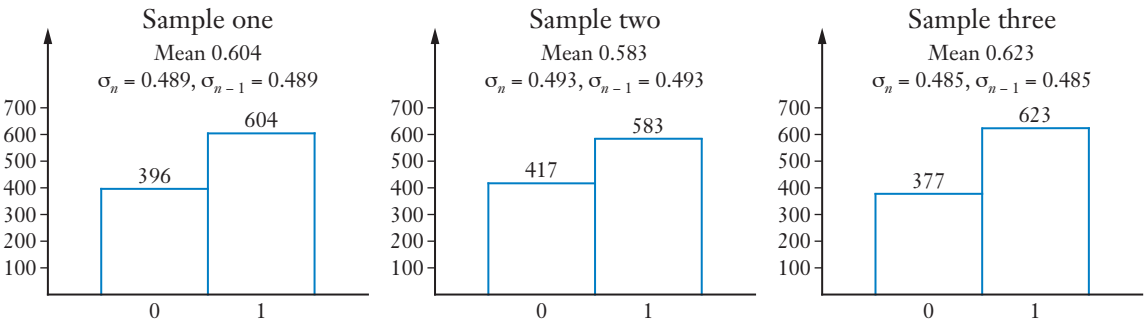
Note: $E(X) = 0.6$, $SD(X) = \sqrt{0.6(1-0.6)} \approx 0.49$.



The following graphs each show the distribution of a sample of 100 numbers generated from a Bernoulli distribution with P(0) = 0.4 and P(1) = 0.6.



The following graphs each show the distribution of a sample of 1000 numbers generated from a Bernoulli distribution with P(0) = 0.4 and P(1) = 0.6.



Use the graphs of this page and the previous two pages to answer the following:

- Are all samples of the same size, and from the same distribution, identical?
- Are the means of samples from the same distribution the same as each other? If not the same are they 'close'? Are the sample means close to the population mean? How close?
- Are the standard deviations of samples from the same distribution the same as each other? If not the same are they 'close'? Are the sample standard deviations close to the population standard deviation?
- As sample size increases does the 'shape' of the graph get closer to the 'shape' of the population distribution?

ISBN 9780170395205

Miscellaneous exercise five

This miscellaneous exercise may include questions involving the work of this chapter, the work of any previous chapters, and the ideas mentioned in the Preliminary work section at the beginning of the book.

1 Evaluate $\log(100) - \ln(e^{-3})$ without the use of a calculator.

2 If $P = 9e^{(t+1)}$ find an exact expression for t in terms of P and evaluate it, correct to three decimal places if rounding is necessary, for **a** $P = 180$, **b** $P = 3600$, **c** $P = 9e^3$.

3 If $\log a = p$ and $\log b = q$, express each of the following in terms of p or q or both p and q.

a $\log(ab)$ **b** $\log\left(\frac{a}{b}\right)$ **c** $\log(a^2b^3)$

d $\log\sqrt{a}$ **e** $\ln a$ **f** $\log_5(b^2)$

4 The discrete random variable X is binomially distributed with $X \sim \text{Bin}(n, p)$.
If $\text{E}(X) = 60$ and $\text{SD}(X) = 6$ find n, p and $\text{P}(X \leq 50)$ giving the last of these correct to three decimal places.

Differentiate each of the following with respect to x.

5 $y = x\ln(5x)$ **6** $y = (\log_e x)^2$ **7** $y = x^2 \ln x$

8 $y = (3 + \ln x)^2$ **9** $y = \frac{2}{x} + 2\ln x$ **10** $y = \frac{1}{\ln x}$

11 If $f(x) = x^3 \ln x$ determine $f''(x)$, the second derivative of the function with respect to x.

12 A particle moves in a straight line such that its displacement from a fixed point O, at time t seconds ($t \geq 0$), is x metres where $x = 9\ln(1 + t) - 4t$.
Find t when **a** the velocity is zero,
b the velocity (in m/s) is numerically equal to the acceleration (in m/s^2).

13 A continuous random variable, X, has pdf:

$$f(x) = \begin{cases} k(4 - x) & \text{for } 1 \leq x \leq 3 \\ 0 & \text{elsewhere.} \end{cases}$$

Determine **a** the value of k,
b $\text{E}(X)$, the expected value, or long-term mean, of X,
c $\text{Var}(X)$, the variance of X,
d $\text{SD}(X)$, the standard deviation of X.
e Define $\text{P}(X \leq x)$, the cumulative distribution function for X, for $-\infty < x < \infty$.

ISBN 9780170395205

14 As part of a population estimation exercise, 127 birds of a particular species visiting a swamp region favoured by migrating birds of this species are caught, tagged and released back into the region.

A few days later a second capture of a sample of this species of bird from the same swamp region caught 89 birds and it was found that amongst these 89 birds there were 3 carrying tags from the first group of 127 birds.

Releasing these 89 birds back into the region and then, a few days later, carrying out a third capture, saw 99 being caught of which 4 carried tags from the first group of 127.

Use these figures to estimate the number of birds of this species living in the area.

What might be a problem with using capture-recapture techniques on migratory birds visiting a particular region?

15 Which of the following statements are true for all $p > 0$ and $q > 0$?

a $\log_p q = \log_q p$

b $\log(p + q) = \log p + \log q$

c $\log(p - q) = \log p - \log q$

d $\log(pq) = \log p \times \log q$

e $(\log p)^q = q \log p$

f $\dfrac{\log p}{\log q} = \log p - \log q$

g $\log\left(\dfrac{p}{q}\right) = \dfrac{\log p}{\log q}$

h $\dfrac{1}{\log p} = \log\left(\dfrac{1}{p}\right)$

i $\log(pq) = \log p + \log q$

j $\log\left(\dfrac{p}{q}\right) = \log p - \log q$

k $\log(p^q) = q \log p$

l $\log_p q \times \log_q p = 1$

16 The random variable, X, is normally distributed with a mean of 1240 and a variance of 56^2, i.e. $X \sim \text{N}(1240, 56^2)$. Determine $\text{P}(1200 < X < 1300)$.

17 If $X \sim \text{N}(50, 10^2)$ determine, to three significant figures:

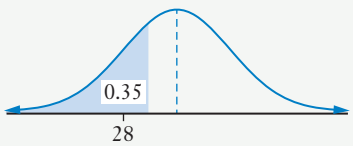
a the 0.34 quantile,

b the 0.82 quantile,

c the 43rd percentile,

d the lower quartile.

18 The continuous random variable, X, is normally distributed with $\text{P}(X < 28) = 0.35$.



a How many standard deviations from the mean is a score of 28?

b If the standard deviation of X is 5.74, find the mean of the distribution, giving your answer correct to two decimal places.

ISBN 9780170395205