Solutions to the Exercises

Chapter 1

Solution 1.1

(a) Your computer may be programmed to allocate borderline cases to the next group down, or the next group up; and it may or may not manage to follow this rule consistently, depending on its handling of the numbers involved. Following a rule which says 'move borderline cases to the next group up', these are the five classifications.

(i)	1.0–1.2	1.2–1.4	1.4–1.6	1.6–1.8	1.8–2.0	2.0–2.2	2.2–2.4
	6	6	4	8	4	3	4
	2.4–2.6	2.6–2.8	2.8–3.0	3.0–3.2	3.2–3.4	3.4–3.6	3.6–3.8
	6	3	2	2	0	1	1

(ii)	1.0–1.3	1.3–1.6	1.6–1.9	1.9–2.2	2.2–2.5
	10	6	10	5	6
	2.5–2.8	2.8–3.1	3.1–3.4	3.4–3.7	
	7	3	1	2	

(iii)	0.8–1.1	1.1–1.4	1.4–1.7	1.7–2.0	2.0–2.3
	2	10	6	10	7
	2.3–2.6	2.6–2.9	2.9–3.2	3.2–3.5	3.5–3.8
	6	4	3	1	1

(iv)	0.85–1.15	1.15–1.45	1.45–1.75	1.75–2.05	2.05–2.35
	4	9	8	9	5
	2.35–2.65	2.65–2.95	2.95–3.25	3.25–3.55	3.55–3.85
	7	3	3	1	1

(v)	0.9–1.2	1.2–1.5	1.5–1.8	1.8–2.1	2.1–2.4
	6	7	11	7	4
	2.4–2.7	2.7–3.0	3.0–3.3	3.3–3.6	3.6–3.9
	7	4	2	1	1

(b) Computer graphics: the diagrams are shown in Figures 1.9 to 1.11.

Solution 1.2

(a) Computer graphics: see Figure 1.12.

(b) Computer graphics: see Figure 1.13.

If your computer gives graphics that are text-character based (otherwise known as low-resolution graphics) then the scatter plots you obtain will not be as precise as those appearing in the text and the fitted line will not be displayed. However, the main message of the data should still be apparent.

Solution 1.3

(a) In order of decreasing brain weight to body weight ratio, the species are as follows.

Species	Body weight	Brain weight	Ratio
Rhesus Monkey	6.800	179.000	26.32
Mole	0.122	3.000	24.59
Human	62.000	1320.000	21.29
Mouse	0.023	0.400	17.39
Potar Monkey	10.000	115.000	11.50
Chimpanzee	52.160	440.000	8.44
Hamster	0.120	1.000	8.33
Cat	3.300	25.600	7.76
Rat	0.280	1.900	6.79
Mountain Beaver	1.350	8.100	6.00
Guinea Pig	1.040	5.500	5.29
Rabbit	2.500	12.100	4.84
Goat	27.660	115.000	4.16
Grey Wolf	36.330	119.500	3.29
Sheep	55.500	175.000	3.15
Donkey	187.100	419.000	2.24
Gorilla	207.000	406.000	1.96
Asian Elephant	2547.000	4603.000	1.81
Kangaroo	35.000	56.000	1.60
Jaguar	100.000	157.000	1.57
Giraffe	529.000	680.000	1.29
Horse	521.000	655.000	1.26
Pig	192.000	180.000	0.94
Cow	465.000	423.000	0.91
African Elephant	6654.000	5712.000	0.86
Triceratops	9400.000	70.000	0.007
Diplodocus	11700.000	50.000	0.004
Brachiosaurus	87000.000	154.500	0.002

(b) (i) Computer graphics: see Figure 1.14.

(ii) Computer graphics: see Figure 1.15.

Solution 1.4

There were 23 children who survived the condition. Their birth weights are 1.130, 1.410, 1.575, 1.680, 1.715, 1.720, 1.760, 1.930, 2.015, 2.040, 2.090, 2.200, 2.400, 2.550, 2.570, 2.600, 2.700, 2.830, 2.950, 3.005, 3.160, 3.400, 3.640. The median birth weight for these children is 2.200 kg (the 12th value in the sorted list).

There were 27 children who died. The sorted birth weights are 1.030, 1.050, 1.100, 1.175, 1.185, 1.225, 1.230, 1.262, 1.295, 1.300, 1.310, 1.500, 1.550, 1.600, 1.720, 1.750, 1.770, 1.820, 1.890, 1.940, 2.200, 2.270, 2.275, 2.440, 2.500, 2.560, 2.730. The middle value is the 14th (thirteen either side) so the median birth weight for these children who died is 1.600 kg.

Solution 1.5

The ordered differences are 3.8, 10.3, 11.8, 12.9, 17.5, 20.5, 20.6, 24.4, 25.3, 28.4, 30.6. The median difference is 20.5.

Solution 1.6

Once the data are entered, most computers will return the sample median at a single command. It is 79.7 inches.

Solution 1.7

(a) The mean birth weight of the 23 infants who survived SIRDS is

$$\overline{x}_S = \frac{1.130 + 1.575 + \cdots + 3.005}{23} = \frac{53.070}{23} = 2.307\,\text{kg};$$

the mean birth weight of the 27 infants who died is

$$\overline{x}_D = \frac{1.050 + 1.175 + \cdots + 2.730}{27} = \frac{45.680}{27} = 1.692\,\text{kg}.$$

The mean birth weight of the entire sample is

$$\overline{x}_T = \frac{1.130 + 1.575 + \cdots + 2.730}{50} = \frac{98.75}{50} = 1.975\,\text{kg}.$$

Notice the subscripts S, D and T used in this solution to label and distinguish the three sample means. It was not strictly necessary to do this here, since we will not be referring to these numbers again in this exercise, but it is a convenient labelling system when a statistical analysis becomes more complicated.

Solution 1.8

The mean 'After – Before' difference in Table 1.11 is

$$\overline{x} = \frac{25.3 + 20.5 + \cdots + 28.4}{11} = \frac{206.1}{11} = 18.74\,\text{pmol/l}.$$

Solution 1.9

The mean snowfall over the 63 years was 80.3 inches.

Solution 1.10

(a) The lower quartile birth weight for the 27 children who died is given by

$$q_L = x_{(\frac{1}{4}(n+1))} = x_{(7)} = 1.230\,\text{kg};$$

the upper quartile birth weight is

$$q_U = x_{(\frac{3}{4}(n+1))} = x_{(21)} = 2.200\,\text{kg}.$$

(b) For these silica data, the sample size is $n = 22$. The lower quartile is

$$q_L = x_{(\frac{1}{4}(n+1))} = x_{(\frac{23}{4})} = x_{(5\frac{3}{4})}$$

which is three-quarters of the way between $x_{(5)} = 26.39$ and $x_{(6)} = 27.08$. This is

$$q_L = 26.39 + \tfrac{3}{4}(27.08 - 26.39) = \tfrac{1}{4}(26.39) + \tfrac{3}{4}(27.08) = 26.908;$$

say, 26.9. The sample median is

$$m = x_{(\frac{1}{2}(n+1))} = x_{(\frac{23}{2})} = x_{(11\frac{1}{2})},$$

which is midway between $x_{(11)} = 28.69$ and $x_{(12)} = 29.36$. This is 29.025; say, 29.0.

The upper quartile is

$$q_U = x_{(\frac{3}{4}(n+1))} = x_{(\frac{69}{4})} = x_{(17\frac{1}{4})},$$

one-quarter of the way between $x_{(17)} = 33.28$ and $x_{(18)} = 33.40$. This is

$$q_U = 33.28 + \tfrac{1}{4}(33.40 - 33.28) = \tfrac{3}{4}(33.28) + \tfrac{1}{4}(33.40) = 33.31;$$

say, 33.3.

Solution 1.11

For the snowfall data the lower and upper quartiles are $q_L = 63.6$ inches and $q_U = 98.3$ inches respectively. The interquartile range is $q_U - q_L = 34.7$ inches.

Solution 1.12

Answering these questions might involve delving around for the instruction manual that came with your calculator! The important thing is not to use the formula—let your calculator do all the arithmetic. All you should need to do is key in the original data and then press the correct button. (There might be a choice, one of which is when the divisor in the 'standard deviation' formula is n, the other is when the divisor is $n - 1$. Remember, in this course we use the second formula.)

(a) You should have obtained $s = 8.33$, to two decimal places.

(b) The standard deviation for the silica data is $s = 4.29$.

(c) For the collapsed runners' β endorphin concentrations, $s = 98.0$.

Solution 1.13

(a) The standard deviation s is 0.66 kg.

(b) The standard deviation s is 23.7 inches.

Solution 1.14

Summary measures for this data set are

$$x_{(1)} = 23, \quad q_L = 34, \quad m = 45, \quad q_U = 62, \quad x_{(11)} = 83.$$

The sample median is $m = 45$; the sample mean is $\overline{x} = 48.4$; the sample standard deviation is 18.1. The range is $83 - 23 = 60$; the interquartile range is $62 - 34 = 28$.

Solution 1.15

The first group contains 19 completed families. Some summary statistics are

$$m = 10, \quad \overline{x} = 8.2, \quad s = 5.2, \quad \text{interquartile range } = 10.$$

For the second group of 35 completed families, summary statistics are

$$m = 4, \quad \overline{x} = 4.8, \quad s = 4.0, \quad \text{interquartile range } = 4.$$

The differences are very noticeable between the two groups. Mothers educated for the longer time period would appear to have smaller families. In each case the mean and median are of comparable size. For the smaller group, the interquartile range is much greater than the standard deviation. If the three or four very large families are removed from the second data set, the differences become even more pronounced.

Solution 1.16

(a) The five-figure summary for the silica data is given by

$$(20.77, 26.91, 29.03, 33.31, 34.82).$$

A convenient scale sufficient to cover the extent of the data is from 20 to 40. The i.q.r. is $33.31 - 26.91 = 6.40$. Then

$$q_U + \text{ i.q.r } = 33.31 + 6.40 = 39.71$$

and this exceeds the sample maximum, so the upper adjacent value is the sample maximum itself, 34.82. Also

$$q_L - \text{ i.q.r. } = 26.91 - 6.40 = 20.51.$$

This value is less than the sample minimum, so the lower adjacent value is the sample minimum itself. For these data there are no extreme values. The boxplot is shown in Figure S1.1.

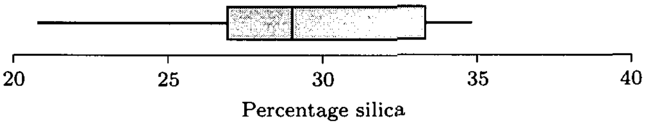


Figure S1.1

(b) For the snowfall data the lower adjacent value is 39.8; the minimum is 25.0. The upper adjacent value is equal to the maximum, 126.4. The boxplot is shown in Figure S1.2.

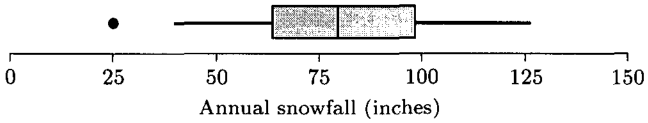


Figure S1.2

Solution 1.17

The sample skewness for the first group of mothers is -0.29.

Solution 1.18

(a) The five-figure summaries for the three groups are

normal:	(14, 92, 124.5, 274.75, 655)
alloxan-diabetic:	(13, 70.25,139.5, 276, 499)
insulin-treated:	(18, 44, 82, 133, 465).

The normal group has one very high recording at 655; the next highest is 455, which is more consistent with the other two groups.

(b) The mean and standard deviation for each group are

normal:	$\overline{x} = 186.1$, $s = 158.8$
alloxan-diabetic:	$\overline{x} = 181.8$, $s = 144.8$
insulin-treated:	$\overline{x} = 112.9$, $s = 105.8$.

The mean reading in the third group seems noticeably less than that for the first two groups, and has a reduced standard deviation.

(c) The sample skewness for each group is

normal:	1.47
alloxan-diabetic:	1.01
insulin-treated:	2.07.

All the samples are positively skewed: the third group has one substantial outlier at 465. Eliminating that outlier reduces the skewness to 1.02.

(d) The comparative boxplot in Figure S1.3 does not suggest any particular difference between the groups. The first two groups are substantially skewed with some evidence of extreme observations to the right; apart from three very extreme observations contributing to a high skewness, observations in the third group are more tightly clustered around the mean.

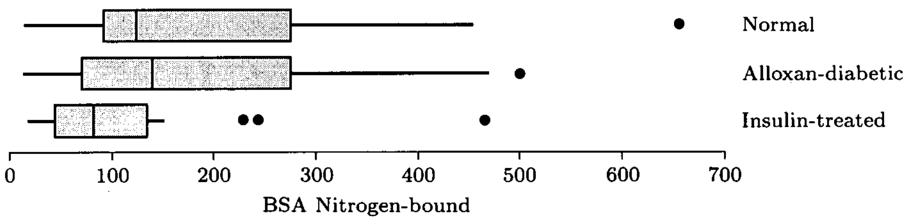


Figure S1.3

Of course, a computer makes detailed exploration of data sets relatively easy, quick and rewarding. You might find it interesting to pursue the story the data have to tell after, say, removing the extreme observations from each group.

Chapter 2

Solution 2.1

In this kind of study it is essential to state beforehand the population of interest. If this consists of rail travellers and workers then the location of the survey may be reasonable. If, on the other hand, the researcher wishes to draw some conclusions about the reading habits of the entire population of Great Britain then this sampling strategy omits, or under-represents, car users and people who never, or rarely, visit London.

A sample drawn at 9 am on a weekday will consist very largely of commuters to work, and if the researcher is interested primarily in their reading habits then the strategy will be a very useful one. On a Saturday evening there will possibly be some overrepresentation of those with the inclination, and the means, to enjoy an evening out.

Solution 2.2

This is a practical simulation. It is discussed in the text following the exercise.

Solution 2.3

A typical sequence of 40 coin tosses, and the resulting calculations and graph, follow.

Table S2.1 The results of 40 tosses of a coin

Toss number	1	2	3	4	5	6	7	8	9	10
Observed result	1	1	1	0	0	0	0	0	0	1
Total so far	1	2	3	3	3	3	3	3	3	4
Proportion (P)	1.00	1.00	1.00	0.75	0.60	0.50	0.43	0.38	0.33	0.40
Toss number	11	12	13	14	15	16	17	18	19	20
Observed result	0	1	0	1	1	0	1	0	1	1
Total so far	4	5	5	6	7	7	8	8	9	10
Proportion (P)	0.36	0.42	0.38	0.43	0.47	0.44	0.47	0.44	0.47	0.50
Toss number	21	22	23	24	25	26	27	28	29	30
Observed result	0	0	1	1	1	0	1	1	1	1
Total so far	10	10	11	12	13	13	14	15	16	17
Proportion (P)	0.48	0.45	0.48	0.50	0.52	0.50	0.52	0.54	0.55	0.57
Toss number	31	32	33	34	35	36	37	38	39	40
Observed result	1	1	1	1	0	0	1	0	0	1
Total so far	18	19	20	21	21	21	22	22	22	23
Proportion (P)	0.58	0.59	0.61	0.62	0.60	0.58	0.59	0.58	0.56	0.58

The graph of successive values of P plotted against the number of tosses is shown in Figure S2.1.

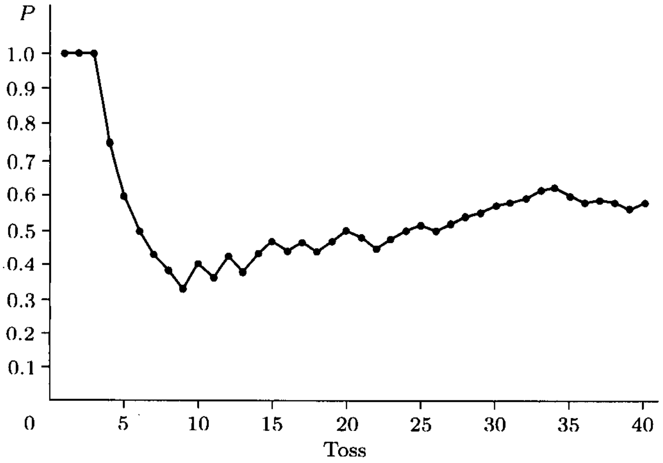


Figure S2.1 Proportion P, 40 tosses of a coin

The same phenomenon is evident here as was seen in Figures 2.2 and 2.3. In this case P seems to be tending to a value close to $\frac{1}{2}$. Did your experiment lead to similar results?

Solution 2.4

(a) The estimate of the probability that a male will be given help is

$$\frac{71}{71+29} = 0.71.$$

(b) The estimate for a female is $89/(89 + 16) = 0.85$.

(c) Since the number 0.85 is greater than the number 0.71, the experiment has provided some evidence to support the hypothesis that people are more helpful to females than to males. However, two questions arise. First, is the difference between the observed proportions sufficiently large to indicate a genuine difference in helping behaviour, or could it have arisen simply as a consequence of experimental variation when in fact there is no underlying difference in people's willingness to help others, whether male or female? Second, is the design of the experiment adequate to furnish an answer to the research question? There may have been differences (other than gender differences) between the eight students that have influenced people's responses. One matter not addressed in this exercise, but surely relevant to the investigation, is the gender of those approached.

Solution 2.5

(a) A count of yeast cells in each square is bound to result in an integer observation: you could not have 2.2 or 3.4 cells. The random variable is discrete.

(b) The data have evidently been recorded to the nearest 0.1 mm, but the actual lengths of kangaroo jawbones are not restricted in this way—within a reasonable range, any length is possible. The random variable is continuous.

(c) The lifetimes have been measured to the nearest integer and recorded as such. However, lifetime is a continuous random variable: components (in general, anyway) would not fail only 'on the hour'. A useful model would be a continuous model.

(d) Rainfall is a continuous random variable.

(e) The number of loans is an integer—the random variable measured here is discrete.

(Data might also be available on the times for which books are borrowed before they are returned. Again, this would probably be measured as integer numbers of days, even though a book could be returned at any time during a working day.)

Solution 2.6

(a) Following the same approach as that adopted in Example 2.8, we can show on a diagram the shaded region corresponding to the required proportion (or probability) $P(T > 5)$. The area of the shaded triangle is given by

$$\tfrac{1}{2} \times \text{(base)} \times \text{(height)}$$
$$= \tfrac{1}{2} \times (20 - 5) \times f(5) = \tfrac{1}{2} \times 15 \times \frac{20-5}{200} = 0.5625.$$

So, according to the model, rather more than half such gaps will exceed 5 seconds. Actually the data suggest that only one-quarter might be so long: our model is showing signs that it could be improved!

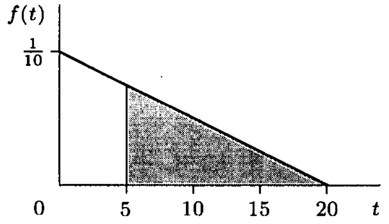


Figure S2.2 The probability $P(T > 5)$

(b) This part of the question asks for a general formula for the probability $P(T \leq t)$. The corresponding shaded region is shown in Figure S2.3. The area of the shaded region is given by

$$\begin{aligned}&(\text{average height}) \times (\text{width})\\ &= \tfrac{1}{2}(\text{long side} + \text{short side}) \times (\text{width}) = \tfrac{1}{2}(f(0) + f(t)) \times (t-0)\\ &= \tfrac{1}{2}\left(\frac{20-0}{200} + \frac{20-t}{200}\right) \times t = \frac{(40-t)t}{400} = \frac{40t-t^2}{400}.\end{aligned}$$

This formula can now be used for all probability calculations based on this model.

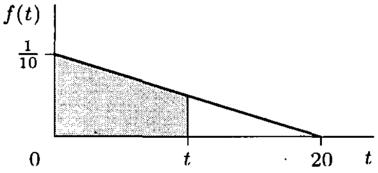


Figure S2.3 The probability $P(T \leq t)$

Solution 2.7

The probability mass function for the score on a *Double-Five* has already been established (see page 66). Summing consecutive terms gives Table S2.2.

Table S2.2 The probability distribution for a *Double-Five*

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

Solution 2.8

(a) Using the c.d.f.

$$P(T \leq t) = F(t) = \frac{40t - t^2}{400},$$

it follows that

$$P(T \leq 10) = F(10) = \frac{(40)(10) - 10^2}{400} = \frac{300}{400} = 0.75.$$

(b) The proportion of waiting times exceeding 5 seconds is given by 1− (the proportion of waiting times that are 5 seconds or less):

$$\begin{aligned}&P(T > 5)\\ &= 1 - P(T \leq 5) = 1 - F(5) = 1 - \frac{(40)(5) - 5^2}{400} = 1 - \frac{175}{400} = 0.5625\end{aligned}$$

(see Solution 2.6(a)).

Solution 2.9

(a) The probability that a woman randomly selected from the population of 6503 has passed the menopause is

$$\frac{591}{6503} = 0.091.$$

(b) Let the random variable X take the value 1 if a woman has passed the menopause and 0 otherwise. The random variable X is Bernoulli(0.091), so

$$p(x) = (0.091)^x(0.909)^{1-x}, \quad x = 0, 1.$$

(It is very important to remember to specify the *range* of the random variable.)

Solution 2.10

This is a Bernoulli trial with

$$P(X=0)=0.22, \quad P(X=1)=0.78.$$

That is, $X \sim$ Bernoulli(0.78).

The probability mass function of X is

$$p(x)=(0.78)^x(0.22)^{1-x}, \quad x=0,1.$$

Solution 2.11

It is possible that the experiment will result in a sequence of 15 failures. Each of these scores 0. Then the random variable Y (the total number of successes) takes the value

$$y=0+0+\cdots+0=0.$$

At the other extreme, the experiment might result in a sequence of 15 successes. Then

$$y=1+1+\cdots+1=15.$$

Any sequence of failures and successes (0s and 1s) between these two extremes is possible, with y taking values $1,2,\ldots,14$. The range of the random variable Y is therefore

$$\{0,1,2,\ldots,15\}.$$

Of course, it is unnecessary to be quite so formal. Your answer might have been a one-line statement of the range, which is all that is required.

Solution 2.12

Obviously the 100 people chosen have not been chosen independently: if one chosen person is female it very strongly influences the probability that the spouse will be male! Indeed, you can see that the distribution of the number of females is not binomial by considering the expected frequency distribution. If it was binomial there would be a non-zero probability of obtaining 0 females, 1 female and so on, up to 100 females. However, in this case you are certain to get exactly 50 females and 50 males. The probability that any other number will occur is zero.

Solution 2.13

(a) (i) The number dropping out in the placebo group is binomial $B(6,0.14)$. The probability that all six drop out is

$$p^6=(0.14)^6=7.53\times10^{-6}.$$

(ii) The probability that none of the six drop out is

$$(1-p)^6=(1-0.14)^6=0.86^6=0.4046.$$

(iii) The probability that exactly two drop out is

$$\binom{n}{2}p^2(1-p)^{n-2}=\binom{6}{2}(0.14)^2(0.86)^4=15(0.14)^2(0.86)^4=0.1608.$$

(b) The assumption of independence reduces, in this case, to saying that whether a patient drops out of the placebo group is unaffected by what happens to other patients in the group. Sometimes patients are unaware of others' progress in this sort of trial; but otherwise, it is at least possible that a large drop in numbers would discourage others from continuing in the study. Similarly, even in the absence of obvious beneficial effects, patients might offer mutual encouragement to persevere. In such circumstances the independence assumption breaks down.

Solution 2.14

(a)
$$\begin{aligned} P(V=2) &= \binom{8}{2}(0.3)^2(1-0.3)^{8-2} \\ &= \frac{8!}{2!\,6!}(0.3)^2(0.7)^6 = 28(0.3)^2(0.7)^6 = 0.2965. \end{aligned}$$

(b)
$$\begin{aligned} P(W=8) &= \binom{12}{8}(0.5)^8(1-0.5)^{12-8} \\ &= \frac{12!}{8!\,4!}(0.5)^8(0.5)^4 = 495(0.5)^{12} = 0.1208. \end{aligned}$$

(c)
$$\begin{aligned} P(X>4) &= P(X=5)+P(X=6) \\ &= \binom{6}{5}(0.8)^5(0.2)^{6-5}+\binom{6}{6}(0.8)^6(0.2)^{6-6} \\ &= 6(0.8)^5(0.2)+(0.8)^6 = 0.6554. \end{aligned}$$

(d)
$$\begin{aligned} P(Y\le 2) &= P(Y=0)+P(Y=1)+P(Y=2) \\ &= \binom{6}{0}\left(\frac{1}{3}\right)^0\left(\frac{2}{3}\right)^{6-0}+\binom{6}{1}\left(\frac{1}{3}\right)^1\left(\frac{2}{3}\right)^{6-1}+\binom{6}{2}\left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)^{6-2} \\ &= \left(\frac{2}{3}\right)^6+6\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)^5+15\left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)^4 = 0.6804. \end{aligned}$$

(e) Writing

$$P(Z\le 7) = P(Z=0)+P(Z=1)+\cdots+P(Z=7)$$

involves calculating eight probabilities and adding them together. It is easier to say

$$\begin{aligned} P(Z\le 7) &= 1-P(Z\ge 8) \\ &= 1-[P(Z=8)+P(Z=9)+P(Z=10)] \\ &= 1-\left[\binom{10}{8}\left(\frac{1}{4}\right)^8\left(\frac{3}{4}\right)^2+\binom{10}{9}\left(\frac{1}{4}\right)^9\left(\frac{3}{4}\right)+\binom{10}{10}\left(\frac{1}{4}\right)^{10}\left(\frac{3}{4}\right)^0\right] \\ &= 1-\left[45\left(\tfrac{1}{4}\right)^8\left(\tfrac{3}{4}\right)^2+10\left(\tfrac{1}{4}\right)^9\left(\tfrac{3}{4}\right)+\left(\tfrac{1}{4}\right)^{10}\right] \\ &= 1-0.000\,416 = 0.999\,584. \end{aligned}$$

(Actually, it is even easier to use your computer for binomial probability calculations.)

Solution 2.15

(a) The distribution of wrinkled yellow peas amongst a 'family' of eight is $B(8,\frac{3}{16})$.

(b) The probability that all eight are wrinkled and yellow is

$$\left(\tfrac{3}{16}\right)^8 = 1.53 \times 10^{-6}.$$

(c) The distribution of wrinkled green peas amongst eight offspring is binomial $B\left(8, \frac{1}{16}\right)$. The probability that there are no wrinkled green peas is

$$\binom{8}{0}\left(\tfrac{1}{16}\right)^0\left(1-\tfrac{1}{16}\right)^{8-0} = \left(\tfrac{15}{16}\right)^8 = 0.597.$$

Solution 2.16

You should find that your computer gives you the following answers. (These answers are accurate to six decimal places.)

(a) 0.200 121 (b) 0.068 892 (c) 0.998 736 (d) 0.338 529

(e) If four dice are rolled simultaneously, then the number of 6s to appear is a binomial random variable $M \sim B\left(4, \frac{1}{6}\right)$. The probability of getting at least one 6 is

$$P(M \geq 1) = 1 - P(M = 0) = 1 - \left(\tfrac{5}{6}\right)^4 = 0.5177.$$

If two dice are rolled, the probability of getting a double-6 is $\frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$. The number of double-6s in twenty-four such rolls is a binomial random variable $N \sim B\left(24, \frac{1}{36}\right)$. The probability of getting at least one double-6 is

$$P(N \geq 1) = 1 - P(N = 0) = 1 - \left(\tfrac{35}{36}\right)^{24} = 0.4914.$$

So it is the first event of the two that is the more probable.

(f) If X is $B(365, 0.3)$ then

$$P(X \geq 100) = 0.8738.$$

(This would be very time-consuming to calculate other than with a computer.) In answering this question the assumption has been made that rain occurs independently from day to day; this is a rather questionable assumption.

Solution 2.17

(a) A histogram of the data looks like the following. The sample mean and standard deviation are:

$$\overline{x} = 18.11\,\text{mm}, \quad s = 8.602\,\text{mm}.$$

The average book width appears to be about 18.11 mm, so for 5152 books the required shelving would be 5152 × 18.11 mm = 93.3 m.

(b) This is a somewhat subjective judgement, since no formal tests have been developed for a 'bell-shaped' appearance, or lack of it. The histogram suggests the data are rather skewed. It is worth observing that the width of the widest book in the sample is about 3.5 standard deviations above the mean; the narrowest book measures only 1.5 standard deviations below the mean.

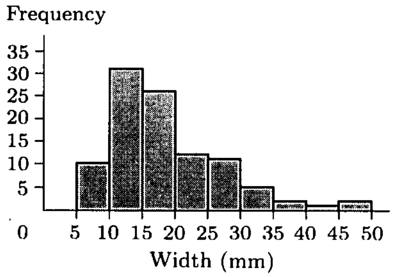


Figure S2.4 Widths of 100 books

Solution 2.18

(a) You might have obtained a sequence of 0s and 1s as follows.

0 0 1 1 0 0 0 0 0 0

The number of 1s in the ten trials is 2. A single observation from $B(10, 0.2)$ was then obtained: it was 3. The sum of ten independent Bernoulli random variables Bernoulli(0.2) is binomial $B(10, 0.2)$. The two observations, 2 and 3, are independent observations, each from $B(10, 0.2)$.

(b) Figure S2.5 shows three bar charts similar to those you might have obtained.

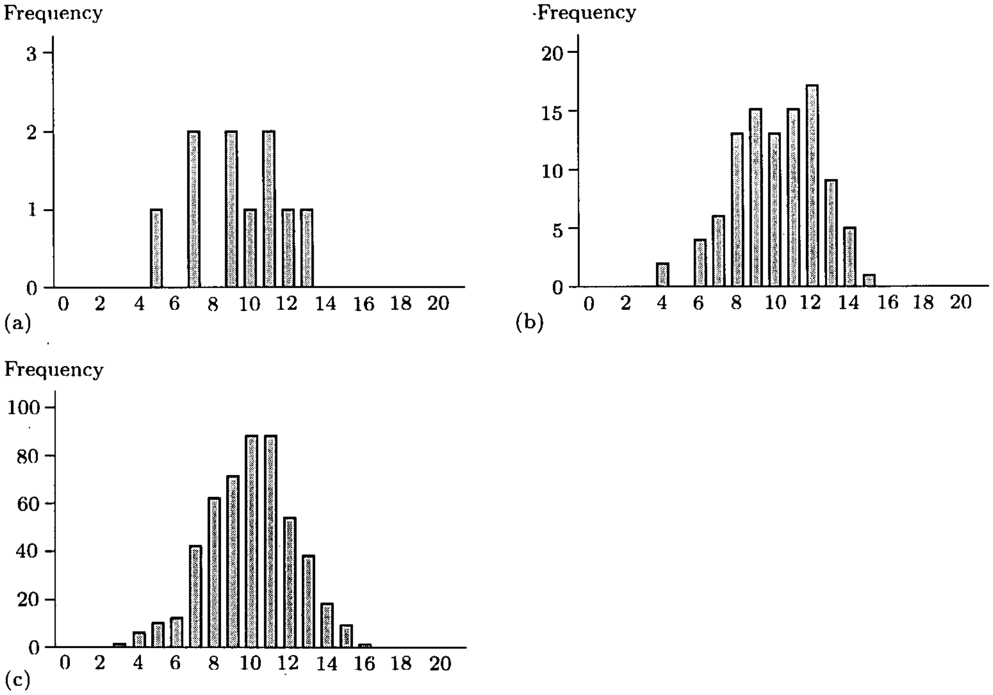


Figure S2.5 (a) 10 values from $B(20, 0.5)$ (b) 100 values from $B(20, 0.5)$ (c) 500 values from $B(20, 0.5)$

Notice that, as the sample size increases, the bar charts for the observed frequencies become less jagged. Even in the case of a sample of size 100, however, the bar chart can be very irregular: this is bimodal. When the sample is of size 500, the observed frequencies are very suggestive of the underlying probability distribution, whose probability mass function is shown in Figure S2.6.

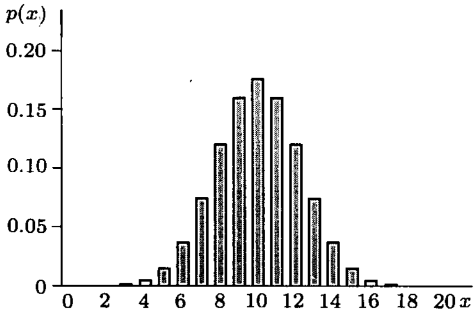


Figure S2.6 The binomial probability distribution $B(20, 0.5)$

(c) Here are three typical bar charts.

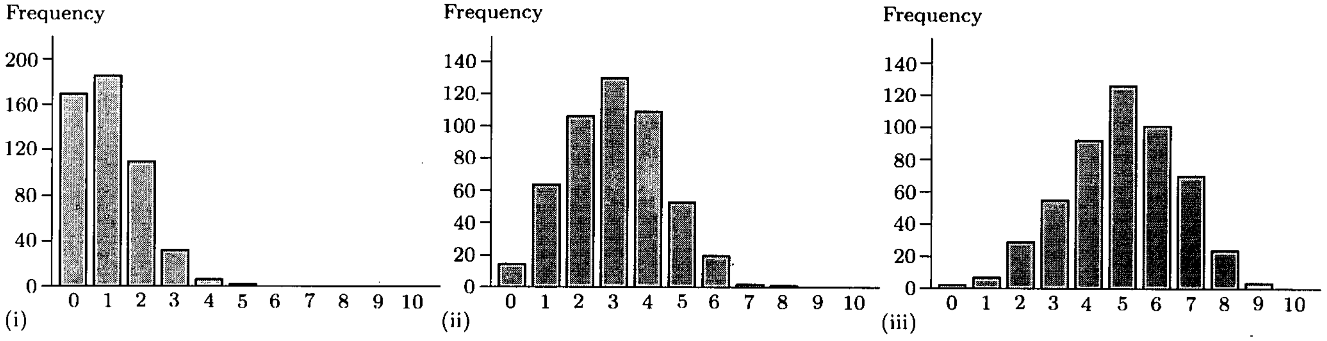


Figure S2.7 (i) 500 values from $B(10, 0.1)$ (ii) 500 values from $B(10, 0.3)$ (iii) 500 values from $B(10, 0.5)$

You can see that the value of the parameter p affects the skewed nature of the sample data.

(d) The following diagrams show three summaries of the data.

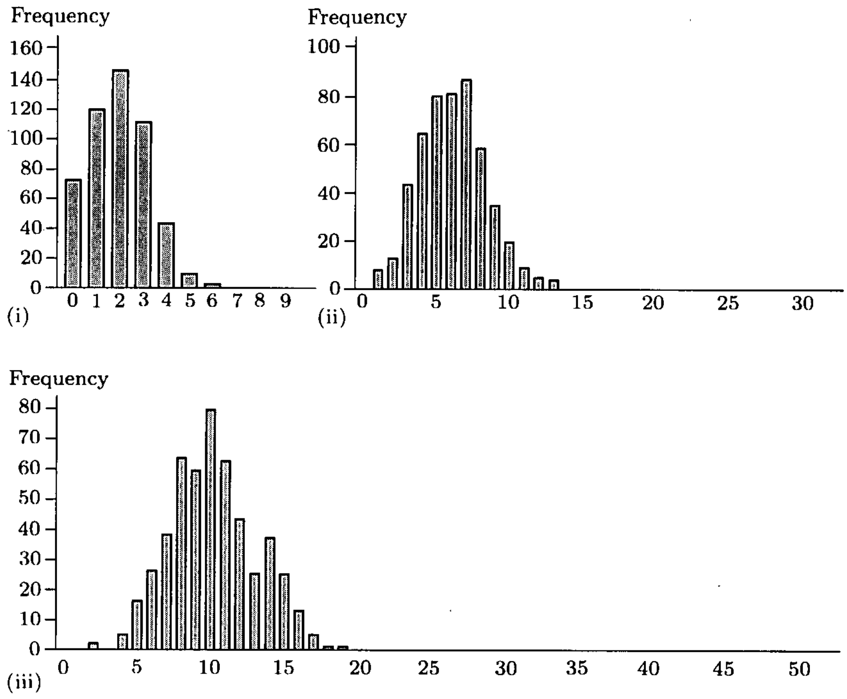


Figure S2.8 (i) 500 values from $B(10, 0.2)$ (ii) 500 values from $B(30, 0.2)$ (iii) 500 values from $B(50, 0.2)$

Even for a value as low as 0.2 for the parameter p, you should have observed from your data, rather as is evident here, that as the parameter n increases the sample histograms become less skewed. This will be further discussed in *Chapter 5*.

Solution 2.19

Out of interest, this experiment was repeated three times, thus obtaining the frequencies in Table S2.3.

Table S2.3 'Opening the bag' three times

Number of defective fuses	Frequency		
0	95	94	93
1	5	4	6
2	0	2	1
3	0	0	0
4	0	0	0
5	0	0	0
6	0	0	0
7	0	0	0
8	0	0	0

You can see that there is some variation in the results here.

Solution 2.20

(a) In 6 rolls of the die, the following results were obtained.

Table S2.4 Rolling a die 6 times

Roll number	1	2	3	4	5	6
Frequency	0	0	1	1	1	3
Relative frequency	0	0	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{2}$

You can see that the sample relative frequencies are widely disparate and do not always constitute very good estimates of the theoretical probabilities: in all cases, these are $\frac{1}{6} = 0.1667$.

(b) In 600 rolls, the following frequencies were obtained.

Table S2.5 Rolling a die 600 times

Roll number	1	2	3	4	5	6
Frequency	80	95	101	111	97	116
Relative frequency	0.1333	0.1583	0.1683	0.1850	0.1617	0.1933

Even in a sample as large as 600, probability estimates can be quite wrong in the second decimal place! But these are generally more consistent and closer to the theoretical values than is the case with the sample of just 6 rolls.

Solution 2.21

One hundred observations on the binomial distribution $B(33, 0.1)$ were generated. Three observations were 8 or more, giving an estimated probability of 0.03 that a sample as extreme as that reported could occur. For interest, the number of left-handed people in each of a 100 groups of 33 individuals was counted. The frequencies were as listed in Table S2.6.

Actually, if X is binomial $B(33, 0.1)$, then

$$P(X \geq 8) = 0.014.$$

This makes it seem very unlikely that the circumstance observed could have arisen by mere chance.

Table S2.6 Left-handedness in 100 groups of 33 individuals

Number of left-handed people	Frequency
0	4
1	9
2	20
3	26
4	23
5	12
6	3
7	0
8	2
9	1
⋮	⋮

Solution 2.22

(a) If the random variable V follows a triangular distribution with parameter 60, then the c.d.f. of V is given by

$$F(v) = P(V \leq v) = 1 - \left(1 - \frac{v}{60}\right)^2, \quad 0 \leq v \leq 60.$$

Then (either directly from your computer, or by using this formula together with your calculator) the following values will be obtained.

(i) $P(V \leq 20) = F(20) = 1 - \left(1 - \frac{20}{60}\right)^2 = 1 - \left(\frac{2}{3}\right)^2 = \frac{5}{9} = 0.556$

(ii) $P(V > 40) = 1 - F(40) = \left(1 - \frac{40}{60}\right)^2 = \left(\frac{1}{3}\right)^2 = \frac{1}{9} = 0.111$

(iii) The probability $P(20 \leq V \leq 40)$ is equal to the area of the shaded region in Figure S2.9. It is given by

$$P(V \leq 40) - P(V \leq 20) = F(40) - F(20) = \tfrac{8}{9} - \tfrac{5}{9} = \tfrac{1}{3} = 0.333$$

(using your answers to parts (i) and (ii)).

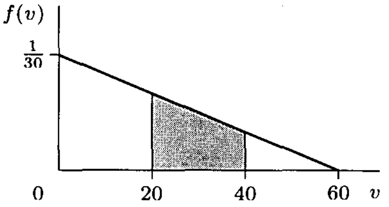


Figure S2.9 The probability $P(20 \leq V \leq 40)$

(b) (i) A histogram of these data is shown in Figure S2.10.

(ii) The data are skewed, with long waiting times apparently less likely than shorter waiting times. The sample is very small, but in the absence of more elaborate models to consider, the triangular model is a reasonable first attempt. The longest waiting time observed in the sample is 171 hours. Any number higher than this would be a reasonable guess at the model parameter—say, 172 or 180 or even 200, without going too high (300, perhaps). Try 180.

(iii) With θ set equal to 180, and denoting by W the waiting time (in hours), then

$$P(W > 100) = 1 - F(100) = \left(1 - \tfrac{100}{180}\right)^2 = 0.198.$$

In the sample of 40 there are 5 waiting times longer than 100 hours (102, 116.5, 122, 144, 171), so the sample-based estimate for the proportion of waiting times exceeding 100 hours is $\frac{5}{40} = 0.125$.

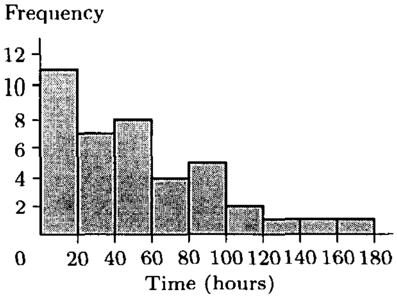


Figure S2.10 Waiting times between admissions

Chapter 3

Solution 3.1

The mean score on a *Double-Five* is given by

$$\mu = 1 \times \tfrac{1}{6} + 3 \times \tfrac{1}{6} + 4 \times \tfrac{1}{6} + 5 \times \tfrac{1}{3} + 6 \times \tfrac{1}{6} = 4.$$

Hence an effect of replacing the 2-face of a fair die by a second 5 is to increase the mean from its value of 3.5 (see Example 3.3) to 4.

Solution 3.2

From the given probability distribution of X, the mean number of members of the family to catch the disease is

$$\begin{aligned}\mu &= 1 \times \tfrac{3}{90} + 2 \times \tfrac{8}{90} + 3 \times \tfrac{15}{90} + 4 \times \tfrac{20}{90} + 5 \times \tfrac{24}{90} + 6 \times \tfrac{20}{90} \\ &= \tfrac{1}{90}(3 + 16 + 45 + 80 + 120 + 120) = \tfrac{384}{90} = 4.3.\end{aligned}$$

Solution 3.3

(a) For a fair coin, $P(\text{Heads}) = p(1) = \frac{1}{2}$. So $p = \frac{1}{2}$ and the mean of the Bernoulli random variable is p, i.e. $\frac{1}{2}$.

(b) As in *Chapter 2*, Exercise 2.2, $p(1) = P(3 \text{ or } 6) = \frac{1}{3}$. Thus $\mu = p = \frac{1}{3}$.

Solution 3.4

The expected value of Y is given by

$$E(Y) = \sum_{y=0}^{2} yp(y)$$
$$= 0p(0) + 1p(1) + 2p(2) = p^3(1-p)^2\left(2p^2 + 2p + 5\right) + 2p(1-p)^6.$$

(a) When $p = 0.1$,

$$E(Y) = 0.1^3 \times 0.9^2 \times 5.22 + 2 \times 0.1 \times 0.9^6 = 0.1105.$$

(b) When $p = 0.4$,

$$E(Y) = 0.4^3 \times 0.6^2 \times 6.12 + 2 \times 0.4 \times 0.6^6 = 0.1783.$$

(c) When $p = 0.6$,

$$E(Y) = 0.6^3 \times 0.4^2 \times 6.92 + 2 \times 0.6 \times 0.4^6 = 0.2441.$$

(d) When $p = 0.8$,

$$E(Y) = 0.8^3 \times 0.2^2 \times 7.88 + 2 \times 0.8 \times 0.2^6 = 0.1615.$$

You can see that when the chain is very fragile or very robust, the expected number of quads is low; only for intermediate p is the expected number of quads more than about 0.2.

Solution 3.5

(a) In one experiment the results in Table S3.1 were obtained. The sample mean is 5.63.

(b) The mean of the first sample drawn in an experiment was 6.861. Together with nine other samples, the complete list of sample means is shown in Table S3.2.

(c) In one experiment the following results were obtained: (i) 9.974; (ii) 97.26; (iii) 198.5.

(d) These findings suggest that the mean of the Triangular(θ) distribution is $\frac{1}{3}\theta$.

Table S3.1

9.72	3.37	12.99	6.92	1.35
2.38	2.08	8.75	7.79	0.95

Table S3.2

6.861	6.468	6.532	6.713	6.667
6.628	6.744	6.586	6.808	6.671

Solution 3.6

Using the information given, the probability required is

$$P(T > \mu) = P\left(T > \tfrac{1}{3}\theta\right) = 1 - P\left(T \leq \tfrac{1}{3}\theta\right) = 1 - F\left(\tfrac{1}{3}\theta\right)$$
$$= 1 - \left[1 - \left(1 - \frac{\theta/3}{\theta}\right)^2\right] = \left(1 - \tfrac{1}{3}\right)^2 = \tfrac{4}{9}.$$

So in any collection of traffic waiting times (assuming the triangular model to be an adequate representation of the variation in waiting times) we might expect just under half the waiting times to be longer than average. Notice that this result holds irrespective of the actual value of the parameter θ.

Solution 3.7

The probability distribution for the *Double-Five* outcome is shown in Table 3.4.

The population mean is 4 (see solution to Exercise 3.1).

The calculation of the variance is as follows.

j	1	3	4	5	6
$j-\mu$	-3	-1	0	1	2
$(j-\mu)^2$	9	1	0	1	4

$$\sigma^2 = \sum_j (j-\mu)^2 p(j)$$

$$= 9 \times \tfrac{1}{6} + 1 \times \tfrac{1}{6} + 0 \times \tfrac{1}{6} + 1 \times \tfrac{1}{3} + 4 \times \tfrac{1}{6} = 2.67$$

The variance of the score on a fair die is 2.92. So, while the mean score on a *Double-Five* is greater than that on a fair die, the variance of the *Double-Five* outcome is smaller. This is not unreasonable since, by replacing the 2 by another 5, one can intuitively expect a little more 'consistency', that is, less variability, in the outcomes.

Solution 3.8

To check for independence, we shall work out $p_{X,Y}(x,y)$ assuming independence, and compare the outcome with Table 3.8. For instance, $p_{X,Y}(0,-1)$ would be the product $p_X(0)p_Y(-1) = 0.4 \times 0.3 = 0.12$, $p_{X,Y}(2,-1)$ would be the product $p_X(2)p_Y(-1) = 0.2 \times 0.3 = 0.06$, and so on. In this way, we produce Table S3.3 of the joint p.m.f. of X and Y *under independence*.

Table S3.3 The joint p.m.f. of X and Y under independence

x	0	1	2
$y=-1$	$0.4 \times 0.3 = 0.12$	$0.4 \times 0.3 = 0.12$	$0.2 \times 0.3 = 0.06$
$y=1$	$0.4 \times 0.7 = 0.28$	$0.4 \times 0.7 = 0.28$	$0.2 \times 0.7 = 0.14$

These values are shown more clearly in Table S3.4.

These values are not the same as those in Table 3.8. For instance, under independence we would require $p_{X,Y}(1,1)$ to equal 0.28, whereas $p_{X,Y}(1,1)$ is 0.30. Hence X and Y are *not* independent.

Table S3.4

			x	
		0	1	2
y	-1	0.12	0.12	0.06
	1	0.28	0.28	0.14

Solution 3.9

(a) The random variable N takes the value 1 if the first trial results in a 'success': $P(N=1)=p$.

(b) Success occurs for the first time only at the second trial if initially there is a failure, followed immediately by a success: $P(N=2)=qp$.

(c) Here, there are two failures followed by a success: $P(N=3)=q^2p$.

(d) A clear pattern is emerging. The random variable N takes the value n only if $(n-1)$ failures are followed at the nth trial by a success:

$$P(N=n)=q^{n-1}p.$$

(e) The range of possible values N can take is $1, 2, 3, \ldots$, the set of positive integers (which you might also know as the set of natural numbers).

Solution 3.10

(a) The proportion of families comprising at least 4 children is found from $P(N \geq 4) = 1 - P(N \leq 3)$.

$$\begin{aligned} 1 - P(N \leq 3) &= 1 - (p(1) + p(2) + p(3)) = 1 - (p + qp + q^2p) \\ &= 1 - (0.514)(1 + 0.486 + 0.486^2) = 1 - (0.514)(1.722) \\ &= 1 - 0.885 = 0.115. \end{aligned}$$

(b) Denoting by 'success' the identification of a defective chip, $p = 0.012$. The size of the inspector's sample of chips is a random variable N where $N \sim G(0.012)$. Then

$$\begin{aligned} P(N < 6) = P(N \leq 5) &= p + qp + q^2p + q^3p + q^4p \\ &= (0.012)(1 + 0.988 + 0.988^2 + 0.988^3 + 0.988^4) \\ &= (0.012)(4.8814) = 0.0586, \end{aligned}$$

so about 6% of daily visits involve a halt in production.

Solution 3.11

In this case, the random variable N follows a geometric distribution with parameter $p = 0.02$. So

$$P(N > 20) = q^{20} = (0.98)^{20} = 0.668.$$

The probability that the inspector will have to examine at least 50 chips is

$$P(N \geq 50) = P(N > 49) = q^{49} = (0.98)^{49} = 0.372.$$

Notice that it is much easier to use the formula $P(N > n) = q^n$ to calculate tail probabilities for the geometric distribution than to add successive terms of the probability function as in Solution 3.10.

Solution 3.12

(a) 2 seems intuitively correct.

(b) If the probability of throwing a 5 is $\frac{1}{3}$, this suggests that the average number of throws necessary to achieve a 5 will be 3.

(c) 6.

(d) By the same argument, guess $\mu = 1/p$.

Solution 3.13

The number N of rolls necessary to start playing is a geometric random variable with parameter $p = 1/6$.

(a) $P(N = 1) = p = 1/6 = 0.167$.

(b) $P(N = 2) = qp = 5/36 = 0.139$; $P(N = 3) = q^2p = 25/216 = 0.116$.

(c) The probability that at least six rolls will be necessary to get started is given by $P(N \geq 6) = P(N > 5) = q^5 = 3125/7776 = 0.402$.

(d) The expected number of rolls for a geometric random variable is $1/p$; which is 6 in this case. The standard deviation is $\sqrt{q}/p = 6\sqrt{5/6} = 5.48$.

Solution 3.14

Your results should not be too different from the following, which were obtained on a computer.

(a) A frequency table for the 1200 rolls summarizes the data as follows.

Outcome	1	2	3	4	5	6
Frequency	195	202	227	208	181	187

(b) The corresponding bar chart is shown in Figure S3.1. The bar chart shows some departures from the theoretical expected frequencies (200 in each of the six cases): these departures may be ascribed to random variation.

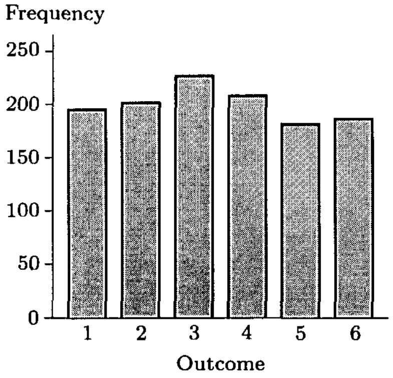


Figure S3.1 Bar chart for 1200 rolls of the die

(c) The computer gave

$$\bar{x} = 3.45, \quad s^2 = 2.798\,08,$$

so $s = 1.67$.

This may be compared with the theoretical sample moments for a discrete uniform distribution:

$$\mu = \tfrac{1}{2}(n+1) = \tfrac{1}{2}(6+1) = 3.5, \quad \sigma^2 = \tfrac{1}{12}\left(n^2-1\right) = \tfrac{1}{12}(36-1) = 2.917,$$

so $\sigma = 1.71$.

The sample gave results that were on average slightly lower than the theoretical scores, and that are slightly less dispersed. These differences are scarcely perceptible and can be ascribed to random variation.

Solution 3.15

A sketch of the p.d.f. of X when $X \sim U(a, b)$ is shown in Figure S3.2.

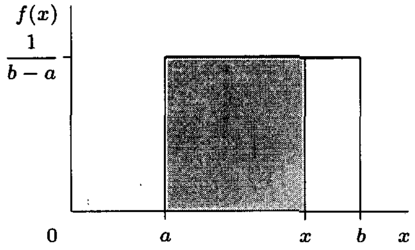


Figure S3.2 The p.d.f. of X, $X \sim U(a, b)$

(a) By symmetry, the mean of X is $\mu = \frac{1}{2}(a+b)$.

(b) The probability $P(X \leq x)$ is equal to the area of the shaded rectangle in the diagram. This is $(x-a) \times \dfrac{1}{b-a}$. So the c.d.f. of X is given by

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

Solution 3.16

The formula for the variance of a continuous uniform random variable $U(a,b)$ is

$$\sigma^2 = \frac{(b-a)^2}{12}.$$

For the standard continuous uniform distribution $U(0,1)$, $a = 0$ and $b = 1$, so the variance is

$$\sigma^2 = \tfrac{1}{12} = 0.083,$$

and the standard deviation is

$$\sigma = \sqrt{\tfrac{1}{12}} = 0.289.$$

Solution 3.17

(a) From Solution 3.16, the c.d.f. of the $U(a,b)$ distribution is

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

To solve $F(m) = \frac{1}{2}$, we need to solve the equation

$$\frac{m-a}{b-a} = \frac{1}{2}$$

or

$$m - a = \frac{b-a}{2}.$$

This gives

$$m = \frac{a+b}{2},$$

and is the median of the $U(a,b)$ distribution. You might recall that this is also the value of the mean of the $U(a,b)$ distribution, and follows immediately from a symmetry argument.

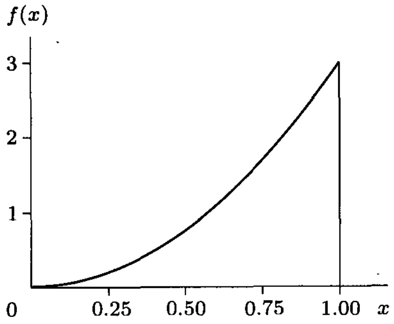


Figure S3.3 $f(x) = 3x^2$, $0 \leq x \leq 1$

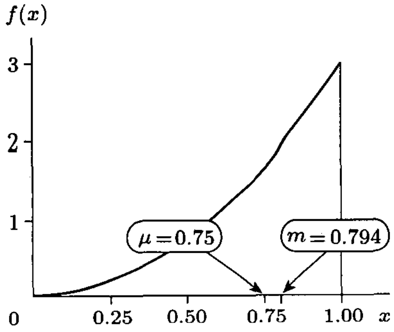


Figure S3.4 The mean and median of X

(b) (i) The density function $f(x) = 3x^2$, $0 \leq x \leq 1$, is shown in Figure S3.3.

(ii) The mean and median are shown in Figure S3.4.

(iii) From $F(x) = x^3$, it follows that the median is the solution of the equation

$$x^3 = \tfrac{1}{2}.$$

This is

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

The mean $\mu = 0.75$ and the median $m = 0.794$ are shown in Figure S3.4.

Solution 3.18

(a) The c.d.f. of this distribution is

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

To obtain the interquartile range, we need both q_U and q_L. To obtain q_L, we solve

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

and obtain

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

Likewise,

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

hence

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

So, the interquartile range is

$$q_U - q_L = 0.909 - 0.630 = 0.279.$$

Solution 3.19

(a) For the binomial distribution $B(10, 0.5)$, $F(4) = 0.3770$, $F(5) = 0.6230$, so the median is 5.

(b) For the binomial distribution $B(17, 0.7)$, $F(11) = 0.4032$, $F(12) = 0.6113$, so the median is 12.

(c) For the binomial distribution $B(2, 0.5)$, $F(1) = 0.75$, therefore the upper quartile is 1. (So is the median!)

(d) For the binomial distribution $B(19, 0.25)$, $F(5) = 0.6678$, $F(6) = 0.8251$, so $q_{0.75} = 6$.

Since $F(2) = 0.1113$ and $F(3) = 0.2631$, $q_{0.25} = 3$.

Hence the interquartile range is $q_{0.75} - q_{0.25} = 6 - 3 = 3$.

(e) For the binomial distribution $B(15, 0.4)$, $F(7) = 0.7869$, $F(8) = 0.9050$, so $q_{0.85} = 8$.

Chapter 4

Solution 4.1

(a) If $X \sim B(50, 1/40)$, then $P(X = 0) = (39/40)^{50} = (0.975)^{50} = 0.2820$.

(b) The probability that the cyclist gets wet twice is

$$P(X = 2) = \binom{50}{2}(0.025)^2(0.975)^{48} = 0.2271.$$

(c) Values of $p(x)$ for $x = 0, 1, 2, 3$, are $p(0) = 0.2820$, $p(1) = 0.3615$, $p(2) = 0.2271$, $p(3) = 0.0932$; so the probability that she gets wet at least four times is

$$\begin{aligned} 1 - (p(0) + p(1) + p(2) + p(3)) &= 1 - (0.2820 + 0.3615 + 0.2271 + 0.0932) \\ &= 1 - 0.9638 = 0.0362. \end{aligned}$$

Solution 4.2

When $X \sim B(60, 1/48)$, $P(X = 0) = (47/48)^{60} = 0.2827$. Rounding to, say, $(0.979)^{60} = 0.28$ would induce rather serious rounding errors. Continuing in this way, obtain the table of probabilities as follows.

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

The last value was obtained by subtraction. In fact, if you use a computer you would find that the probability $P(X \geq 4)$ when X is $B(60, 1/48)$ is 0.0366, to 4 decimal places.

Solution 4.3

(a) In this case $X \sim B(360, 0.01)$.

(b) Including also the probabilities calculated in the text for $B(320, 0.011\,25)$, the results are as listed in the table below.

	$P(X=0)$	$P(X=1)$	$P(X=2)$	$P(X=3)$	$P(X \geq 4)$
$B(320, 0.01125)$	0.0268	0.0975	0.1769	0.2134	0.4854
$B(360, 0.01)$	0.0268	0.0976	0.1769	0.2133	0.4854

In this case the results are close, identical to three decimal places. (Again, the last column was found by subtraction. To 4 decimal places, when $X \sim B(320, 0.011\,25)$, the probability $P(X \geq 4)$ is 0.4855.)

Solution 4.4

Using the given recursion,

(a) $p_X(1) = \frac{\mu}{1}p_X(0) = \frac{\mu}{1}e^{-\mu} = \mu e^{-\mu}$,

(b) $p_X(2) = \frac{\mu}{2}p_X(1) = \frac{\mu}{2}\mu e^{-\mu} = \frac{\mu^2}{2!}e^{-\mu}$,

(c) $p_X(3) = \frac{\mu}{3}p_X(2) = \frac{\mu}{3}\frac{\mu^2}{2!}e^{-\mu} = \frac{\mu^3}{3!}e^{-\mu}$,

where the notation $k!$ means the number $1 \times 2 \times \ldots \times k$.

(d) There is an evident pattern developing here: a general formula for the probability $p_X(x)$ is

$$p_X(x) = \frac{\mu^x}{x!}e^{-\mu}.$$

Solution 4.5

The completed table is as follows.

	$P(X=0)$	$P(X=1)$	$P(X=2)$	$P(X=3)$	$P(X \geq 4)$
$B(320, 0.01125)$	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

(Probabilities in the last column are found by subtraction: to 4 decimal places, the probability $P(X \geq 4)$ when X is Poisson(3.6) is 0.4848.)

Solution 4.6

(a) The exact probability distribution for the number of defectives in a box is $B(50, 0.05)$ which (unless you have access to very extensive tables!) will need calculation on a machine as follows (i.e. recursively, retaining displayed values on the machine):

$$p_X(0) = (0.95)^{50} = 0.0769$$
$$p_X(1) = 50 \times \frac{0.05}{0.95} \times p_X(0) = 0.2025$$
$$p_X(2) = \frac{49}{2} \times \frac{0.05}{0.95} \times p_X(1) = 0.2611$$
$$p_X(3) = \frac{48}{3} \times \frac{0.05}{0.95} \times p_X(2) = 0.2199$$
$$p_X(4) = \frac{47}{4} \times \frac{0.05}{0.95} \times p_X(3) = 0.1360$$

and, by subtraction,

$$P(X > 4) = 1 - (0.0769 + 0.2025 + \cdots + 0.1360) = 0.1036.$$

(b) The approximating probability distribution is Poisson(2.5). The probabilities are shown for comparison in the following table.

	$P(X=0)$	$P(X=1)$	$P(X=2)$	$P(X=3)$	$P(X=4)$	$P(X \geq 5)$
$B(50, 0.05)$	0.0769	0.2025	0.2611	0.2199	0.1360	0.1036
Poisson(2.5)	0.0821	0.2052	0.2565	0.2138	0.1336	0.1088

(c) The probabilities are 'similar', but are not really very close—certainly, not as close as in some previous exercises and examples. The parameter $p = 0.05$ is at the limit of our 'rule' for when the approximation will be useful (and, in some previous examples, n has been counted in hundreds, not in tens).

Solution 4.7

(a) You should have observed something like the following. The computer gave the random sample

9 7 6 8 6.

The sample mean is

$$\frac{9+7+6+8+6}{5} = \frac{36}{5} = 7.2,$$

resulting in an estimate of 7.2 for the population mean μ (usually unknown, but in this case known to be equal to 8).

(b) From 100 repetitions of this experiment, the observed sample means ranged from as low as 4.9 to as high as 11.6, with frequencies as follows.

[4, 5)	1
[5, 6)	12
[6, 7)	14
[7, 8)	25
[8, 9)	25
[9, 10)	19
[10, 11)	3
[11, 12)	1

(c) A histogram of the distribution of sample means is shown in Figure S4.2. The data vector had mean 7.96 and variance 1.9.

(d) Repeating the experiment for samples of size 50 gave the following results. Observed sample means ranged from 6.90 to 9.46, with frequencies

$[6.5, 7)$	2
$[7, 7.5)$	10
$[7.5, 8)$	41
$[8, 8.5)$	39
$[8.5, 9)$	7
$[9, 9.5)$	1

and corresponding histogram as shown in Figure S4.1. The data vector had mean 7.9824 and variance 0.2. What has happened is that the sample means based on samples of size 50 (rather than 5) are much more contracted about the value $\mu = 8$. A single experiment based on a sample of size 50 is likely to give an estimate of μ that is closer to 8 than it would have been in the case of an experiment based on a sample of size 5.

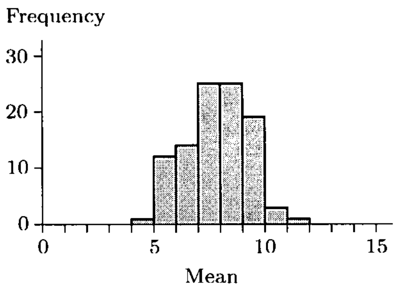


Figure S4.1

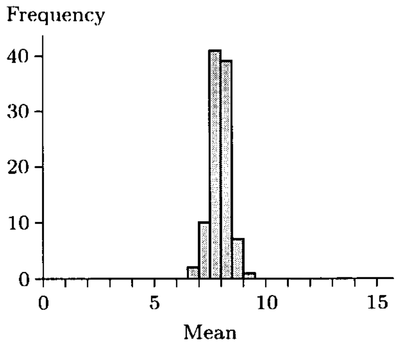


Figure S4.2

Solution 4.8

(a) If X (chest circumference measured in inches) has mean 40, then the random variable $Y = 2.54X$ (chest circumference measured in cm) has mean

$$E(Y) = E(2.54X) = 2.54E(X) = 2.54 \times 40 = 101.6.$$

(Here, the formula $E(aX + b) = aE(X) + b$ is used, with $a = 2.54$ and $b = 0$.)

(b) If X (water temperature measured in degrees Celsius) has mean 26, then the random variable $Y = 1.8X + 32$ (water temperature measured in °F) has mean

$$E(Y) = E(1.8X + 32) = 1.8E(X) + 32 = 1.8 \times 26 + 32 = 78.8.$$

Solution 4.9

If X (finger length in cm) has mean 11.55 and standard deviation 0.55, and if the random variable finger length (measured in inches) is denoted by Y, then $Y = X/2.54$, hence

$$\mu_Y = \frac{\mu_X}{2.54} = \frac{11.55}{2.54} = 4.55, \qquad \sigma_Y = \frac{\sigma_X}{2.54} = \frac{0.55}{2.54} = 0.22.$$

Solution 4.10

The probability distribution for the outcome of throws of a *Double-Five* is as follows.

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

The expected value of X^2 is given by

$$\begin{aligned}E(X^2) &= 1^2 \times \tfrac{1}{6} + 3^2 \times \tfrac{1}{6} + 4^2 \times \tfrac{1}{6} + 5^2 \times \tfrac{1}{3} + 6^2 \times \tfrac{1}{6}\\ &= 1 \times \tfrac{1}{6} + 9 \times \tfrac{1}{6} + 16 \times \tfrac{1}{6} + 25 \times \tfrac{1}{3} + 36 \times \tfrac{1}{6}\\ &= \tfrac{1}{6} + \tfrac{9}{6} + \tfrac{16}{6} + \tfrac{25}{3} + \tfrac{36}{6}\\ &= \tfrac{112}{6},\end{aligned}$$

and so from the formula (4.16) the variance of X is given by

$$V(X) = E(X^2) - (E(X))^2 = \frac{112}{6} - 4^2 = \frac{16}{6} = 2.67$$

as before.

Solution 4.11

If X is binomial with parameters $n = 4$, $p = 0.4$ then according to (4.17) the mean of X is

$$E(X) = np = 4 \times 0.4 = 1.6$$

and the variance of X is

$$V(X) = npq = 4 \times 0.4 \times 0.6 = 0.96.$$

From the individual probabilities for X, it follows that

$$\begin{aligned}E(X^2) &= 0^2 \times 0.1296 + 1^2 \times 0.3456 + \cdots + 4^2 \times 0.0256\\ &= 0 + 0.3456 + 1.3824 + 1.3824 + 0.4096 = 3.52,\end{aligned}$$

and so

$$V(X) = E(X^2) - (E(X))^2 = 3.52 - 1.6^2 = 3.52 - 2.56 = 0.96,$$

confirming the result obtained previously.

Solution 4.12

A time interval of four years includes one leap year—1461 days altogether. The probability of a lull exceeding 1461 days is

$$\left(1 - \frac{62}{27\,107}\right)^{1461} \simeq 0.0352;$$

so, in a list of 62 waiting times one might expect about two of them to exceed 1461 days. In this case there were exactly two such lulls, one of which lasted 1617 days, and the other, already identified, was of 1901 days' duration.

Solution 4.13

Set the parameter λ equal to $1/437$.

(a) A time interval of three years including one leap year will last 1096 days altogether. The probability that no earthquake occurs during this interval is

$$P(T > t) = e^{-\lambda t} = e^{-1096/437} = e^{-2.508} = 0.0814.$$

(b) The equation $F(x) = \frac{1}{2}$ may be written

$$1 - e^{-x/437} = \tfrac{1}{2},$$

or

$$e^{-x/437} = \tfrac{1}{2},$$

or

$$x = -437 \log \tfrac{1}{2} = 437 \log 2 = 303 \text{ days}.$$

(c) The proportion of waiting times lasting longer than expected is

$$P(T > 437) = e^{-437/437} = e^{-1} = 0.368;$$

thus just over one-third of waiting times are longer than average!

Solution 4.14

If $X \sim \text{Poisson}(8.35)$ then $p(0) = 0.0002$, $p(1) = 0.0020$, $p(2) = 0.0082$ and $p(3) = 0.0229$. So,

(a) the probability of exactly two earthquakes is 0.0082;

(b) the probability that there will be at least four earthquakes is

$$1 - (0.0002 + 0.0020 + 0.0082 + 0.0229) = 1 - 0.0333 = 0.9667.$$

Solution 4.15

The general median waiting time is the solution of the equation

$$F(x) = 1 - e^{-\lambda x} = \tfrac{1}{2},$$

or

$$x = \frac{-\log \frac{1}{2}}{\lambda} = \frac{\log 2}{\lambda} = \mu_T \times \log 2 = 0.6931\mu_T,$$

where μ_T is the mean waiting time. So for an exponential random variable the median is approximately 70% of the mean.

Solution 4.16

(a) You will probably have got something not too different to this. The simulation can be shown on a table as follows. There are 7300 days in twenty years, so the simulation has to be extended up to or beyond 7300 days. When we start, we do not know how many random numbers that will take, so we just have to keep going. Waiting times are drawn from the exponential distribution with mean 437.

Earthquake	Waiting time	Cumulative time
1	695.4	695.4
2	279.2	974.6
3	685.2	1659.8
4	420.8	2080.6
5	758.3	2838.9
6	385.3	3224.2
7	156.0	3380.2
8	681.7	4061.9
9	334.5	4396.4
10	370.1	4766.5
11	557.7	5324.2
12	381.2	5705.4
13	901.8	6607.2
14	179.0	6786.2
15	92.9	6879.1
16	851.6	7730.7
17		
18		

The 16th earthquake happened just after the twenty years time limit. A diagram of the incidence of earthquakes with passing time is shown in Figure S4.3.

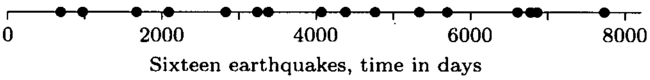


Figure S4.3 Incidence of earthquakes (simulated)

(b) There are 15 earthquakes counted in the simulation. The expected number was

$$\lambda t = \left(\frac{1}{437 \text{ days}}\right) \times (7300 \text{ days}) = 16.7.$$

The number of earthquakes is an observation on a Poisson random variable with mean 16.7. The median of the Poisson(16.7) distribution is 17. (For, if $X \sim \text{Poisson}(16.7)$, then $F(16) = P(X \leq 16) = 0.4969$, while $F(17) = 0.5929$.)

Solution 4.17

(a) A histogram of the data is given in Figure S4.4. The data are very skewed and suggest that an exponential model might be plausible.

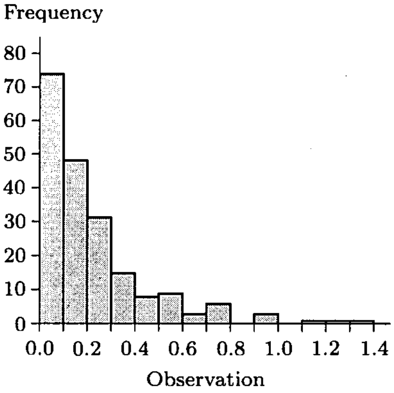


Figure S4.4

(b) (i) The sample mean is 0.224 and the sample median is 0.15. So the sample median is about 67% of the sample mean, mimicking corresponding properties of the exponential distribution (69%).

(ii) The sample standard deviation is 0.235 which is close to the sample mean. (For the exponential distribution, the mean and standard deviation are equal.)

(c) The c.d.f. of the exponential distribution is given by

$$F(x) = 1 - e^{-\lambda t} = 1 - e^{-t/\mu}, \quad t \geq 0$$

and so the lower quartile is the solution of the equation

$$F(x) = 1 - e^{-x/\mu} = 0.25.$$

That is,

$$e^{-x/\mu} = 0.75,$$

so

$$x/\mu = -\log 0.75;$$

so

$$q_L = -\mu \log 0.75 = 0.29\mu.$$

Similarly,

$$q_U = 1.39\mu.$$

For these data, the sample lower quartile is 0.06, which is 0.27 times the sample mean, and the sample upper quartile is 0.29, which is 1.29 times the sample mean. The similarity to corresponding properties of exponential distribution is fairly marked.

(d) During the first quarter-minute there were 81 pulses; during the second there were 53.

(e) It looks as though it might be reasonable to model the incidence of nerve pulses as a Poisson process, with mean waiting time between pulses estimated by the sample mean

$$\bar{t} = 0.2244 \text{ seconds.}$$

Then the pulse rate λ may be estimated by

$$\frac{1}{\bar{t}} = 4.456 \text{ per second.}$$

Over quarter-minute (15-second) intervals the expected number of pulses is

$$(4.456 \text{ per second}) \times (15 \text{ seconds}) = 66.8,$$

and so our two observations 81 and 53 are observations on the Poisson distribution with mean 66.8.

Solution 4.18

(a) Your simulation may have given something like the following. Twenty observations from the Poisson distribution Poisson(3.2) were

4	6	2	3	3	2	8	1	3	4
0	4	4	4	6	3	4	3	7	2

with frequencies as follows.

Count	Frequency
0	1
1	1
2	3
3	5
4	6
5	0
6	2
7	1
8	1

For a random sample of size 50 a typical frequency table is given by

Count	Frequency
0	3
1	8
2	11
3	13
4	6
5	4
6	4
7	1

and for a random sample of size 100 the frequencies are as follows.

Count	Frequency
0	6
1	13
2	26
3	16
4	19
5	13
6	4
7	1
8	2

(b) For a sample of size 1000 the sample relative frequencies were as shown below. These may be compared with the probability mass function

$$p(n) = \frac{e^{-3.2}3.2^n}{n!}, \quad n = 0, 1, 2, 3, \ldots.$$

Count	Frequency	Relative frequency	Probability
0	31	0.031	0.0408
1	146	0.146	0.1304
2	194	0.194	0.2087
3	236	0.236	0.2226
4	168	0.168	0.1781
5	124	0.124	0.1140
6	61	0.061	0.0608
7	25	0.025	0.0278
8	11	0.011	0.0111
9	2	0.002	0.0040
10	2	0.002	0.0013
11	0	0.000	0.0004
12	0	0.000	0.0001
13	0	0.000	0.0000

(Notice the small rounding error in the assessment of the probabilities in the fourth column. They add to 1.0001.)

Solution 4.19

(a) One simulation gave $x = 59$ for the number of males, and therefore (by subtraction) 41 females.

(b) The number of colour-deficient males present is therefore a random observation from $B(59, 0.06)$: this simulation gave $y_1 = 3$. The number of colour-deficient females is a random observation from $B(41, 0.004)$. This simulation gave $y_2 = 0$.

(c) The resulting observation on the random variable W is

$$w = y_1 + y_2 = 3 + 0 = 3.$$

(d) The expected number of males is 50, equal to the expected number of females. Intuitively, the expected number of colour-deficient males is $50 \times 0.06 = 3$; the expected number of colour-deficient females is $50 \times 0.004 = 0.2$. The expected number of colour-deficient people is $3 + 0.2 = 3.2$. This result is, as it happens, correct, though quite difficult to confirm formally: no attempt will be made to do so here.

(e) Repeating the exercise gave a data vector of 1000 observations on W with the following frequencies.

Count	Frequency
0	30
1	137
2	186
3	243
4	182
5	115
6	63
7	28
8	12
9	3
10	0
11	0
12	1

This data set has mean $\overline{w} = 3.25$ and standard deviation $s = 1.758$.

Solution 4.20

(a) If the mean arrival rate is $\lambda = 12$ claims per week this is equivalent to

$$\frac{12}{7 \times 24} = \frac{1}{14}$$

claims per hour. So the mean waiting time between claim arrivals is 14 hours. By adding together 20 successive observations from the exponential distribution with mean 14, the twenty arrival times may be simulated. You might have got something like the following.

Claim number	Waiting time	Arrival time	Approximation
1	4.0	4.0	4 am, Mon
2	13.2	17.2	5 pm, Mon
3	3.3	20.5	9 pm, Mon
4	44.3	64.8	5 pm, Wed
5	17.3	82.1	10 am, Thu
6	6.0	88.1	4 pm, Thu
7	4.7	92.8	9 pm, Thu
8	4.0	96.8	1 am, Fri
9	3.2	100.0	4 am, Fri
10	11.7	111.7	4 pm, Fri
11	25.5	137.2	5 pm, Sat
12	33.3	170.5	3 am, Mon
13	1.3	171.8	4 am, Mon
14	0.5	172.3	4 am, Mon
15	4.9	177.2	9 am, Mon
16	2.7	179.9	12 *noon*, Mon
17	5.5	185.4	5 pm, Mon
18	3.7	189.1	9 pm, Mon
19	30.7	219.8	4 am, Wed
20	3.6	223.4	7 am, Wed

(b) Ten weeks of simulated claims gave 8 claims in the first week, 18 in the second and 14 in the third. You should have observed a continuing sequence with similar numbers. These are all observations on a Poisson random variable with mean 14.

Chapter 5

Solution 5.1

In Figure 5.4(a), $\mu = 100$; it looks as though $\mu + 3\sigma$ is about 150; so σ is about 17. In Figure 5.4(b), $\mu = 100$ and $\mu + 3\sigma$ is about 115: therefore, the standard deviation σ looks to be about 5. In Figure 5.4(c), $\mu = 72$ and σ is a little more than 1; and in Figure 5.4(d), $\mu = 1.00$ and σ is about 0.05.

Solution 5.2

(a) $P(Z \leq 2)$.

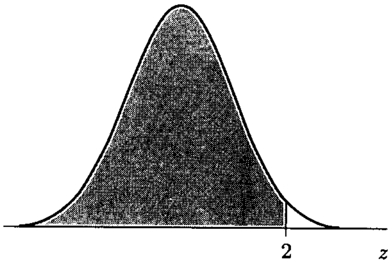


Figure S5.1

(b) $P(Z > 1)$

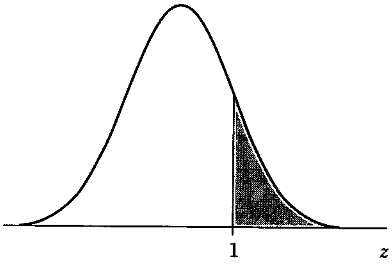


Figure S5.2

(c) $P(-1 < Z \leq 1)$

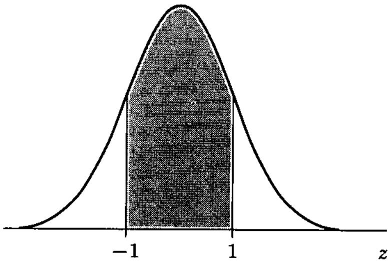


Figure S5.3

(d) $P(Z \leq -2)$

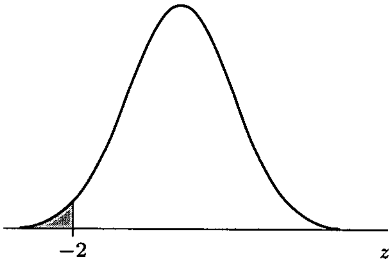


Figure S5.4

Solution 5.3

(a) Writing $X \sim N(2.60, 0.33^2)$, where X is the enzyme level present in individuals suffering from acute viral hepatitis, the proportion of sufferers whose measured enzyme level exceeds 3.00 is given by

$$P(X > 3.00) = P\left(Z > \frac{3.00 - 2.60}{0.33}\right)$$

(writing $z = (x - \mu)/\sigma$). This probability reduces to

$$P\left(Z > \frac{0.40}{0.33}\right) = P(Z > 1.21)$$

and is represented by the shaded area in Figure S5.5.

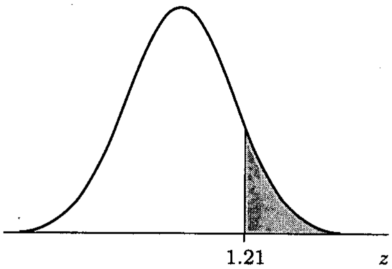


Figure S5.5

(b) Writing $Y \sim N(2.65, 0.44^2)$, where Y is the enzyme level in individuals suffering from aggressive chronic hepatitis, the proportion required is given by the probability

$$\begin{aligned} P(Y < 1.50) &= P\left(Z < \frac{1.50 - 2.65}{0.44}\right) \\ &= P\left(Z < -\frac{1.15}{0.44}\right) \\ &= P(Z < -2.61); \end{aligned}$$

this (quite small) proportion is given by the shaded area in Figure S5.6.

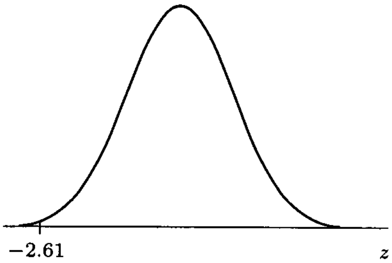


Figure S5.6

(c) The sample mean and sample standard deviation are

$$\overline{x} = 1.194 \quad \text{and} \quad s = 0.290.$$

The lower extreme (0.8 mm) may be written in standardized form as

$$z = \frac{x - \mu}{\sigma} = \frac{0.8 - 1.194}{0.290} = -1.36;$$

and the upper extreme (1.2 mm) as

$$z = \frac{x - \mu}{\sigma} = \frac{1.2 - 1.194}{0.290} = 0.02.$$

The proportion of ball-bearings whose diameter is between 0.8 mm and 1.2 mm can be shown on a sketch of the standard normal density as in Figure S5.7.

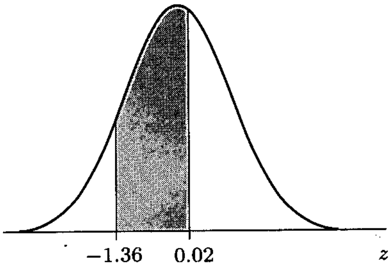


Figure S5.7

Solution 5.4

(a) The probability $P(Z \leq 1.00)$ is to be found in the row for $z = 1.0$ and in the column headed 0: this gives $P(Z \leq 1.00) = 0.8413$. This is shown in Figure S5.8.

(b) The probability $P(Z \leq 1.96)$ is given in the row for $z = 1.9$ and in the column headed 6: $P(Z \leq 1.96) = 0.9750$. This is illustrated in Figure S5.9.

(c) The probability $P(Z \leq 2.25)$ is to be found in the row for $z = 2.2$ and in the column headed 5: that is, $P(Z \leq 2.25) = 0.9878$. This probability is given by the shaded area in Figure S5.10.

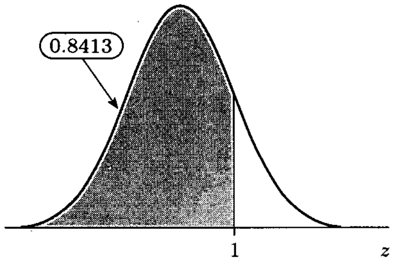


Figure S5.8

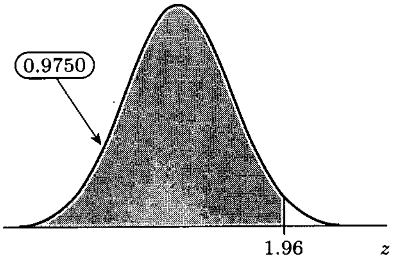


Figure S5.9

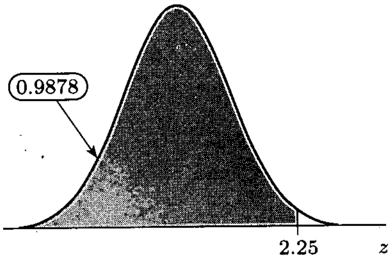


Figure S5.10

Solution 5.5

(a) First, sketch the standard normal density, showing the critical points $z = -1.33$ and $z = 2.50$. From the tables, $P(Z \leq 2.50) = 0.9938$ and so $P(Z > 2.50) = 0.0062$; by symmetry, $P(Z \leq -1.33) = P(Z \geq 1.33) = 1 - 0.9082 = 0.0918$. By subtraction, the probability required is

$$1 - 0.0062 - 0.0918 = 0.9020.$$

(b) From the tables, $P(Z \geq 3.00) = 1 - 0.9987 = 0.0013$. By symmetry,

$$P(-3.00 \leq Z \leq 3.00) = 1 - 0.0013 - 0.0013 = 0.9974.$$

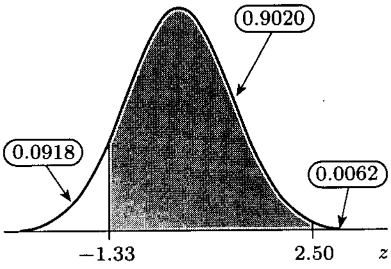


Figure S5.11

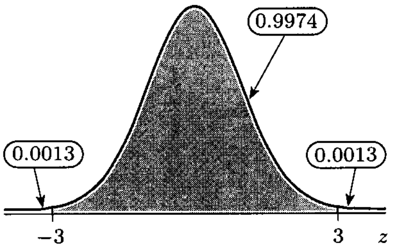


Figure S5.12

(c) First, sketch the standard normal density, showing the critical points $z = 0.50$ and $z = 1.50$. The probability $P(Z \leq 0.50)$ is 0.6915; the probability $P(Z \leq 1.50)$ is 0.9332. By subtraction, therefore, the probability required is

$$P(0.50 \leq Z \leq 1.50) = 0.9332 - 0.6915 = 0.2417.$$

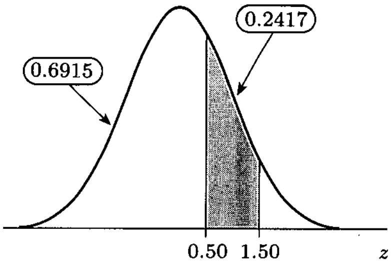


Figure S5.13

Solution 5.6

(a) The probability $P(|Z| \leq 1.62)$ is given by the shaded area in Figure S5.14. From the tables, $P(Z \geq 1.62) = 1 - P(Z \leq 1.62) = 1 - 0.9474 = 0.0526$, so the probability required is

$$P(|Z| \leq 1.62) = 1 - 0.0526 - 0.0526 = 0.8948.$$

(b) The probability $P(|Z| \geq 2.45)$ is given by the sum of the two shaded areas in Figure S5.15. From the tables, $P(Z \geq 2.45) = 1 - P(Z \leq 2.45) = 1 - 0.9929 = 0.0071$, so the total probability is $2 \times 0.0071 = 0.0142$.

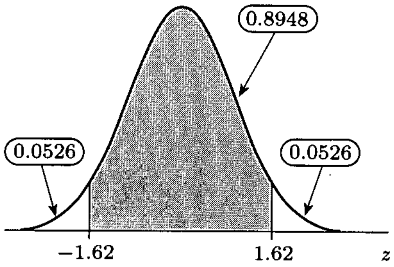


Figure S5.14

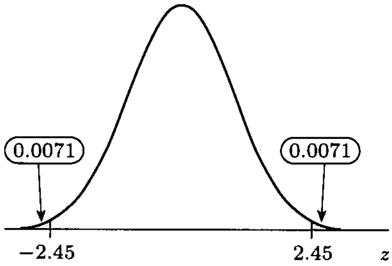


Figure S5.15

Solution 5.7

(a) The proportion within one standard deviation of the mean is given by $P(|Z| \leq 1)$, shown in Figure S5.16. Since

$$P(Z > 1) = 1 - P(Z \leq 1) = 1 - 0.8413 = 0.1587,$$

the answer required is $1 - 0.1587 - 0.1587 = 0.6826$: that is, nearly 70% of a normal population are within one standard deviation of the mean.

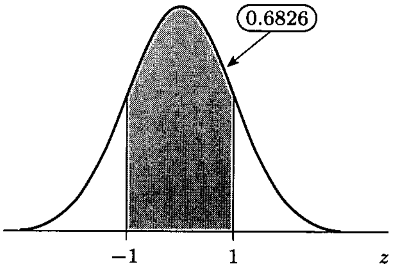


Figure S5.16

(b) Here we require the probability $P(|Z| > 2)$. Since

$$P(Z > 2) = 1 - P(Z \leq 2) = 1 - 0.9772 = 0.0228,$$

this is $2 \times 0.0228 = 0.0456$.

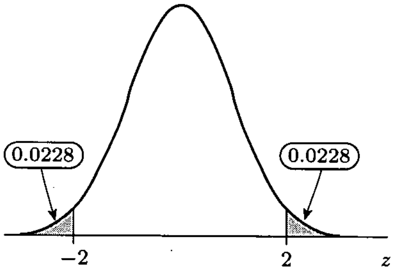


Figure S5.17

Less than 5% of a normal population are more than two standard deviations away from the mean.

Solution 5.8

We are told that $X \sim N(40, 4)$ ($\mu = 40, \sigma^2 = 4$; so $\sigma = 2$). The probability required is

$$P(37 \le X \le 42) = P\left(\frac{37-40}{2} \le Z \le \frac{42-40}{2}\right) = P(-1.50 \le Z \le 1.00)$$

shown in Figure S5.18. The probability required is

$$P(-1.50 \le Z \le 1.00) = 0.7745.$$

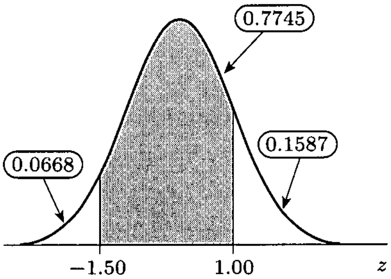


Figure S5.18

Solution 5.9

Writing $A \sim N(0, 2.75)$, the probability required is $P(0 < A < 2)$.

Subtracting the mean μ and dividing by the standard deviation σ, using $\mu = 0$ and $\sigma = \sqrt{2.75}$, this may be rewritten in terms of Z as

$$P\left(\frac{0-0}{\sqrt{2.75}} < Z < \frac{2-0}{\sqrt{2.75}}\right) = P(0 < Z < 1.21).$$

From the tables, $\Phi(1.21) = 0.8869$, so the probability required is

$$0.8869 - 0.5 = 0.3869.$$

Solution 5.10

(a) If T is $N(315, 17\,161)$ then a sketch of the distribution of T is given in Figure S5.19.

(b) Standardizing gives

$$P(T < 300) = P\left(Z < \frac{300-315}{131}\right) = P(Z < -0.11) = 0.4562.$$

This is shown in Figure S5.20.

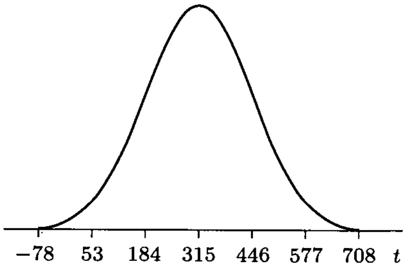


Figure S5.19

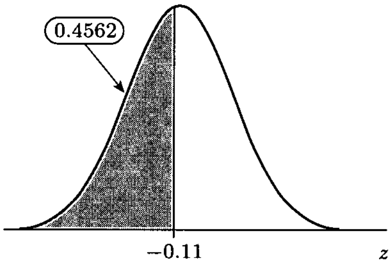


Figure S5.20

(c) $$P(300 \le T \le 500) = P\left(\frac{300-315}{131} \le Z \le \frac{500-315}{131}\right)$$
$$= P(-0.11 \le Z \le 1.41).$$

This is shown in Figure S5.21.The area of the shaded region is 0.4645.

(d) First, we need

$$P(T > 500) = P(Z > 1.41) = 1 - 0.9207 = 0.0793.$$

The number of smokers with a nicotine level higher than 500 in a sample of 20 smokers has a binomial distribution $B(20, 0.0793)$. The probability that at most one has a nicotine level higher than 500 is

$$p_0 + p_1 = (0.9207)^{20} + 20(0.9207)^{19}(0.0793) = 0.1916 + 0.3300 = 0.52.$$

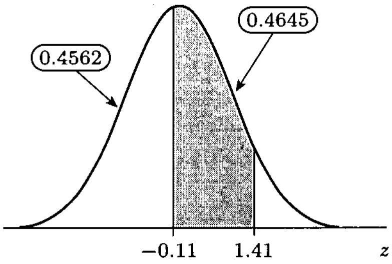


Figure S5.21

Solution 5.11

By symmetry, $q_{0.2} = -q_{0.8} = -0.842$ for the standard normal distribution, and $q_{0.4} = -q_{0.6} = -0.253$. Assuming IQ scores to be normally distributed with mean 100 and standard deviation 15, then

$$q_{0.2} = 100 - 0.842 \times 15 = 87.4$$
$$q_{0.4} = 100 - 0.253 \times 15 = 96.2$$
$$q_{0.6} = 100 + 0.253 \times 15 = 103.8$$
$$q_{0.8} = 100 + 0.842 \times 15 = 112.6$$

and these quantiles are illustrated in Figure S5.22.

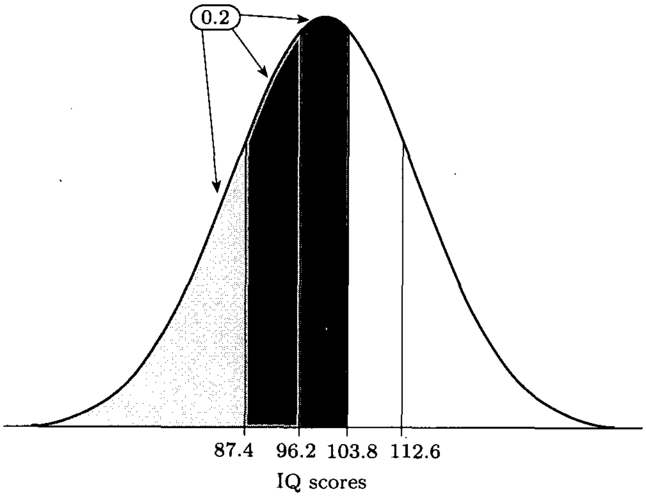


Figure S5.22

Solution 5.12

(a) Most statistical computer programs should be able to furnish standard normal probabilities and quantiles. The answers might be different in the fourth decimal place to those furnished by the tables when other than simple calculations are made.

(i) 0.0446 (ii) 0.9641 (iii) 0.9579 (iv) 0.0643

(v) $q_{0.10} = -q_{0.90} = -1.2816$ (vi) $q_{0.95} = 1.6449$

(vii) $q_{0.975} = 1.9600$ (viii) $q_{0.99} = 2.3263$

(b) The distribution of X is normal $N(100, 225)$. Most computers should return non-standard normal probabilities routinely, taking the distribution parameters as function arguments, and insulating users from the requirements to re-present problems in terms of the standard normal distribution.

(i) 0.0478 (ii) 0.1613 (iii) 100 (iv) 119.2 (v) 80.8

(c) (i) 0.1587 (ii) 166.22 cm

(iii) The first quartile is $q_L = 155.95$; the third quartile is $q_U = 164.05$; the interquartile range is given by the difference $q_U - q_L = 8.1$ cm.

(iv) 0.3023

(d) (i) 0.1514 (ii) 530.48 (iii) 0.6379

(iv) This question asks 'What proportion of smokers have nicotine levels within 100 units of the mean of 315?'. Formally,

$$P(|T-315| \leq 100) = P(-100 \leq T - 315 \leq 100) = P(215 \leq T \leq 415)$$

which is 0.5548.

(v) $q_{0.80} = 425.25$

(vi) The range of levels is that covered by the interval $(q_{0.04}, q_{0.96})$ allowing 4% either side. This is (85.7, 544.3).

(vii)

$$\begin{aligned} &P(215 < T < 300) + P(350 < T < 400) \\ &= 0.231\,794 + 0.136\,451 = 0.3682. \end{aligned}$$

Solution 5.13

Your solution might have gone something like the following.

(a) The first sample of size 5 from Poisson(8) consisted of the list 6, 7, 3, 8, 4. This data set has mean $\overline{x}_5 = \frac{1}{5}(6+7+3+8+4) = 5.6$. When a vector of 100 observations on $\overline{X}_5$ was obtained, the observed frequencies of different observations were as follows.

[5, 6)	4
[6, 7)	25
[7, 8)	27
[8, 9)	28
[9, 10)	10
[10, 11)	6

So there were 90 observed in [6, 10).

(b) The 100 observations on $\overline{X}_{20}$ were distributed as follows. (Your results will be somewhat different.)

[6, 7)	8
[7, 8)	42
[8, 9)	43
[9, 10)	6
[10, 11)	1

So all the observations but one were in [6, 10), and 85 of the 100 were in [7, 9).

(c) All the 100 observations were in [7, 9).

(d) The larger the sample size, the less widely scattered around the population mean $\mu = 8$ the observed sample means were. In non-technical language, 'larger samples resulted in sample means that were more precise estimates of the population mean'.

Solution 5.14

The exponential distribution is very skewed, and you might have expected more scatter in the observations. This was apparent in the distributions of the sample means. For samples of size 5 the following observations were obtained on $\overline{X}_5$. (Remember, $\overline{X}_5$ estimates the population mean, $\mu = 8$.)

$[0, 5)$	18
$[5, 10)$	56
$[10, 15)$	22
$[15, 20)$	3
$[20, 25)$	1

The largest observation was $\overline{x}_5 = 21.42$. Nevertheless, it is interesting to observe that the distribution of observations on $\overline{X}_5$ peaks not at the origin but somewhere between 5 and 10.

For samples of size 20 the following distribution of observations on $\overline{X}_{20}$ was obtained.

$[4, 6)$	9
$[6, 8)$	37
$[8, 10)$	39
$[10, 12)$	12
$[12, 14)$	3

These observations are peaked around the point 8.

Finally, for samples of size 80 the following observations on $\overline{X}_{80}$ were obtained.

$[6, 7)$	11
$[7, 8)$	42
$[8, 9)$	33
$[9, 10)$	12
$[10, 11)$	2

Solution 5.15

(a) (i) The following typical 100 observations on $\overline{X}_2$ resulted in a histogram almost as skewed as the distribution of X itself.

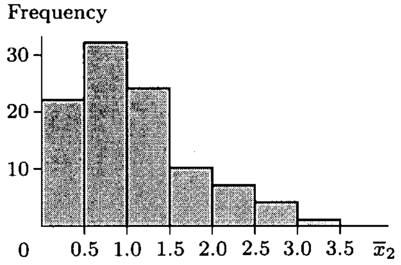


Figure S5.23

(ii) The histogram for 100 observations on $\overline{X}_{30}$ is given in Figure S5.24.

(iii) The histogram of part (ii) is evidently more symmetric than that of part (i), and it appears that a normal density might provide a usable approximation to the distribution of $\overline{X}_{30}$.

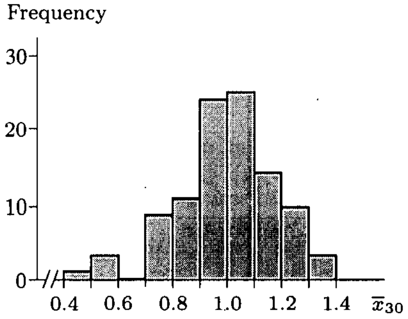


Figure S5.24

(b) (i) The simulation of 100 observations on $\overline{X}_2$ resulted in a histogram that was roughly triangular over $[0, 2]$, and very different to that obtained at part (a)(i).

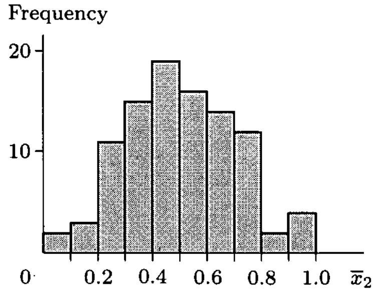


Figure S5.25

(ii) For means of samples of size 30, the following histogram was obtained.

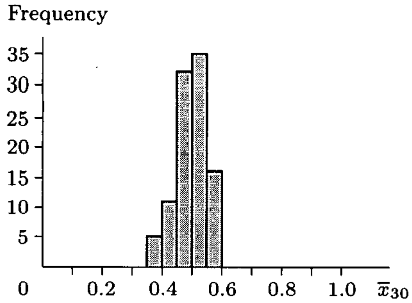


Figure S5.26

(iii) The differences between the histograms of parts (b)(i) and (b)(ii) are not so apparent. The distribution in both cases peaks at 1 and is roughly symmetric. However, notice that for samples of size 30 there is much less variation in the distribution of the sample mean.

Solution 5.16

It does not matter that there is no proposed model for the duration of individual patients' visits: the numbers $\mu = 20$ minutes and $\sigma = 15$ minutes provide sufficient information. By the central limit theorem, the total time T required of the dentist by the 12 patients is approximately normally distributed with mean

$$12 \times 20 = 240$$

and variance

$$12 \times 15^2 = 2700.$$

The probability that T will be less than 3 hours (that is, 180 minutes) is

$$P(T < 180) = P\left(Z < \frac{180 - 240}{\sqrt{2700}}\right) = P(Z < -1.15)$$

or $1 - 0.8749 = 0.1251$ from the tables. So she will only be free at 12 with probability about $\frac{1}{8}$.

(If you used a computer for the normal probability without the intermediate calculation $Z = -1.15$, you would have obtained the solution 0.1241.)

Solution 5.17

If the error in any individual transaction is written W, then $W \sim U(-\frac{1}{2}, \frac{1}{2})$. The mean of W is $E(W) = 0$, by symmetry, and the variance of W is $V(W) = \frac{1}{12}\left(\frac{1}{2} - (-\frac{1}{2})\right)^2 = \frac{1}{12}$.

The accumulated error in 400 transactions is given by the sum

$$S = W_1 + W_2 + \cdots + W_{400}.$$

By the central limit theorem, S has mean

$$\mu = 400 \times 0 = 0,$$

and variance

$$\sigma^2 = \frac{400}{12} = 33.333,$$

and is approximately normally distributed.

The probability that after 400 transactions her estimate of her bank balance is less than ten pounds in error is

$$P(-10 < S < 10) = P\left(\frac{-10-0}{\sqrt{33.333}} < Z < \frac{10-0}{\sqrt{33.333}}\right) = P(-1.73 < Z < 1.73).$$

This probability is given by the shaded area in Figure S5.27.

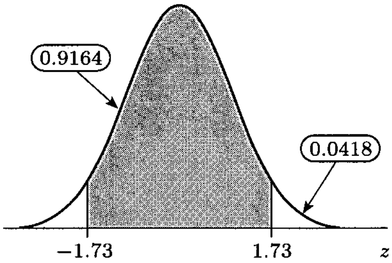


Figure S5.27

Since $P(Z > 1.73) = 1 - 0.9582 = 0.0418$, the probability required is

$$1 - 2 \times 0.0418 = 0.9164$$

(or rather more than 90%).

Solution 5.18

The probability

$$P(12 \leq X \leq 15)$$

can be approximated by the probability

$$P\left(11\tfrac{1}{2} \leq Y \leq 15\tfrac{1}{2}\right),$$

where $Y \sim N(8, 4)$; this is the same as

$$P\left(\frac{11\frac{1}{2} - 8}{2} \leq Z \leq \frac{15\frac{1}{2} - 8}{2}\right) = P(1.75 \leq Z \leq 3.75)$$

and (from the tables) this is $0.9999 - 0.9599 = 0.0400$. The actual binomial probability is 0.0384.

Solution 5.19

(a) (i) 0.164 538 (ii) 0.182 820 (iii) 0.165 408 (iv) 0.124 056

(b) 0.472 284

(c) Since $np = \frac{25}{4} = 6.25$, $npq = \frac{75}{16} = 4.6875$, then $\mu = 6.25$, $\sigma^2 = 4.6875$.

(d) (i) $P(X = 6) \simeq P\left(5\frac{1}{2} \leq Y \leq 6\frac{1}{2}\right) = 0.181\,446$. This may be compared with the binomial probability 0.182 820.

(ii) $P(X = 7) \simeq P\left(6\frac{1}{2} \leq Y \leq 7\frac{1}{2}\right) = 0.172\,185$. This may be compared with the exact binomial probability, which is 0.165 408.

(iii) $P(X = 8) \simeq P\left(7\frac{1}{2} \leq Y \leq 8\frac{1}{2}\right) = 0.132\,503$. The exact binomial probability is 0.124 056.

(iv) $P(6 \leq X \leq 8) \simeq P\left(5\frac{1}{2} \leq Y \leq 8\frac{1}{2}\right) = 0.486\,134$. The corresponding binomial probability is 0.472 284.

Solution 5.20

A computer gives the probability $P(30 \leq X \leq 45)$ when $X \sim$ Poisson(40) as 0.766 421. The central limit theorem permits the approximation

$$P(30 \leq X \leq 45) = P\left(29\tfrac{1}{2} \leq Y \leq 45\tfrac{1}{2}\right)$$

where Y is normal with mean $\mu = 40$ and variance $\sigma^2 = 40$. The right-hand side is given by

$$P\left(\frac{29\frac{1}{2} - 40}{\sqrt{40}} \leq Z \leq \frac{45\frac{1}{2} - 40}{\sqrt{40}}\right)$$
$$= P(-1.66 \leq Z \leq 0.87) = 0.8078 - (1 - 0.9515) = 0.7593$$

from the tables. Directly from a computer, we would obtain 0.759 310. This approximation is reasonable.

Chapter 6

Solution 6.1

For any random sample taken from a population with mean μ and variance σ^2, the sample mean $\overline{X}$ has mean and variance

$$E(\overline{X}) = \mu, \qquad V(\overline{X}) = \frac{\sigma^2}{n},$$

where n is the sample size. When the population is Poisson, the variance σ^2 is equal to the mean μ, so

$$E(\overline{X}) = \mu, \qquad V(\overline{X}) = \frac{\mu}{n}.$$

The random variable $\overline{X}_{(1)}$ was based on samples of size 103, the random variable $\overline{X}_{(2)}$ on samples of size 48. So

$$E(\overline{X}_{(1)}) = E(\overline{X}_{(2)}) = \mu;$$

but

$$V(\overline{X}_{(1)}) = \frac{\mu}{103}, \qquad V(\overline{X}_{(2)}) = \frac{\mu}{48}.$$

The larger the sample taken, the smaller the variance in the sample mean.

Solution 6.2

The mean of the random variable Y_i is given by

$$E(Y_i) = E(\alpha + \beta x_i + W_i),$$

which is of the form $E(\text{constant} + W_i)$, since α, β and x_i are all constants. This is therefore

$$E(Y_i) = \alpha + \beta x_i + E(W_i),$$

and from our assumption that W_i has mean 0 it follows that

$$E(Y_i) = \alpha + \beta x_i,$$

for all $i = 1, 2, \ldots, 6$.

Similarly, the variance of Y_i is given by

$$V(Y_i) = V(\text{constant} + W_i) = V(W_i) = \sigma^2,$$

for all $i = 1, 2, \ldots, 6$.

Solution 6.3

(a) The midpoint of the coordinates (x_1, Y_1) and (x_2, Y_2) is the point $\left(\frac{1}{2}(x_1 + x_2), \frac{1}{2}(Y_1 + Y_2)\right)$.

The midpoint of the coordinates (x_5, Y_5) and (x_6, Y_6) is the point $\left(\frac{1}{2}(x_5 + x_6), \frac{1}{2}(Y_5 + Y_6)\right)$.

The slope of the line joining the two midpoints is

$$\widehat{\beta}_2 = \frac{\frac{1}{2}(Y_5 + Y_6) - \frac{1}{2}(Y_1 + Y_2)}{\frac{1}{2}(x_5 + x_6) - \frac{1}{2}(x_1 + x_2)} = \frac{Y_5 + Y_6 - Y_1 - Y_2}{x_5 + x_6 - x_1 - x_2}.$$

The centre of gravity of the points (x_1, Y_1), (x_2, Y_2) and (x_3, Y_3) is $\left(\frac{1}{3}(x_1 + x_2 + x_3), \frac{1}{3}(Y_1 + Y_2 + Y_3)\right)$.

The centre of gravity of the points (x_4, Y_4), (x_5, Y_5) and (x_6, Y_6) is $\left(\frac{1}{3}(x_4 + x_5 + x_6), \frac{1}{3}(Y_4 + Y_5 + Y_6)\right)$.

The slope of the line joining the two centres of gravity is

$$\widehat{\beta}_3 = \frac{\frac{1}{3}(Y_4 + Y_5 + Y_6) - \frac{1}{3}(Y_1 + Y_2 + Y_3)}{\frac{1}{3}(x_4 + x_5 + x_6) - \frac{1}{3}(x_1 + x_2 + x_3)} = \frac{Y_4 + Y_5 + Y_6 - Y_1 - Y_2 - Y_3}{x_4 + x_5 + x_6 - x_1 - x_2 - x_3}.$$

(b) Consequently

$$\begin{aligned} E(\widehat{\beta}_2) &= E\left(\frac{Y_5 + Y_6 - Y_1 - Y_2}{x_5 + x_6 - x_1 - x_2}\right) = \frac{1}{x_5 + x_6 - x_1 - x_2} E(Y_5 + Y_6 - Y_1 - Y_2) \\ &= \frac{1}{x_5 + x_6 - x_1 - x_2}((\alpha + \beta x_5) + (\alpha + \beta x_6) - (\alpha + \beta x_1) - (\alpha + \beta x_2)) \\ &= \frac{1}{x_5 + x_6 - x_1 - x_2}(\beta x_5 + \beta x_6 - \beta x_1 - \beta x_2) = \beta; \end{aligned}$$

and $E(\widehat{\beta}_3)$ reduces to β in a similar way.

Solution 6.4

The variance of the second estimator $\widehat{\beta}_2$ is given by

$$\begin{aligned}V(\widehat{\beta}_2) &= V\left(\frac{Y_5+Y_6-Y_1-Y_2}{x_5+x_6-x_1-x_2}\right)\\ &= \frac{1}{(x_5+x_6-x_1-x_2)^2}V(Y_5+Y_6-Y_1-Y_2)\\ &= \frac{1}{(x_5+x_6-x_1-x_2)^2}(V(Y_5)+V(Y_6)+V(Y_1)+V(Y_2))\\ &= \frac{4\sigma^2}{8^2} = 0.0625\sigma^2.\end{aligned}$$

The variance of $\widehat{\beta}_3$ is

$$\begin{aligned}V(\widehat{\beta}_3) &= V\left(\frac{Y_4+Y_5+Y_6-Y_1-Y_2-Y_3}{x_4+x_5+x_6-x_1-x_2-x_3}\right)\\ &= \frac{1}{(x_4+x_5+x_6-x_1-x_2-x_3)^2}V(Y_4+Y_5+Y_6-Y_1-Y_2-Y_3)\\ &= \frac{1}{(x_4+x_5+x_6-x_1-x_2-x_3)^2}(V(Y_4)+V(Y_5)+V(Y_6)+V(Y_1)+V(Y_2)+V(Y_3))\\ &= \frac{6\sigma^2}{9^2} = 0.0741\sigma^2.\end{aligned}$$

Solution 6.5

(a) There is one unknown parameter here, μ, which is also the mean of the Poisson distribution. Matching sample and population moments gives the estimate $\widehat{\mu} = \overline{x}$; the corresponding estimator for μ is $\widehat{\mu} = \overline{X}$.

(b) In this case the population mean is $1/p$. Matching moments gives $\overline{X} = 1/\widehat{p}$; so $\widehat{p} = 1/\overline{X}$.

(c) Here, there are two unknown parameters, so we shall need to use two sample moments. These are $\overline{X}$, the sample mean, and S^2, the sample variance. Notice the use of the upper-case letter S, implying that like the sample mean, the sample variance is a random variable. Matching moments gives $\widehat{\mu} = \overline{X}$, $\widehat{\sigma}^2 = S^2$.

(d) The mean of the exponential distribution is $1/\lambda$: matching moments gives $\overline{X} = 1/\widehat{\lambda}$; so $\widehat{\lambda} = 1/\overline{X}$.

(e) There is one unknown parameter here. Matching the sample mean to the binomial mean gives $\overline{X} = m\widehat{p}$, so

$$\widehat{p} = \frac{\overline{X}}{m} = \frac{X_1+X_2+\cdots+X_n}{mn}.$$

(This was the 'intuitive' estimate $\widehat{p}$ of p that was used in Example 6.1.)

Solution 6.6

(a) You might have obtained something like this: the 1000 samples of size 2 may be represented as a rectangular array

$$\begin{bmatrix} 0.156 & 0.093 \\ 0.183 & 0.203 \\ 0.066 & 0.168 \\ & \vdots & \\ 0.679 & 0.218 \end{bmatrix}$$

with sample means

$$\begin{bmatrix} 0.124 \\ 0.193 \\ 0.117 \\ \vdots \\ 0.449 \end{bmatrix}.$$

Taking reciprocals gives

$$\begin{bmatrix} 8.05 \\ 5.19 \\ 8.56 \\ \vdots \\ 2.23 \end{bmatrix}$$

which is a data vector of 1000 different independent estimates of λ.

(b) The mean of this data vector is 9.20, close to twice the true value of λ.

(For interest, the experiment was repeated four more times, resulting in four further estimates 11.4, 11.1, 11.8 and 9.6.)

Solution 6.7

(a) The method of moments says, simply, that $\widehat{\mu} = \overline{X}$.

(b) For any random sample, the sample mean $\overline{X}$ has expectation μ, the population mean, so in this case

$$E(\widehat{\mu}) = E(\overline{X}) = \mu;$$

it follows that $\widehat{\mu}$ is unbiased for μ.

(c) Using the same set of 1000 random samples of size 2 as was used in Solution 6.6, our data vector of 1000 different independent estimates of μ is

$$\begin{bmatrix} 0.124 \\ 0.193 \\ 0.117 \\ \vdots \\ 0.449 \end{bmatrix}.$$

(d) This data vector has mean 0.202, close to the true value $\mu = 0.2$.

(For interest, the experiment was repeated four more times, resulting in estimates 0.202 (again), 0.196, 0.195 and 0.201.)

Solution 6.8

The average run length is the sample mean

$$\overline{x} = \frac{1 \times 71 + 2 \times 28 + 3 \times 5 + 4 \times 2 + 5 \times 2 + 6 \times 1}{71 + 28 + 5 + 2 + 2 + 1} = \frac{166}{109} = 1.523;$$

assuming a geometric model with mean $1/p$, the moment estimate of p is

$$\widehat{p} = \frac{1}{\overline{x}} = \frac{109}{166} = 0.657.$$

Solution 6.9

(a) The mean of these data is $\overline{t} = 437.21$, and so the moment estimator for the exponential parameter λ is

$$\widehat{\lambda} = 1/\overline{t} = 0.0023.$$

The units of $\widehat{\lambda}$ are 'earthquakes per day'. We know that the moment estimator is biased:

$$E(\widehat{\lambda}) = \lambda\left(1 + \frac{1}{n-1}\right).$$

However, in this case $n = 62$. The moment estimator may be expected to overestimate the true value of λ by a factor $1 + 1/61 = 1.016$, which is very small.

(b) The moment estimate of μ is $\widehat{\mu} = \overline{t} = 437.21$; the estimator $\widehat{\mu}$ is unbiased. The units of $\widehat{\mu}$ are 'days between earthquakes'.

Solution 6.10

In this case $n = 3$ and $x_{\max} = 13.1$, so

$$\widehat{\theta} = \left(1 + \tfrac{1}{n}\right) x_{\max} = \tfrac{4}{3}(13.1) = 17.5.$$

Solution 6.11

(a) The mean of the Pareto$(100, \theta)$ probability distribution is

$$\mu = \frac{100\theta}{\theta - 1}.$$

(b) The method of moments says that the moment estimator of θ for a sample from the Pareto distribution where $K = 100$ may be found using

$$\overline{X} = \frac{100\widehat{\theta}}{\widehat{\theta} - 1};$$

so

$$\widehat{\theta}\overline{X} - \overline{X} = 100\widehat{\theta};$$

thus, finally,

$$\widehat{\theta} = \frac{\overline{X}}{\overline{X} - 100}.$$

In this case the sample total is $\sum x_i = 3624$ and so the sample mean is $\overline{x} = 3624/30 = 120.8$. The moment estimate of θ is

$$\widehat{\theta} = \frac{\overline{x}}{\overline{x} - 100} = \frac{120.8}{20.8} = 5.81.$$

Solution 6.12

The maximum likelihood estimate of θ is (after a little calculation) $\widehat{\theta}_{ML} = 5.59$; this is to be compared with the moment estimate $\widehat{\theta}_{MM} = 5.81$ obtained in the previous exercise. Numerically there is little to choose between either estimate. We know that maximum likelihood estimators possess good properties. (However, $\widehat{\theta}_{MM}$ was in this case a great deal easier to obtain, and to calculate.)

Solution 6.13

(a) The total number of mice tested is $12 + 20 + \cdots + 20 = 505$; the total number afflicted is $0 + 0 + \cdots + 4 = 43$. The maximum likelihood estimate of p is $\widehat{p} = 43/505 = 0.085$.

(b) The total number of normal *Drosophila* is 419; the total number of vestigial *Drosophila* is 68. The maximum likelihood estimate of the proportion normal is given by

$$\widehat{p} = \frac{419}{419 + 68} = \frac{419}{487} = 0.86.$$

Solution 6.14

The likelihood for the sample observed is

$$\left(\tfrac{1}{2}(1-r)\right)^{147}\left(\tfrac{1}{2}r\right)^{65}\left(\tfrac{1}{2}r\right)^{58}\left(\tfrac{1}{2}(1-r)\right)^{133}$$
$$= \left(\tfrac{1}{2}(1-r)\right)^{147+133}\left(\tfrac{1}{2}r\right)^{65+58} = \left(\tfrac{1}{2}(1-r)\right)^{280}\left(\tfrac{1}{2}r\right)^{123} = \frac{(1-r)^{280}r^{123}}{2^{403}}.$$

This expression is maximized where $(1-r)^{280}r^{123}$ is maximized. This occurs at $\widehat{r} = 0.3052$.

You might have found this using numerical procedures or—perhaps not quite so precisely—by scanning a graph of the function $(1-r)^{280}r^{123}$. Differentiation gives the exact fractional answer, $\widehat{r} = 123/403$.

Solution 6.15

The likelihood of p for the sample of 1469 cars is given by

$$p_1^{902}p_2^{403}p_3^{106}p_4^{38}p_5^{16}P(X \geq 6)^4,$$

where $p_j = P(X = j) = (1-p)^{j-1}p, \quad j = 1, 2, \ldots$. This is

$$p^{902}((1-p)p)^{403}\left((1-p)^2p\right)^{106}\left((1-p)^3p\right)^{38}\left((1-p)^4p\right)^{16}\left((1-p)^5\right)^4$$
$$= p^{1465}(1-p)^{813}.$$

This is maximized at $\widehat{p} = 1465/2278 = 0.6431$. (The exact fraction 1465/2278 was found using differentiation—numerical and graphical techniques should provide an answer close to 0.6431.)

Notice that for these data the sample mean is at least $2282/1469 = 1.553$ (that is, the sample mean if all four of the fullest cars only contained six passengers); so the maximum likelihood estimate for p is going to be just under $1469/2282 = 0.6437$, as indeed it is. This small calculation is a useful check on your answer. Notice that the censoring has not in fact influenced the calculation unduly.

Solution 6.16

For these data (and assuming a Poisson model) the likelihood is given by

$$p_0^{11} p_1^{37} p_2^{64} p_3^{55} p_4^{37} p_5^{24} P_6^{12},$$

where p_j is the probability $P(X = j)$ when X is Poisson(μ), and where P_j is the probability $P(X \geq j)$. The likelihood is maximized at $\widehat{\mu} = 2.819$. (Again, the sample mean assuming at most 6 colonies per quadrat would be $670/240 = 2.792$. The estimate is not very different for the censored data.)

Solution 6.17

The average time between pulses is given by the sample mean $\bar{t} = 0.2244$. The units are hundredths of a second. Consequently, the maximum likelihood estimate of the pulse rate (per second) is $100/0.2244 = 446$.

Solution 6.18

All that is required here is the sample mean $\widehat{\mu} = \overline{x} = 159.8$ (measurements in cm).

Solution 6.19

(a) In one simulation the observations

$$x_1 = 96.59, \qquad x_2 = 99.87, \qquad x_3 = 107.15$$

were obtained, having sample variance $s^2 = 29.2$, which is fairly close to 25. However, this sequence was immediately followed by the sample

$$x_1 = 100.82, \qquad x_2 = 99.30, \qquad x_3 = 100.91,$$

having sample variance $s^2 = 0.82$, which is very far from 25!

(b) Your collected samples may have looked something like

$$\begin{bmatrix} 108.08 & 96.18 & 89.85 \\ 102.19 & 97.58 & 106.97 \\ 98.52 & 99.23 & 96.88 \\ & \vdots & \\ 106.01 & 95.45 & 96.58 \end{bmatrix}$$

with sample variances

$$\begin{bmatrix} 85.68 \\ 22.04 \\ 1.44 \\ \vdots \\ 33.60 \end{bmatrix}.$$

(i) The mean of this vector of sample variances is 24.95 which is very close to 25; but you can see from the four elements listed that the variation is very considerable. The highest sample variance recorded in the 100 samples was 133.05; the lowest was 0.397.

(ii) The variation in the sample variances is evident from a frequency table, and from the histogram shown in Figure S6.1.

Observation	Frequency
0–20	48
20–40	32
40–60	15
60–80	1
80–100	2
100–120	1
120–140	1

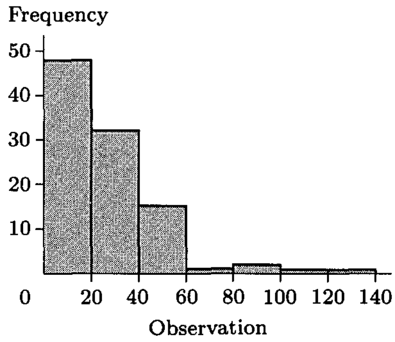


Figure S6.1

(iii) The variance in the recorded sample variances is 524.3!

Your results will probably have shown similarly gross variation in the observed sample variances.

Solution 6.20

(a) The mean of the 100 observed sample variances, based on samples of size 10 was, in one particular experiment, 24.22.

(b) The results are summarized below. A histogram is given in Figure S6.2.

Observation	Frequency
0–10	10
10–20	27
20–30	36
30–40	15
40–50	10
50–60	2

The distribution is very much less dispersed.

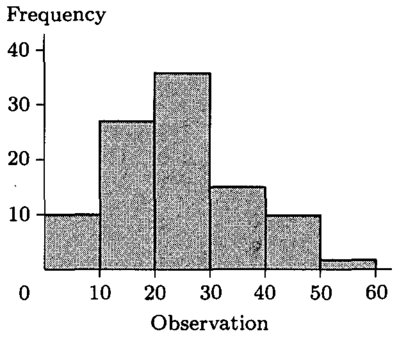


Figure S6.2

(c) The variance in the sample variances is now only 128. (The maximum observation was 55.6, the minimum was 3.49.)

Solution 6.21

(a) (i) $P(-1 \leq Z \leq 1) = 0.6827$. (ii) $P(-\sqrt{2} \leq Z \leq \sqrt{2}) = 0.8427$.

(iii) $P(-\sqrt{3} \leq Z \leq \sqrt{3}) = 0.9167$. (iv) $P(-2 < Z \leq 2) = 0.9545$.

(b) (i) $P(W \leq 1) = P(Z^2 \leq 1)$ (by definition)
$= P(-1 \leq Z \leq 1) = 0.6827$.

(ii) $P(W \leq 2) = P(-\sqrt{2} \leq Z \leq \sqrt{2}) = 0.8427$.

(iii) $P(W \leq 3) = 0.9167$.

(iv) $P(W \leq 4) = 0.9545$.

(c) In one particular run, the following results were obtained:

(i) proportion less than 1 = 677/1000 = 0.677;

(ii) proportion less than 2 = 831/1000 = 0.831;

(iii) proportion less than 3 = 908/1000 = 0.908;

(iv) proportion less than 4 = 955/1000 = 0.955.

(d) The corresponding histogram is shown in Figure S6.3. Notice that this histogram is exceedingly skewed.

(e) From this simulation, an estimate of μ_W is

$$\overline{w} = 1.032,$$

and an estimate for σ_W^2 is

$$s_W^2 = 2.203.$$

(For interest, the experiment was repeated a further four times. Estimates of μ_W and σ_W^2 were

$$1.009, 2.287; \quad 0.975, 2.033; \quad 1.001, 1.782; \quad 1.044, 1.982.)$$

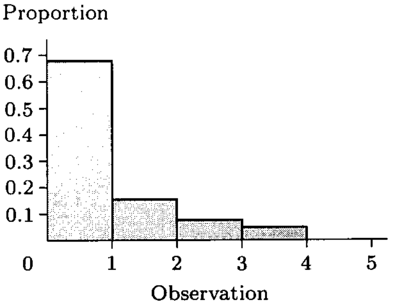


Figure S6.3

Solution 6.22

If Z^2 has mean 1 and variance 2, then the sum of r independent observations on Z^2,

$$W = Z_1^2 + Z_2^2 + \cdots + Z_r^2,$$

will have mean and variance

$$E(W) = r, \qquad V(W) = 2r.$$

Solution 6.23

(a) 0.5697 (b) 0.1303 (c) 0.0471 (d) 0.0518

Solution 6.24

(a) 4.594 (b) 15.507 (c) 11.651 (d) 18.338 (e) 36.191

The last three results are summarized in Figure S6.4.

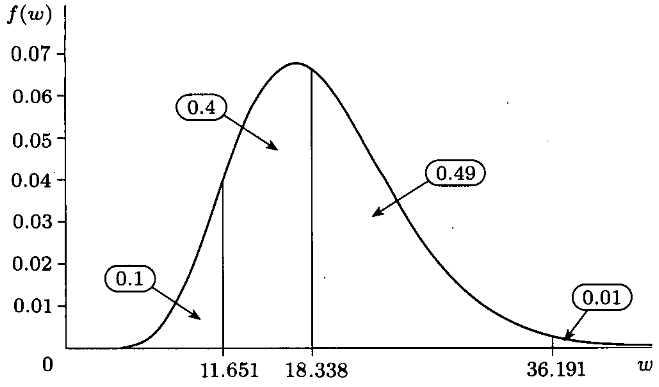


Figure S6.4

Solution 6.25

You should have obtained the following quantiles from the table.

(a) 13.091 (b) 4.168 (c) 11.340 (d) 24.769 (e) 3.841

Solution 6.26

Writing

$$W = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$$

it follows that

$$E(W) = n-1, \quad V(W) = 2(n-1).$$

Also,

$$S^2 = \frac{\sigma^2}{n-1}W$$

so

$$E(S^2) = E\left(\frac{\sigma^2}{n-1}W\right) = \frac{\sigma^2}{n-1}E(W) = \frac{\sigma^2}{n-1}(n-1) = \sigma^2;$$

and

$$V(S^2) = V\left(\frac{\sigma^2}{n-1}W\right) = \frac{\sigma^4}{(n-1)^2}V(W) = \frac{\sigma^4}{(n-1)^2}2(n-1) = \frac{2\sigma^4}{n-1}.$$

Chapter 7

Solution 7.1

The lower confidence limit for μ is found by solving the equation

$$P(T \geq t) = e^{-t/\mu} = 0.05$$

or

$$\frac{t}{\mu} = -\log 0.05$$

or

$$\mu = \frac{t}{-\log 0.05} = \frac{157}{2.996} = 52.4 \text{ days}$$

(to 3 significant figures). The confidence statement may be expressed thus: 'A 90% confidence interval for the mean number of days between disasters, assuming an exponential model and based on the single observation 157 days, is from $\mu_- = 52.4$ days to $\mu_+ = 3060$ days.'

Solution 7.2

To obtain the upper confidence limit p_+ for p in this case, it is necessary to solve the equation

$$P(N \geq 13) = 0.025$$

or

$$(1-p)^{12} = 0.025.$$

This has solution

$$p_+ = 1 - 0.025^{1/12} = 1 - 0.735 = 0.265.$$

Consequently, a 95% confidence interval for p, assuming a geometric model and based on the single observation 13, is given by

$$(p_-, p_+) = (0.002, 0.265).$$

Notice that the confidence interval contains the maximum likelihood estimate $\widehat{p} = 0.077$. The approach has usefully provided a range of plausible values for p.

Solution 7.3

To obtain a 90% confidence interval for p, it is necessary to solve the two equations

$$P(N \leq 13) = 0.05 \quad \text{and} \quad P(N \geq 13) = 0.05.$$

The first may be written

$$1 - (1-p)^{13} = 0.05$$

and has solution

$$p_{-} = 1 - (1 - 0.05)^{1/13} = 1 - 0.996 = 0.004.$$

The second may be written

$$(1-p)^{12} = 0.05$$

and has solution

$$p_{+} = 1 - 0.05^{1/12} = 1 - 0.779 = 0.221.$$

Thus a 90% confidence interval for p is given by

$$(p_{-}, p_{+}) = (0.004, 0.221).$$

This interval is narrower than the 95% confidence interval (0.002, 0.265) and in this sense is more useful; but less confidence may be attached to it. The only way to reduce the width of a confidence interval while maintaining a high confidence level is to increase the sample size.

Solution 7.4

The confidence level required is 99%, so $\frac{1}{2}\alpha = 0.005$. Writing

$$P(N \leq n) = 1 - \left(1 - \frac{1}{\mu}\right)^{n} = \tfrac{1}{2}\alpha,$$

we need to solve, in this case, the equation

$$1 - \left(1 - \frac{1}{\mu}\right)^{4} = 0.005$$

for μ. From

$$\left(1 - \frac{1}{\mu}\right)^{4} = 1 - 0.005 = 0.995$$

it follows that

$$\mu = \frac{1}{1 - 0.995^{1/4}} = 798.5,$$

and this is the upper confidence limit μ_{+}. Similarly, writing

$$P(N \geq n) = \left(1 - \frac{1}{\mu}\right)^{n-1} = \tfrac{1}{2}\alpha,$$

we need to solve, in this case, the equation

$$\left(1-\frac{1}{\mu}\right)^3 = 0.005.$$

This has solution

$$\mu_- = \frac{1}{1-0.005^{1/3}} = 1.206.$$

So a 99% confidence interval for the average length of runs of diseased trees, based on the observation 4 and assuming a geometric model, is given by

$$(\mu_-, \mu_+) = (1.206, 798.5).$$

Notice the width of this confidence interval, and particularly the extent of the upper confidence limit! This is due to the inherent skewed nature of the geometric distribution, but also to the dearth of data. The only way to reduce the width of the confidence interval is to collect more data.

Solution 7.5

If T has a triangular density with parameter $\theta\,(T \sim \text{Triangular}(\theta))$ then the c.d.f. of T is given by

$$F(t) = P(T \leq t) = 1 - \left(1-\frac{t}{\theta}\right)^2, \quad 0 \leq t \leq \theta.$$

(a) Writing

$$P(T \leq t) = 1 - \left(1-\frac{t}{\theta}\right)^2 = \tfrac{1}{2}\alpha,$$

it follows that

$$\left(1-\frac{t}{\theta}\right)^2 = 1 - \tfrac{1}{2}\alpha$$

or

$$\frac{t}{\theta} = 1 - \sqrt{1-\tfrac{1}{2}\alpha};$$

so, finally,

$$\theta = \frac{t}{1-\sqrt{1-\tfrac{1}{2}\alpha}}.$$

This is the upper confidence limit θ_+: it is high values of θ that render low values of t unlikely.

The lower confidence limit θ_- is found by solving the equation

$$P(T \geq t) = \left(1-\frac{t}{\theta}\right)^2 = \tfrac{1}{2}\alpha$$

for θ; writing

$$\frac{t}{\theta} = 1 - \sqrt{\tfrac{1}{2}\alpha}$$

it follows that

$$\theta_- = \frac{t}{1-\sqrt{\tfrac{1}{2}\alpha}}.$$

So a $100(1-\alpha)\%$ confidence interval for the triangular parameter θ, based on a single observation t, is given by

$$(\theta_-, \theta_+) = \left(\frac{t}{1-\sqrt{\frac{1}{2}\alpha}}, \frac{t}{1-\sqrt{1-\frac{1}{2}\alpha}} \right).$$

(b) For instance, if $t = 5$, then a 95% confidence interval ($\frac{1}{2}\alpha = 0.025$) is given by

$$(\theta_-, \theta_+) = \left(\frac{5}{1-\sqrt{0.025}}, \frac{5}{1-\sqrt{0.975}} \right) = (5.94, 397).$$

Again, the confidence interval is extremely wide. But it makes sense: the parameter θ specifies the right-hand edge of the range of T. If the value $t = 5$ has been observed, the value of θ must be at least 5.

Solution 7.6

First, a model is required. Assuming the unknown number of *Firefly* dinghies manufactured to date to be equal to θ then, in the absence of any information to the contrary, we could assume that any one of the dinghies is as likely to have been observed as any other. That is, denoting by X the sail number observed, the random variable X has a discrete uniform distribution

$$P(X = x) = \frac{1}{\theta}, \quad x = 1, 2, 3, \ldots, \theta.$$

Then X has c.d.f.

$$P(X \leq x) = \frac{x}{\theta}, \quad x = 1, 2, 3, \ldots, \theta.$$

The confidence level required is 90%: so $\frac{1}{2}\alpha = 0.05$. Writing

$$P(X \leq 3433) = \frac{3433}{\theta} = 0.05,$$

we obtain the upper confidence limit $\theta_+ = 3433/0.05 = 68\,660$.

Now, the probability $P(X \geq 3433)$ is given by

$$P(X \geq 3433) = 1 - P(X \leq 3432) = 1 - \frac{3432}{\theta},$$

and so the lower confidence limit θ_- for θ is given by the solution of the equation

$$P(X \geq 3433) = 1 - \frac{3432}{\theta} = 0.05;$$

this solution is

$$\theta_- = \frac{3432}{0.95} = 3612.6.$$

The unknown number θ is indubitably an integer. Erring a little on the safe side, we can conclude from the one sighting made (3433) that a 90% confidence interval for the number θ of *Firefly* dinghies manufactured to date is given by

$$(\theta_-, \theta_+) = (3612, 68\,660).$$

Again, the interval is so wide as to be of questionable use. We shall see in Subsection 7.2.5 the very useful consequences of taking a larger sample.

Solution 7.7

(a) (i) A 90% confidence interval for p, based on observing 4 successes in 11 trials, is given by

$$(p_-, p_+) = (0.1351, 0.6502).$$

(ii) The corresponding 95% confidence interval is

$$(p_-, p_+) = (0.1093, 0.6921).$$

(b) Confidence intervals based on observing 8 successes in 22 trials are

90%: $(p_-, p_+) = (0.1956, 0.5609)$;
95%: $(p_-, p_+) = (0.1720, 0.5934)$.

In both cases, the larger sample size has led to narrower confidence intervals. The reason is the increase in information. Just as larger samples lead to reduced variation in parameter estimates, so they permit narrower (more precise) confidence intervals.

(c) A 99% confidence interval for p based on observing 4 successes in 5 trials is given by

$$(p_-, p_+) = (0.1851, 0.9990).$$

(d) In one experiment (your results might have been similar) the sequence of 10 observations on $B(20, 0.3)$ is

1 9 8 7 3 9 6 4 8 5.

The corresponding confidence limits for p are

1: $(p_-, p_+) = (0.0026, 0.2161)$;
9: $(p_-, p_+) = (0.2587, 0.6531)$;
8: $(p_-, p_+) = (0.2171, 0.6064)$;
7: $(p_-, p_+) = (0.1773, 0.5580)$;
3: $(p_-, p_+) = (0.0422, 0.3437)$;
9: $(p_-, p_+) = (0.2587, 0.6531)$;
6: $(p_-, p_+) = (0.1396, 0.5078)$;
4: $(p_-, p_+) = (0.0714, 0.4010)$;
8: $(p_-, p_+) = (0.2171, 0.6064)$;
5: $(p_-, p_+) = (0.1041, 0.4556)$.

For discrete random variables, the calculations underlying the construction of confidence intervals conceal an interesting feature, exemplified here. Only the confidence intervals for $x = 3, 4, \ldots, 9$ contain the value 0.3: other values of x are 'too low' or 'too high'. If X is binomial $B(20, 0.3)$, then $P(3 \leq X \leq 9) = 0.9166 > 0.90$. That is, an average of about 92% (more than 90%) of confidence intervals generated in this way will contain the (usually unknown) parameter value. The procedure is 'conservative'.

Of these ten intervals, only the first one does not contain the known value of p, 0.3. Remember the interpretation of a confidence interval—in *repeated experiments*, a proportion $100(1-\alpha)\%$ of confidence intervals obtained may be expected to contain the (usually) unknown value of the parameter.

Here, the confidence level set was 90%; and, as it happened, exactly nine out of the ten calculated intervals contained the known parameter value $p = 0.3$—just as expected. What happened in your experiment?

The observation '1 success in 20 trials' is so low, that it reduces our confidence that the underlying success probability is, or could be, as high as 0.3.

(Incidentally, an observed success count as high as 10 in 20 would have resulted in the confidence interval

$$(p_-, p_+) = (0.3020, 0.6980),$$

which does not contain the value $p = 0.3$, either.)

Solution 7.8

(a) (i) A 90% confidence interval for a Poisson mean μ, based on the single observation 3, is given by

$$(\mu_-, \mu_+) = (0.8177, 7.754).$$

(ii) The corresponding 95% confidence interval is

$$(\mu_-, \mu_+) = (0.6187, 8.767).$$

(b) (i) An estimate of the mean underlying accident rate μ is given by the sample mean

$$\widehat{\mu} = \overline{x} = \frac{4+4+3+0+5+3+2}{7} = \frac{21}{7} = 3.0.$$

So the estimate of μ is the same as it was in part (a), but this time it is based on seven observations rather than on one.

(ii) Confidence intervals for μ based on these data are given by

$$90\%: \quad (\mu_-, \mu_+) = (2.010, 4.320);$$
$$95\%: \quad (\mu_-, \mu_+) = (1.857, 4.586).$$

Notice that the increased information has resulted in narrower confidence intervals.

(c) (i) The estimated mean annual accident rate for girls of this age is

$$\widehat{\mu} = \frac{20}{6} = 3.333.$$

(ii) In this case we are only told the sample total $t = 20$; but we know that the random variable T, on which t is a single observation, is Poisson(6μ). All that is required is that we obtain confidence limits for the mean of T, and then divide these limits by 6. This approach gives

$$90\%: \quad (\mu_-, \mu_+) = (2.209, 4.844);$$
$$95\%: \quad (\mu_-, \mu_+) = (2.036, 5.148).$$

Solution 7.9

(a) The mean of the 62 time intervals is 437.21 days, about 14 months. Assuming that times between earthquakes are exponentially distributed, a 90% confidence interval for the mean time interval between serious earthquakes world-wide is given by

$$(\mu_-, \mu_+) = (359.06, 546.06),$$

or from about twelve months to eighteen months.

(b) (i) The mean waiting time between vehicles is

$$\widehat{\mu} = \frac{212}{50} = 4.24 \text{ seconds},$$

so the estimated traffic rate is

$$\widehat{\lambda} = \frac{1}{\mu} = 0.2358 \text{ vehicles per second} = 14.15 \text{ vehicles per minute}.$$

(ii) A 90% confidence interval for the mean traffic rate is given by

$$(\lambda_-, \lambda_+) = (11.03, 17.60).$$

(c) When this experiment was tried the ten resulting confidence intervals were as follows. (The sample mean is also shown in brackets.)

0.7821	(1.090)	1.645
0.9361	(1.305)	1.969
0.7864	(1.096)	1.654
0.7149	(0.997)	1.504
0.6848	(0.955)	1.440
0.8603	(1.199)	1.810
0.7142	(0.996)	1.502
0.8163	(1.138)	1.717
0.7145	(0.996)	1.503
0.6423	(0.895)	1.351

Interestingly, an eleventh experiment gave

0.4114 (0.5735) 0.8654

which does not contain the number 1.0, but the sample had extraordinarily low numbers in it. They were

0.11	0.75	0.80	0.07	0.09	1.54	0.54	0.34	0.15	1.67
0.01	0.60	0.36	0.66	0.72	0.44	0.57	0.06	1.86	0.13.

Note that only two of these numbers exceed the mean, 1.

A twelfth experiment gave

1.152 (1.606) 2.423

which also does not contain the number 1; in this case the numbers sampled from $M(1)$ were unusually high.

0.91	0.13	3.71	1.23	0.56	2.45	0.03	2.51	0.42	2.09
0.56	1.09	1.05	3.13	4.29	1.96	2.10	1.59	0.35	1.95

Solution 7.10

(a) One simulation gave the following ten observations on N.

20 6 3 7 4 2 8 10 3 15

The corresponding 90% confidence interval for p based on these data is

$$(p_-, p_+) = (0.0712, 0.1951).$$

The width of the confidence interval is $0.1951 - 0.0712 = 0.124$. It contains both the values $p = 1/10$ and $p = 1/6$.

(b) For interest, this part of the exercise was performed three times. The corresponding confidence intervals were

$$(0.0890, 0.1217), \quad \text{width } = 0.033;$$
$$(0.0825, 0.1130), \quad \text{width } = 0.030;$$
$$(0.0940, 0.1284), \quad \text{width } = 0.034.$$

In all three cases the confidence interval contained the value $p = 1/10$ and *not* the value $p = 1/6$, providing quite strong evidence that the die is loaded (as we know; but you can imagine that there are circumstances where one might not know this, but merely suspect it).

(The experiment was repeated on the computer a further 997 times, making a total of 1000 experiments altogether. The number 1/10 was included in the resulting confidence intervals 892 times, roughly as expected—90% of 1000 is 900. The number 1/6 was not included in any of the intervals.)

Solution 7.11

Either from tables or your computer, you should find (a) $t = 1.699$; (b) $t = 1.697$; (c) $t = -3.365$; (d) $t = 2.262$.

Solution 7.12

For the five data points the sample mean and sample standard deviation are $\overline{x} = 3.118$ and $s = 0.155$. For a confidence level of 95%, the corresponding critical t-value is obtained from the t-distribution $t(4)$: it is the 97.5% point of $t(4)$, $q_{0.975} = 2.776$. This is shown in Figure S7.1. A 95% confidence interval for μ, the mean coal consumption in pounds per draw-bar horse-power hour, is given by

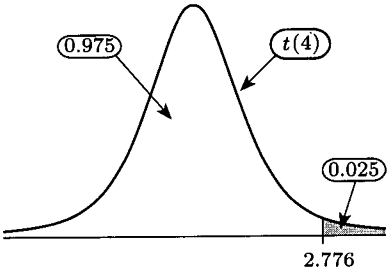


Figure S7.1 Critical values from $t(4)$

$$\begin{aligned}(\mu_-, \mu_+) &= \left(\overline{x} - \frac{ts}{\sqrt{n}}, \overline{x} + \frac{ts}{\sqrt{n}}\right)\\ &= \left(3.118 - \frac{2.776 \times 0.155}{\sqrt{5}}, 3.118 + \frac{2.776 \times 0.155}{\sqrt{5}}\right)\\ &= (3.118 - 0.192, 3.118 + 0.192) = (2.93, 3.31).\end{aligned}$$

Solution 7.13

(a) For the data

1.2 2.4 1.3 1.3 0 1.0 1.8 0.8 4.6 1.4

the sample mean is $\overline{x} = 1.58$ and the sample standard deviation is $s = 1.23$. A 95% confidence interval is required for the mean difference μ between the two treatments. Reference to tables of $t(9)$ gives the critical value $t = 2.262$. Thus the corresponding confidence interval is

$$\begin{aligned}(\mu_-, \mu_+) &= \left(\overline{x} - \frac{ts}{\sqrt{n}}, \overline{x} + \frac{ts}{\sqrt{n}}\right)\\ &= \left(1.58 - \frac{2.262 \times 1.23}{\sqrt{10}}, 1.58 + \frac{2.262 \times 1.23}{\sqrt{10}}\right)\\ &= (1.58 - 0.88, 1.58 + 0.88) = (0.70, 2.46).\end{aligned}$$

(b) Since the number 0 (indicating no difference between the two treatments) is not included in this confidence interval, there is considerable evidence that the treatment L-hyoscyamine hydrobromide is a more effective hypnotic than the alternative.

Solution 7.14

This is a 'self-help' exercise included to encourage you to use your computer.

Solution 7.15

In this case the sample size was $n = 8$; the sample variance was $s^2 = 37.75$.

(a) For a 90% confidence interval we will require the 5% and the 95% points of $\chi^2(7)$. These are $c_L = 2.167$ and $c_U = 14.067$ respectively. Consequently, a 90% confidence interval for the population variance σ^2, based on the data provided, is

$$(\sigma_-^2, \sigma_+^2) = \left(\frac{7 \times 37.75}{14.067}, \frac{7 \times 37.75}{2.167}\right) = (18.79, 121.9).$$

(b) For a 95% confidence interval we need

$$c_L = q_{0.025} = 1.690, \quad c_U = q_{0.975} = 16.013.$$

The interval is

$$(\sigma_-^2, \sigma_+^2) = \left(\frac{7 \times 37.75}{16.013}, \frac{7 \times 37.75}{1.69}\right) = (16.50, 156.4).$$

Solution 7.16

(a) An estimate for σ is given by the sample standard deviation

$$s = \sqrt{37.75} = 6.1.$$

(b) We found in Solution 7.15(a) that a 90% confidence interval for the population variance σ^2 is given by (18.79, 121.9). Taking square roots, the corresponding 90% confidence interval for the population standard deviation σ is given by

$$(\sigma_-, \sigma_+) = (4.3, 11.0).$$

Solution 7.17

For the earthquake data, the mean time between quakes is given by $\widehat{\mu} = \overline{x} = 437.21$ days. There were 62 waiting times recorded; to establish a 95% confidence interval for the mean time between serious earthquakes world-wide, it is necessary to use the fact that the 97.5% point of the standard normal distribution is $z = 1.96$. The confidence interval required is

$$\begin{aligned}(\mu_-, \mu_+) &= \left(\frac{\widehat{\mu}}{1 + z/\sqrt{n}}, \frac{\widehat{\mu}}{1 - z/\sqrt{n}}\right) \\ &= \left(\frac{437.21}{1 + 1.96/\sqrt{62}}, \frac{437.21}{1 - 1.96/\sqrt{62}}\right) \\ &= (350.1, 582.1).\end{aligned}$$

This is an interval of between about twelve and nineteen months. Here, it is assumed that the sum of 62 exponential waiting times is approximately normally distributed.

Solution 7.18

(a) For the Kwinana traffic data the mean waiting time between vehicles was $\widehat{\mu} = 4.24$ seconds (corresponding to a traffic rate of $\widehat{\lambda} = 14.15$ vehicles per minute); the sample size was $n = 50$. Assuming the sample total to be approximately normally distributed, a 90% confidence interval for the mean waiting time (taking $z = 1.645$) is given by

$$\begin{aligned}(\mu_-, \mu_+) &= \left(\frac{\widehat{\mu}}{1+z/\sqrt{n}}, \frac{\widehat{\mu}}{1-z/\sqrt{n}}\right)\\ &= \left(\frac{4.24}{1+1.645/\sqrt{50}}, \frac{4.24}{1-1.645/\sqrt{50}}\right)\\ &= (3.44, 5.53).\end{aligned}$$

By taking reciprocals, confidence limits for the mean traffic rate are $\lambda_- = 1/\mu_+ = 0.18$ vehicles per second, or 10.9 vehicles per minute; and $\lambda_+ = 1/\mu_- = 0.29$ vehicles per second, or 17.4 vehicles per minute.

(b) The confidence interval in the solution to Exercise 7.9 was

$$(\lambda_-, \lambda_+) = (11.03, 17.60).$$

The differences, which are very slight, are due to the approximation induced by assuming a normal distribution for the sample total.

Solution 7.19

The stated constraints imply that the following inequalities must hold:

$$\frac{\widehat{\mu}}{1+z/\sqrt{n}} \geq 0.97\widehat{\mu} \quad \text{and} \quad \frac{\widehat{\mu}}{1-z/\sqrt{n}} \leq 1.03\widehat{\mu}$$

where $z = 1.96$. The first inequality gives

$$\frac{1}{1+z/\sqrt{n}} \geq 0.97,$$

so

$$1 + \frac{z}{\sqrt{n}} \leq \frac{1}{0.97}.$$

This gives

$$\frac{z}{\sqrt{n}} \leq \frac{0.03}{0.97},$$

so

$$\sqrt{n} \geq \frac{97z}{3} = \frac{97 \times 1.96}{3};$$

that is, $n \geq 4016.2$. The second inequality gives

$$\frac{1}{1-z/\sqrt{n}} \leq 1.03,$$

so

$$\frac{z}{\sqrt{n}} \leq \frac{0.03}{1.03}.$$

This gives

$$\sqrt{n} \geq \frac{103z}{3} = \frac{103 \times 1.96}{3};$$

that is, $n \geq 4528.4$. For both inequalities to hold, the sample size must be 4529 or more. This is an extremely large sample!

Solution 7.20

The sample mean is

$$\frac{0\times 101+1\times 143+2\times 120+\cdots+10\times 2}{101+143+120+\cdots+2}=\frac{1522}{621}=2.451.$$

An approximate 95% confidence interval for the underlying mean accident rate over the whole of the eight-year period (assuming a Poisson model) is

$$\begin{aligned}(\mu_-,\mu_+)&=\left(\widehat{\mu}-z\sqrt{\frac{\widehat{\mu}}{n}},\widehat{\mu}+z\sqrt{\frac{\widehat{\mu}}{n}}\right)\\&=\left(2.451-1.96\sqrt{\frac{2.451}{621}},2.451+1.96\sqrt{\frac{2.451}{621}}\right)\\&=(2.451-0.123,2.451+0.123)\\&=(2.33,2.57).\end{aligned}$$

Solution 7.21

There were 109 runs observed. The mean length of a run was

$$\widehat{\mu}=166/109=1.523.$$

For a 99% confidence interval the 99.5% point of Z is required. This is $z=2.576$. The resulting approximate 99% confidence interval for the mean length of runs of diseased trees is

$$\begin{aligned}(\mu_-,\mu_+)&=\left(\widehat{\mu}-z\sqrt{\frac{\widehat{\mu}(\widehat{\mu}-1)}{n}},\widehat{\mu}+z\sqrt{\frac{\widehat{\mu}(\widehat{\mu}-1)}{n}}\right)\\&=\left(1.523-2.58\sqrt{\frac{1.523\times 0.523}{109}},1.523+2.58\sqrt{\frac{1.523\times 0.523}{109}}\right)\\&=(1.523-0.221,1.523+0.221)\\&=(1.30,1.74).\end{aligned}$$

Here, it is assumed that the sum of 109 geometric run lengths is approximately normally distributed.

Solution 7.22

(a) In this case, the estimate of p is $\widehat{p}=286/5387=0.0531$. An approximate 90% confidence interval for the underlying proportion of red-haired children in Scotland is

$$\begin{aligned}(p_-,p_+)&=\left(\widehat{p}-z\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}},\widehat{p}+z\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right)\\&=\left(0.0531-1.645\sqrt{\frac{0.0531\times 0.9469}{5387}},0.0531+1.645\sqrt{\frac{0.0531\times 0.9469}{5387}}\right).\\&=(0.0531-0.0050,0.0531+0.0050)\\&=(0.048,0.058).\end{aligned}$$

Here (as well as the usual assumption of normality) it has been assumed that the school children of Caithness are typical of all those in Scotland, which may not necessarily be the case.

(b) Here, the estimated proportion of fair-haired children who are blue-eyed is $\widehat{p} = 1368/5789 = 0.236\,31$. An approximate 95% confidence interval for this proportion is

$$\begin{aligned}(p_-, p_+) &= \left(\widehat{p} - z\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}, \widehat{p} + z\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right)\\ &= \left(0.236\,31 - 1.96\sqrt{\frac{0.236\,31 \times 0.763\,69}{5789}}, 0.236\,31 - 1.96\sqrt{\frac{0.236\,31 \times 0.763\,69}{5789}}\right)\\ &= (0.236\,31 - 0.010\,94, 0.236\,31 + 0.010\,94)\\ &= (0.225, 0.247).\end{aligned}$$

Again, it has been assumed here that Aberdeen school children are representative of all school children in Scotland.

Solution 7.23

For these data the sample mean is $\overline{x} = 1.992$ and the sample standard deviation is $s = 1.394$. A 90% confidence interval for the average number of books borrowed in a year is given by

$$\begin{aligned}(\mu_-, \mu_+) &= \left(\overline{x} - z\frac{s}{\sqrt{n}}, \overline{x} + z\frac{s}{\sqrt{n}}\right)\\ &= \left(1.992 - 1.645 \times \frac{1.394}{\sqrt{122}}, 1.992 + 1.645 \times \frac{1.394}{\sqrt{122}}\right)\\ &= (1.992 - 0.208, 1.992 + 0.208)\\ &= (1.78, 2.20).\end{aligned}$$

Solution 7.24

For these data the sample mean is $\overline{x} = 0.3163$ and the sample standard deviation is $s = 0.0805$. Making no assumptions at all about the distribution of eggshell thicknesses for Anacapa pelicans, a 95% confidence interval for the mean thickness is given by

$$\begin{aligned}(\mu_-, \mu_+) &= \left(\overline{x} - z\frac{s}{\sqrt{n}}, \overline{x} + z\frac{s}{\sqrt{n}}\right)\\ &= \left(0.3163 - 1.96 \times \frac{0.0805}{\sqrt{65}}, 0.3163 + 1.96 \times \frac{0.0805}{\sqrt{65}}\right)\\ &= (0.3163 - 0.020, 0.3163 + 0.020)\\ &= (0.30, 0.34).\end{aligned}$$

Chapter 8

Solution 8.1

For this test the null hypothesis is

$$H_0 : p = \tfrac{2}{3};$$

the alternative hypothesis is

$$H_1 : p \neq \tfrac{2}{3}.$$

An appropriate model for the number of 1s in a sequence of 25 trials is binomial $B(25, p)$. In this experiment the number of 1s observed is 10.

(a) A 95% confidence interval for p, based on observing 10 successes in 25 trials is given by

$$(p_-, p_+) = (0.2113, 0.6133).$$

(b) The hypothesized value $p_0 = \frac{2}{3} = 0.6667$ is not contained in this confidence interval. The conclusions of the test may be stated as follows.

On the basis of these data, there is evidence at the significance level 0.05 to reject the null hypothesis $p = \frac{2}{3}$ in favour of the alternative hypothesis that p differs from $\frac{2}{3}$. (In fact, there is evidence from the sample that $p < \frac{2}{3}$.)

(c) A 99% confidence interval for p is given by

$$(p_-, p_+) = (0.1679, 0.6702).$$

The interval contains the hypothesized value $p_0 = \frac{2}{3}$: at the 1% level of significance there is no evidence, from these data, to reject the hypothesis.

Solution 8.2

The Kwinana Freeway data consist of 50 observations on times assumed to be exponentially distributed. The mean waiting time μ is unknown. The hypothesis that the mean traffic flow rate is 10 vehicles per minute is equivalent to hypothesizing a mean waiting time of $\frac{1}{10}$ minute, or 6 seconds:

$$H_0 : \mu = 6.$$

An appropriate alternative hypothesis is furnished by

$$H_1 : \mu \neq 6.$$

Confidence intervals at levels 90%, 95% and 99% for μ, based on the data, are

$$\begin{aligned} 90\%: &\quad (3.41, 5.44); \\ 95\%: &\quad (3.27, 5.71); \\ 99\%: &\quad (3.02, 6.30). \end{aligned}$$

Only the last of these contains the hypothesized value $\mu_0 = 6$. One may conclude the test as follows. Based on these data, the hypothesis that the mean traffic flow rate is 10 vehicles per minute is rejected at the 5% level of significance; at the 1% level the evidence is insufficient to reject the hypothesis.

There are two additional points to notice here. The first is that no mention is made of the conclusion of the test at the 10% significance level: this is because rejection at 5% implies rejection at 10%. Second, at some significance level between 1% and 5% it is clear that the hypothesized value μ_0 will itself be at the very boundary of the decision rule. This idea will be explored in Section 8.3.

Solution 8.3

Large-sample confidence intervals for the Bernoulli parameter p, based on approximate normal distribution theory, are of the form

$$(p_-, p_+) = \left(\widehat{p} - z\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}, \widehat{p} + z\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right)$$

(a) In this case $n = 1064$ and $\widehat{p} = 787/1064 = 0.740$. A specified 10% significance level for the test implies a 90% confidence level; this in turn implies $z = q_{0.95} = 1.645$, the 95% quantile of the standard normal distribution. The confidence interval required is given by

$$\begin{aligned}(p_-, p_+) &= \left(0.740 - 1.645\sqrt{\frac{0.740 \times 0.260}{1064}}, 0.740 + 1.645\sqrt{\frac{0.740 \times 0.260}{1064}}\right)\\ &= (0.740 - 0.022, 0.740 + 0.022) = (0.718, 0.762).\end{aligned}$$

The confidence interval contains the hypothesized value $p_0 = 0.75$. Thus, at the 10% level of significance, these data offer no evidence to reject the hypothesis that the proportion of yellow peas is equal to $\frac{3}{4}$.

(b) Here, $n = 100$ and $\widehat{p} = \frac{60}{100} = 0.6$. A 95% confidence interval for p is given by

$$\begin{aligned}(p_-, p_+) &= \left(0.6 - 1.96\sqrt{\frac{0.6 \times 0.4}{100}}, 0.6 + 1.96\sqrt{\frac{0.6 \times 0.4}{100}}\right)\\ &= (0.6 - 0.096, 0.6 + 0.096) = (0.504, 0.696).\end{aligned}$$

The hypothesized value $p_0 = \frac{2}{3} = 0.667$ is contained in the interval: on the basis of these data, there is no evidence at this level to reject H_0.

Solution 8.4

The observed mean ratio is

$$\overline{r} = \frac{0.693 + 0.662 + \cdots + 0.933}{20} = 0.6605.$$

The sample standard deviation is $s = 0.0925$. Consequently, the observed value t of the test statistic T under the null hypothesis is

$$t = \frac{\overline{r} - \mu}{s/\sqrt{n}} = \frac{0.6605 - 0.618}{0.0925/\sqrt{20}} = 2.055.$$

To work out the rejection region for the test, we need the 2.5% and 97.5% quantiles for $t(19)$. These are

$$q_{0.025} = -2.093, \qquad q_{0.975} = 2.093.$$

The rejection region is shown in Figure S8.1, together with the observed value t of T.

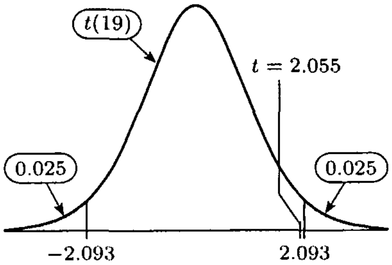


Figure S8.1

As you can see, the observed value $t = 2.055$ is very close to the boundary of the rejection region (suggesting Shoshoni rectangles are somewhat 'square'); but strictly according to the predetermined significance level, there is insufficient evidence, on the basis of these data, to reject the null hypothesis that $\mu = 0.618$.

Solution 8.5

The null hypothesis here is

$$H_0 : \mu = 0$$

and, as for L-hyoscyamine hydrobromide, the alternative hypothesis is

$$H_1 : \mu > 0.$$

Assuming a normal model for the variation in sleep gain after a dose of D-hyoscyamine hydrobromide, then under the null hypothesis

$$T = \frac{\overline{D}}{S/\sqrt{n}} \sim t(n-1).$$

To determine the rejection region at level 0.05, the critical quantile is $q_{0.95} = 1.833$. The observed value t of the test statistic T is

$$t = \frac{\overline{d}}{s/\sqrt{n}} = \frac{0.75}{1.79/\sqrt{10}} = 1.33$$

which is less than $q_{0.95} = 1.833$, and therefore falls outside the rejection region. On the basis of these data, and at this level, there is no reason to suspect that the hypnotic D-hyoscyamine hydrobromide has any measurable effect in prolonging sleep.

Solution 8.6

(a) The null and alternative hypotheses are

$$H_0 : \mu = 0, \qquad H_1 : \mu \neq 0,$$

where μ is the mean difference between the heights of a pair of cross-fertilized and self-fertilized plants whose parents were grown from the same seed. An appropriate test statistic is

$$T = \frac{\overline{D}}{S/\sqrt{n}}$$

with null distribution $t(n-1)$.

You will notice here the very precise statement of the meaning of the parameter μ. Sometimes it is important to be pedantic in this way.

(b) The test is two-sided at 10%: the rejection region is defined by the boundary points

$$q_{0.05} = -1.761; \qquad q_{0.95} = 1.761,$$

the 5% and 95% quantiles of $t(14)$. If the observed value t of T is less than -1.761 or more than 1.761, the null hypothesis will be rejected in favour of the alternative.

(c) For this data set

$$t = \frac{\overline{d}}{s/\sqrt{n}} = \frac{20.93}{37.74/\sqrt{15}} = 2.15.$$

This exceeds $q_{0.95} = 1.761$, so the hypothesis of zero difference is rejected in favour of the alternative hypothesis that there is a difference in the mean height of cross-fertilized and self-fertilized plants. (In fact, on the basis of the data, it appears that cross-fertilized plants are taller than self-fertilized plants.)

Solution 8.7

(a) Assume in this case that the test statistic is $X \sim \text{Poisson}(\mu)$. Under the null hypothesis, the distribution of X is Poisson(3.0). We require to find values x_1 and x_2 such that, as closely as can be attained,

$$P(X \leq x_1) \simeq 0.05, \quad P(X \geq x_2) \simeq 0.05.$$

This is shown in Figure S8.2.

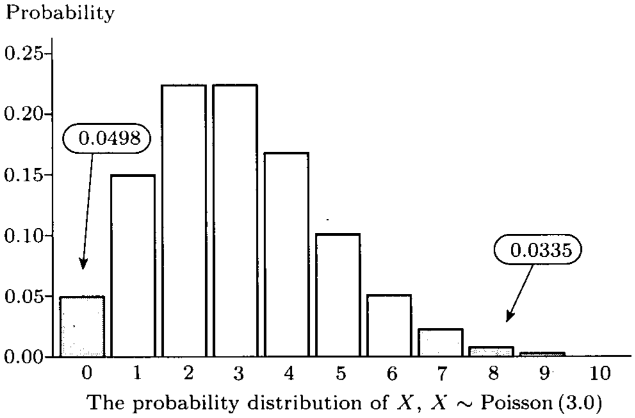


Figure S8.2 Identifying the rejection region

In this case

$$P(X \leq 0) = 0.0498 \text{ and } P(X \geq 7) = 0.0335$$

so the rejection region for the test, based on the single observation x, is

$$x = 0 \text{ or } x \geq 7.$$

Otherwise the hypothesis $H_0 : \mu = 3$ is not rejected. The actual level of this test is $0.0498 + 0.0335 = 0.0833$.

(b) A sensible test statistic is the sample total $V = X_1 + X_2 + \cdots + X_5$, which under the null hypothesis has a Poisson distribution with mean 15. In this case

$$P(V \leq 8) = 0.0374 \text{ and } P(V \geq 22) = 0.0531$$

and the null hypothesis will not be rejected if the sample total is between 9 and 21 inclusive. The level of the test is $0.0374 + 0.0531 = 0.0905$.

(c) For a sample of size 10 the sample total $W = X_1 + X_2 + \cdots + X_{10}$ has the null distribution Poisson(30). Useful quantiles are given by

$$P(W \leq 21) = 0.0544 \text{ and } P(W \geq 40) = 0.0463,$$

and so the rejection region for the test based on the sample total W is

$$w \leq 21 \text{ or } w \geq 40.$$

The level of the test is $0.0544 + 0.0463 = 0.1007$.

Solution 8.8

(a) The forecast relative frequency of light-greys is $\frac{1}{8}$: if this is true, the distribution of the number of light-greys in samples of size 18 is binomial $N \sim B(18, \frac{1}{8})$.

(b) The probability of observing four light-greys is

$$p_N(4) = 0.1152.$$

The diagram shows all the values in the range of the null distribution $B(18, \frac{1}{8})$ that are as extreme as that observed.

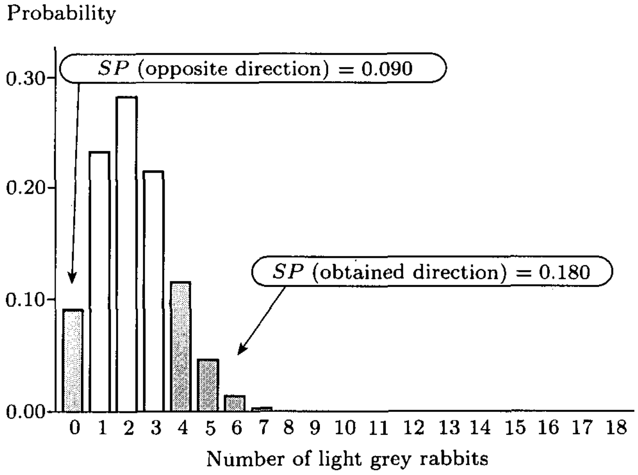


Figure S8.3 Counts as extreme as 4 from $B(18, \frac{1}{8})$

The SP may be calculated as follows.

$$SP(\text{obtained direction}) = P(N \geq 4) = 0.180$$
$$SP(\text{opposite direction}) = P(N = 0) = 0.090$$
$$SP(\text{total}) = 0.270$$

This completes the assessment of the null hypothesis in the light of the sample. There is no substantial evidence that the null hypothesis is flawed: were it true, more than a quarter of future samples would, in fact, offer less support for it than did the sample collected.

Solution 8.9

If the number of insects caught in a trap is denoted by $N \sim \text{Poisson}(\mu)$, then the total number caught in 33 traps is

$$T = N_1 + N_2 + \cdots + N_{33} \sim \text{Poisson}(33\mu).$$

Under the null hypothesis $H_0 : \mu = 1$, the distribution of T is Poisson(33).

In fact, the total number of insects counted was

$$t = 0 \times 10 + 1 \times 9 + 2 \times 5 + 3 \times 5 + 4 \times 1 + 5 \times 2 + 6 \times 1$$
$$= 0 + 9 + 10 + 15 + 4 + 10 + 6 = 54.$$

Counts more extreme than $t = 54$ are shown in Figure S8.4.

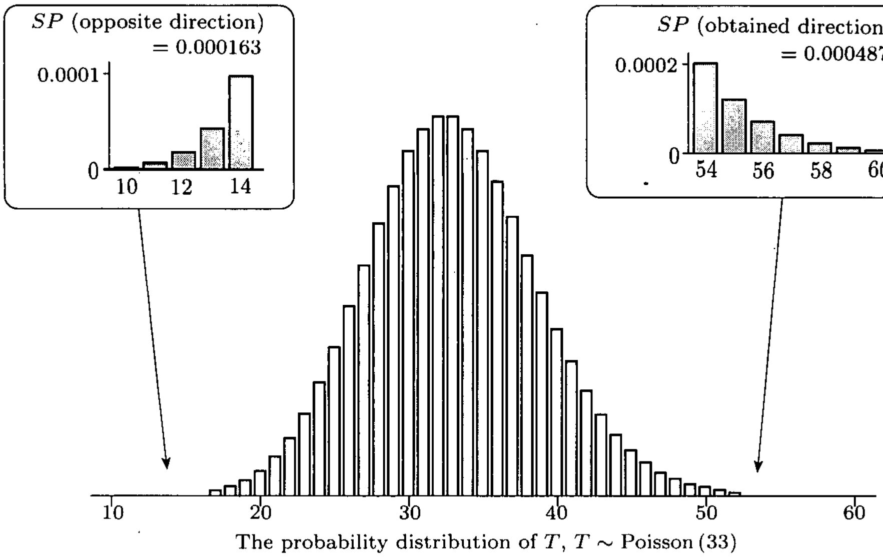


Figure S8.4 Counts more extreme than $t = 54$ when $T \sim$ Poisson(33)

The significance probabilities for the test are given by

$$SP(\text{obtained direction}) = P(T \geq 54) = 0.000\,487$$
$$SP(\text{opposite direction}) = P(T \leq 14) = 0.000\,163$$
$$SP(\text{total}) = 0.000\,487 + 0.000\,163 = 0.000\,650.$$

These significance probabilities are extremely small, offering strong evidence that the hypothesis $\mu = 1$ is false. The value obtained suggests that in fact μ is rather greater than 1.

The observed sample mean catch is $54/33 = 1.64$, which does not at first glance appear to be so very different from the hypothesized value. Evidently the difference is very considerable.

Solution 8.10

(a) Neither the boxplots nor the histograms suggest that either sample is skewed, or that the variances are substantially different. The histograms suggest that a normal model would be an extremely useful representation of the variability in the measurements.

(b) For the Etruscan skulls,

$$s_1^2 = 35.65;$$

for the modern Italian skulls,

$$s_2^2 = 33.06.$$

These are remarkably close: there is no question of having to forgo the test on the grounds that the test assumptions are not satisfied.

(c) Further summary statistics are

$$n_1 = 84, \qquad \overline{x}_1 = 143.77; \qquad n_2 = 70, \qquad \overline{x}_2 = 132.44.$$

The pooled estimate of the variance σ^2 is given by

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{83 \times 35.65 + 69 \times 33.06}{84 + 70 - 2} = 34.47.$$

The observed value of the test statistic T is

$$t = \frac{\overline{x}_1 - \overline{x}_2}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} = \frac{143.77 - 132.44}{\sqrt{34.47}\sqrt{\dfrac{1}{84} + \dfrac{1}{70}}} = 11.92.$$

The test statistic needs to be assessed against Student's t-distribution with $n_1 + n_2 - 2 = 84 + 70 - 2 = 152$ degrees of freedom. The shape of $t(152)$ is not markedly different from the standard normal density. Without recourse to tables, or to a computer, the total SP for this test is very close to 0: differences in the mean maximum skull breadth for Etruscans and the modern Italian male are very marked indeed.

Solution 8.11

It may be that at a single command your computer permits a comparison of the two diets returning, for instance, the value of the test statistic t and associated SP. The details are as follows (so you can check your program!). For those rats given the restricted diet,

$$n_1 = 106, \quad \overline{x}_1 = 968.745, \quad s_1^2 = 80\,985.7;$$

and for those rats given the *ad libitum* diet,

$$n_2 = 89, \quad \overline{x}_2 = 684.011, \quad s_2^2 = 17\,978.6.$$

Then the pooled estimate of σ^2 is given by

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{105 \times 80\,985.7 + 88 \times 17\,978.6}{106 + 89 - 2} = 52\,257.1.$$

This estimate of σ^2 is rather different to either of s_1^2 or s_2^2; and, in fact, the ratio of the sample variances is given by

$$s_1^2/s_2^2 = 4.5$$

which exceeds 3, and so suggests that the assumption of equal variances underlying the two-sample t-test is untenable. This suggestion may be confirmed formally using an appropriate test: in fact, there turns out to be considerable evidence that the variances are different.

Both samples are also considerably skewed. The software on your computer may blindly perform a two-sample t-test if you ask it to do so, with or without a warning message that the assumptions of the test may be seriously adrift. If so, then the resulting value of the test statistic T is

$$t = \frac{\overline{x}_1 - \overline{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{968.745 - 684.011}{\sqrt{52\,257.1}\sqrt{\frac{1}{106} + \frac{1}{89}}} = 8.66.$$

Against the null distribution $t(193)$ (or, essentially, against the normal distribution $N(0,1)$) the SP is negligible. Assuming the t-test to be viable, there is very considerable evidence that the mean lifelengths under the different diet regimes are different (and that, in fact, a restricted diet leads to increased longevity).

The question of whether the t-test is a valid procedure in this case is a real one, and also a worrying one: however, lacking any other procedure for comparing two means, the t-test has provided some 'feel' for the extent to which the data indicate a difference between the two diet regimes. Other tests for a comparison of two means, with less restrictive assumptions, are available in the statistical literature.

Solution 8.12

(a) The data $\widehat{p}_1 = r_1/n_1 = 71/100 = 0.71$ and $\widehat{p}_2 = r_2/n_2 = 89/105 = 0.85$ suggest that females are more likely to be successful in a request for help. Fisher's exact test applied to the data $r_1 = 71, n_1 = 100, r_2 = 89$, $n_2 = 105$, yields the results

$$SP(\text{obtained direction}) = 0.013$$
$$SP(\text{opposite direction}) = 0.006$$
$$SP(\text{total}) = 0.019.$$

The total SP is less than 2%; in a directed test of the hypothesis $H_0 : p_1 = p_2$ against the alternative $H_1 : p_1 < p_2$ the SP is 1.3%. There is strong evidence of a difference in proportions, and that females are more likely to be given help when it is requested.

(b) In this case the proportions to be compared are $\widehat{p}_1 = r_1/n_1 = 8/20 = 0.40$ and $\widehat{p}_2 = r_2/n_2 = 11/20 = 0.55$. A reasonable null hypothesis might be

$$H_0 : p_1 = p_2.$$

No particular suggestion has been offered that certain types of brain damage might reduce a person's facility in handling syllogisms: we might write

$$H_1 : p_1 \neq p_2.$$

However, the data suggest the possibility that p_1 is less than p_2. In fact, Fisher's test gives

$$SP(\text{obtained direction}) = 0.264$$
$$SP(\text{opposite direction}) = 0.264$$
$$SP(\text{total}) = 0.527.$$

There is no serious evidence (whatever alternative hypothesis one might care to suggest) on the basis of these data that the individuals tested showed a significantly different capability.

Solution 8.13

(a) In this case the data available to test the null hypothesis $H_0 : \mu_1 = \mu_2$ (assuming a Poisson model) are

$$n_1 = n_2 = 1, \quad t = 3 + 6 = 9$$

and consequently the aim is to test the observed value $t_1^* = 3$ against the binomial distribution $B(t, n_1/(n_1 + n_2))$ or $B\left(9, \frac{1}{2}\right)$. By the symmetry of the binomial distribution

$$\begin{aligned} SP(\text{obtained direction}) &= SP(\text{opposite direction}) \\ &= p(3) + p(2) + p(1) + p(0) = 0.254 \end{aligned}$$

and so

$$SP(\text{total}) = 2 \times 0.254 = 0.508.$$

The SPs in either direction are not small: there is no strong evidence to reject the hypothesis that the two underlying mean densities are equal.

(b) In this case

$$n_1 = 4, \quad n_2 = 8;$$
$$t_1 = 77 + 61 + 157 + 52 = 347;$$
$$t_2 = 17 + 31 + 87 + 16 + 18 + 26 + 77 + 20 = 292;$$
$$t = t_1 + t_2 = 639.$$

In fact $\widehat{\mu}_1$ (the estimate of the mean plant density under Treatment 1) is 347/4=86.75; and $\widehat{\mu}_2$ is 292/8=36.50. There seems to be a considerable difference here. Formally, we need to test the observed value $t_1^* = 347$ against the binomial distribution $B(639, \frac{1}{3})$. The null distribution peaks at the mean $(\frac{1}{3} \times 639 = 213)$ and so fairly evidently

$$SP(\text{obtained direction}) = P(T_1^* \geq 347) \simeq 0;$$

this was calculated using a computer.

Working out the SP in the opposite direction strictly involves a scan of the binomial distribution $B(639, \frac{1}{3})$ to ascertain those values in the range occurring with probability less than $p(347)$.

A normal approximation has $T_1^* \simeq N(213, 142)$ (for the binomial mean is $np = 639 \times \frac{1}{3} = 213$; the variance is $npq = 639 \times \frac{1}{3} \times \frac{2}{3} = 142$). The SP in the obtained direction is again

$$P(T_1^* \geq 347);$$

using a continuity correction this is approximately

$$1 - \Phi\left(\frac{346\frac{1}{2} - 213}{\sqrt{142}}\right) = 1 - \Phi(11.2) \simeq 0;$$

by the symmetry of the normal distribution the SP in the opposite direction is the same as the SP in the obtained direction. There is very considerable evidence that the mean plant densities under the two treatments differ and, in fact, that Treatment 1 leads to a higher density than Treatment 2.

Chapter 9

Solution 9.1

Since $n = 7$, we determine the points for which $\Phi(x_i) = i/8$.

i	$y_{(i)}$	$i/8$	x_i
1	5.1	0.125	−1.150
2	5.3	0.250	−0.674
3	5.5	0.375	−0.319
4	5.6	0.500	0.000
5	5.8	0.625	0.319
6	5.8	0.750	0.674
7	6.2	0.875	1.150

A probability plot for the points $y_{(i)}$ against x_i is given in Figure S9.1.

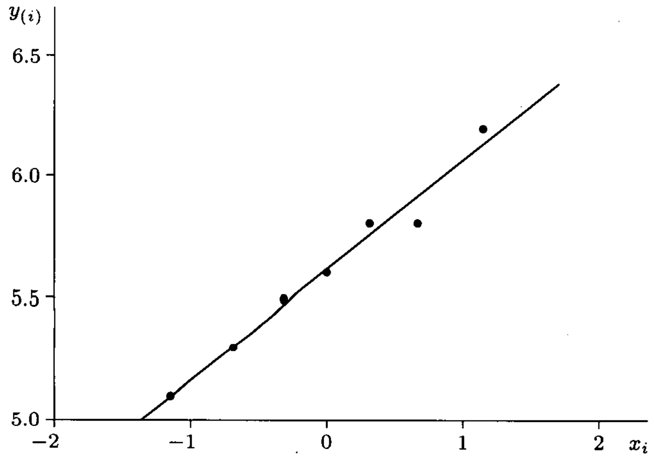


Figure S9.1 Silver content against normal scores (fourth coinage)

A straight line fits the points quite closely and we can conclude that the normal distribution provides an adequate model for the variation in the data.

Solution 9.2

The differences (in microns) are given in Table S9.1.

Table S9.1 Corneal thickness in patients with glaucoma (microns)

Patient	1	2	3	4	5	6	7	8
Glaucomatous eye	488	478	480	426	440	410	458	460
Normal eye	484	478	492	444	436	398	464	476
Difference	4	0	−12	−18	4	12	−6	−16

Since $n = 8$, we determine the points for which $\Phi(x_i) = i/9$. The points $y_{(i)}$ and x_i are shown in the table in the margin. The points $(x_i, y_{(i)})$ are shown plotted in Figure S9.2.

i	$y_{(i)}$	$i/9$	x_i
1	−18	1/9	−1.221
2	−16	2/9	−0.765
3	−12	3/9	−0.431
4	−6	4/9	−0.140
5	0	5/9	0.140
6	4	6/9	0.431
7	4	7/9	0.765
8	12	8/9	1.221

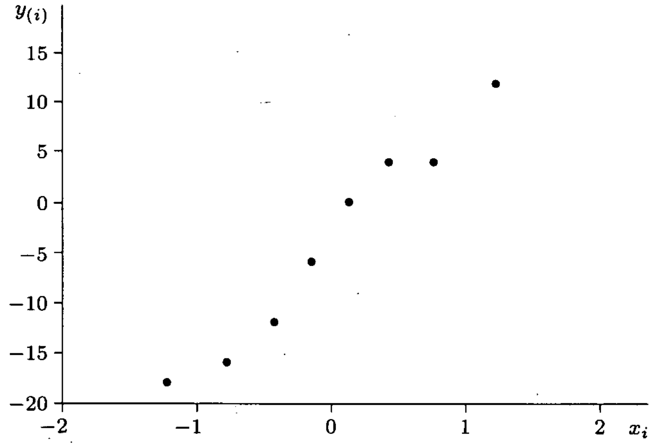


Figure S9.2 Corneal thickness differences against normal scores

The points do not appear to lie on a straight line and the evidence in favour of a normal modelling distribution for the differences in corneal thickness is not strong. However, there is no systematic pattern to the points.

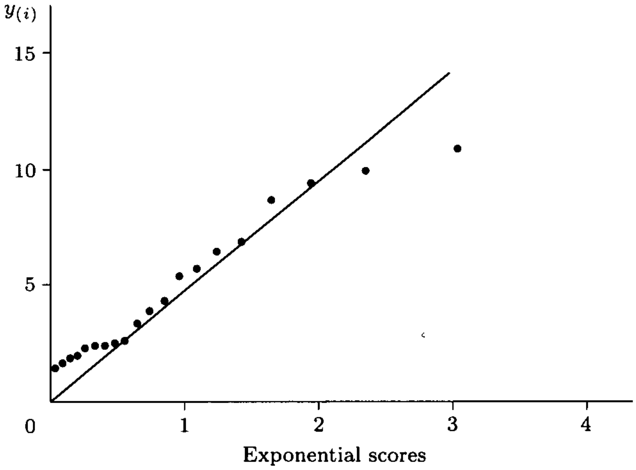
Solution 9.3

Since $n = 20$, we plot $y_{(i)}$ against $x_i = -\log((21 - i)/21)$.

Normally you would need a calculator for the logarithms, but in this case the x_is are the same as they are in Table 9.5.

i	$y_{(i)}$	x_i	i	$y_{(i)}$	x_i
1	1.45	0.049	11	3.87	0.742
2	1.67	0.100	12	4.33	0.847
3	1.90	0.154	13	5.35	0.965
4	2.02	0.211	14	5.72	1.099
5	2.32	0.272	15	6.48	1.253
6	2.35	0.336	16	6.90	1.435
7	2.43	0.405	17	8.68	1.658
8	2.47	0.480	18	9.47	1.946
9	2.57	0.560	19	10.00	2.351
10	3.33	0.647	20	10.93	3.045

The exponential probability plot ($y_{(i)}$ against x_i) is given in Figure S9.3.



Remember that for an exponential probability plot, the fitted straight line must pass through the origin.

Figure S9.3 Unpleasant memory recall times against exponential scores

The points do not lie on a straight line and the evidence does not support an exponential modelling distribution for the variation in recall times.

Solution 9.4

The normal probability plot, together with a fitted straight line, for the data on the 84 Etruscan skulls is shown in Figure S9.4.

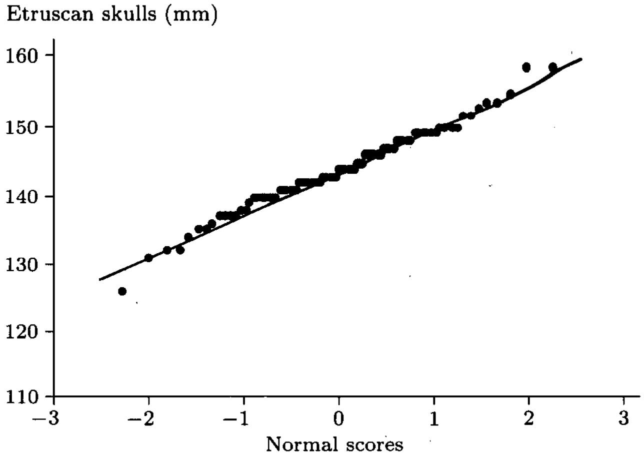


Figure S9.4 Normal probability plot, Etruscan skulls

The fit appears to be very good, supporting an earlier assertion that the variation in the data may plausibly be modelled by a normal distribution. Notice also that the fitted line has intercept at about 144 ($\overline{x}_E = 143.8$) and slope about 6 ($s_E = 5.97$).

The corresponding plot for the 70 modern Italian skulls is given in Figure S9.5. Again, a straight line appears to fit the points very well (though there is a departure from the fitted line at both extremes—small skulls are surprisingly small and large skulls are surprisingly large). The fitted line in the diagram has intercept at about 132 ($\overline{x}_I = 132.4$) and slope again about 6 ($s_I = 5.75$).

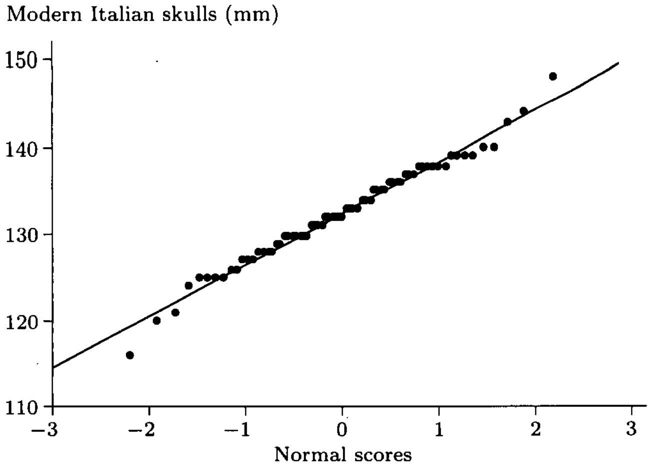


Figure S9.5 Normal probability plot, modern Italian skulls

Solution 9.5

(a) The exponential probability plot for the 62 waiting times between earthquakes, together with a fitted straight line through the origin, is given in Figure S9.6. The fit looks very good. (The slope of the line appears to be about 450: the sample mean for the data set is 437.)

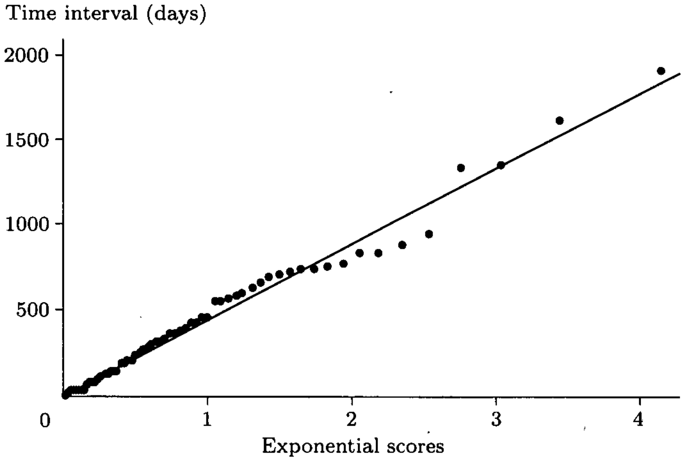


Figure S9.6 Exponential probability plot, waiting times between earthquakes

(b) In this case (see Figure S9.7) there is a clear departure from linearity, suggesting that the variation in the waiting times between successive coal-mining disasters is other than exponential. Nevertheless, the plot does suggest some sort of systematic variation in these data.

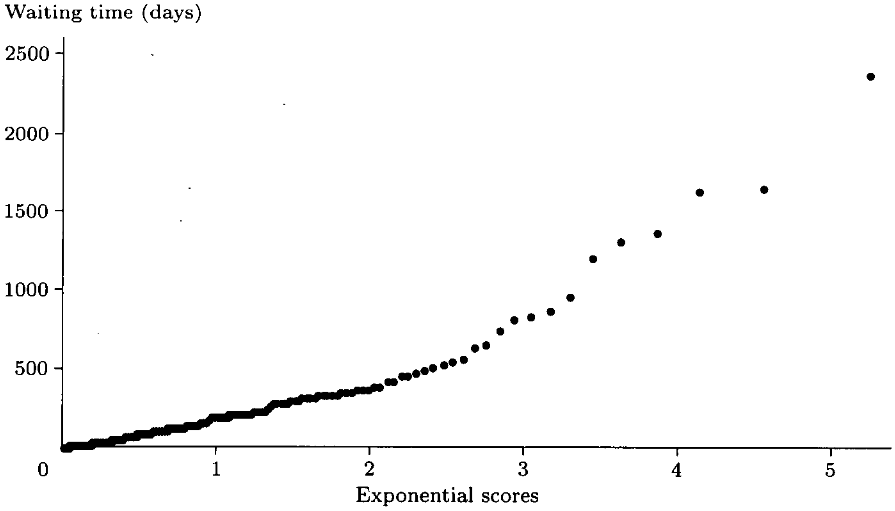


Figure S9.7 Exponential probability plot, waiting time between coal-mining disasters

Solution 9.6

Your computer may have provided you with 30 observations from the Pareto(8, 4) distribution similar to those shown in Table S9.2.

Table S9.2 Thirty observations from Pareto(8, 4)

8.18	10.91	8.73	8.17	12.90	8.88	10.31	8.19	9.59	13.73
11.68	8.52	13.14	10.10	8.44	10.72	8.18	8.12	8.33	9.20
8.78	10.41	11.49	9.54	12.55	12.28	17.26	9.07	8.05	9.99

Writing $W \sim \text{Pareto}(8, \theta)$, then the random variable $Y = \log(W/8)$ has an exponential distribution. The ordered sample $y_{(1)}, y_{(2)}, \ldots, y_{(30)}$ and the associated exponential scores $x_1, x_2, \ldots, x_{30}$ are shown in Table S9.3.

The plot of the points y_i against x_i, together with a fitted straight line through the origin, is given in Figure S9.8.

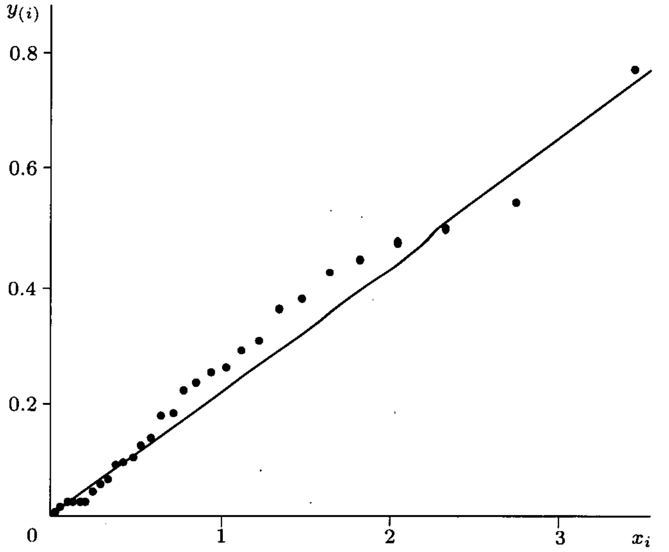


Figure S9.8 Probability plot, simulated Pareto data

Table S9.3 Exponential scores for Pareto data

i	$w_{(i)}$	$y_{(i)}$	x_i
1	8.05	0.006	0.033
2	8.12	0.015	0.067
3	8.17	0.021	0.102
4	8.18	0.022	0.138
5	8.18	0.022	0.176
6	8.19	0.023	0.215
7	8.33	0.040	0.256
8	8.44	0.054	0.298
9	8.52	0.063	0.343
10	8.73	0.087	0.389
11	8.78	0.093	0.438
12	8.88	0.104	0.490
13	9.07	0.126	0.544
14	9.20	0.140	0.601
15	9.54	0.176	0.661
16	9.59	0.181	0.726
17	9.99	0.222	0.795
18	10.10	0.233	0.869
19	10.31	0.254	0.949
20	10.41	0.263	1.036
21	10.72	0.293	1.131
22	10.91	0.310	1.237
23	11.49	0.362	1.355
24	11.68	0.378	1.488
25	12.28	0.429	1.642
26	12.55	0.450	1.825
27	12.90	0.478	2.048
28	13.14	0.496	2.335
29	13.73	0.540	2.741
30	17.26	0.769	3.434

If the maximum point of the 30 is ignored, there is a pronounced curve to the probability plot, suggesting in this case some systematic error with the Pareto generator.

Did your plot provide a more convincing straight line?

Solution 9.7

(a) There are four categories with expected frequencies

$$E_i = n\theta_i = 290\theta_i, \quad i = 1, 2, 3, 4,$$

where

$$\theta_1 = \tfrac{9}{16},\ \theta_2 = \tfrac{3}{16},\ \theta_3 = \tfrac{3}{16},\ \theta_4 = \tfrac{1}{16}.$$

A table for calculating the chi-squared test statistic is given in Table S9.4.

Table S9.4 *Pharbitis nil*, simple theory

i	O_i	E_i	$(O_i - E_i)$	$(O_i - E_i)^2/E_i$
1	187	163.125	23.875	3.49
2	35	54.375	−19.375	6.90
3	37	54.375	−17.375	5.55
4	31	18.125	12.875	9.15

The chi-squared value for the test is

$$\chi^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = 3.49 + 6.90 + 5.55 + 9.15 = 25.09.$$

Measured against $\chi^2(3)$ ($k = 4$; no model parameters were estimated from the data), this gives a SP of about 0.000 015. This is exceptionally small: there is very considerable evidence that the simple theory is flawed.

(b) Allowing for genetic linkage, then the expected frequencies are considerably changed, as shown in Table S9.5.

For instance,
$E_1 = n\theta_1 = 290 \times 0.6209 = 180.061.$

Table S9.5 *Pharbitis nil*, genetic linkage

i	O_i	E_i	$(O_i - E_i)$	$(O_i - E_i)^2/E_i$
1	187	180.061	6.939	0.27
2	35	37.439	−2.439	0.16
3	37	37.439	−0.439	0.01
4	31	35.061	−4.061	0.47

The chi-squared value for this test is

$$\chi^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = 0.27 + 0.16 + 0.01 + 0.47 = 0.91.$$

Measured against $\chi^2(2)$ ($k = 4$; one model parameter was estimated from the data so $p = 1$), this gives a SP of 0.63. There is no evidence to reject the genetic linkage.

Solution 9.8

Pielou assumed a geometric modelling distribution with parameter $p = 0.657$ for the following data; expected frequencies obtained by multiplying the hypothesized probability mass function by 109 are included in Table S9.6. For instance, $E_4 = 109\theta_4 = 109(0.343)^3(0.657) = 2.89$.

Run lengths of diseased trees in an infected plantation

Run length	1	2	3	4	5	6	> 6
Observed number of runs	71	28	5	2	2	1	0
Estimated number of runs	71.61	24.56	8.43	2.89	0.99	0.34	0.18

Pooling runs of length 3 or more and performing the chi-squared test calculation gives Table S9.6.

In fact, pooling runs of length 4 or more gives an expected frequency of $2.89 + 0.99 + 0.34 + 0.18 = 4.40$ which is less than 5; but the conclusions of the chi-squared test would not be seriously adrift.

Table S9.6 Diseased trees: testing a geometric fit

Run length	O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2/E_i$
1	71	71.61	−0.61	0.005
2	28	24.56	3.44	0.482
≥ 3	10	12.83	−2.83	0.624

The test statistic is

$$\chi^2 = \sum_{i=1}^{3} \frac{(O_i - E_i)^2}{E_i} = 0.005 + 0.482 + 0.624 = 1.1.$$

One parameter ($p = 0.657$) was estimated from these data, so the chi-squared null distribution has $(3 - 1 - 1) = 1$ degree of freedom for a SP of 0.29. The geometric distribution is not rejected as a model and this confirms that Pielou's assumptions were not unreasonable.

Solution 9.9

The suggested grouping gives the following observed frequencies.

Level	< 150	$150 - 250$	$250 - 350$	$350 - 450$	≥ 450
Observed frequency	4	14	19	7	11

Working to full computer accuracy, the expected frequencies and chi-squared test calculations are as shown in Table S9.7, using $\overline{x} = 314.91$ and $s = 131.16$.

Table S9.7

O_i	E_i	$(O_i - E_i)$	$(O_i - E_i)^2/E_i$
4	5.74	−1.74	0.53
14	11.33	2.67	0.63
19	16.23	2.77	0.47
7	13.37	−6.37	3.03
11	8.33	2.67	0.86

The value of the test statistic is

$$\chi^2 = \sum_{i=1}^{5} \frac{(O_i - E_i)^2}{E_i} = 0.53 + 0.63 + \cdots + 0.86 = 5.52.$$

Two parameters were estimated from these data, so the chi-squared distribution has $(5 - 2 - 1) = 2$ degrees of freedom for a SP of 0.061. There is some evidence for rejection of the null hypothesis that the data are fitted by a normal distribution.

Solution 9.10

The details of observed and expected frequencies, and of the chi-squared test calculations are given in Table S9.8, using $\overline{x} = 1369.1$ and $s = 693.7$.

Table S9.8 Chi-squared calculations, rainfall data

Rainfall	O_i	E_i	$(O_i - E_i)$	$(O_i - E_i)^2/E_i$
< 600	5	6.29	−1.29	0.26
600–1000	14	7.69	6.31	5.18
1000–1400	10	10.36	−0.36	0.01
1400–1800	8	10.10	−2.10	0.44
1800–2200	5	7.13	−2.13	0.64
≥ 2200	5	5.43	−0.43	0.03

The value of the chi-squared test statistic is

$$\chi^2 = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} = 0.26 + 5.18 + \cdots + 0.03 = 6.56.$$

Two parameters were estimated from these data, so the chi-squared distribution has $(6 - 2 - 1) = 3$ degrees of freedom for a SP of 0.09. There is insufficient evidence for rejection of the null hypothesis that the data are fitted by a normal distribution, although the fit is not good and one should have reservations about it.

Solution 9.11

The sample skewness of birth weights for the group of children who survived is 0.229; and that for the group of children who died is 0.491.

The skewness of the group of children who died is rather higher than one would like and it is worth trying to transform the data. One possible transformation is to take logarithms: this reduces the sample skewnesses to −0.291 and 0.177 respectively.

It is worth noticing that this transformation not only gives skewnesses for the two groups that are similar in size and opposite in sign but also gives approximately equal variances to the two groups.

Solution 9.12

A two-sample t-test for equal means may now be carried out. This gives a t-statistic of −3.67 on 48 degrees of freedom, with a total SP equal to 0.0006. It may be concluded that there is a significant difference, provided the residuals are plausibly normal. Subtract 0.482 from the transformed data of the first group (that is, those who died) and subtract 0.795 from the transformed data of the second group (survivors); pool the data and construct a normal probability plot. This is shown in Figure S9.9.

It is possible to fit an acceptable straight line and there is strong evidence for rejection of the null hypothesis that there is no difference between the groups.

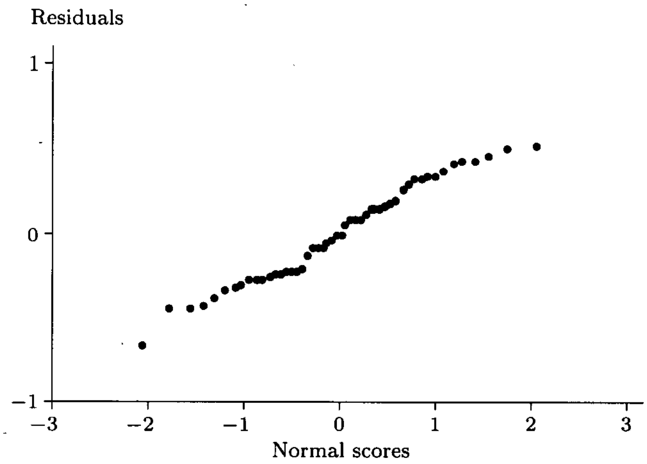


Figure S9.9 Residuals against normal scores

Solution 9.13

The analysis is very similar to that in Example 9.8. It turns out that the same transformation works well.

As before, we can carry out a two-sample t-test for equal means. This gives a t-statistic of 2.46 on 52 degrees of freedom, which has a SP equal to 0.017. It may be concluded that there is evidence of a significant difference, provided the residuals are plausibly normal. Subtract 4.565 from the transformed data of the group with less formal education and subtract 3.029 from the transformed data of the group with more formal education, pool the data and construct a normal probability plot. This is shown in Figure S9.10.

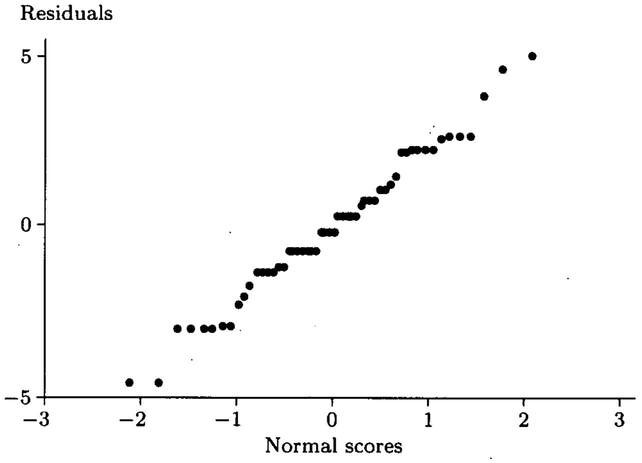


Figure S9.10 Residuals against normal scores

This does not suggest a good straight line, although it may be just about acceptable. This casts doubt on our assumption of normality.

Solution 9.14

(a) Subtracting 0.618 from each entry in Table 8.4 and allocating signed ranks produces the following table.

Difference	0.075	0.044	0.072	−0.012	−0.048	0.131	0.054	0.010	−0.009	0.226
Sign	+	+	+	−	−	+	+	+	−	+
Rank	17	10	16	$5\frac{1}{2}$	11	18	14	4	3	19

Difference	0.036	−0.003	0.050	−0.017	−0.042	0.052	−0.012	−0.007	−0.065	0.315
Sign	+	−	+	−	−	+	−	−	−	+
Rank	8	1	12	7	9	13	$5\frac{1}{2}$	2	15	20

There are no 0s and there are only two tied differences.

(b) The sums of the ranks are 151 for the positive differences and 59 for the negative differences, thus the Wilcoxon signed rank statistic is 151. This gives

$$SP(\text{obtained direction}) = SP(\text{opposite direction}) = 0.044;$$

$$SP(\text{total}) = 0.088.$$

There is some evidence for rejection of the null hypothesis of zero difference; in other words, there is some evidence that the rectangles do not conform to the Greek standard.

(c) A t-test for zero difference gives a total SP of 0.054, which may be interpreted as giving some evidence, although not strong evidence, for rejection of the hypothesis. There must be doubt about such a result because of the lack of normality of the data.

Solution 9.15

The sample size is 20, so that

$$E(S) = \frac{n(n+1)}{4} = \frac{20 \times 21}{4} = 105,$$

$$V(S) = \frac{n(n+1)(2n+1)}{24} = \frac{20 \times 21 \times 41}{24} = 717.5.$$

Therefore, we have

$$z = \frac{151 - 105}{\sqrt{717.5}} = 1.717$$

which is the 0.957 quantile of the standard normal distribution. The SP is therefore 0.086, which is very close to that given by the exact test in Exercise 9.14.

Solution 9.16

The ranks are as follows.

Pleasant memory	Rank	Unpleasant memory	Rank
1.07	1	1.45	5
1.17	2	1.67	7
1.22	3	1.90	8
1.42	4	2.02	10
1.63	6	2.32	$12\frac{1}{2}$
1.98	9	2.35	14
2.12	11	2.43	15
2.32	$12\frac{1}{2}$	2.47	16
2.56	17	2.57	18
2.70	19	3.33	25
2.93	20	3.87	27
2.97	21	4.33	28
3.03	22	5.35	31
3.15	23	5.72	33
3.22	24	6.48	35
3.42	26	6.90	36
4.63	29	8.68	37
4.70	30	9.47	38
5.55	32	10.00	39
6.17	34	10.93	40
	$345\frac{1}{2}$		$474\frac{1}{2}$

Labelling the pleasant memory recall times as group A, $u_A = 345\frac{1}{2}$. An exact test gives

$$SP(\text{total}) = 0.082.$$

Alternatively, using a normal approximation,

$$E(U_A) = \frac{n_A(n_A + n_B + 1)}{2} = \frac{20 \times 41}{2} = 410,$$

$$V(U_A) = \frac{n_A n_B(n_A + n_B + 1)}{12} = \frac{20 \times 20 \times 41}{12} = 1366.667$$

giving

$$z = \frac{345.5 - 410}{\sqrt{1366.667}} = -1.745.$$

Normal tables give a total SP of 0.081 for the two-sided test. Therefore, the conclusion is that the evidence for rejection of the null hypothesis, namely that memory recall times are different for pleasant and unpleasant memories, is not very strong.

Solution 9.17

The data are rather interesting. A naive estimate of the age of the site is the sample mean, 2622 years. However, Figure S9.11 shows (a) a boxplot for these data; (b) a normal probability plot; and (c) a normal probability plot with the single high outlier (3433 years) removed.

A straight line fits the seven points in Figure S9.11(c) very well. A good estimate of the age of the site would seem to be provided by the mean of the trimmed sample, 2506 years, nearly 1000 years less than the untrimmed estimate.

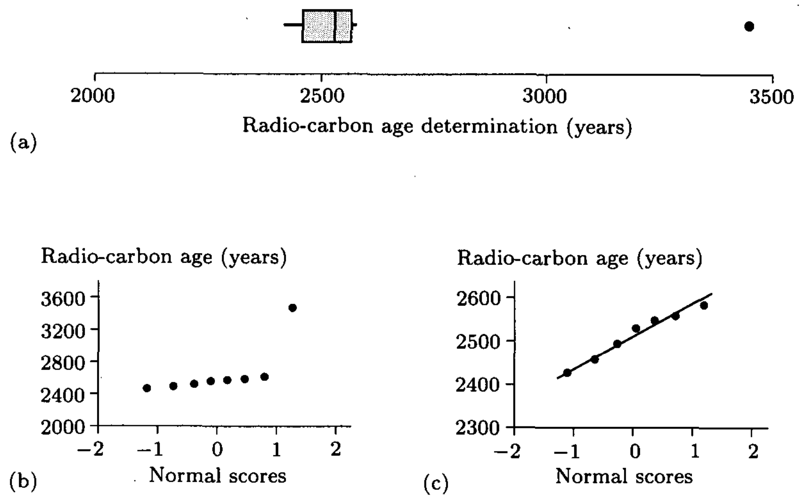


Figure S9.11 (a) Boxplot (b) Normal probability plot (c) Trimmed normal probability plot

Chapter 10

Solution 10.1

(a) The problem has been set up as a prediction problem, with bracket weight the explanatory variable (x) and beetle count the response variable (y). A scatter plot of beetle count against bracket weight is given in Figure S10.1.

Figure S10.1 also shows the fitted straight line found in part (b).

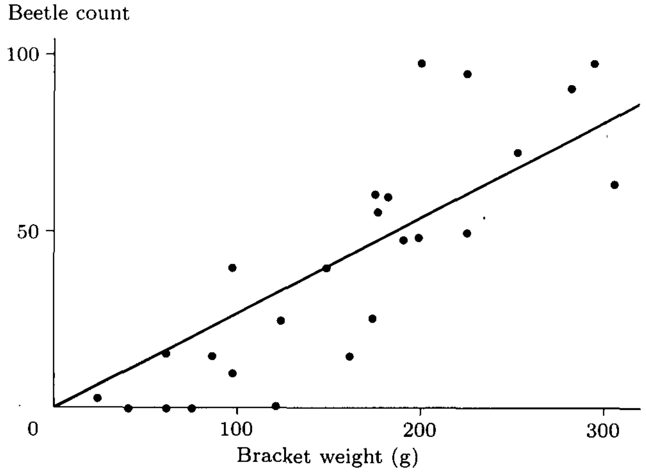


Figure S10.1 Beetle count against bracket weight

(b) From the scatter plot in part (a), it appears that a straight line through the origin would provide a useful regression model for these data:

$$Y_i = \gamma x_i + W_i, \quad i = 1, 2, \ldots, 25.$$

Using

$$\sum x_i y_i = 62 \times 16 + 226 \times 50 + \cdots + 162 \times 15$$
$$= 992 + 11\,300 + \cdots + 2430 = 219\,817$$

You will probably be able to obtain these sums directly from your calculator after keying in the data, without having to record subtotals.

and

$$\sum x_i^2 = 62^2 + 226^2 + \cdots + 162^2$$
$$= 3844 + 51\,076 + \cdots + 26\,244 = 796\,253,$$

the least squares estimate for the slope γ is given by

$$\widehat{\gamma} = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{219\,817}{796\,253} = 0.276.$$

(c) For a fungus weighing 240 g, the predicted beetle count is

$$240\widehat{\gamma} = 240 \times 0.276 = 66.3 \text{ beetles;}$$

say, 66 beetles.

(d) There are two useful representations for the residual sum of squares (see (10.4)). In either case we need the result $\sum y_i^2 = 68\,918$. Either say

$$\sum (y_i - \widehat{y}_i)^2 = \sum y_i^2 - \frac{\sum (x_i y_i)^2}{\sum x_i^2} = 68\,918 - \frac{219\,817^2}{796\,253} = 8234.4$$

or say

$$\sum (y_i - \widehat{y}_i)^2 = \sum y_i^2 - \widehat{\gamma}^2 \sum x_i^2 = 68\,918 - (0.276)^2(796\,253) = 8262.6.$$

Notice that the rounding error induced by using $\widehat{\gamma}$ at the calculator keypad (0.276 instead of 0.276 064 266) has been considerable.

A properly programmed computer will give you all these answers, including the scatter plot, from a few keyboard commands.

Solution 10.2

The summary statistics in this case (writing x for specific gravity and y for strength) are

$$n = 10, \quad \sum x_i = 4.951, \quad \sum y_i = 118.77, \quad \sum x_i^2 = 2.488\,995,$$
$$\sum x_i y_i = 59.211\,61,$$

so the slope estimate is

$$\widehat{\beta} = \frac{10 \times 59.211\,61 - 4.951 \times 118.77}{10 \times 2.488\,995 - 4.951^2}$$
$$= 10.8220,$$

and the estimate of the constant term is

$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\overline{x} = \tfrac{1}{10}(118.77 - 10.8220 \times 4.951) = 6.5190.$$

So the fitted model is

$$y = 6.52 + 10.82x$$

or

$$\text{Strength} = 6.52 + 10.82 \times \text{Specific gravity.}$$

Solution 10.3

For the finger-tapping data (tapping frequency y against caffeine dose x), the summary statistics are

$$n = 30, \quad \sum x_i = 3000, \quad \sum y_i = 7395, \quad \sum x_i^2 = 500\,000,$$
$$\sum x_i y_i = 743\,000,$$

so the slope estimate is

$$\begin{aligned}\widehat{\beta} &= \frac{30 \times 743\,000 - 3\,000 \times 7395}{30 \times 500\,000 - 3000^2} \\ &= \frac{105\,000}{6\,000\,000} = 0.0175,\end{aligned}$$

and the constant term is estimated by

$$\begin{aligned}\widehat{\alpha} &= \overline{y} - \widehat{\beta}\overline{x} \\ &= \tfrac{1}{30}(7395 - 0.0175 \times 3000) \\ &= 244.75.\end{aligned}$$

So the fitted model is

$$y = 244.75 + 0.0175x$$

or

$$\text{Tapping frequency} = 244.75 + 0.0175 \times \text{Caffeine dose.}$$

Solution 10.4

(a) With the relevant data keyed in (or the appropriate data file accessed), most statistical software would provide the equation of the fitted straight line at a single command. For Forbes' data, the estimators are $\widehat{\alpha} = 155.30$ and $\widehat{\beta} = 1.90$, so the equation of the fitted line is given by

$$\text{Boiling point} = 155.30 + 1.90 \times \text{Atmospheric pressure,}$$

where temperature is measured in °F and atmospheric pressure in inches Hg.

(b) For Hooker's data, $\widehat{\alpha} = 146.67$, $\widehat{\beta} = 2.25$; the fitted line has equation

$$\text{Boiling point} = 146.67 + 2.25 \times \text{Atmospheric pressure.}$$

Solution 10.5

(a) The problem here was to predict morbidity rates from mortality rates, so in the following scatter diagram (see Figure S10.2) morbidity rates (y) are plotted against mortality rates (x).

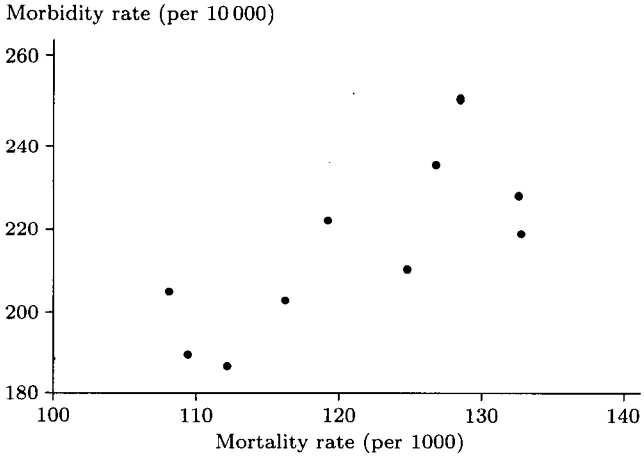


Figure S10.2 Morbidity rates against mortality rates

(b) The points are suggestive of a useful straight line fit. From $\widehat{\alpha} = 16.5478$ and $\widehat{\beta} = 1.6371$ it follows that the least squares regression line equation is given by

$$\text{Morbidity rate (per 1000)} = 16.55 + 1.64 \times \text{Mortality rate (per 10 000)}.$$

Solution 10.6

(a) A heavy car uses a lot of fuel because it is heavy: it is possible that under unusual circumstances one might wish to predict kerb weight from fuel consumption figures, but in general the problem would be to estimate fuel consumption, given the size of the car. Therefore plot consumption (miles per gallon, y) against kerb weight (kilograms, x). The appropriate scatter diagram is shown in Figure S10.3.

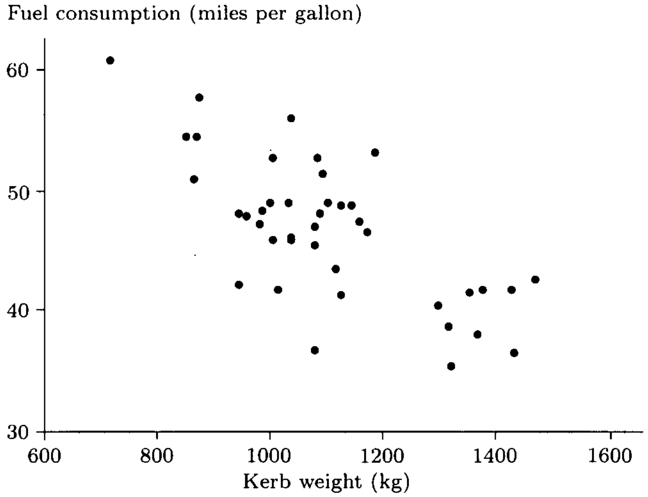


Figure S10.3 Consumption against kerb weight

(b) There is a lot of scatter: there is less evidence of an underlying 'formula' relating fuel consumption to kerb weight than there is in some other contexts. However, there is a pronounced downward trend. The least squares regression line has equation

$$\text{Fuel consumption} = 73.48 - 0.024 \times \text{Kerb weight},$$

where fuel consumption is measured in miles per gallon, and kerb weight in kilograms.

Solution 10.7

A plot of temperature against chirping frequency for these data is given in Figure S10.4.

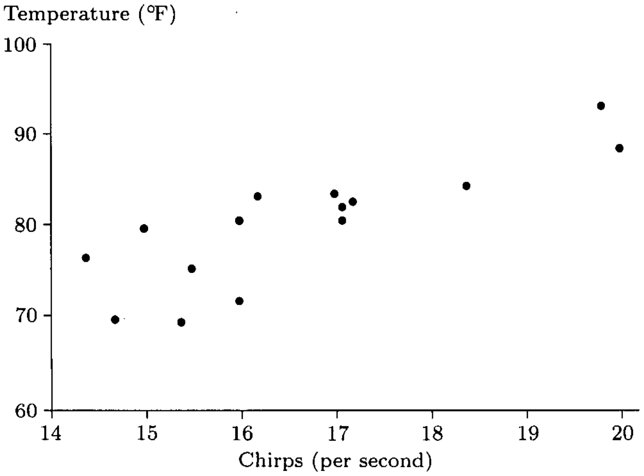


Figure S10.4 Temperature against chirping frequency

The least squares fitted line through the scattered points has equation

$$\text{Temperature} = 25.23 + 3.29 \times \text{Chirping frequency},$$

where temperature is measured in ℉ and chirps are counted every second. If $x_0 = 18$ chirps per second, the corresponding estimate of temperature, y_0, is given by

$$y_0 = 25.23 + 3.29x_0 = 25.23 + 3.29 \times 18 = 84.5\,℉.$$

Solution 10.8

Taking examination score as the response variable (Y) and time taken as the explanatory variable (x), then we fit the model

$$Y_i = \alpha + \beta x_i + W_i, \quad i = 1, 2, \ldots, 134,$$

where the random terms W_i are independent and normally distributed random variables with mean 0 and variance σ^2. The proposition that x is valueless as a predictor for Y is covered by the hypothesis

$$H_0 : \beta = 0.$$

For these data,

$$\widehat{\alpha} = 56.7333, \qquad \widehat{\beta} = -0.0012,$$

and the residual sum of squares is

$$\sum(y_i - \widehat{y}_i)^2 = \sum(y_i - \overline{y})^2 - \widehat{\beta}^2 \sum(x_i - \overline{x})^2 = 12\,809.38 - 41.82 = 12\,767.56.$$

So our estimator of variance is given by

$$s^2 = \frac{\sum(y_i - \widehat{y})^2}{n-2} = \frac{12\,767.56}{132} = 96.724.$$

The observed value of the test statistic

$$\frac{\widehat{\beta} - 0}{s/\sqrt{\sum(x_i - \overline{x})^2}} = \frac{-0.0012 - 0}{\sqrt{96.724}/\sqrt{28\,651\,067.56}} = -0.658$$

is at the 26% quantile of $t(132)$. So we have

$SP(\text{obtained direction}) = SP(\text{opposite direction}) = 0.26;$

$SP(\text{total}) = 0.52.$

The total SP is not small: the hypothesis that $\beta = 0$ is not rejected. This confirms our suspicion that examination time is not a good predictor for eventual score.

Solution 10.9

(a) A scatter diagram for percentage of non-contaminated peanuts against aflatoxin level is given in Figure S10.5.

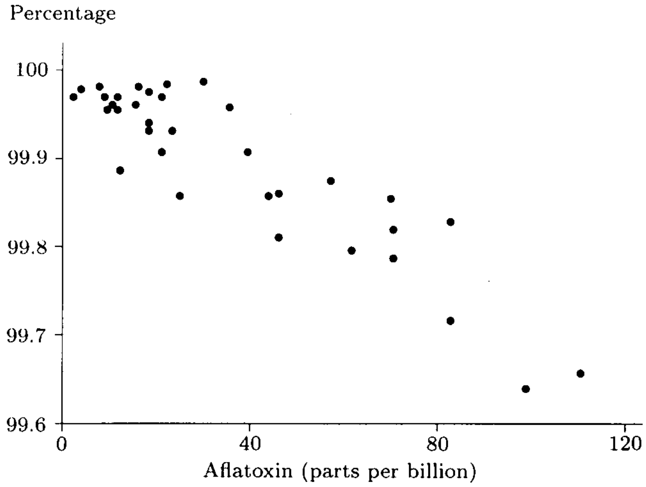


Figure S10.5 Percentage of non-contaminated against aflatoxin level

(b) The least squares regression line has equation

Percentage non-contaminated = 100.002 − 0.003 × Aflatoxin level

and the scatter plot suggests that a straight line would be a good fit to the data.

However, on the grounds of these results, and on the basis of what we know of the problem—that zero aflatoxin level would indicate 100% non-contamination—a better model would be to fit the constrained line

$$\text{Percentage non-contaminated} = 100 - \gamma \times \text{Aflatoxin level},$$

or, equivalently

$$y = 100 - \gamma x.$$

This model has one parameter: the equivalent hypothesis under test is $H_0 : \gamma = 0$.

(c) Proceeding, however, with the two-parameter model, the residual sum of squares is

$$\sum(y_i - \widehat{y}_i)^2 = \sum(y_i - \overline{y})^2 - \widehat{\beta}^2 \sum(x_i - \overline{x})^2 = 0.288\,645 - 0.239\,150$$
$$= 0.049\,495.$$

Our estimate of variance is

$$s^2 = \frac{0.049\,495}{32} = 0.001\,547,$$

and our test statistic is

$$\frac{\widehat{\beta} - 0}{s/\sqrt{\sum(x_i - \overline{x})^2}} = \frac{-0.003 - 0}{\sqrt{0.001\,547}/\sqrt{28\,367.7}} = -12.434.$$

The SP is negligible. Despite the slope estimate being a small number in absolute terms, it represents a significant downward trend (as indicated by the scatter plot in part (a)).

Solution 10.10

In this context, breathing resistance is the response variable (y) and height is the explanatory variable (x). A scatter plot of the data (see Figure S10.6) was not asked for in this question, but it is always a useful first step. (Some would say it is an essential first step of a regression analysis.)

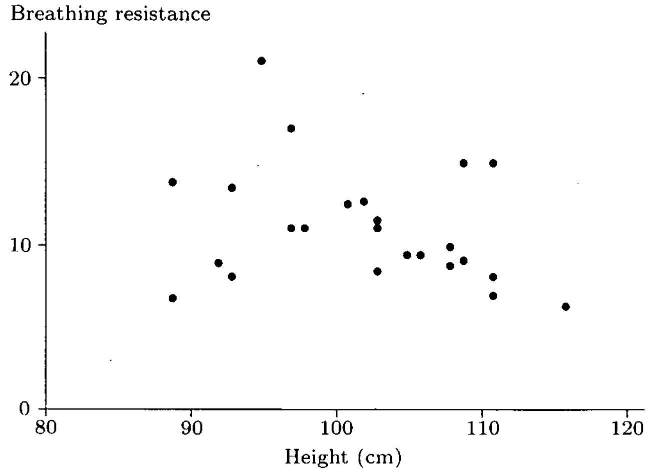


Figure S10.6 Breathing resistance against height

The fitted line through the scattered points has equation

$$y = 23.807 - 0.125x$$

and so the predicted breathing resistance for a child $x_0 = 100\,\mathrm{cm}$ tall is 11.346. From the data given in Table 10.15, the sample size is $n = 24$, the sample mean is $\overline{x} = 102.042$, $\sum(x_i - \overline{x})^2 = 1352.96$ and

$$\begin{aligned}\sum(y_i - \widehat{y}_i)^2 &= \sum(y_i - \overline{y})^2 - \widehat{\beta}^2 \sum(x_i - \overline{x})^2 \\ &= 284.64 - (0.125)^2(1352.96) = 263.6.\end{aligned}$$

Using a computer package, intermediate results are not really necessary: the residual sum of squares is a standard summary statistic in the context of regression.

So our estimate of σ^2 is

$$s^2 = \frac{263.6}{22} = 11.98.$$

The 97.5% quantile of $t(22)$ is 2.074, and consequently the 95% confidence interval for the mean breathing resistance for children 100 cm tall, based on these data, is

$$\begin{aligned}&\left(\widehat{\alpha} + \widehat{\beta}x_0 \pm q_{0.975}s\sqrt{\frac{(x_0 - \overline{x})^2}{\sum(x_i - \overline{x})^2} + \frac{1}{n}}\right) \\ &= \left(11.307 \pm 2.074\sqrt{11.98}\sqrt{\frac{(-2.042)^2}{1352.96} + \frac{1}{24}}\right) \\ &= (11.307 \pm 1.519) = (9.8, 12.8).\end{aligned}$$

The units of measurement are those for breathing resistance, not stated in the original table.

Solution 10.11

The important measures obtained from the data in Table 10.16 are

$$n = 42, \quad \overline{x} = 124.31, \quad \sum(x_i - \overline{x})^2 = 7058.98, \quad \widehat{\alpha} = 27.516,$$

$$\widehat{\beta} = -0.136\,06, \quad \sum(y_i - \widehat{y}_i)^2 = 796.055, \quad s^2 = 19.9014.$$

If all your calculations are done on a computer, these intermediate results are unnecessary.

Also, the 97.5% quantile of $t(40)$ is 2.021.

The predicted mean when x_0 is 100 is

$$\widehat{\alpha} + \widehat{\beta}x_0 = 27.516 - 0.136\,06 \times 100 = 13.91,$$

and the 95% prediction interval for this particular child's breathing resistance is given by

$$\begin{aligned}&\left(13.91 \pm 2.021\sqrt{19.9014}\sqrt{\frac{(-24.31)^2}{7058.98} + \frac{1}{42} + 1}\right) \\ &= (13.91 \pm 9.49) = (4.4, 23.4).\end{aligned}$$

The prediction interval is extremely wide: the reason for this is that there is a great deal of scatter in the data.

Solution 10.12

From the data given in Table 10.13

$$n = 15, \quad \overline{x} = 16.653, \quad \sum(x_i - \overline{x})^2 = 40.557, \quad \widehat{\alpha} = 25.2323,$$
$$\widehat{\beta} = 3.2911, \quad \sum(y_i - \widehat{y}_i)^2 = 190.547, \quad s^2 = 14.6575.$$

Some of these were calculated in Exercise 10.7, though not to as many places of decimals.

Also, the 99.5% quantile of $t(13)$ is 3.012. The predicted mean is

$$\widehat{\alpha} + \widehat{\beta}x_0 = 84.472.$$

The 99% prediction interval is

$$\left(84.472 \pm 3.012\sqrt{14.6575}\sqrt{\frac{(18 - 16.653)^2}{40.557} + \frac{1}{15} + 1}\right)$$
$$= (84.472 \pm 12.157) = (72.3, 96.6).$$

Solution 10.13

From the data given in Table 10.12,

$$n = 42, \quad \overline{x} = 1104.69, \quad \sum(x_i - \overline{x})^2 = 1\,229\,650.98, \quad \widehat{\alpha} = 73.480,$$
$$\widehat{\beta} = -0.0242, \quad \sum(y_i - \widehat{y}_i)^2 = 689.33, \quad s^2 = 17.2333.$$

Some of these were calculated in Exercise 10.6, though less precisely.

The 97.5% quantile of $t(40)$ is 2.021. The predicted mean consumption is

$$\widehat{\alpha} + \widehat{\beta}x_0 = 47.344.$$

The 95% prediction interval is

$$\left(47.354 \pm 2.021\sqrt{17.2333}\sqrt{\frac{(1080 - 1104.69)^2}{1\,229\,650.98} + \frac{1}{42} + 1}\right)$$
$$= (47.354 \pm 8.491) = (38.9, 55.8).$$

Chapter 11

Solution 11.1

In scatter plot (a) the variables are negatively related.

In scatter plot (b) the variables are positively related.

In scatter plot (c) the variables do not appear to be related at all. Knowing the value of one of them tells you nothing about the value of the other.

Solution 11.2

(a) The random variables X and Y are related, because the conditional probability that $Y = 10$ given $X = 4$ is not the same as the unconditional probability that $Y = 10$.

(b) In this case it is not possible to say whether W and Z are related. The question does not give enough information. Knowing that $W = 4$ tells us nothing new about the probability that Z took the value 5. However, we do not know what would happen if we knew, say, that W took the value 6. Would that change the probability distribution of Z? If knowing the values of W does not change the probability distribution of Z, then W and Z are not related, but otherwise they are related.

Solution 11.3

(a) Altogether, out of 2484 people, 110 provided a value of Yes for the random variable X. Thus an estimate for the probability $P(X = \text{Yes})$ is 110/2484 or 0.044.

(b) There are 254 people in Table 11.2 for whom the value of the random variable Y is Snore every night. Of these, 30 provided a value of Yes for X (heart disease). Thus an estimate for the conditional probability $P(X = \text{Yes}|Y = \text{Snore every night})$ is 30/254 or 0.118. This is getting on for three times the unconditional probability that $X = \text{Yes}$. That is, knowing that someone snores every night tells you something about how likely it is that they have heart disease: snorers are more likely to have it. (Note that this does not tell you that snoring causes heart disease, or for that matter that heart disease causes snoring. More on this point is discussed in Section 11.3.)

This could be put more formally, if you prefer. Unless for some reason these estimates are very inaccurate, it appears that

$$P(X = \text{Yes}|Y = \text{Snore every night}) \neq P(X = \text{Yes});$$

therefore, using (11.1), X and Y are related.

Solution 11.4

Your answer may be as follows. The relationship is negative, therefore the Pearson correlation r will be negative, and the amount of scatter in the data in Figure 11.3 is not too different from that in Figure 11.7(b). The value of r for Figure 11.7(b) is 0.787, so the value of r for Figure 11.3 might be around -0.7 or -0.8. Perhaps a guess of anything between -0.6 and -0.9 would be reasonable. In fact, the value of r for Figure 11.3 is -0.767.

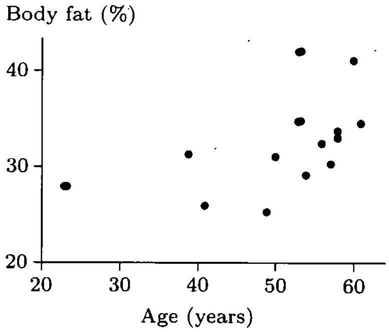


Figure S11.1 Body fat percentage against age

Solution 11.5

(a) The scatter plot is as shown in Figure S11.1. This clearly shows that the two variables are positively related, though the association between them is not particularly strong.

(b) For these data,

$$n = 14, \quad \sum x_i = 712, \quad \sum y_i = 452.5, \quad \sum x_i^2 = 37\,600,$$
$$\sum y_i^2 = 14\,937.57, \quad \sum x_i y_i = 23\,346.5.$$

So using (11.5),

$$r = \frac{14 \times 23\,346.5 - 712 \times 452.5}{\sqrt{(14 \times 37\,600 - 712^2)(14 \times 14\,937.57 - 452.5^2)}}$$
$$= \frac{4671}{\sqrt{19\,456 \times 4369.73}} = 0.507.$$

This value indicates a moderate positive association between the two variables.

Solution 11.6

(a) The correlation is 0.743. This matches the impression given by the scatter plot of a reasonably strong positive relationship between the variables.

If you want to check your computer output, the exact value of r using (11.5) is

$$r = \frac{37\,880}{\sqrt{2\,600\,135\,980}}.$$

(b) The scatter plot for these data is as shown in Figure S11.2. This shows a reasonably strong positive association between the variables. The Pearson correlation coefficient is 0.716, which gives the same impression of the strength of the relationship. The exact value of r is

$$r = \frac{193\,000}{\sqrt{72\,610\,213\,900}}.$$

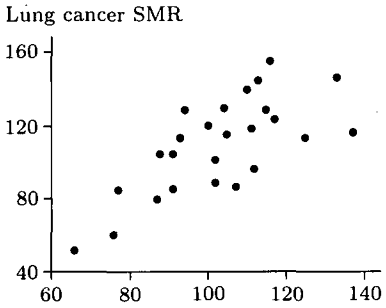


Figure S11.2 Lung cancer SMR against smoking ratio

Solution 11.7

(a) The scatter plot is as shown in Figure S11.3.

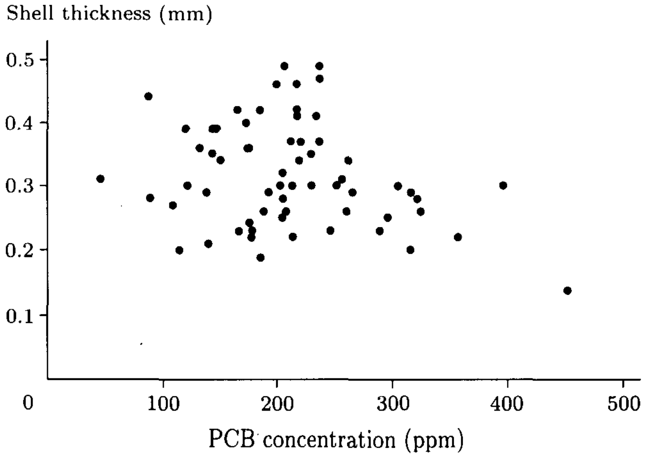


Figure S11.3 Shell thickness against PCB concentration

There seems to be a rather weak negative relationship between the two variables; the more the PCB, the thinner the shell. But remember that we cannot conclude from these data that the PCB causes the shells to become thin. (However, there is evidence from other sources that this causal explanation is true.)

Several points appear to be some distance from the main pattern; but the most obvious is the point at the bottom right. It is the thinnest shell with the highest concentration of PCB.

(b) The Pearson correlation coefficient for the full data set is −0.253. This is in line with the interpretation (a weak negative relationship) in part (a).

(c) Omitting the most extreme point at the bottom right from the calculation, the Pearson correlation becomes −0.157. This value is considerably nearer 0 than the correlation coefficient for the full data set. This indicates that much of the impression that these two variables are negatively correlated stems from this single egg, which seems to be rather atypical.

Solution 11.8

The Pearson correlation coefficient for the untransformed data is −0.005. For the log transformed data, it is 0.779.

Solution 11.9

(a) For these data, $r_S = 0.716$. This fairly large value corresponds to the fairly strong linear association seen in the scatter plot of the data after they had been transformed.

(b) The Spearman rank correlation coefficient would be the same as for the original data, which is 0.716. This is because the logarithmic transformation does not change the order in which the data come; consequently, the ranks of the log-transformed data are the same as the ranks of the original data. (If you do not believe this, check it!)

Solution 11.10

The computed SP is

$$SP(\text{obtained direction}) = 0.0323;$$

so there is moderate evidence of a positive relationship between these two variables in the population.

Solution 11.11

(a) In this case you should calculate $r\sqrt{\frac{n-2}{1-r^2}}$ and compare it against a t-distribution with $64 - 2 = 62$ degrees of freedom. The value of this quantity is

$$-0.157\sqrt{\frac{64-2}{1-(-0.157)^2}} = -1.256.$$

Hence the total SP is 0.214. There is no evidence of a relationship between the variables. Our previous impression, that the apparent relationship between the variables depends on the single extreme egg, is confirmed.

(b) The calculations are much the same as for part (a). You need to calculate the test statistic and compare it against a t-distribution on $28 - 2 = 26$ degrees of freedom. The value of this quantity is

$$r_S\sqrt{\frac{n-2}{1-r_S^2}} = 0.716\sqrt{\frac{28-2}{1-0.716^2}} = 5.232.$$

It is not really necessary to turn on your computer or open your statistical tables to see that the obtained SP is approximately zero. There is very strong evidence that the two variables are related.

Solution 11.12

Using Fisher's exact test, $SP = 0.0001$. There is very strong evidence that the two variables are related. It is clear from the original table that the relationship works in such a way that patients with impaired sulphoxidation are more likely to exhibit a toxic reaction to the drug.

Solution 11.13

The expected frequencies are as follows.

Season	Colour pattern		Row totals
	Bright red	Not bright red	
Spring	280.918	223.082	504
Summer	93.082	73.918	167
Column totals	374	297	671

In all four cells, the absolute difference between observed and expected frequencies is 21.082. This gives

$$\chi^2 = \frac{21.082^2}{280.918} + \frac{21.082^2}{223.082} + \frac{21.082^2}{93.082} + \frac{21.082^2}{73.918} = 14.36$$

and the SP is 0.000 15. There is strong evidence of association between season and colour pattern. Comparing the expected and observed frequencies, we see that there are more bright red beetles in spring and fewer in summer than is expected under the null hypothesis of no association.

Solution 11.14

(a) Looking at the observed and expected frequencies, Compressor 1 seems to fail relatively frequently in the centre leg and relatively rarely in the south leg, while Compressor 4 fails relatively rarely in the centre leg and relatively frequently in the south leg. However, the value of the chi-squared test statistic is 11.72, and since there are 4 rows and 3 columns, the number of degrees of freedom is $(4-1)(3-1) = 6$. The significance probability is $SP = 0.068$. There is only very weak evidence of association between the two variables: the data are consistent with the null hypothesis that the pattern of location of failures is the same in all four compressors.

(b) The value of the chi-squared test statistic is 7.885. The number of degrees of freedom is $(3-1)(2-1) = 2$, and the significance probability is $SP = 0.019$. There is fairly strong evidence of a relationship between tonsil size and carrier status. Comparing the observed and expected frequencies, it appears that carriers are less likely than non-carriers to have normal tonsil size, and carriers are more likely than non-carriers to have very large tonsils. In short, on average, carriers have larger tonsils than do non-carriers.

Solution 11.15

Your scatter plot should look like the scatter plot in Figure S11.4.

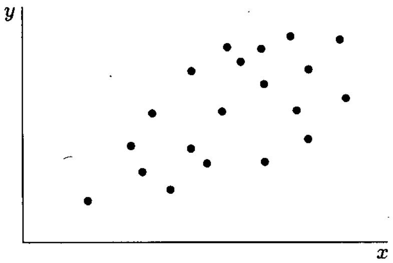


Figure S11.4

Chapter 12

Solution 12.1

(a) The adequacy of the Bernoulli model would rather depend on what service was being provided. For instance, queues at a bank or a post office might consist largely of individuals who have arrived at the service point independently of one another, and alone. On the other hand, queues at a cinema box office will often include male–female pairs, implying a strong dependence between consecutive individuals.

(b) Rather as for sequences of wet and dry days, it is likely that there will be some noticeable association between the characteristics of consecutive days, and a Bernoulli model would not be appropriate.

(c) Some card-players strenuously maintain a belief in runs of luck (good and bad) and if the phenomenon exists, then the Bernoulli process will not be an appropriate model here. For games involving some skill, it is probable that players perform better on some occasions than on others, over relatively long periods. For games involving chance alone there may be some association (the order of cards dealt) but it would be difficult to demonstrate and very hard to quantify.

Solution 12.2

(a) Using the probabilities at (12.1), three different weekly weather sequences (starting with a wet day) were simulated. They were as follows.

1100000
1001100
1000000

You should have generated a sequence similar to these.

(b) Four different families of size 4 were generated (starting with a girl). They were as follows.

0101
0111
0101
0111

Again, you should have generated a similar sequence.

Solution 12.3

(a) Using (12.3), the transition matrix M is given by

$$M = \begin{array}{c} 0 \\ 1 \end{array} \begin{bmatrix} 0.4567 & 0.5433 \\ 0.5007 & 0.4993 \end{bmatrix},$$

and the overall probability of a boy is given by

$$p = \frac{\alpha}{\alpha + \beta} = \frac{0.5433}{0.5433 + 0.5007} = 0.5204.$$

(b) Three typical families of 5 children were generated. They were

0 0 1 1 0
0 1 0 1 1
1 1 0 1 1.

Solution 12.4

(a) The matrix of transition frequencies is

$$N = \begin{array}{c} 0 \\ 1 \end{array} \left[\begin{array}{cc} 9 & 8 \\ 9 & 13 \end{array}\right] \begin{array}{c} 17 \\ 22 \end{array};$$

the corresponding matrix of estimated transition probabilities is

$$\widehat{M} = \begin{array}{c} 0 \\ 1 \end{array} \left[\begin{array}{cc} 9/17 & 8/17 \\ 9/22 & 13/22 \end{array}\right] = \begin{array}{c} 0 \\ 1 \end{array} \left[\begin{array}{cc} 0.529 & 0.471 \\ 0.409 & 0.591 \end{array}\right].$$

(b) $$N = \begin{array}{c} 0 \\ 1 \end{array} \left[\begin{array}{cc} 15 & 10 \\ 10 & 4 \end{array}\right] \begin{array}{c} 25 \\ 14 \end{array}; \quad \widehat{M} = \begin{array}{c} 0 \\ 1 \end{array} \left[\begin{array}{cc} 15/25 & 10/25 \\ 10/14 & 4/14 \end{array}\right] = \begin{array}{c} 0 \\ 1 \end{array} \left[\begin{array}{cc} 0.600 & 0.400 \\ 0.714 & 0.286 \end{array}\right].$$

(c) $$N = \begin{array}{c} 0 \\ 1 \end{array} \left[\begin{array}{cc} 12 & 2 \\ 3 & 22 \end{array}\right] \begin{array}{c} 14 \\ 25 \end{array}; \quad \widehat{M} = \begin{array}{c} 0 \\ 1 \end{array} \left[\begin{array}{cc} 12/14 & 2/14 \\ 3/25 & 22/25 \end{array}\right] = \begin{array}{c} 0 \\ 1 \end{array} \left[\begin{array}{cc} 0.857 & 0.143 \\ 0.120 & 0.880 \end{array}\right].$$

(d) $$N = \begin{array}{c} 0 \\ 1 \end{array} \left[\begin{array}{cc} 25 & 4 \\ 3 & 7 \end{array}\right] \begin{array}{c} 29 \\ 10 \end{array}; \quad \widehat{M} = \begin{array}{c} 0 \\ 1 \end{array} \left[\begin{array}{cc} 25/29 & 4/29 \\ 3/10 & 7/10 \end{array}\right] = \begin{array}{c} 0 \\ 1 \end{array} \left[\begin{array}{cc} 0.862 & 0.138 \\ 0.300 & 0.700 \end{array}\right].$$

(e) $$N = \begin{array}{c} 0 \\ 1 \end{array} \left[\begin{array}{cc} 9 & 14 \\ 15 & 1 \end{array}\right] \begin{array}{c} 23 \\ 16 \end{array}; \quad \widehat{M} = \begin{array}{c} 0 \\ 1 \end{array} \left[\begin{array}{cc} 9/23 & 14/23 \\ 15/16 & 1/16 \end{array}\right] = \begin{array}{c} 0 \\ 1 \end{array} \left[\begin{array}{cc} 0.391 & 0.609 \\ 0.938 & 0.063 \end{array}\right].$$

(f) $$N = \begin{array}{c} 0 \\ 1 \end{array} \left[\begin{array}{cc} 12 & 13 \\ 13 & 1 \end{array}\right] \begin{array}{c} 25 \\ 14 \end{array}; \quad \widehat{M} = \begin{array}{c} 0 \\ 1 \end{array} \left[\begin{array}{cc} 12/25 & 13/25 \\ 13/14 & 1/14 \end{array}\right] = \begin{array}{c} 0 \\ 1 \end{array} \left[\begin{array}{cc} 0.480 & 0.520 \\ 0.929 & 0.071 \end{array}\right].$$

Solution 12.5

For Exercise 12.4(c) the matrix of transition frequencies is

$$N = \begin{array}{c} 0 \\ 1 \end{array} \left[\begin{array}{cc} 12 & 2 \\ 3 & 22 \end{array}\right] \begin{array}{c} 14 \\ 25 \end{array};$$

the number of runs is

$$r = 2 + 3 + 1 = 6.$$

For Exercise 12.4(e) the matrix of transition frequencies is

$$N = \begin{array}{c} 0 \\ 1 \end{array} \left[\begin{array}{cc} 9 & 14 \\ 15 & 1 \end{array}\right] \begin{array}{c} 23 \\ 16 \end{array};$$

the number of runs is

$$r = 14 + 15 + 1 = 30.$$

Solution 12.6

(a) The total SP is 0.421; there is no evidence to reject a Bernoulli model here.

(b) The total SP is 0.853; there is no evidence to reject a Bernoulli model here.

(c) The total SP is 1.4×10^{-6}, which is very small indeed. The Bernoulli model is firmly rejected. (In fact, the realization shows a very small number of long runs.)

(d) The total SP is 0.004; the realization shows a small number of long runs inconsistent with a Bernoulli model.

(e) The SP is very small (0.001); but here there are many short runs inconsistent with a Bernoulli model.

(f) The total SP is 0.016; there is evidence to reject the Bernoulli model in the light of many short runs.

Solution 12.7

(a) The matrix of transition frequencies is

$$N = \begin{array}{c} 0 \\ 1 \end{array} \begin{bmatrix} 7 & 4 \\ 5 & 7 \end{bmatrix} \begin{array}{c} 11 \\ 12 \end{array};$$

the corresponding matrix of estimated transition probabilities is

$$\widehat{M} = \begin{array}{c} 0 \\ 1 \end{array} \begin{bmatrix} 7/11 & 4/11 \\ 5/12 & 7/12 \end{bmatrix} = \begin{array}{c} 0 \\ 1 \end{array} \begin{bmatrix} 0.636 & 0.364 \\ 0.417 & 0.583 \end{bmatrix}$$

(b) The number of runs in the data is

$$r = n_{01} + n_{10} + 1 = 4 + 5 + 1 = 10.$$

(c) The total SP against a hypothesized Bernoulli model is $SP = 0.301$. There is no evidence to reject a Bernoulli model for the sequence of 0s and 1s here.

(d) In this case, $n_0 = n_1 = 12$ and so from (12.7)

$$E(R) = \frac{2 \times 12 \times 12}{12 + 12} + 1 = 13, \qquad V(R) = \frac{288 \times 264}{24^2 \times 23} = 5.739.$$

The corresponding z-score is

$$z = \frac{r - E(R)}{SD(R)} = \frac{10 - 13}{\sqrt{5.739}} = -1.252;$$

the corresponding SP is $2 \times \Phi(-1.252) = 0.211$. The number of runs is not noticeably extreme (high or low). Again, there is no evidence to reject the Bernoulli model here.

Solution 12.8

(a) The accumulated times are given in Table S12.1.

Table S12.1

13.2	14.7	16.5	24.1	32.3
43.0	52.4	60.5	62.0	64.3
67.3	68.9	73.4	80.7	90.1
103.9	134.0	135.2	136.0	142.2
149.2	156.6	162.9	164.1	168.3
172.8	175.1	176.8	179.4	183.0
190.5	200.8	204.1	232.3	255.4
257.1	259.5	263.0	279.2	280.7
293.6	299.2	303.5	316.7	319.7
325.3	329.7	344.6	345.8	349.8

Observation ceased with the passing of the 50th vehicle at time $\tau = 349.8$. This leaves 49 vehicle observations:

$$w_1 = \frac{t_1}{\tau} = \frac{13.2}{349.8} = 0.038, \quad w_2 = \frac{t_2}{\tau} = \frac{14.7}{349.8} = 0.042, \quad \ldots,$$

$$w_{49} = \frac{t_{49}}{\tau} = \frac{345.8}{349.8} = 0.989;$$

the observed value of the Kolmogorov test statistic D is $d = 0.090$ with $n = 49$. The corresponding SP exceeds 0.2. We can conclude that the Poisson process is a reasonable model for the traffic flow along Burgess Road during the observation period.

(b) Here τ is known to be 1608; we therefore have 56 observations

$$w_1 = \frac{t_1}{\tau} = \frac{28}{1608} = 0.017, \quad w_2 = \frac{t_2}{\tau} = \frac{38}{1608} = 0.024, \quad \ldots,$$

$$w_{56} = \frac{t_{56}}{\tau} = \frac{1591}{1608} = 0.989;$$

the observed value of D is $d = 0.159$ with $n = 56$. The corresponding SP is between 0.1 and 0.2: there is some small evidence that volcano eruptions are not well fitted by a Poisson process.

Chapter 13

Solution 13.1

Use of the word *to* provides possibly the most difficult discriminant of the three: *by* would have been easier! But we are obliged to use the data available, not the data we wish had been available. Histograms showing Hamilton's and Madison's use of the word *to* are given in Figure S13.1.

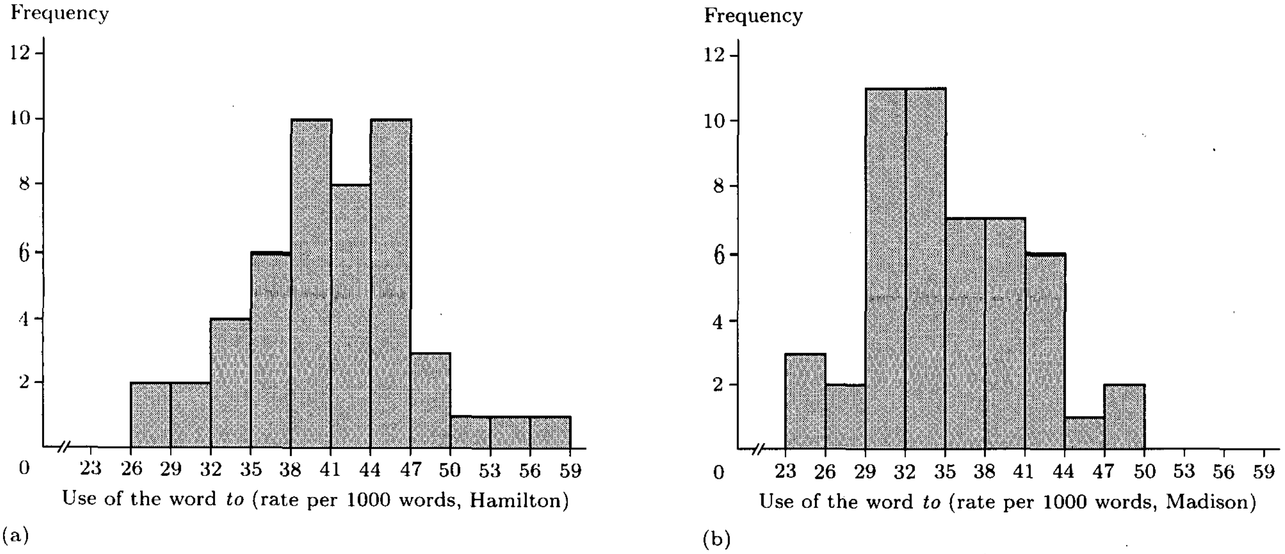


Figure S13.1 Distribution of rates of occurrence of *to* in (a) 48 Hamilton papers and (b) 50 Madison papers

If we were sure that all twelve disputed papers were by the same author, then by comparison with Figure 13.1, the histograms might possibly suggest Madison as the author.

Solution 13.2

The completed table of observed and expected frequencies, showing the chi-squared calculations, is given in Table S13.1. The last four cells were pooled to ensure that all expected frequencies were at least 5.

Table S13.1 Weldon's data: chi-squared calculations

Number of 5s or 6s	Observed frequency	Expected frequency	$O_i - E_i$	$\frac{(O_i - E_i)^2}{E_i}$
0	45	54.0	−9.0	1.50
1	327	324.0	3.0	0.03
2	886	891.0	−5.0	0.03
3	1475	1484.9	−9.9	0.07
4	1571	1670.6	−99.6	5.94
5	1404	1336.4	67.6	3.42
6	787	779.6	7.4	0.07
7	367	334.1	32.9	3.24
8	112	104.4	7.6	0.55
9	29	23.2	5.0	0.93
10	2	3.5		
11	1	0.3		
12	0	0.0		
Total	7006	7006	0	15.78

The observed chi-squared value

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 1.50 + 0.03 + \cdots + 0.93 = 15.78$$

is compared against the χ^2 distribution with $10 - 1 = 9$ degrees of freedom. (There are ten cells; the binomial model $B(12, \frac{1}{3})$ was specified completely, with no parameters requiring estimation.) The corresponding SP is 0.072. There is some evidence of a poor binomial fit, but the evidence is not overwhelming—Pearson's objections seem somewhat exaggerated.

Solution 13.3

(a) The likelihood of μ for the data, writing

$$p(x; \mu) = P(X = x) = \frac{e^{-\mu}\mu^x}{x!}, \quad x = 0, 1, 2, \ldots$$

is given by

$$\begin{aligned} &(p(0;\mu) + p(1;\mu))^1 \times (p(2;\mu))^2 \times (p(3;\mu))^5 \times (p(4;\mu))^9 \\ &\quad \times (p(5;\mu))^{10} \times (p(6;\mu))^5 \times (p(7;\mu))^2 \times (p(8;\mu))^3 \times (p(9;\mu))^1 \\ &\quad \times (P(10;\mu))^1 \end{aligned}$$

where $P(10; \mu) = P(X \geq 10)$. This is maximized at $\mu = \widehat{\mu} = 4.9621$.

(b) The table of observed and expected frequencies, showing the chi-squared calculations, is given in Table S13.2.

The observed value of the test statistic $\chi^2 = 3.51$ compared against $\chi^2(4)$ (six cells, one parameter estimated from the data) gives a SP of 0.476. There is no evidence to reject the hypothesis of a Poisson model for the goal frequency. It is reasonable to suppose that the incidence of goals occurs at random during the course of play.

Table S13.2

Number of goals	Observed frequency	Expected frequency	$O_i - E_i$	$\frac{(O_i - E_i)^2}{E_i}$
$0 - 1$	1	1.63	-1.99	0.79
2	2	3.36		
3	5	5.56	-0.56	0.06
4	9	6.89	2.11	0.65
5	10	6.84	3.16	1.46
6	5	5.66	-0.66	0.08
7	2	4.01	-2.06	0.47
8	3	2.49		
9	1	1.37		
≥ 10	1	1.19		
Total	39	39	0	3.51

Solution 13.4

(a) The seven differences are

$$33, 2, 24, 27, 4, 1, -6$$

with mean $\overline{d} = 12.143$ and standard deviation $s = 15.378$. The corresponding value of test statistic t in a test of zero mean difference is

$$t = \frac{\overline{d}}{s/\sqrt{n}} = 2.089.$$

(b) The t-distribution against which the test statistic is compared has 6 degrees of freedom, and $P(T_6 > 2.089) = 0.041$.

No indication has been given whether there should be an expected increase or decrease in the CO transfer factors. For a two-sided test, the corresponding SP is 0.082. This provides little evidence that there is any significant difference between CO transfer factors in smokers at entry and one week later.

Solution 13.5

The normal probability plot is shown in Figure S13.2.

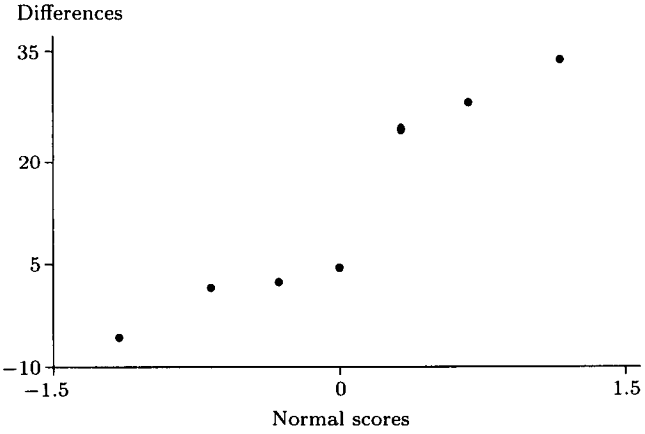


Figure S13.2 Normal probability plot for the differences between CO transfer factors

The plot shows that the data are split into two smaller groups—a group of four where there is little change in the transfer factor and a group of three where the change is comparatively large. It is clear that the normality assumption is not tenable.

Solution 13.6

Wilcoxon's signed rank test involves ranking the absolute differences, as shown below.

Table S13.3 Ranked differences

Patient	Entry	One week	Difference	Rank
1	40	73	33	7
2	50	52	2	2
3	56	80	24	5
4	58	85	27	6
5	60	64	4	3
6	62	63	1	1
7	66	60	−6	4

The sum over the positive ranks is 24 and the sum over the negative ranks is 4. The observed value of the test statistic is $w_+ = 24$. Using a computer, the total SP is 0.109, and there is no reason to reject the null hypothesis of no difference between the CO transfer factors.

Using a normal approximation, the corresponding z-score is

$$z = \frac{w_+ - E(W_+)}{SD(W_+)} = \frac{24 - 14}{\sqrt{35}} = 1.690,$$

and since $P(Z > z) = 0.0455$, the total SP is 0.091.

Solution 13.7

(a) The normal probability plot for the differences is shown in Figure S13.3.

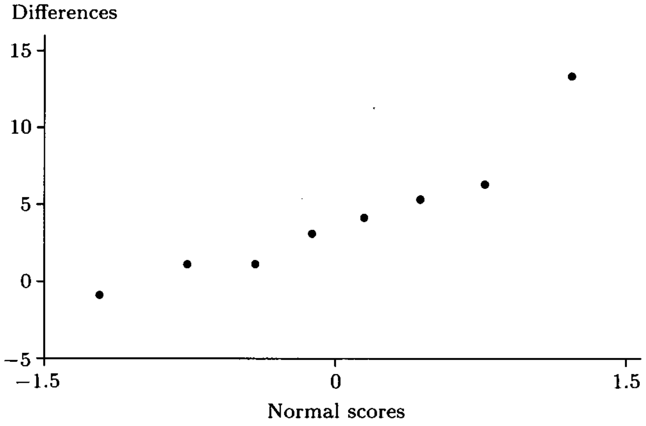


Figure S13.3 Normal probability plot

Clearly there is an outlier and this is confirmed in Table S13.4, which shows the differences. Leaf 1 is atypical.

Table S13.4 Viral lesions on tobacco leaves

Leaf	Preparation 1	Preparation 2	Difference
1	31	18	13
2	20	17	3
3	18	14	4
4	17	11	6
5	9	10	−1
6	8	7	1
7	10	5	5
8	7	6	1

However, the other seven points appear to lie on a straight line and removal of the outlier should leave data which are plausibly normal.

(b) A t-test on the seven points gives a t-value of 2.875 which, tested against $t(6)$, results in a total SP of 0.028. There is evidence for rejecting the null hypothesis, and for concluding that the two preparations have a significantly different effect. (The experimental results suggest that the first preparation leads to a substantially higher average number of lesions.)

Solution 13.8

(a) First, let us consider the two groups at Day 0. Sample statistics are

$$n_1 = 20, \quad \overline{x}_1 = 0.3945, \quad s_1^2 = 0.0089,$$
$$n_2 = 10, \quad \overline{x}_2 = 0.3520, \quad s_2^2 = 0.0093.$$

Notice that the sample variances are very close, well within a factor of 3 of one another. There is no evidence that the assumption of equal variances for the two populations is broken. The pooled sample variance is given by

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = 0.009\,031,$$

and the value of the test statistic t for a test of equal population means is

$$t = \frac{\overline{x}_1 - \overline{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = 1.155.$$

Compared against Student's t-distribution with $n_1 + n_2 - 2 = 28$ degrees of freedom, this gives a total SP of $2 \times 0.129 = 0.258$.

There is no evidence that there is any difference between the mean urea levels at admission for the two groups of patients.

(b) For the two groups at Day 6, the analysis proceeds as follows. Sample statistics are

$$n_1 = 20, \quad \overline{x}_1 = 0.5390, \quad s_1^2 = 0.0168,$$
$$n_2 = 10, \quad \overline{x}_2 = 0.6830, \quad s_2^2 = 0.0254.$$

Again, the sample variances do not suggest any significant difference between the two population variances.

The pooled sample variance is

$$s_p^2 = 0.0196,$$

and the value of the test statistic t is

$$t = \frac{\overline{x}_1 - \overline{x}_2}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} = -2.655.$$

Compared against $t(28)$, this gives a total SP of $2 \times 0.0065 = 0.013$.

In this case, there is substantial evidence that by Day 6 after admission there is a significant difference between mean serum urea levels for the surviving and non-surviving patients.

Solution 13.9

The regression coefficients, computed separately for each rat, appear in Table S13.5. In fact, the difference between the groups is striking—the second group have the larger growth rates. On the other hand, there are only four values in the second group, and one of these is similar in size to the values from the first group. So, with such small sample sizes, could these results have arisen easily by chance?

Table S13.5 Regression slopes for each rat separately

Group 1:	4.3	1.2	1.0	0.8	2.3	0.4	3.8	3.4
Group 2:	7.4	8.8	1.4	8.4				

The means for the first and second groups are respectively 2.15 and 6.50 and their respective standard deviations are 1.514 and 3.451. At first glance there seems to be a difference, but we should not jump to conclusions at this stage.

Solution 13.10

Group A:	4.3	1.2	1.0	0.8	2.3	0.4	3.8	3.4
Rank	9	4	3	2	6	1	8	7
Group B:	7.4	8.8	1.4	8.4				
Rank	10	12	5	11				

In this solution the groups have been relabelled A and B to achieve a convenient notation, consistent with previous work.

The value of the Mann–Whitney–Wilcoxon test statistic is $u_A = 40$ with a computed total SP of 0.048. There is some evidence that the rates of growth in the two groups differ.

(The normal approximation gives

$$z = \frac{u_A - E(U_A)}{SD(U_A)} = \frac{40 - 52}{\sqrt{34.667}} = -2.038$$

with a total SP of 0.042.)

Solution 13.11

The result of performing Fisher's exact test on the two sample proportions 17/31 and 16/28 is

$$SP(\text{obtained direction}) = 0.534$$

using a one-tailed test. For a two-sided test exploring merely whether there is a significant difference, the total SP is 1. There is no evidence for a difference in either direction.

Solution 13.12

The expected values for each cell are shown in brackets. For instance, the expected value in the top left-hand cell is found from

$$\frac{90 \times 47}{367} = 11.5.$$

	Hospital					Total
	A	B	C	D	E	
No improvement	13 (11.5)	5 (7.6)	8 (19.4)	21 (31.4)	43 (20.1)	90
Partial	18 (19.1)	10 (12.6)	36 (32.1)	56 (52.0)	29 (33.3)	149
Complete	16 (16.4)	16 (10.8)	35 (27.5)	51 (44.6)	10 (28.6)	128
Total	47	31	79	128	82	367

The chi-squared test statistic is calculated from

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 56.7,$$

summed over all 15 cells. For the chi-squared test of independence, we need the number of degrees of freedom for the null distribution. This parameter is calculated as $(r-1)(c-1)$, where r is the number of rows in the table and c the number of columns. In our case these are 3 and 5 respectively, so that the distribution we need is the chi-squared distribution with $(3-1)(5-1) = 8$ degrees of freedom.

The probability of obtaining a value as high as 56.7 from a chi-squared distribution with 8 degrees of freedom is very low indeed, about 2×10^{-9}. So we conclude from this test that there are real differences between the distributions of outcomes across the hospitals.

Solution 13.13

The chi-squared test statistic is 20.85, and relating this to a chi-squared distribution with $(4-1)(4-1) = 9$ degrees of freedom gives a SP equal to 0.013. There is thus a low probability that such a distribution would be obtained by chance if the two variables really were independent. We conclude that serum cholesterol level and systolic blood pressure are associated.

Solution 13.14

Depending on your software, you might be able to calculate directly that for a bivariate sample of size 10 with an underlying correlation of zero, the probability of obtaining a sample correlation of $r = -0.72$ or less is

$$P(R \leq -0.72) = P(R \geq 0.72) = 0.0098$$

and so the total SP for the test is $2 \times 0.0098 = 0.02$.

This offers considerable evidence that there is an underlying association between PEF and the S:C ratio; in fact, there is evidence that there is a negative association between the two measures.

(Alternatively, use the fact that

$$R\sqrt{\frac{n-2}{1-R^2}} \sim t_{(n-2)}$$

and compare

$$t = r\sqrt{\frac{n-2}{1-r^2}} = -0.72\sqrt{\frac{10-2}{1-(-0.72)^2}} = -2.910$$

against $t(8)$. Again, the SP for the test is given by $2 \times 0.0098 = 0.02$.)

Solution 13.15

The correlation coefficient is given by $r = 0.971$. This is a high value and we see that the number of finger ridges in identical twins are highly correlated. The SP for a test that the underlying correlation is zero is given by 1.5×10^{-7}: this is very low!

Solution 13.16

Writing x for the explanatory variable (wind speed) and y for the response (race time), summary statistics are

$$n = 21, \quad \sum x_i = 4.7, \quad \sum y_i = 279.36, \quad \sum x_i^2 = 45.11,$$
$$\sum x_i y_i = 58.796,$$

and the estimated slope is

$$\widehat{\beta} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{21 \times 58.796 - 4.7 \times 279.36}{21 \times 45.11 - (4.7)^2}$$
$$= \frac{-78.276}{925.22} = -0.084\,602\,6.$$

Also,

$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\overline{x} = 13.321\,792.$$

Consequently, the fitted regression model is

Race time = 13.32 − 0.085 × Wind speed,

where race time is measured in seconds and wind speed in metres per second. This reflects the fact, suggested in the scatter plot of the data, that stronger following winds tend to lead to reduced race times. However, the model should not be extrapolated too far. For instance, the current world record for 110 m Hurdles (men) is 12.91 s (held by Jackson). The model suggests that with wind speeds much above 4.9 m/s, he would routinely race inside world record times! In fact, some hurdlers are hampered by severe following winds: it gets them too close to the next hurdle to jump, and therefore destroys their rhythm.

Solution 13.17

The sampling distribution of the estimator, assuming the scatter to be normally distributed about the model $y = \alpha + \beta x$, is given by

$$\frac{\widehat{\beta} - \beta}{S/\sqrt{\sum(x_i - \overline{x})^2}} \sim t_{(n-2)}.$$

Under the null hypothesis $H_0 : \beta = 0$, the value of the test statistic is

$$\frac{\widehat{\beta}}{s/\sqrt{\sum(x_i - \overline{x})^2}}$$

where $\widehat{\beta} = -0.085, \quad s^2 = \dfrac{\sum(y_i - \widehat{y})^2}{n-2} = \dfrac{0.4665}{19} = 0.024\,55,$

and $\sum(x_i - \overline{x})^2 = 44.058\,095$. So

$$t = -\frac{0.085}{\sqrt{0.024\,55}\,/\sqrt{44.058\,095}} = -3.584.$$

The obtained SP for the test is $P(T_{19} \leq -3.584) = 0.001$. There is very considerable evidence from these data to reject the hypothesis that, in fact, $\beta = 0$, and so wind speed effect is significant.

Solution 13.18

Writing $y = \log(n_t)$, $x = t$, then the regression line of y on x has slope

$$\widehat{\beta} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = -0.0364$$

and intercept

$$\widehat{\alpha} = 10.578.$$

The fitted model is therefore

$$y = \widehat{\alpha} + \widehat{\beta}x = 10.578 - 0.0364x$$

or

$$\log(n_t) = 10.578 - 0.0364t$$

or, taking exponentials,

$$n_t = e^{10.578 - 0.0364t} = 39\,270e^{-0.0364t}.$$

The estimated value of n_0 is 39 270 (or, say, 40 000).