Chapter 3 Exploratory Tools for Relationships

Exercises for Section 3.1.1

1. In all cases Y is random.

   (b), **(e)**, **(f)**, **(g)**, **(h)**: both random. **(a)**, **(c)**, **(d)**, **(i)**: X nonrandom and Y random.

   Note about part (a): In an experiment in which the experimenter deliberately sets the levels of fertilizer to be used, X is nonrandom (controlled). In an observational study in which we just observed the fertilizer levels that individual farmers had happened to apply, however, X is random.

2. For the reader.

Exercises for Section 3.1.2

1. This exercise emphasizes the idea summarizing trend (via a of a trend curve) and scatter in a scatter plot. If you compare your curves with those of fellow students, you will see differences due to the subjective nature of curves drawn by eye.

2. **(b)** Predict X. **(e)** A symmetric comparison makes sense. We might also want to see how well screening test results Y predict actual alcohol dependence levels X. **(f)** The choice is unclear, though you would probably want to predict Y. **(g)** A symmetric comparison makes sense. You might also want to see, for example, if large litters tend to produce mice with smaller brains (use X values to predict or explain the behavior of Y values). **(h)** Predict Y.

3. **(a)**
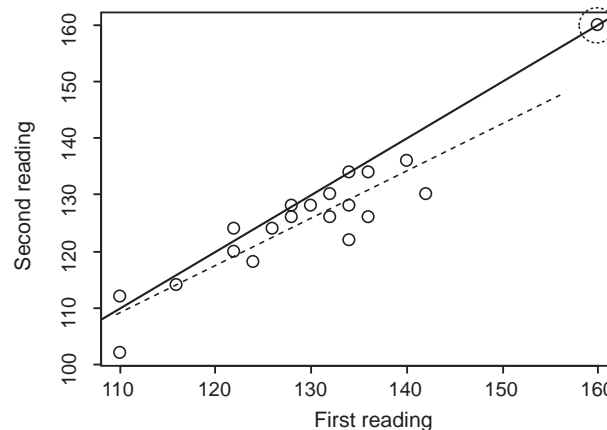
   Apart from the point $(160, 160)$ (circled), we see a linear trend which we have summarized with the dashed line). There is a reasonable amount of scatter. First and second readings on the same person can differ by as much as 8 units. The point $(160, 160)$ is an outlier if we look just at 1st readings (alone) or just at 2nd readings. If it was close to the trend in the scatter plot, we would be less inclined to think it was a mistake and more inclined to think it was just a person with an

unusually high blood pressure. The twin facts that it is moderately off the trend and that the x and y readings are identical makes us suspicious of the reliability of this point.

(c) Most of the points lie below the line $y = x$ (solid line). This indicates that the y-value (2nd reading) is usually less than the x-value (1st reading). This is not surprising; we might expect the blood pressure to be less the second time it is read, as the person interviewed will then be more relaxed.

4. (a)
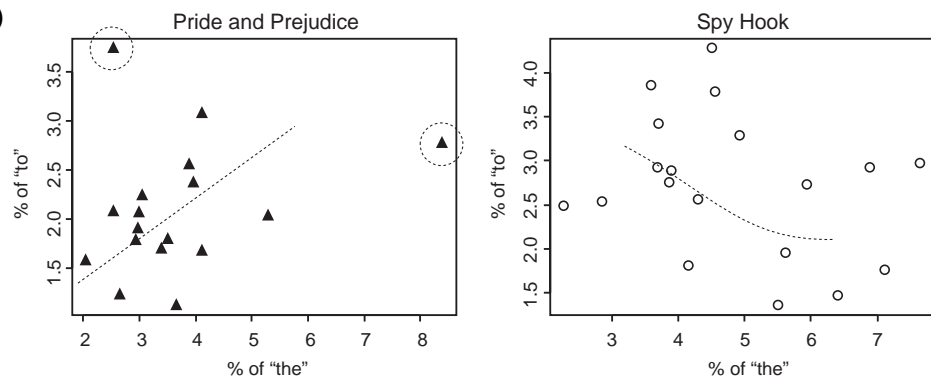
(b) Apart from two "unusual looking" points (circled), the plot for *Pride and Prej-udice* seems to have quite a strong upward trend. The relationship for *Spy Hook* looks weaker and seems to have a curved trend initially heading downward.

(c)
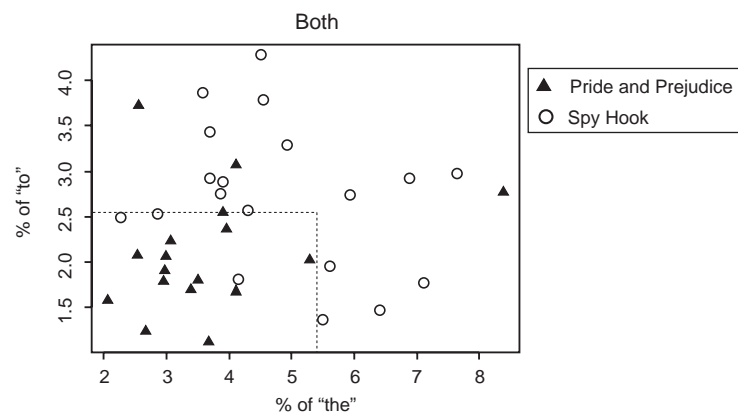
There is a tendency for *Pride and Prejudice* to use both fewer "to"s and fewer "the"s. The points for *Pride and Prejudice* tend to lie in the bottom left-hand corner. We have placed a dotted box on the plot to show this idea. Most observations for *Pride and Prejudice* are inside the box and those for *Spy Hook* outside. We would not make many errors classifying these pages as coming from one book or the other just using numbers of "to"s and "the"s.

(d) Given a few more points from one of the two books, it might be possible to make an educated guess as to which book they come from. It would be interesting to have more data both for this purpose and also to see if we could confirm the trends we saw in (a).

Exercises on Section 3.1.3

The predictions to follow are all subjective so will probably differ slightly from your answers.

1.
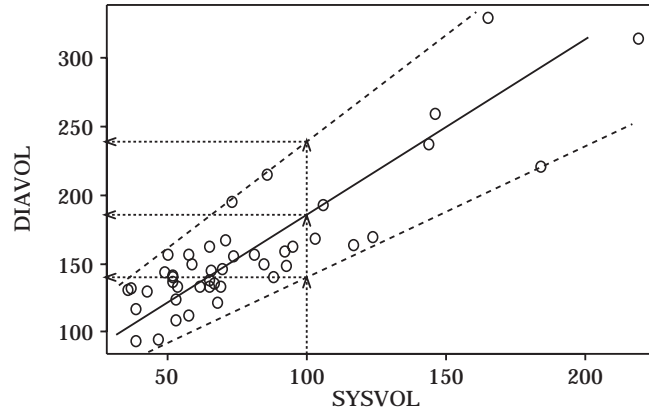
We see a linear trend with increasing scatter from left to right and have drawn our trend line and "envelope" (dashed) to capture this. Our resulting single-number estimate is approximately 180. Our interval extends from approximately 140 to 240.

2.
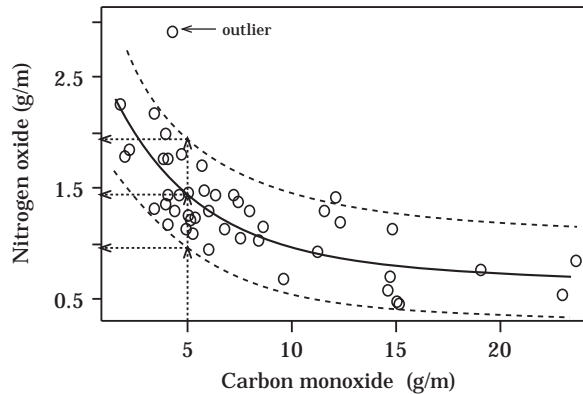
Apart form an outlier, we see a trend curving downwards from left to right and have drawn our trend curve and "envelope" to capture this and the usual level of scatter (dashed). Our resulting single-number estimate is approximately 1.4. Our interval extends from approximately 0.9 to 1.9.

3. The plot follows. We see a downward and roughly linear trend until about 1970 where it seems to flatten out. Our resulting single-number estimate is approximately 216. Our interval extends from approximately 211 to 221.
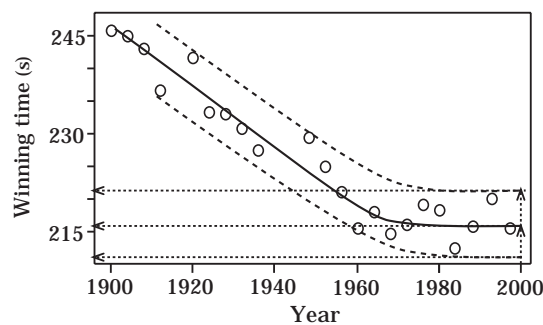
Exercises on Section 3.2
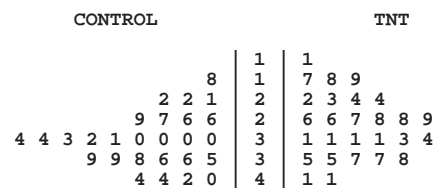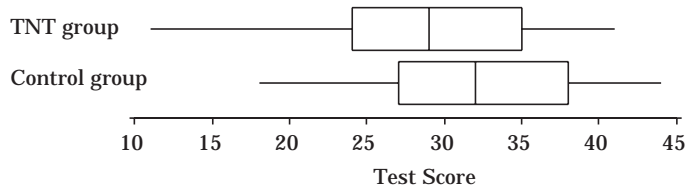
1. (a)

(i) **Dot plots**
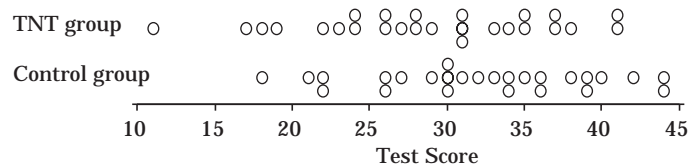
(ii) **Back-to-back stem-and-leafs**

```
Units  3 | 5 = 35

        CONTROL                      TNT

                     1 | 1
                  8  1 | 7 8 9
              2 2 1  2 | 2 3 4 4
            9 7 6 6  2 | 6 6 7 8 8 9
  4 4 3 2 1 0 0 0 0  3 | 1 1 1 1 3 4
      9 9 8 6 6 5    3 | 5 5 7 7 8
          4 4 2 0    4 | 1 1
```

(iii) **Box plots**

(b) The three plots indicate that the two data sets have similar spreads and "shapes" (slightly negatively skewed). The median of the control group is greater than the median of the TNT group. The control scores are higher on average than the TNT scores.

(c) The three plots show the same features. It is tempting to read too much into such plots. This issue is dealt with in later chapters (particularly Chapter 10). It is useful to assess a plot by thinking about how much your impressions would change if you changed one value slightly.

2. (a)
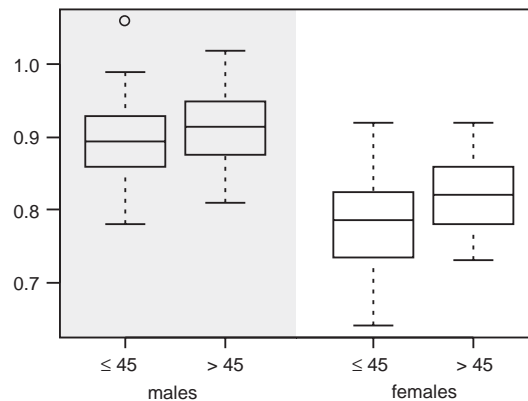
(b) It is clear that the women usually have lower waist-to-hip ratios than the men. There is also an effect of age. The older men have slightly larger ratios on average than the younger men. The same effect is evident for women. All box plots indicate approximate symmetry. For the men, the spreads are similar, whereas for women the spread for the younger women is slightly greater than that for the older women. Whether this is true for the whole population is not clear, as the difference is small.

Exercises on Section 3.3

1. The two-way table follows:

	SURG					
BETA	0	1	2	3	4	Total
1	7	1	0	0	0	8
2	25	4	4	2	2	37
Total	32	5	4	2	2	45

2. (a) $1151/39859 = 0.0289$.

(b) $7447/39859 = 0.1868$.

(c) $6732/39859 = 0.1689$ [same as $(0 + 1205 + 3163 + 1213 + 1151)/39859$].

(d) $10952/39859 = 0.2748$ [same as $(1205 + 2300 + 7447)/39859$].

(e) $1205/10952 = 0.1100$.

(f) $2373/13178 = 0.1801$.
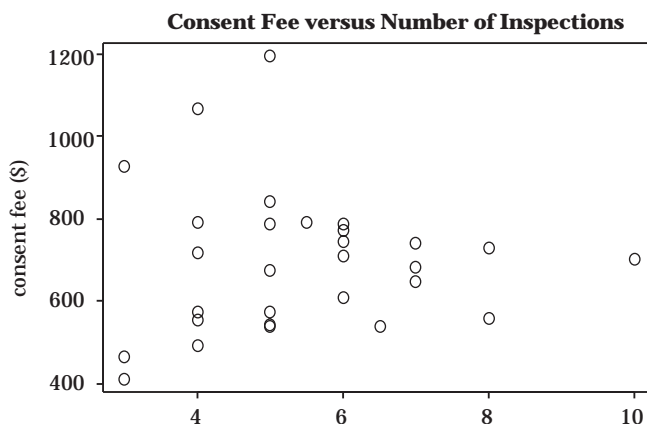
(g) $6552/25369 = 0.2583$.

Review Exercises 3

1. **(a)** `Units 4 | 1 = $410`

   ```
    4 | 1 6 9
    5 | 4 4 4 6 6 7 8
    6 | 1 5 7 8
    7 | 0 1 2 3 4 4 7 9 9 9 9
    8 | 4
    9 | 3
   10 | 7
   11 | 9
   ```

   (The plot shows a great deal of variability in cost. The distribution has a long upper tail.)

   (Min, Q_1, Med, Q_3, Max) = (412, 555.5, 701, 786.5, 1193).

   [Uses book definition of quartiles. Some packages give slightly different values.]

   (b) Max/Min = 2.9 times greater (to one decimal place).

   (c) We use the following scatter plot.

   Consent Fee versus Number of Inspections

   There is no relationship between the fee charged and number of inspections performed (some of the highest charges come from cities that perform the fewest inspections). The spread is greatest for cities with the lowest numbers of inspections.

3. **(a)**
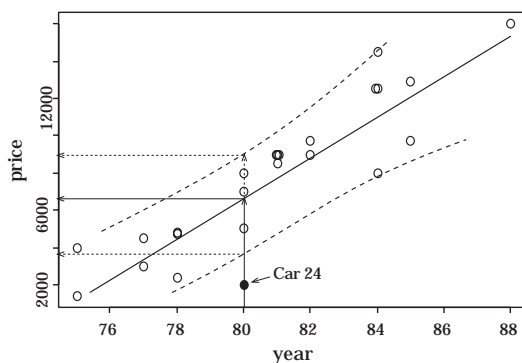
The plot is approximately linear with quite a lot of variation in prices for trucks of the same year (scatter).

(b) Note that the prices of the three 1981 trucks we observed seem to be on the high side. It is the rest of the relationship that suggests this to us. You may be able to see from the plot that because of the strong relationship between the variables year and price, we would have substantial information about likely prices in a given year (say 1981) even if we had not collected any information at all on cars of that year.

(c) Approximately $6800 (reading from the plot as presented).

(d) Approximately $3600 to $8900 (reading from the plot as presented).

[Our answers to (c) and (d) are subjective, so your answers will probably be slightly different.]

(e) Buy it! [Our friend did buy it and it was a great truck for many years.]

(f) Distance travelled, number of owners, condition of the truck, and so on. Distance travelled can be read from the odometer (provided you trust it hasn't been tampered with) and is often included in advertisements. In New Zealand, number of owners is included on the ownership papers for the vehicle. This would be hard to get for a large number of cars if you had to ask each seller. For "condition" you would need some sort of subjective rating, taking into account such things as rust and appearance. To add such information to the plot, we could for example use number of owners as our plotting symbol, and add condition information via a color code.

5. (a) *Homicides*: A wide range of positively skewed ages with the 20s being the predominant (or modal) age group.
Accidents: Ages between about 15 and the mid 30s.
Suicides: Strongly bimodal with a drop in the 40 to 64 age group. Possibly three different groups to be considered here, namely 15 to 45, 45 to 64 and 65 to 80. The peak at 65 suggests possibly retirement-related problems.
Self-defense: A wide range of positively skewed ranges (2-56).
Overall: Homicides and suicides are by far the most common causes with similar numbers of deaths for both. Accidents are concentrated in younger age groups whereas self-defense shooting are concentrated near the middle of the age range.

(b) Have all four plots arranged side by side with their scales coinciding to highlight the age differences between the different types of death.

(c) The five number summaries are: homicides $(2, 22, 28, 38, 79)$;
accidents $(9, 17, 22, 29, 37)$; suicides $(14, 27, 38, 65, 87)$; self-defense $(21, 26, 34, 46, 56)$.

[These all use the book definition of quartiles. Some packages give slightly different values.]

The box plot for suicides does not reveal the bimodal nature of the data. No, as the median is a single measure of location. We need a measure for each of the two groups.

(d) The combined stem-and-leaf plot will tend to be dominated by the homicide and suicide data. The peak of the homicide data and the lower peak of the suicides come in approximately the same place and should "add" to produce a large peak. We would also expect a small 2nd peak around 65 from the upper mode of the suicide data. A plot of all the ages follows.

```
The decimal point is at the |
 2 | 0
 4 |
 6 |
 8 | 0
10 |
12 | 0
14 | 0 0 0 0 0 0 0
16 | 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
18 | 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
20 | 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
22 | 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
24 | 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
26 | 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
28 | 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
30 | 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
32 | 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
34 | 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
36 | 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
38 | 0 0 0 0 0 0 0 0 0 0 0 0 0 0
40 | 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
42 | 0 0 0 0 0 0 0 0 0 0 0 0
44 | 0 0 0 0 0 0 0 0 0
46 | 0 0 0 0 0 0 0 0 0 0 0 0 0
48 | 0 0 0 0 0 0
50 | 0 0 0 0 0
52 | 0 0 0 0 0 0 0 0
54 | 0 0 0 0
56 | 0 0 0 0 0 0 0 0
58 | 0 0 0 0 0
60 | 0 0 0 0 0 0
62 | 0 0 0 0
64 | 0 0 0 0 0 0 0 0 0 0 0
66 | 0 0 0 0 0 0 0 0 0 0
68 | 0 0 0 0 0 0
70 | 0 0 0 0 0 0 0 0
72 | 0 0 0 0
74 | 0 0 0 0
76 | 0 0 0 0 0 0
78 | 0 0 0 0
80 | 0 0
82 | 0 0
84 | 0
86 | 0 0 0 0
```

(e) No, patterns of gun ownership are quite different in these other countries.

7. (a) Figure (a) suggests that the measurements are not easy to make reliably. From (b) we see that there are a lot of lighter perch and a few heavy pike. The weights for bream are less variable than for perch and pike. The perch data seem bimodal, and there is a hint of bimodality for both bream and pike. From (c), pike are longer on average, and bream are less variable in length. The weights of the few heavy pike are due to their length. From (d) the widths for perch are clearly bimodal, indicating the presence of two distinct groups. Figure (e) confirms (d). Figures (f), (g), and (h) indicate that pike tend to be longer with a smaller cross-sectional area than bream and perch. Figures (b) to (e) now suggest that all forms of the pike data are unimodal but positively skewed with a long tail to the right. Bream seem to be substantially higher and slightly narrower than perch for the same length. The presence of 2 groups in the perch (seen in some plots as bimodality) shows up most clearly in (h). The plots are curved in (f) and approximately linear in (h).

(b) Height versus length completely separates the species.

9. Except for pie charts, we produced the same types of graphs for the GDP figures as those above for gold production. First we produced graphs for the actual GDP figures of which two examples are given below.

The overwhelming impression from these graphs is just of overall growth of the economy in dollar terms through time (best shown by the segmented bars). This growth effect was so large that it obscured more detailed changes. [Some of this growth could have been just due to inflation, so it would have been better to work in constant (inflation adjusted) dollars.]

To see changes in the contributions of the various sectors to overall GDP, we decided to work with percentages of GDP as in the two graphs to follow. The bar graph has been arranged so that the major blocks are economic sectors; from left to right within a block we see how the percentage changes through time. This is good for tracking time changes within a sector. We see rapid growth in the service sector. In the latest figures given, it has become the largest sector. There has been strong growth in the financial, insurance and real estate sector (ranked number 2 by 1994). The contribution of manufacturing of durable goods (originally number 1) has fallen considerably. Retail, wholesale trade, and transportation have all been fairly flat, as has state and local government after an initial growth period. The contribution of the manufacturing of nondurable goods, and of agriculture and fisheries (a comparatively small sector of the economy) has been falling. The line-linked scatter plot tells the

same stories, but make it easier to see, for example, where the services sector first overtook the financial and real estate sector.
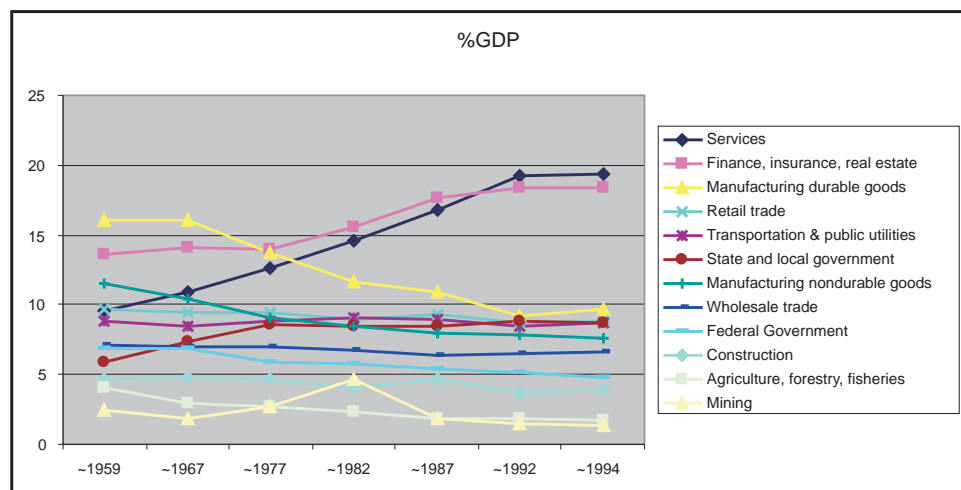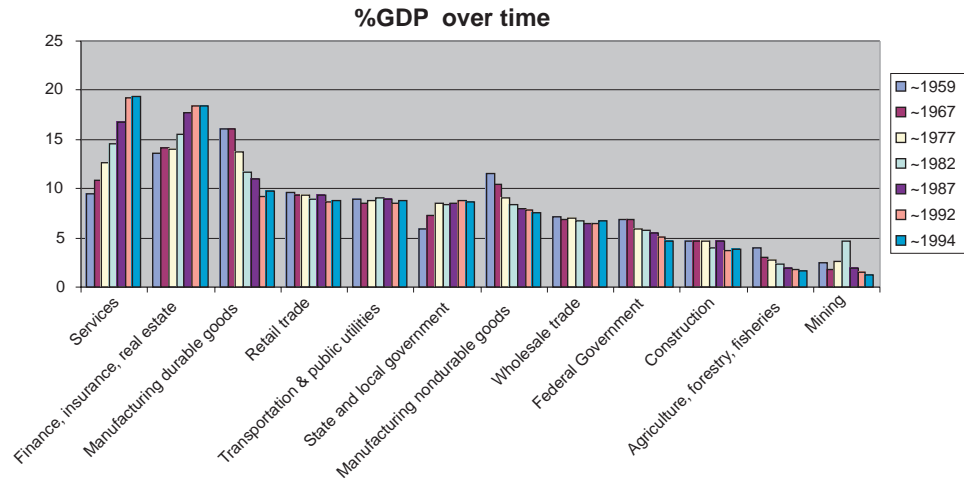
Chapter 4 Probabilities and Proportions

Exercises for Section 4.3

1. Spin the pointer a large number of times and count the proportion that fall in the grey sector. Yes, only if the spinner is balanced and the angle of the grey sector is known.

2. Yes, by tossing the thumb tack a large number of times and calculating the proportion landing point down. No, as the outcomes are not equally likely. We cannot argue from symmetry conditions as we could with the pointer.

3. **(a)** F (the number of 9s in a column will be random).
   (b) T.
   (c) F (previous outcomes do not affect the current outcome).
   (d) F (all sequences are equally likely).

Exercises for Section 4.4.2

1. $A = \{(1,3), (2,2), (3,1)\}$.

2. $A = \{HHT, HHHHT, HHHHHHT, \ldots\}$. $\overline{A} = \{T, HT, HHHT, \ldots\}$, $\overline{A} =$ "odd (or zero) number of tosses before the first tail."

3. $S = \{TT, THT, HTT, THHT, HTHT, THHH, HHH, HTHH, HHTT, HHTH\}$, $A = \{THT, HTT, HHH\}$.

Exercises for Section 4.4.3

1. $S = \{(T1), (T2), \ldots, (T6), (H1), (H2), \ldots, (H6)\}$.
   (i) $A = \{(T3), (H3)\}$. **(ii)** $B = \{(H1), (H2), (H3), (H4), (H5), (H6)\}$.
   (iii) A and $B = \{(H3)\}$.
   (iv) A or $B = \{(T3), (H1), (H2), (H3), (H4), (H5), (H6)\}$.

2. $S = \{(A, A), (A, B), (A, O), (A, AB), (B, A), (B, B), (B, O), (B, AB),$
   $(O, A), (O, B), (O, O), (O, AB), (AB, A), (AB, B), (AB, O), (AB, AB)\}$.

   (a) $C = \{(A, A), (B, B), (O, O), (AB, AB)\}$.
   (b) $D = \{(A, A), (A, B), (A, O), (A, AB), (B, A), (O, A), (AB, A)\}$.
   (c) C and $D = \{(A, A)\}$.

Exercises for Section 4.4.4

1. **(i)** Four of the 36 equally-likely outcomes add to 9 [namely (3,6), (4,5), (5,4) and (6,3)] so the probability is $\frac{4}{36} = \frac{1}{9}$.
   (ii) Arguing similarly we get $\frac{18}{36} = \frac{1}{2}$.

2. **(a)** **(i)** $\frac{363}{5584} = 0.0650$. **(ii)** $\frac{911}{5584} = 0.1631$.

    (b) $S = \{MW, MS, MP, FM, FS, FP\}$. The probability distribution is found by dividing each number in the table by 5,584, namely : 0.305, 0.214, 0.098, 0.217, 0.101, 0.065. (See Table 4.4.2.)

    (c) $S = \{W, S, P\}$. The probability distribution is found by dividing the column sums by 5,584, namely : 0.522, 0.315, 0.163.

    (d) It is easier to compute any probability that we might be interested in.

3. (a) People may lose more than one job.

    (b) No, as there may be fewer females with jobs (which is likely).

    (c) We need to compare the proportions of those with jobs losing their jobs for both males and females.

Exercises for Section 4.5.1

1. Let $A =$ "wet" and $B =$ "windy". We enter the numbers given to us into the following two-way table

	Wet (A)	Dry ($\overline{A}$)	Total
Windy (B)	0.2		0.4
Windless ($\overline{B}$)			
Total		0.7	1.0

and complete it to obtain

	Wet (A)	Dry ($\overline{A}$)	Total
Windy (B)	0.2	0.2	0.4
Windless ($\overline{B}$)	0.1	0.5	0.6
Total	0.3	0.7	1.0

Reading from the table, we get:

    (a) $\mathrm{pr}(A) = 1 - 0.7 = 0.3$.

    (b) $\mathrm{pr}(A \text{ or } B) = 0.2 + 0.2 + 0.1 = 0.5$.

    (c) $\mathrm{pr}(\overline{A} \text{ and } \overline{B}) = 0.5$. We can also get this from the complement of (b).

2. (a) $0.353 + 0.062 = 0.415$.

    (b) $0.195 + 0.016 = 0.211$.

    (c) Sum the entries in the first three rows and columns to get 0.895.

    (d) 1 - 0.895 = 0.105.

    (e) $0.111+0.195+0.008+0.021+0.022 +0.003+0.016= 0.376$.

    (f) $0.111 + 0.008 + 0.022 = 0.141$.

3. We construct the following two-way table by entering the numbers given to us

	Too tall	Not too tall	Total
Attractive	0.18		
Not attractive			0.16
Total	0.24		1.00

and complete it to obtain

	Too tall	Not too tall	Total
Attractive	0.18	0.66	0.84
Not attractive	0.06	0.10	0.16
Total	(.24)	.76	1.00

Reading from the table, we get:

    (a) 0.06.

    (b) 0.66.

    (c) 0.1.

4. We can use either Table 4.4.2 or, for more accuracy, Table 4.4.1 (as used below).

    (a) $1 - \frac{1703}{5584} = 0.6950$.

    (b) $\frac{1703+1196+548+1210}{5584} = 0.8340$.

    (c) $\frac{1196+548}{5584} = 0.3123$.

Exercises for Section 4.6.1

1. (a) $\frac{19}{68} = 0.2794$.

    (b) $\frac{19+33}{125} = \frac{52}{125} = 0.416$.

2. $\text{pr}(S \mid W) = \frac{\text{pr}(S \text{ and } W)}{\text{pr}(W)} = \frac{0.5}{0.7} = \frac{5}{7}$.

3. (a) (i) $\frac{564}{5584} = 0.1010$.

    (ii) $\frac{564}{2137} = 0.2639$.

    (b) (i) $\frac{1196}{5584} = 0.2142$.

    (ii) $\frac{1196}{3447} = 0.3470$.

    (iii) $\frac{1760}{5584} = 0.3152$.

    (c) (i) $\frac{363}{2137} = 0.1699$.

    (ii) $\frac{548}{3447} = 0.1590$.

    (iii) $\frac{911}{5584} = 0.1631$.

4. (a) (i) 0.353. (ii) $\frac{0.195}{0.353} = 0.5524$. (iii) $\frac{0.195}{0.336} = 0.5804$. (iv) 0.195.

    (b) (i) Divorced (from the female divorced column).

    (ii) To get the conditional probabilities, we divide each entry in column two by the total 0.336. This does not affect the relative magnitudes of the entries.

    (iii) Never married. $\frac{0.401}{0.554} = 0.7238$.

    (c) (i) $\frac{0.401}{0.554} = 0.7238$.

    (ii) 0.401.

    (d) $\frac{0.195+0.024+0.008+0.016}{0.353+0.031} = 0.6328$.

5. $0.2267 + 0.0366 = 0.2633$.

Exercises for Section 4.6.2

1. Let H = "homeless" and S = "schizophrenic". Then $\mathrm{pr}(H) = 0.008$, $\mathrm{pr}(S) = 0.01$, and $\mathrm{pr}(S \mid H) = \frac{1}{3}$.

2. $0.26 \times 0.77 = 0.2002$.

3. $\frac{2}{6} \times \frac{1}{5} = \frac{1}{15}$.

Exercises for Section 4.6.3

1. We construct the following two-way table

	Blood type		
Race	AB	Not AB	Total
European	.0324×.85 (= .02754)		.85
Maori	.0177×.15 (= .002655)		.15
Total			1.00

and complete to produce

	Blood type		
Race	AB	Not AB	Total
European	.027540	.822460	.85
Maori	.002655	.147345	.15
Total	.030195	.969805	1.00

From the table we obtain:

$\mathrm{pr}(group\ AB) = 0.0302$, or 3.02%.

$\mathrm{pr}(maori \mid group\ AB) = \frac{0.002655}{0.030195} = 0.0879$, or 8.8%.

2. We construct the following two-way table

	Parents			
Child	Both	One only	Neither	Total
Left-handed	$\frac{1}{2} \times .02$	$\frac{1}{6} \times .2$	$\frac{1}{16} \times .78$	
Right-handed				
Total	.02	.20	.78	1

and complete to produce

	Parents			
Child	Both	One only	Neither	Total
Left-handed	$\frac{1}{2} \times .02$	$\frac{1}{6} \times .2$	$\frac{1}{16} \times .78$	.09208
Right-handed	$\frac{1}{2} \times .02$	$\frac{5}{6} \times .2$	$\frac{15}{16} \times .78$	.90792
Total	.02	.20	.78	1

From the table we obtain:

$\mathrm{pr}(child\ is\ left\text{-}handed) = 0.09208$.

$\mathrm{pr}(neither\ parent\ left\text{-}handed \mid child\ left\text{-}handed) = \frac{0.78/16}{0.09208} = 0.5294$.

3. We construct the following two-way table

	Killed white	Killed black	Total
Death sentence	30		36
No death sentence	184		
Total			326

and complete to produce

	Killed white	Killed black	Total
Death sentence	30	6	36
No death sentence	184	106	290
Total	214	112	326

From the table we obtain:

The probability of getting the death sentence for killing a white is $\frac{30}{214}$ ($= 0.1402$).

The probability for killing a black is $\frac{6}{112}$ ($= 0.0536$). The ratio is about 2.6.

Exercises for Section 4.7.1

1. $0.06 \times 0.14 = 0.0084$ (assuming independence). The events may not be independent. He may think he is safe, as she is on the pill!

2. As the two are chosen at random, we can construct the following table. The information given belongs in the row- and colum-total positions. Independence allows us to obtain the internal entries by multiplying corresponding row- and column-total entries.

	A	B	O	AB	Total
A	$.41 \times .41$	$.41 \times .11$	$.41 \times .45$	$.41 \times .03$	.41
B	$.11 \times .41$	$.11 \times .11$	$.11 \times .45$	$.11 \times .03$	.11
O	$.45 \times .41$	$.45 \times .11$	$.45 \times .45$	$.45 \times .03$	.45
AB	$.03 \times .41$	$.03 \times .11$	$.03 \times .45$	$.03 \times .03$	.03
Total	.41	.11	.45	.03	1.00

We obtain the answer by summing down the diagonal elements.

$\text{pr}(\textit{Same blood phenotype}) = (0.41)^2 + (0.45)^2 + (0.11)^2 + (0.03)^2 = 0.3836.$

Exercises for Section 4.7.3

1. (a) $0.5 \times 0.75 \times 0.25 \times 0.70 = 0.065625.$

   (b) It is hard to tell. Perhaps matching and curing may not be independent (a closer match may be more likely to cure). It depends on what is meant by "matching".

2. Let $A =$ "sirens not working," $B =$ "visual signals not noticed," $C =$ "batteries run down," $D =$ "power failure" and $E =$ "routing switches shut off." If the sirens going off means that the visual warnings will be noticed (i.e., $\overline{A}$ implies $\overline{B}$), then A and B are not independent. Otherwise A and B are likely to be independent. We have the direct causal sequence $C \rightarrow D \rightarrow E$, so that any event that is independent of C is also independent of D and E. If the sirens go off, then C will not occur (i.e., $\overline{A}$ implies $\overline{C}$). The same is true if the visual warnings are noted. Thus A and B are indirectly related to C.

Review Exercises 4

1. (a) $S = \{\text{Yes}, \text{No}\}$.
   (b) $S = \{0, 1, 2, \ldots\}$.
   (c) $S = \{0, 1, 2, \text{unknown}\}$.
   (d) $S = [a, b]$ depends on where you live, e.g., $[-20°C, +50°C]$.
   (e) $S = \{0, 1, 2, \ldots\}$.
   (f) $S = \{h_1, h_2, \ldots, h_{10}\}$.
   (g) $S = \{\text{bus, car, train, boat, two-wheeler, walk, other}\}$. The "other" category would include combinations such as car, then bus.

3. There are several ways of approaching this. One follows: $1 = \text{pr}(A) + \text{pr}(\overline{A}) = \frac{4}{3}\text{pr}(A)$, so that $\text{pr}(A) = \frac{3}{4}$.

5. (a) $\frac{541}{856} = 0.6320$.
   (b) $\frac{143}{856} = 0.1671$.
   (c) $\frac{191}{288} = 0.6632$.
   (d) $\frac{192}{284} = 0.6761$.
   (e) $\frac{191}{541} = 0.3530$.
   (f) The four respective proportions are (TW.sm) $\frac{86}{143} = 0.6014$; (TW.lg) $\frac{191}{288} = 0.6632$; (TS.sm) $\frac{72}{141} = 0.5106$; and (TS.lg) $\frac{192}{284} = 0.6761$. We would prefer to use the typeset large (TS.lg) page format as it had the highest response rate.

7. (a) We construct the following 2-way table.

	Arts	Com	Sci	Eng	Law	Educ	Med
Female	.65×.3	.41×.19	.39 × .18	.15 × .07	.52 × .07	.82 × .06	.49 × .04
Male				.			
Total	.3	.19	.18	.07	.07	.06	.04

	Other	Total
continued	.47 × .09	
	.09	1.0000

and complete to form

	Arts	Com	Sci	Eng	Law	Educ	Med	Other	Total
Female	.195	.0779	.0702	.0105	.0364	.0492	.0196	.0423	.5011
Male	.105	.1121	.1098	.0595	.0336	.0108	.0204	.0477	.4989
Total	.3	.19	.18	.07	.07	.06	.04	.09	1.0000

From the table we obtain:

(b) 50.11%.
(c) (i) $\frac{0.195}{0.5011}$, or 38.91%. (ii) $\frac{0.0364}{0.5011}$, or 7.26%. (iii) $\frac{0.0105}{0.5011}$, or 2.10%. (iv) $\frac{0.0492}{0.5011}$, or 9.82%.

9. 40% of 1% = $0.4 \times 0.01 = 0.004$ or 4 per thousand. Or more formally,

$$\text{pr}(\textit{Baby infec.} \text{ and } \textit{Mother HIV})$$
$$= \text{pr}(\textit{Baby infec.} \,|\, \textit{Mother HIV})\text{pr}(\textit{mother HIV})$$
$$= 0.4 \times 0.01 \quad \text{(i.e, four in every thousand births)}.$$

11. (a) To see how the course-work score related to the distribution of final grades.

    (b) The set of all members of the class.

    (c) 0.101. **(d)** $0.853 \times 0.101 = 0.08615$. **(e)** 0.853.

    (f) $0.100 + 0.004 = 0.104$. **(g)** 0.339. **(h)** $0.339 \times 0.133 = 0.04509$. **(i)** 0.133.

    (j) We use the following two-way table.

	Course-work mark			
	0–5	5^+–10	10^+–15	15^+–20
Failed	0.912×0.027	0.797×0.046	0.539×0.181	0.133×0.339
Passed				
Total	0.027	0.046	0.181	0.339

	20^+–25	25^+–30	Total
continued	0.003×0.306	0×0.101	
	0.306	0.101	1.0000

$\text{pr}(\textit{Failed}) = 0.912 \times 0.027 + 0.797 \times 0.046 + 0.539 \times 0.181 + 0.133 \times 0.339$
$\qquad\qquad +0.003 \times 0.306 + 0 \times 0.101 \ = \ 0.2049.$
$\text{pr}(\textit{Passed}) = 1 - 0.2049 = 0.7951.$

    (k) We use the following two-way table.

	Course-work mark			
	0–5	5^+–10	10^+–15	15^+–20
A grade	0×0.027	0×0.046	0.004×0.181	0.067×0.339
Not A				
Total	0.027	0.046	0.181	0.339

	20^+–25	25^+–30	Total
continued	0.268×0.306	0.853×0.101	
	0.306	0.101	1.0000

$\text{pr}(A \ \textit{grade}) = 0 \times 0.027 + 0 \times 0.046 + 0.004 \times 0.181 + 0.067 \times 0.339$
$\qquad\qquad +0.268 \times 0.306 + 0.853 \times 0.101 \ = \ 0.1916.$

    (l) $\text{pr}(25^+\text{–}30 | A \ \textit{grade}) = \dfrac{\text{pr}(A \ \textit{grade and } 25^+\text{–}30)}{\text{pr}(A \ \textit{grade})} = \dfrac{0.853 \times 0.101}{0.1916} = 0.4497.$

13. We use the following two-way table

	Test Result		
	Positive	Negative	Total
Bowel cancer		$.4 \times .0005 = .0002$	.0005
No bowel cancer	$.4 \times .9995$		.9995
Total			1.0000

and complete it to produce

	Test Result		
	Positive	Negative	Total
Bowel cancer	$.0005 \times .6 = .0003$	$.0005 \times .4 = .0002$	.0005
No bowel cancer	$.9995 \times .4 = .3998$	$.9995 \times .6 = .5997$	.9995
Total	.4001	.5999	1.0000

(a) The proportion of bowel cancer cases overlooked is 0.40. (Information originally given.)

Now using the table, we obtain:

(b) $\text{pr}(Test\ positive) = 0.4001$.

(c) $\text{pr}(Bowel\ cancer\,|\,Test\ positive) = \dfrac{\text{pr}(Bowel\ cancer\ \text{and}\ Test\ positive)}{\text{pr}(Test\ positive)} = \dfrac{0.0003}{0.4001}$
$= 0.00075$.

(d) It is recommended that no animal meat be consumed for at least a day before the test is taken.

15. (a) Let W and L denote a win and a loss respectively for player 1. The sample space is $S = \{WW, WLW, WLL, LWW, LWL, LL\}$.

(b) (i) $\begin{aligned}\text{pr}(A) &= \text{pr}(WW) + \text{pr}(WLW) + \text{pr}(LWW)\\ &= (.5)^2 + (.5)^3 + (.5)^3 = 0.25 + 0.125 + 0.125 = 0.5.\end{aligned}$

(ii) $\text{pr}(B) = \text{pr}(WW) + \text{pr}(LL) = 0.25 + 0.25 = 0.5$.

(iii) $\text{pr}(A\ \text{and}\ B) = \text{pr}(WW) = 0.25$.

(iv) $\begin{aligned}\text{pr}(A\ \text{or}\ B) &= \text{pr}(WW) + \text{pr}(WLW) + \text{pr}(LWW) + \text{pr}(LL)\\ &= 0.25 + 0.125 + 0.125 + 0.25 = 0.75.\end{aligned}$

(v) $\text{pr}(B\,|\,A) = \dfrac{\text{pr}(A\ \text{and}\ B)}{\text{pr}(A)} = \dfrac{0.25}{0.5} = 0.5$.

(c) We have $\text{pr}(A\ \text{and}\ B) = 0.25$ and also $\text{pr}(A) \times \text{pr}(B) = 0.5 \times 0.5 = 0.25$, so A and B are independent.

(d) A and B are not mutually exclusive since they have an outcome, namely WW, in common.

17. (a) Let $M_i =$ "Minor prize on the ith draw" and $\overline{M}$ be its complement (no minor prize on ith draw).

(i) $\begin{aligned}\text{pr}(No\ minor\ prizes) &= \text{pr}(\overline{M}_1\overline{M}_2\overline{M}_3)\\ &= \text{pr}(\overline{M}_1\ \text{and}\ \overline{M}_2\ \text{and}\ \overline{M}_3) \quad \text{(translating notation)}\\ &= \text{pr}(\overline{M}_1)\text{pr}(\overline{M}_2)\text{pr}(\overline{M}_3) \quad \text{(using independence)}\\ &= 0.9^3 = 0.729.\end{aligned}$

Remaining parts use the same types of argument as in (a)(i) but are given with less detail.

(ii) $\text{pr}(all\ 3\ minor\ prizes) = \text{pr}(M_1M_2M_3) = 0.1^3 = 0.001$.

(iii) $\begin{aligned}\text{pr}(exactly\ 1\ minor\ prize) &= \text{pr}(M_1\overline{M}_2\overline{M}_3\ \text{or}\ \overline{M}_1M_2\overline{M}_3\ \text{or}\ \overline{M}_1\overline{M}_2M_3)\\ &= \text{pr}(M_1\overline{M}_2\overline{M}_3) + \text{pr}(\overline{M}_1M_2\overline{M}_3) + \text{pr}(\overline{M}_1\overline{M}_2M_3) \quad \text{(as mutually exclusive)}\\ &= 0.1 \times 0.9^2 + 0.9 \times 0.1 \times 0.9 + 0.9^2 \times 0.1\\ &= 3 \times 0.1 \times 0.9^2 = 0.243.\end{aligned}$

(iv) $\begin{aligned}\text{pr}(exactly\ 2\ minor\ prizes) &= \text{pr}(M_1M_2\overline{M}_3\ \text{or}\ M_1\overline{M}_2M_3\ \text{or}\ M_1M_2\overline{M}_3)\\ &= \text{pr}(M_1M_2\overline{M}_3) + \text{pr}(M_1\overline{M}_2M_3) + \text{pr}(M_1M_2\overline{M}_3) \quad \text{(as mutually exclusive)}\\ &= 3 \times 0.1^2 \times 0.9 = 0.027.\end{aligned}$

(b) Let $L =$ "wins Las Vegas" and $H =$ "wins Hawaii".

    (i) $\text{pr}(\overline{L} \text{ and } \overline{H}) = \frac{349}{350} \times \frac{99}{100} = 0.9872.$

    (ii) $\text{pr}(L \text{ and } \overline{H}) = \frac{1}{100} \times \frac{349}{350} = 0.009971.$

    (iii) $\text{pr}(L \text{ and } H) = \frac{1}{350} \times \frac{1}{100} = 2.857 \times 10^{-5}.$

(c) The process of picking major prizes is independent of that for picking minor prizes.

    (i) $\text{pr}(No\ minor\ prizes$ and $No\ major\ prizes)$
    $= \text{pr}(No\ minor\ prizes) \times \text{pr}(No\ major\ prizes) = 0.729 \times 0.9872 = 0.7197.$
    We have used independence and the answers to (a)(i) and (b)(i).

    (ii) $\text{pr}(1\ or\ more\ minor\ prizes$ and $No\ major\ prizes)$
    $= \text{pr}(1\ or\ more\ minor\ prizes) \times \text{pr}(No\ major\ prizes) = (1 - 0.729) \times 0.9872$
    $= 0.2675.$
    (Note that 1 or more minor prizes is the complement of no minor prizes.)

19. (a) We form the following two-way table

	Low	Medium	High	Total
Claim	0.01×0.2	0.04×0.7	0.1×0.1	
No claim				
Total	0.2	0.7	0.1	1.00

and complete it to form

	Low	Medium	High	Total
Claim	0.002	0.028	0.01	0.04
No claim	0.198	0.672	0.09	0.96
Total	0.200	0.700	0.10	1.00

(b) $\text{pr}(Claim \mid Medium) = 0.04$ (information originally given).

(c) $\text{pr}(Claim$ and $Medium) = 0.028.$

(d) $\text{pr}(Claim) = 0.04.$

(e) $\text{pr}(High \mid Claim) = \dfrac{\text{pr}(High \text{ and } Claim)}{\text{pr}(Claim)} = \dfrac{0.01}{0.04} = 0.25.$

*(f) Arguing as in problem 17(a)(i), the probability a low risk person has no claims in 3 years is 0.99^3. Similarly, for medium risk people it is 0.96^3 and for high risk people 0.9^3. We use this to create the following two-way table.

	Low	Medium	High	Total
No claims in 3 years	$0.99^3 \times 0.2$	$0.96^3 \times 0.7$	$0.9^3 \times 0.1$	
Some claims				
Total	0.2	0.7	0.1	1.00

From the table we see that
$\text{pr}(No\ claim\ in\ 3\ years) = 0.99^3 \times 0.2 + 0.96^3 \times 0.7 + 0.9^3 \times 0.1$
and thus $\text{pr}(Low \mid No\ claims\ in\ 3\ years) = \dfrac{\text{pr}(Low \text{ and } No\ claims\ in\ 3\ years)}{\text{pr}(No\ claim\ in\ 3\ years)}$

$$= \frac{0.99^3 \times 0.2}{0.99^3 \times 0.2 + 0.96^3 \times 0.7 + 0.9^3 \times 0.1} = 0.2190.$$

21. (a) (i) $\frac{97,473}{98,826} = 0.9863.$

    (ii) $\frac{89,099}{98,304} = 0.9064.$

(b) (i) 20 to 25 where the probability of dying is $\frac{97473-96594}{97473} = 0.0090$.

(ii) 35 to 40 where the probability of dying is $\frac{97317-96792}{97317} = 0.0054$.
Young men are risk takers.

(c) (i) $\frac{82690}{97473} \times \frac{89099}{98304} = 0.84834 \times 0.90636 = 0.7689$.

(ii) $\frac{97473-82690}{97473} \times \frac{89099}{98304} = 0.15166 \times 0.90636 = 0.1375$.

(iii) $\frac{97473-82690}{97473} \times \frac{98304-89099}{98304} = 0.15166 \times 0.09364 = 0.0142$.

(d) We have assumed that the life times of the man and the woman are independent. No. Positive. Keep each other going!? Two people in a couple will tend to be more similar in diet and lifestyle factors than two random people and they will have more similar exposures to hazards.

(e) A proportion $\frac{96792}{98304}$ of males who are alive at 20 live to 40. The corresponding proportion for females is $\frac{96792}{98304}$. We construct the following two-way table that applies only to people still alive at 20.

	Male	Female	Total
Live till 40	$\frac{94407}{97473} \times 0.5$	$\frac{96792}{98304} \times 0.5$	
Die before 40			
Total	0.5	0.5	1.00

From this we obtain $\mathrm{pr}(\textit{Live till 40}) = \frac{94407}{97473} \times 0.5 + \frac{96792}{98304} \times 0.5 = 0.9766$.

(f) We assume that the survival rates haven't changed since the data were compiled. However, we can expect life expectancies, and hence the survival rates, to have increased a little. Still, the answers will give a reasonable approximation. The realities are rather subtle. Life-table data reflects current death rates for people at each age at the time the data was collected. For example, the people now dying or not dying at age 50 were 20-years old 30 years ago. The environment (including diet, disease patterns and many other things) they have lived through may be quite different from the environment current 20-year olds will experience over the next 30 years and such differences may produce different death rates for 50 year-olds. So these predictions depend on death rates for people of every age being the same as they were in 1985–7.

23. Suppose door 1 is the car (C) and doors 2 and 3 are goats (G). If you never switch, $\mathrm{pr}(C) = \frac{1}{3}$. Now,

$$\mathrm{pr}(C) = \mathrm{pr}(C \,|\, \textit{Choose 2 or 3})\mathrm{pr}(\textit{Choose 2 or 3})$$
$$+\mathrm{pr}(C \,|\, \textit{Choose 1})\mathrm{pr}(\textit{Choose 1}).$$

If you always switch, then initially choosing 2 or 3 will guarantee that you get the car when you switch (you will have initially chosen a goat and the host will have chosen the other goat), i.e., $\mathrm{pr}(C \,|\, \textit{Choose 2 or 3}) = 1$. However, if you initially chose 1 then you won't get the car when you switch, i.e., $\mathrm{pr}(C \,|\, \textit{Choose 1}) = 0$. Hence

$$\mathrm{pr}(C) = 1 \times \frac{2}{3} + 0 \times \frac{1}{3} = \frac{2}{3}.$$

25. (a) We construct the following two-way table

Result

Digit	Correct	Not corect	Total
0		0.001×0.4	0.4
1		0.002×0.6	0.6
Total			1.0

and complete it to obtain

	Correct	Not corect	Total
0	$0.999 \times 0.4 = 0.3996$	$0.001 \times 0.4 = 0.0004$	0.4
1	$0.998 \times 0.6 = 0.5988$	$0.002 \times 0.6 = 0.0012$	0.6
Total	0.9984	0.0016	1.0

Reading from the table we obtain $\mathrm{pr}(Correct) = 0.9984$.

(b) (i) $0.9984^{1000} = 0.20164$. **(ii)** $0.9984^{10000} = 1.11 \times 10^{-7}$.

(c) The probability of reversing twice will be negligible so that we can ignore this event. The required probability is then $0.4 \times 0.999 \times 0.998 + 0.6 \times 0.998 \times 0.996 = 0.3988 + 0.5964 = 0.9952$.

*27. Let $\mathrm{pr}(Accident \,|\, Right\ handed) = p$. Then, $\mathrm{pr}(Accident \,|\, Left\ handed) = 1.89p$. We put this information on a two-way table as follows

	Accident	No accident	Total
Left handed	$1.89p \times 0.1$		0.1
Right handed	$p \times 0.9$		0.9
Total			1.0

and partially complete it to form

	Accident	No accident	Total
Left handed	$1.89p \times .1$	*	0.1
Right handed	$p \times .9$	*	0.9
Total	$p \times 1.089$	*	1.0

(We don't need to know the items labeled * as they do not come into our calculations.) Then

$$\mathrm{pr}(Left\ handed \,|\, Accident) = \frac{\mathrm{pr}(Left\ handed\ \text{and}\ Accident)}{\mathrm{pr}(Accident)}$$

$$= \frac{1.89p \times 0.1}{1.089p} = \frac{0.189}{1.089} = 0.1736.$$

Note that the answer does not involve p.