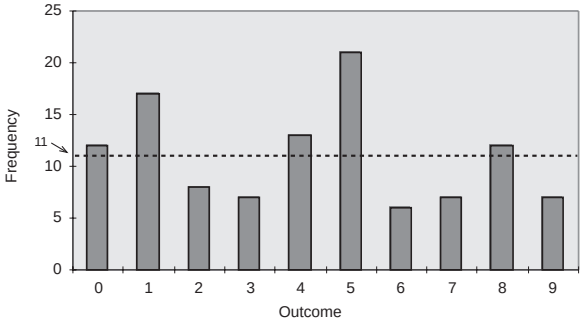
Chapter 11 Tables of Counts

Exercises for Section 11.1

1. **(a)** A bar graph of the observations O_i follows.

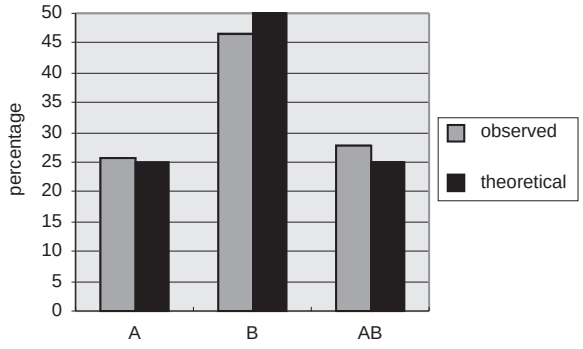


We wish to test $H_0 : p_0 = p_1 = \cdots = p_{10} = \frac{1}{10}$ versus H_1: p_i's not all equal. Under H_0, each expected frequency is $E_i = n\,p_i = 100 \times \frac{1}{10} = 11$. The test statistic is therefore $x_0^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{(12-11)^2}{11} + \frac{(17-11)^2}{11} + \ldots + \frac{(7-11)^2}{11} = 20.36$, with $df = 10 - 1 = 9$, and P-value $= \text{pr}(X^2 \geq 20.36) = 0.016$. There is fairly strong evidence against H_0 and therefore fairly strong evidence that this method of selecting numbers is not random.

(b) We don't have a simple random sample of telephone numbers. However, the sampling method is not unreasonable, particularly if the (same) number used for each page is randomly selected from the first page.

2. **(a)** The null hypothesis is that the probabilities follow the genetic law, i.e., $H_0 : p_A = \frac{1}{4}, p_{AB} = \frac{1}{2}$, and $p_B = \frac{1}{4}$.

(b) The sample proportions are 0.26, 0.46, and 0.28. A side-by-side bar graph comparing the observed percentages with the expected percentages follows. Thedifferences look relatively small.



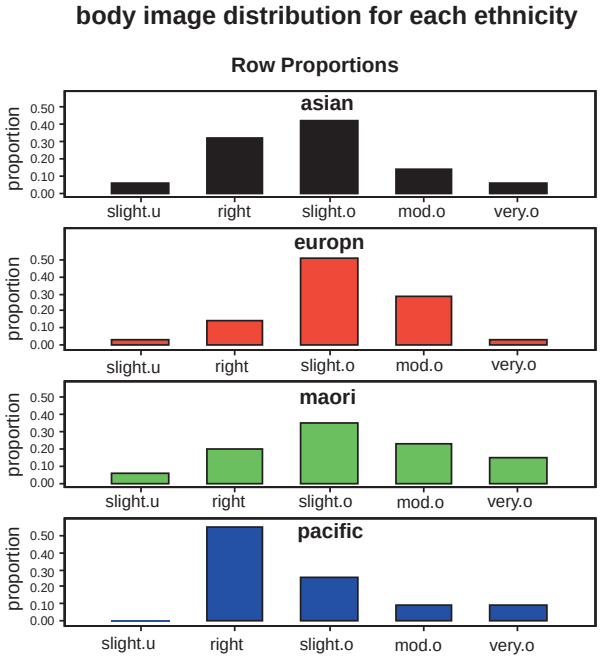
The expected counts when H_0 is true are $E_A = \frac{1}{4} \times 151 = 37.75$; similarly $E_B = 37.75$ and $E_{AB} = \frac{1}{2} \times 151 = 75.5$. The Chi-square test statistic is $x_0^2 = \sum \frac{(O_X - E_X)^2}{E_X} = \frac{(39-37.75)^2}{37.75} + \frac{(70-75.5)^2}{75.5} + \frac{(42-37.75)^2}{37.75} = 0.93$, $df = 3 - 1 = 2$, and P-value $=$

$\text{pr}(X^2 \geq 0.93) = 0.63$. The data provides no evidence against H_0, i.e., no evidence against the genetic law.

Exercises for Section 11.2.1 to 11.2.3

In all answers from now on, we quote and interpret Chi-square statistics and P-values automatically generated by statistical computer packages.

1. Bar graphs for the row proportions follow (the proportions in each row sum to 1).

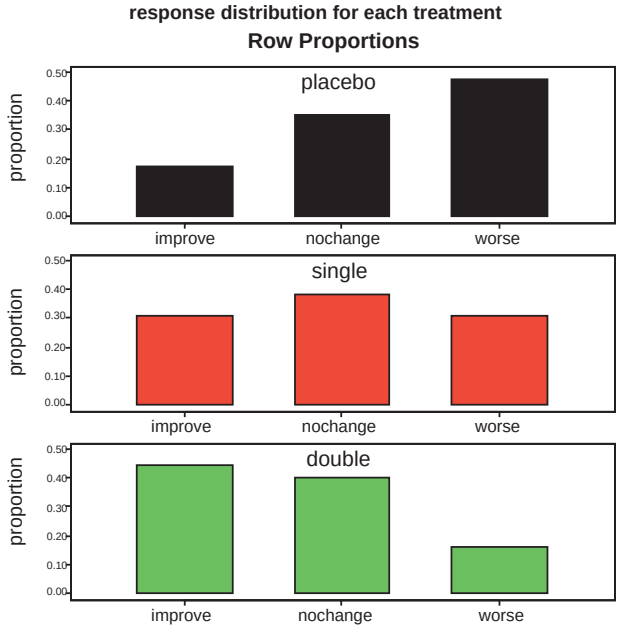


We wish to test H_0: the factors are independent versus H_1: the factors are not independent. The Chi-square test statistic is $x_0^2 = \sum \frac{(O_{ij}-E_{ij})^2}{E_{ij}} = 39.2$, with $df = (4-1)\times(5-1) = 12$, and P-value $= 0.0001$. There is very strong evidence against H_0, i.e. very strong evidence that body image and ethnicity are related. The bar graphs suggest that the Pacific group is quite different from the rest. Deleting this group, $x_0^2 = 20.2$, $df = 8$, and P-value $= 0.0095$, i.e., we still reject H_0 very strongly. There is very strong evidence that body image and ethnicity are related even in the remaining three groups.

Since the sampling of this data was situation (1) of Fig. 11.2.7 we are also allowed think in terms or row proportions (or column proportions) and homogeneity as an alternative to thinking in terms of the situation (1) picture and independence. The row proportions correspond to the distribution of body-image choices for each ethnic group. Comparing body-image distributions between ethnic groups seems to us to be the most natural way of thinking about this data. The biggest visual differences we can see among the 3 remaining groups is that the Asian group seems to have a higher proportion placing themselves in the just right category than the other 2 groups, whereas the European group seems to have higher proportions placing themselves in the slightly overweight and moderately overweight categories. The Maori group seems to be more uniformly spread across the categories than are the other groups. We could use confidence intervals for differences in proportions between ethnic groups to see if

the features we are seeing are likely to be real or whether they can be explained simply in terms of sampling variation.

2. **(a)** Situation (2) with rows corresponding to separate samples. Bar graphs for the three row proportions follow (the proportions in each row sum to 1).



(b) We wish to test for homogeneity of row distributions, which in this case corresponds to testing whether the underlying true distributions amongst the response categories (improve/no change/get worse) are the same for each treatment group. H_0 says the underlying distributions are identical, H_1 says differences exist.

The Chi-square test statistic is $x_0^2 = \sum \frac{(O_{ij}-E_{ij})^2}{E_{ij}} = 55.08$, with $df = (3-1) \times (3-1) = 4$ giving P-value $= \text{pr}(X^2 \geq 55.08) \approx 0$. We very strongly reject H_0. There is clear evidence that real differences exist.

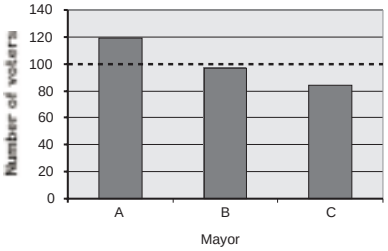
This is not surprising from the bar graphs as they look very different in shape. They show a drift from the bad category towards the better categories as we go from placebo, to a single dose of the drug (fewer get worse, more improve). This trend continues as we go from a single dose to a double dose.

The sample proportions showing improvement with the three treatments are, respectively, 17.5%, 31% and 44%. Appropriate 95% confidence intervals for the true difference are (since we are comparing independent proportions) *single dose*$-$*placebo* : $0.31 - 0.175 \pm 1.96\sqrt{\frac{(.31)(.69)}{200} + \frac{(.175)(.825)}{200}}$, i.e., $[0.05, 0.22]$; and *double dose*$-$*single dose* : $0.44-0.31+1.96\sqrt{\frac{(.44)(.56)}{200} + \frac{(.31)(.69)}{200}}$, i.e., $[0.04, 0.22]$. Neither contains zero so there is evidence that a single dose is more effective than the placebo, and a double dose is more effective than a single dose.

Review Exercises 11

In all answers in these Review Exercises, we quote and interpret Chi-square statistics and P-values automatically generated by statistical computer packages.

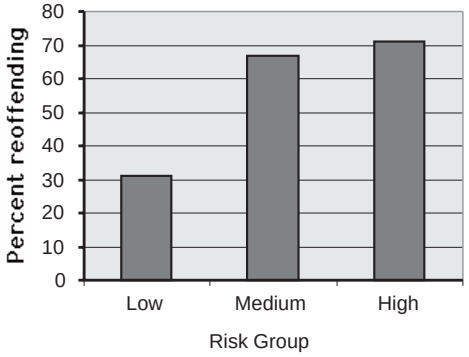
1. We wish to test whether the true proportions of people preferring each of the candidates are the same, i.e., we wish to test $H_0 : p_A = p_B = p_C = \frac{1}{3}$. Each expected value is 100 which we can locate on the following bar graph of the observed counts.



The Chi-square test statistic is $x_0^2 = \sum \frac{(O_X - E_X)^2}{E_X} = 6.26$, $df = 3-1 = 2$, and P-value $= \text{pr}(X^2 \geq 6.26) = 0.044$. There is some evidence that the three candidates are not equally preferred.

3. **(a)** We wish to test H_0: the three conviction rates are all the same versus H_1: the three conviction rates are not all the same.
The Chi-square test statistic is $x_0^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 29.55$, $df = (3-1)\times(2-1) = 2$, and P-value ≈ 0. There is very strong evidence that the rates are different for the three groups.

(b) We use subscripts L, M, and H to denote Low, Medium, and High. The observed proportions reconvicted are $\widehat{p}_L = \frac{23}{75} = 0.3067$, $\widehat{p}_M = \frac{50}{75} = 0.6667$, and $\widehat{p}_H = \frac{53}{75} = 0.7067$. These are represented by the following bar graph.

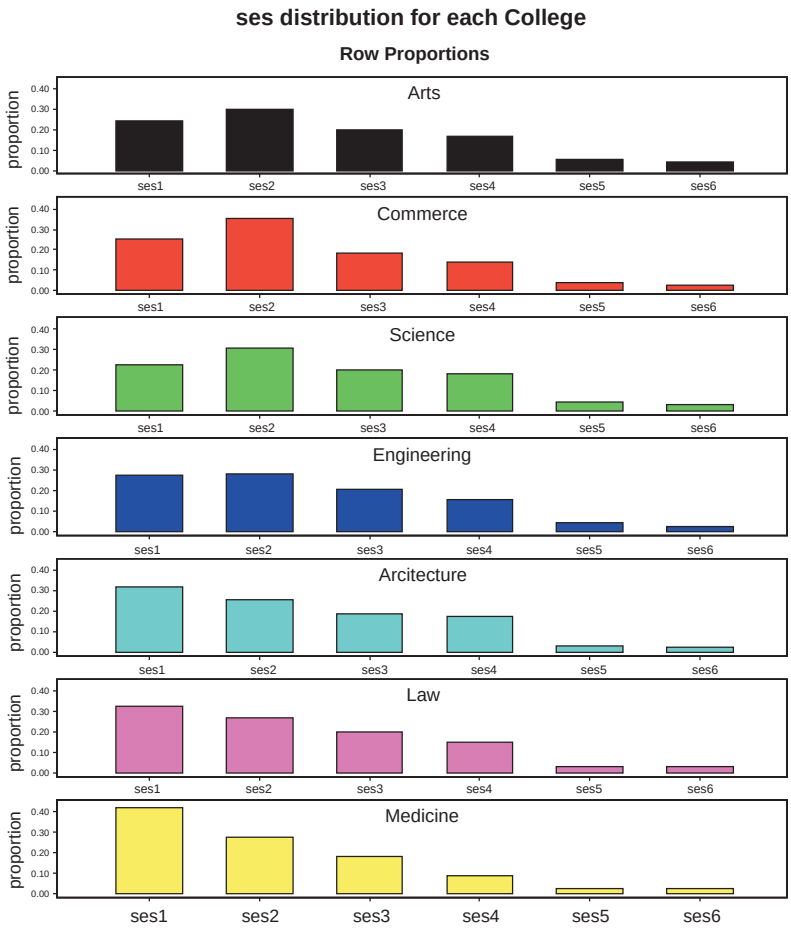


The 95% confidence interval for $p_H - p_M$ is $0.04 \pm 1.96\sqrt{\frac{(.6667)(.3333)}{75} + \frac{(.7067)(.2933)}{75}}$, i.e, $[-0.11, 0.19]$. The corresponding interval for $p_M - p_L$ is $[0.21, 0.51]$, and that for $p_H - p_L$ is $[0.25, 0.55]$. As we might expect from the bar graph, there is a a significant difference between the low and the medium or high rates (of more than 20 percentage points), but no significant difference between the medium and high rates.

(c) No, the number of re-offenders will be greater than those reconvicted as some won't get caught.

5. (a) We wish to test the hypothesis H_0: there is no relationship between the enrollment rate of each college (faculty) and father's Socio-economic status. The Chi-square test statistic is $x_0^2 = \sum \frac{(O_{ij}-E_{ij})^2}{E_{ij}} = 146.2$, $df = (7-1)\times(6-1) = 30$, and P-value ≈ 0. There is very strong evidence against H_0, i.e., we have very strong evidence that College and socio-economic status are related.

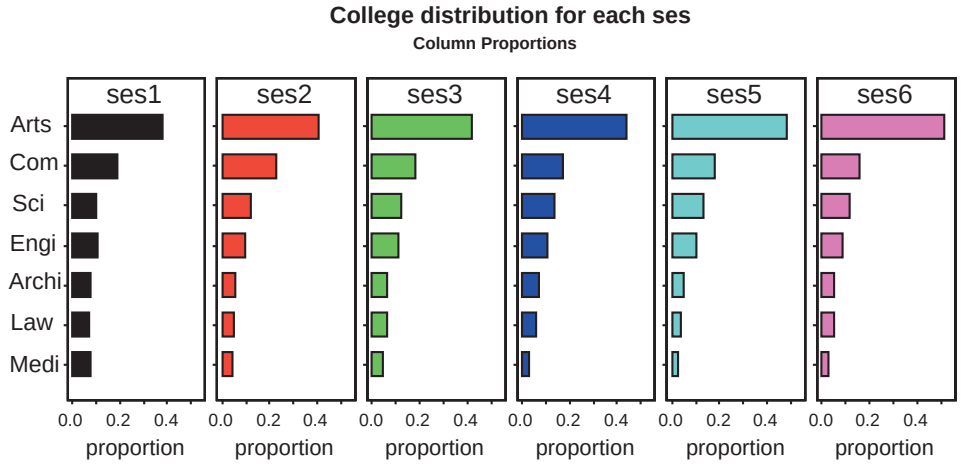
(b) The row proportions, which give us the observed socio-economic status distribution for each College, are plotted in the bar graphs are given below and given (rounded) in the table following the graphs.



	Socio-economic Status						
College	1	2	3	4	5	6	Total
Arts	.24	.30	.20	.17	.05	.04	1
Comm.	.26	.35	.18	.14	.04	.03	1
Sci.	.23	.31	.20	.18	.05	.03	1
Eng.	.28	.28	.21	.16	.04	.03	1
Arch.	.32	.26	.19	.17	.03	.03	1
Law	.32	.27	.20	.15	.03	.03	1
Med.	.42	.27	.18	.09	.02	.02	1

The shapes of the distributions for Arts, Commerce and Science seem similar with ses2 being having the greatest proportion and ses1 having the 2nd greatest. The shapes for the professional schools of Architecture, Law and Medicine also look fairly similar with ses1 being the biggest category and then getting progressively getting smaller as ses levels get lower. (The weighting towards higher categories seems stronger in Medicine.) Engineering falls in between the two groups with the ses1 and ses2 proportions being roughly the same.

For Arts, Commerce and Science there are higher proportions in status 2 than in status 1; in Engineering they are about the same; but in Architecture, Law and Medicine the proportions are higher in status 1. These differences show up clearly in the pairwise 95% confidence intervals for independent proportions. For status 1, all the confidence intervals comparing any one of Arts, Commerce and Science with any one of Architecture, Law and Medicine do not contain zero. This indicates that these two groups of colleges are different with respect to status 1. Engineering tends to fall in the middle. In the same way Medicine is different from the other 6 colleges; there is a significantly higher proportion of medical students among status 1. Apart from Medicine, the proportions for the status categories 3-6 are very similar for the remaining colleges. We also include the bar graphs for the column proportions below, but they are not as informative. They confirm our previous comments.



7. **(a)** Incomplete information. Probably some people did not answer the smoking question. We expect little bias here due to missing data because the proportion of missing data is very small. There have been many warnings about smoking during

pregnancy, however, so we might expect some social-desirability bias consisting of under-reporting of smoking.

(b) (i) Smoke, sleep prone, breast feed. (ii) Socioeconomic status and season. Those in (i). We can use them to address the crib or cot death problem.

(c) In what follows, the cases and controls come from independent populations so that corresponding proportions are independent.

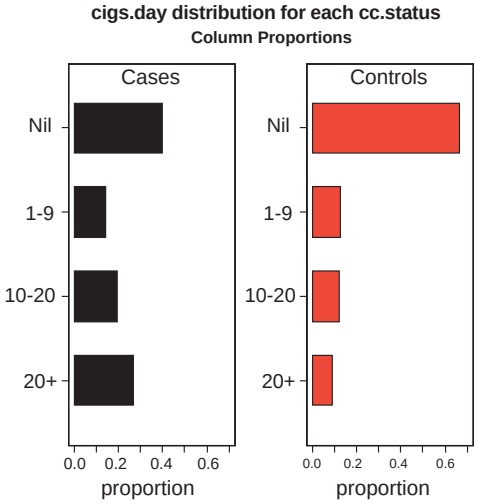
(i) *Smoke versus crib death rate.*
We wish to test H_0: there is no relationship between smoking and the crib death rate. The Chi-square test statistic is $x_0^2 = \sum \frac{(O_{ij}-E_{ij})^2}{E_{ij}} = 35.1$, $df = (2-1)\times(2-1) = 1$, and P-value ≈ 0. There is very strong evidence that smoking is related to crib death. The proportions are $\widehat{p}_{cases} = 0.6320$ and $\widehat{p}_{controls} = 0.3415$. The 95% confidence interval for the true difference in proportions who smoked between case mothers and control mothers, $p_{cases} - p_{controls}$, is $0.632 - 0.3415 \pm 1.96\sqrt{\frac{(.6320)(.3680)}{125} + \frac{(.3415)(.6585)}{492}}$, i.e, $[0.20, 0.38]$. Smoking is clearly a major factor.

(ii) *Sleep prone versus crib death rate.*
We wish to test H_0: there is no relationship between the prone sleeping position and the crib death rate. The Chi-square test statistic is $x_0^2 = \sum \frac{(O_{ij}-E_{ij})^2}{E_{ij}} = 36.1$, $df = (2-1)\times(2-1) = 1$, and P-value ≈ 0. There is very strong evidence that sleeping in the prone position is related to crib death. The 95% confidence interval for the difference, $p_{cases} - p_{controls}$, in proportions of babies who slept in the prone position is $0.7266 - 0.4294 \pm 1.96\sqrt{\frac{(.7266)(.2734)}{128} + \frac{(.4294)(.5706)}{503}}$, i.e., $[0.21, 0.39]$. Case babies are considerably more likely to have slept in a prone position.

(iii) *Breast feed versus crib death rate.*
We wish to test H_0: there is no relationship between breast feed and crib death rate. The Chi-square test statistic is $x_0^2 = \sum \frac{(O_{ij}-E_{ij})^2}{E_{ij}} = 24.1$, $df = (2-1)\times(2-1) = 1$, and P-value ≈ 0. There is very strong evidence that not-breast-feeding is related to crib deaths. The 95% confidence interval for the true difference, $p_{controls} - p_{cases}$, is $0.8529 - 0.6641 \pm 1.96\sqrt{\frac{(.8529)(.1471}{503} + \frac{(.6641)(.3359)}{128}}$, i.e, $[0.10, 0.28]$. Not breast feeding is clearly a major factor.

(iv) *Mother's smoking in the last two weeks versus crib death rate.*
We have already established a smoking relationship in (i). The column proportions give the observed numbers-smoked distributions for the cases and control groups respectively. Rounded column proportions and bar graphs of the column proportions are given below. There seems to be a clear difference between the cases and the controls. This difference is confirmed by the Chi-square test statistic $x_0^2 = 42.0$, $df = 1$, and P-value ≈ 0. The case-mothers appear to be much less likely to fall into the "nill" category and much more likely to fall into the 20+ per day category and somewhat more likely to fall into the 10-20 per day category.

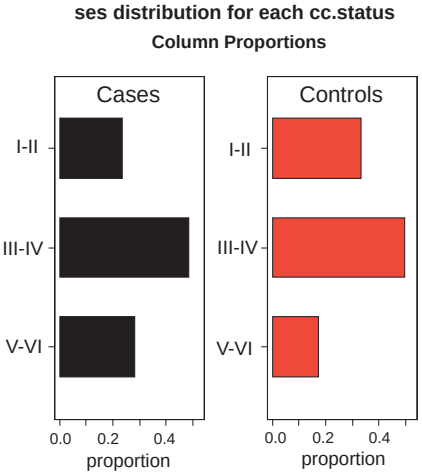
Cigs/day	Cases	Controls
Nil	0.40	0.66
1-9	0.14	0.13
10-19	0.20	0.12
20+	0.27	0.09
Total	1.00	1.00
No. in column	128	503



The 95% confidence intervals for $p_{cases} - p_{controls}$ for the proportions falling into the following categories are as follows: Nil, $[-0.36, -0.17]$; 1-9, $[-.051, .082]$; 10-19, $[-.002, 0.15]$; and 20+, $[0.10, 0.26]$. There are clear differences for the categories Nil and 20+. For example, with 95% confidence, the true percentage of cases falling into the 20+ category is bigger than the corresponding percentage for controls by somewhere between 10 and 26 percentage points. The situation is not so clear cut for the 10–19 group as the left-hand end of this interval just contains 0.

(v) *Socioeconomic status versus crib death rate.* The bar graphs for the column proportions are given below. The column proportions in the table are rounded.

Status	Cases	Controls
I-II	0.23	0.33
III-IV	0.48	0.50
V-VI	0.28	0.17
Total	1.00	1.00

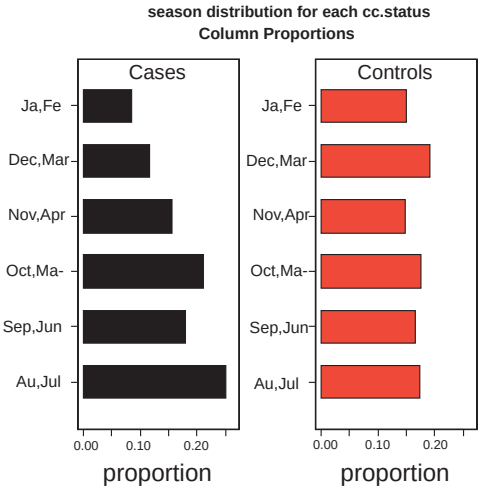


There appears to be some differences, with the cases group being weighted more towards the lower V and VI status groups. To confirm this we test H_0: there is no relationship between socioeconomic status and crib death rate. The Chi-square test statistic is $x_0^2 = \sum \frac{(O_{ij}-E_{ij})^2}{E_{ij}} = 9.3$, $df = (3-1) \times (2-1) = 2$, and P-value $= 0.01$. There is strong evidence relating crib deaths to socioeconomic status.

Of the three pairwise 95% confidence intervals for for differences, $p_{cases} - p_{controls}$,

for the various categories, only the one comparing the proportions of cases and controls in the V and VI groups ($[0.024, 0.193]$) does not contain 0.

(vi) *Season versus crib death rate.*
The bar graphs for the column proportions are given below. There appear to be some seasonal differences. Also, there is a greater seasonal variation for the cases. We wish to test H_0: there is no relationship between season and the crib death rate. The Chi-square test statistic is $x_0^2 = \sum \frac{(O_{ij}-E_{ij})^2}{E_{ij}} = 10.26$, $df = (6-1) \times (2-1) = 5$, and P-value $= 0.068$. There is weak evidence relating crib deaths to season. Considering the pairwise 95% confidence intervals, those for November-April, October-May, September-June and August-July all contain 0, while 0 lies just outside the other two, January-February and December-March. The main differences occur in the N.Z. summer months.



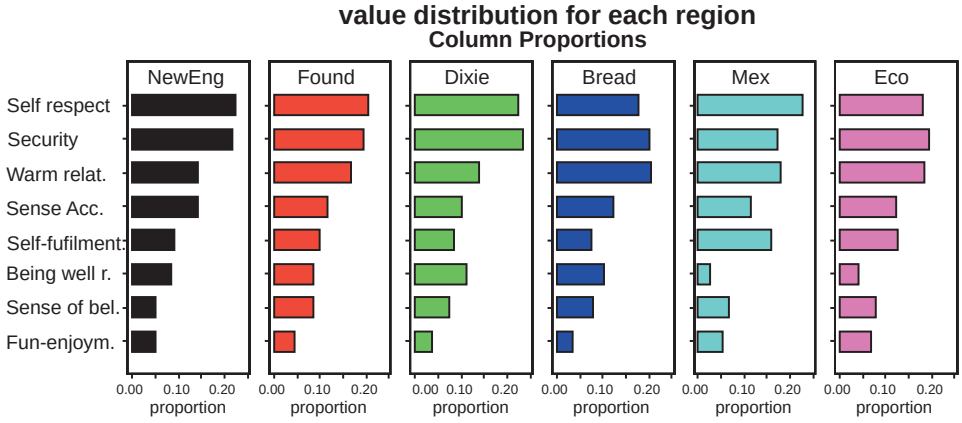
(d) The bar graphs above suggest a seasonal variation for cases but not for controls.

(i) For the cases we test $H_0 : p_1 = p_2 = \cdots = p_6 = \frac{1}{6}$. Each $E_i = \frac{128}{6} = 21.3333$. The Chi-square test statistic is $x_0^2 = \sum \frac{O_i-E_i)^2}{E_i} = 13.94$, $df = 6-1 = 5$, and P-value $= 0.016$. There is moderate evidence against H_0, i.e., cot deaths do not occur uniformly throughout the year. The low point is January-February (N.Z. summer), while the high point is August-July (N.Z. winter).

(ii) For the controls we test $H_0 : p_1 = p_2 = \cdots = p_6 = \frac{1}{6}$. Each $E_i = \frac{503}{6} = 83.83333$. The Chi-square test statistic is $x_0^2 = \sum \frac{(O_i-E_i)^2}{E_i} = 4.18$, $df = (6-1) = 5$, and P-value $= 0.52$. There is no evidence against H_0, i.e., no evidence of a seasonal variation.

(e) We are using collapsed tables which ignore interactions and hidden variables (recall Simpson's paradox). We cannot conclude causality from this evidence alone.

9. **(a)** The bar graphs for the column proportions follow.



There appear to be only minor differences between the regions. To check this we test H_0: there is no relationship between the region and the most important values. The Chi-square test statistic is $x_0^2 = \sum \frac{(O_{ij}-E_{ij})^2}{E_{ij}} = 48.7$, $df = (8-1) \times (6-1) = 35$, and P-value $= 0.062$. There is weak evidence against H_0. A rounded table of proportions follows.

Most Important value	Region: New Engl.	The Foundry	Dixie	Bread-basket	Mex. Amer.	Eco-topia
Self-respect	.23	.21	.23	.18	.23	.18
Security	.22	.20	.23	.20	.17	.20
Warm relat.	.14	.17	.14	.21	.18	.19
Sense Accomp.	.14	.12	.10	.12	.11	.12
Self-fulfillment	.09	.10	.08	.08	.16	.13
Being well resp.	.08	.09	.11	.10	.03	.04
Sense of belong.	.05	.08	.08	.08	.07	.08
Fun-enjoyment	.05	.05	.04	.04	.05	.07
Total	1.00	1.02	1.01	1.01	1.00	1.01

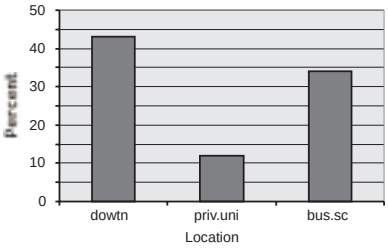
The first few values seem to be more popular than the rest for all the segments. We see many minor differences that might be worth investigating further, e.g., the low rating given to being well respected in "Mex-Amer" and "Ecotopia" compared with the other segments.

(b) You would want to divide up the market into distinct groups (segments) so that the people within each group were as similar as possible. This way you could target each group separately. In the above example the same values appealed to all the segments so that either the values chosen for surveying were not very discriminating, or the segments used have failed to divide the market into groups of people with similar values and aspirations. A marketing strategy which focused on the four most popular values should appeal to all the segments.

11. The appropriate two-way table of counts is:

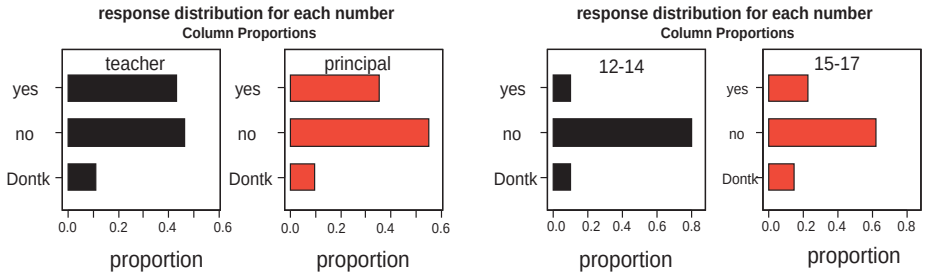
	Downtown	Private university	Bus. school.	Total
Caps back.	174	29	107	310
Other	233	207	212	652
Total	407	236	319	962

The bar graph for the three proportions of students wearing their caps backwards follows.



The private university location appears to be different from the other two (a much smaller proportion). We wish to H_0: there is no relationship between the way a baseball cap is worn and the location. The Chi-square test statistic is $x_0^2 = \sum \frac{(O_{ij}-E_{ij})^2}{E_{ij}} = 63.9$, $df = (2-1)\times(3-1) = 2$, and P-value ≈ 0. There is very strong evidence against H_0, i.e., we have very strong evidence that differences exist between the true proportions. The 95% confidence intervals comparing the three proportions are virtually identical to the two-standard error intervals we obtained in Review Exercises 7 and we reach the same conclusions.

13. **(a)** Bar graphs for the proportions follow.



The teachers and principals have very different perceptions from those of the students. The percentages answering *yes* are: principals (48%), teachers (34%), 15-17 (23%) and 12-14 (10%). Teachers and principals have much higher percentages than the students. Assuming that principals tend to be older than teachers, the percentage seems to increase with age.

(b) Using a little bit of detective work to reconstruct the raw frequencies from the percentages, the appropriate table is:

Response	Teachers	Principals	Total
Yes	355	288	643
No	379	452	831
Don't know	91	82	173
Total	825	822	1647

We wish to test H_0: there is no difference between teachers and principals with respect to beliefs about the ability of students to smoke marijuana every weekend and still do well at school. The Chi-square test statistic is $x_0^2 = \sum \frac{(O_{ij}-E_{ij})^2}{E_{ij}} = 13.9$, $df = (2-1)\times(3-1) = 2$, and P-value $= 0.001$. There is very strong evidence against H_0. Pairwise confidence intervals indicate that the differences are in the *yes* and *no* category. For example, a 95% confidence interval for the difference between the population proportions of teachers and principals answering "yes" is given by [0.033, 0.127].

(c) The appropriate table is:

Response	12-14	15-17	Total
Yes	50	115	165
No	400	310	710
Don't know	50	75	125
Total	500	500	1000

The Chi-square test statistic is $x_0^2 = \sum \frac{(O_{ij}-E_{ij})^2}{E_{ij}} = 42.0$, $df = (2-1)\times(3-1) = 2$, and P-value ≈ 0. There is very strong evidence against H_0. None of the the 95% confidence intervals for each of the three pairwise differences contain 0, indicating that the two age groups are quite different in their responses. For example, a 95% confidence interval for the difference between the population proportions of 15–17 year-olds and 12–14 year-olds answering "yes" is given by [0.085, 0.175].

15. We investigated this data in Example 9.3.6. All we want to do here is apply the Chi-square test. The two-way table is given in Table 9.3.4. We test H_0: there is no relationship between the sex of the first child in a family and the sex of the second child. The Chi-square test statistic is $x_0^2 = \sum \frac{(O_{ij}-E_{ij})^2}{E_{ij}} = 30.2$, $df = (2-1)\times(2-1) = 1$, and P-value ≈ 0. There is very strong evidence against H_0. We investigated the nature of the relationship in Example 9.3.6. Note that the Chi-square test statistic is the square of the t (or z) test statistic we got in Example 9.3.6.

17. The appropriate table is:

	Smoker	Non-smoker	Total
Stroke	171	117	288
No stroke	3264	4320	7584
Total	3435	4437	7872

We wish to test H_0: there is no relationship between smoking and having a stroke. The Chi-square test statistic is $x_0^2 = \sum \frac{(O_{ij}-E_{ij})^2}{E_{ij}} = 30.1$, $df = (2-1)\times(2-1) = 1$, and *P-value* ≈ 0. There is very strong evidence against H_0. A higher proportion of smokers get strokes. We gave a confidence interval for the difference in proportions getting strokes between smokers and nonsmokers in our solution to problem 9 of Review Exercises 9. The Chi-square test is approximately the square of the test statistic in Review Exercise 9.

Chapter 12 Relationships Between Quantitative Variables

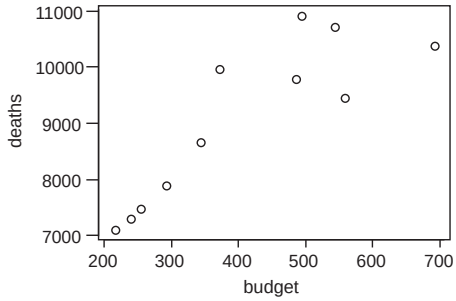
Exercises for Section 12.1.3

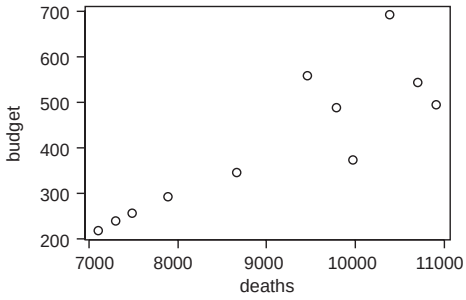
(a) You measure V for different values of I. You can then construct a scatter plot using the pairs of values of I and V using I for the x-axis and V for the y-axis. When I gets smaller so does V, and when $I = 0$, $V = 0$. You can express Ohm's law in the form $y = \beta x$, where $y = V$, $X = I$ and $\beta = r$. This equation represents a straight line through the origin (see Fig. 12.2.2). If the law is true, then the scatter plot should be linear, apart from any random fluctuations due to measurement error. A straight line should therefore fit the plot well. It should pass through the origin.

(b) With a straight-line trend drawn through the origin, an estimate of r is the slope of the fitted line.

Exercises for Section 12.2

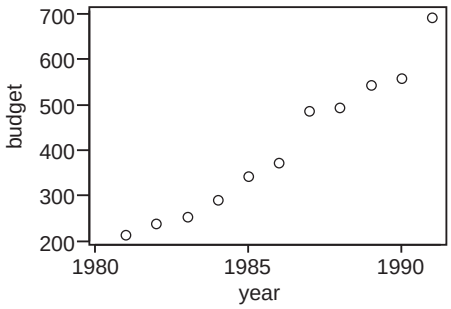
(a) The plot of deaths versus budget follows (below left).





The budget did not cause the deaths as a statistical relationship does not necessarily imply a causal relationship. If drug problem are getting worse over time (reflected in increased numbers of deaths) and budgets are continually increased to try to combat the problem, a pattern similar to this could be expected. The plot of budget versus deaths (above right) is not very helpful, as there does not seem to be a clear statistical relationship after 9000 deaths.

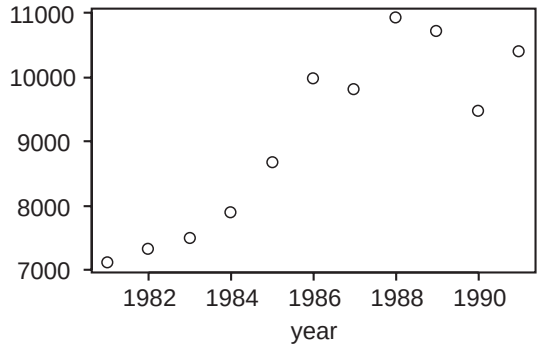
(b)



(i) A plot of budget versus year is shown above. We see that the budget is steadily increasing over time (cf. comments in (a)). A straight line is not an adequate fit.

***(ii)** Another possible trend is an exponential curve, or possibly a quadratic curve.

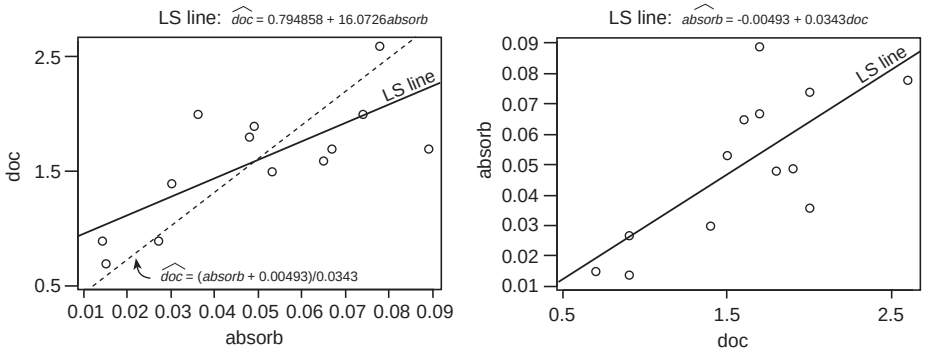
(c) A plot of deaths versus year follows.



There does not seem to be any relationship between budget and deaths other than that they both tend to increase with time.

Exercises for Section 12.3

(a) A plot of DOC versus ABSORBANCE follows (below left). Yes, there is a clear upward trend. It looks slightly curved. [The superimposed lines makes it harder to see the curve in the trend.]



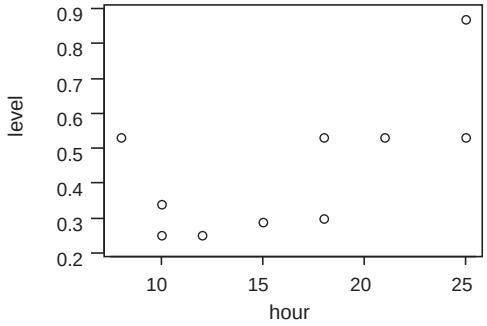
(b) The least squares line is $\widehat{doc} = 0.795 + 16.1\,absorb$. It is the solid (LS) line on the above plot.

(c) A plot of ABSORBANCE versus DOC is shown in (a) above right. From both plots we see a trend of one variable increasing as the other increases. The right-hand trend looks weaker. The least-squares line for the right-hand plot is $\widehat{absorb} = -0.00493 + 0.0343\,doc$. To superimpose this on our previous plot (above left), we need to express *doc* in terms of *absorb*, namely $\widehat{doc} = (absorb + 0.00493)/0.0343$ or, simplifying, $\widehat{doc} = 0.143 + 29.15\,absorb$. The superimposed line is then $y = 0.143 + 29.15x$, the dotted line on the plot for (a).

(d) It depends on the purpose of the study and your beliefs about the system under study. If you wish to see how optical absorbance varies with DOC, then you would use the plot from (c) (right-hand side). This appears to be the more likely use. However, you might want to predict DOC from a measure of Optical Absorbance. You would then use (a). If you thought changes in ABSORBANCE caused changes in DOC you would use (a). If you though the relationship was causal but went the other way, you would use (c).

Exercises for Section 12.4.2

1. (a) The plot follows.



It is a strange-looking plot! Cyclist 1, the cyclist with 8 hours of training could be an outlier. If this point is ignored, there appears to be an increasing trend that looks curved. On the other hand, cyclist 10, with a reading of 0.87, could be an outlier. If this point is ignored, the plot looks somewhat flat. Clearly, these two points have a big effect on what we see.

(b) We will fit a simple linear model and test $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. The test (from standard regression-program output) has $t_0 = 1.38$ and a P-value of 0.047, which indicates that we have some evidence against the no-relationship null hypothesis, i.e., we have some evidence of a trend.

(c) Omitting cyclist 10, $t_0 = 2.35$ with P-value $= 0.211$, which suggests there is no evidence of a trend.

(d) No, as no good reason is given for why this cyclist should be omitted. As we mentioned in (a), the picture would be very different if we omitted cyclist 1.

(e) The data will depend on the training habits of each cyclist, for example, when and where they cycle. Other factors that could affect their exposure to lead are where they live and work. More information is needed about each cyclist.

2. The confidence interval is $[6.44, 25.71]$.

Exercises for Section 12.4.3

The prediction intervals follow (obtained from computer-program output):

(a) When $x = 0.5$ the interval is $[0.170, 0.339]$, with width 0.17.

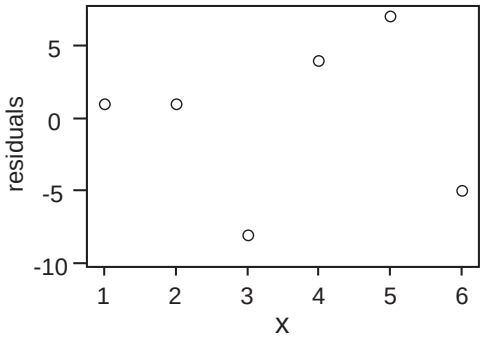
(b) When $x = 6$ the interval is $[2.583, 2.764]$, with width 0.18. The second interval is slightly wider, as 6 is further from $\overline{x}$ than 0.5 (see Fig. 12.4.8).

Exercises for Section 12.4.4

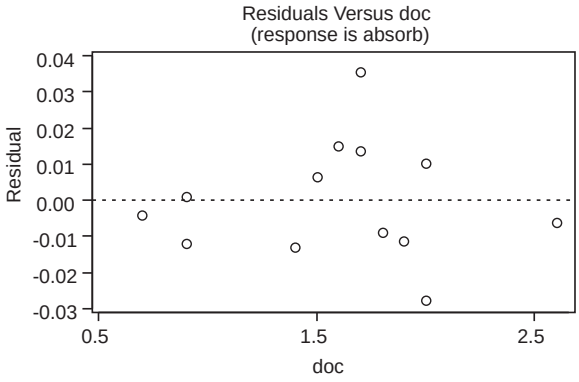
1. The residuals [$residual = y - \widehat{y} = y - (6 + 3\,x)$] are:

x	1	2	3	4	5	6
y	10	13	7	22	28	19
residual	1	1	-8	4	7	-5

The residual plot follows

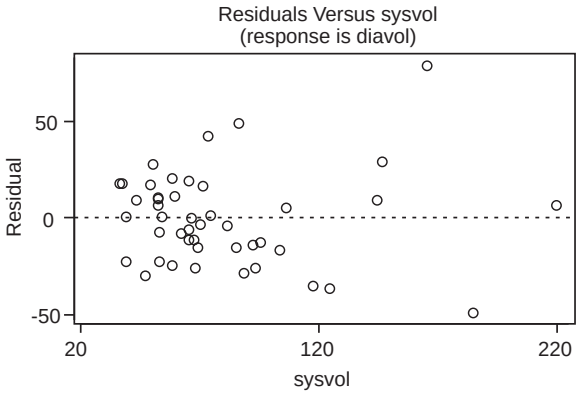


2. The least-squares line is $\widehat{y} = -0.0049 + 0.0343x$. The residual plot below is very similar to the original scatter plot.

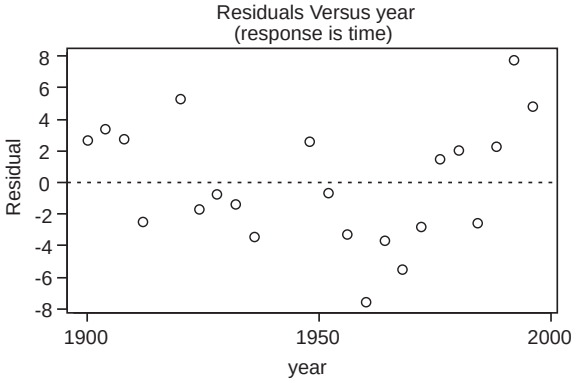


The spread does not seem to be constant. However, there may be two outliers; these need closer investigation.

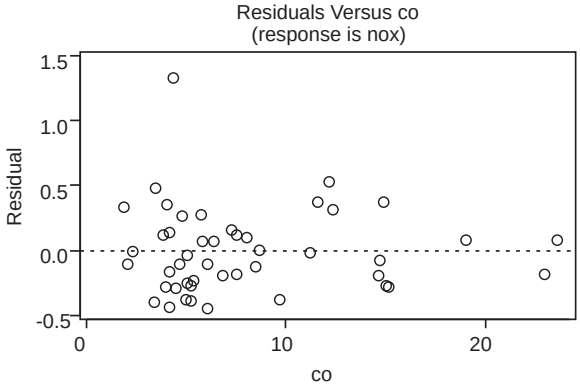
3. The least-squares line is $\hat{y} = 74.713 + 1.060x$. The residual plot below suggests the possibility of a slight fan effect, though this visual effect is caused mainly by only two points – those with the largest residuals.



4. The residual plot below is curved, indicating that a straight-line model is not appropriate.

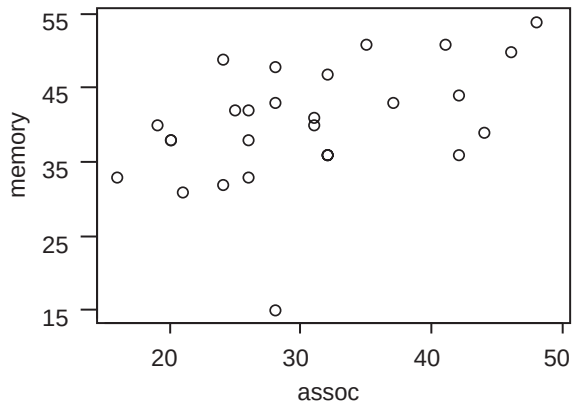


***5.** From the residual plot below, the quadratic model seems to fit reasonably well. However, there is clearly an outlier and a hint of a reducing spread with increasing CO, though the points with large CO are sparse.



Exercises for Section 12.5

(a) From the following scatter plot there seems to be an upward trend that is roughly linear.



(b) It is visible only with the MEMORY scores.

(c) $r = 0.457$ with the outlier, and $r = 0.542$ when the outlier is removed.

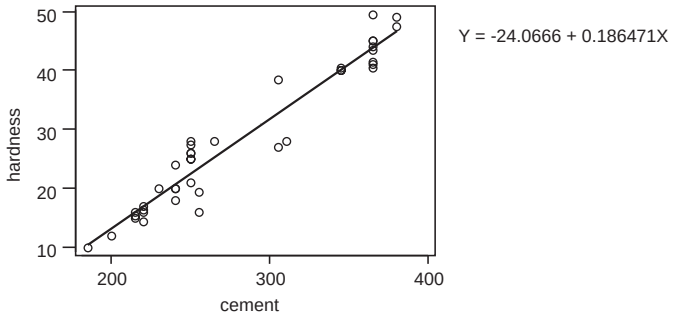
(d) It means that the correlation is significantly different from zero, and relates to H_0 : $\rho = 0$, the hypothesis of no (linear) relationship.

(e) Without the outlier, P-value $= 0.004$, indicating the existence of a relationship.

Review Exercises 12

All answers here were obtained from regression-program output unless otherwise noted.

1. **(a)** Plotting HARDNESS versus CEMENT below, we get an approximate linear relationship.



(b) The least-squares line is $\widehat{y} = -24.1 + 0.186x$.

(c) Once the amount of cement gets below a certain positive level c, say, the mix will not harden at all, i.e., $y = 0$ for all values X below c. The regression line is therefore not relevant to the physical situation below $x = c$, and certainly not at $x = 0$, which gives a negative hardness of -24.1!

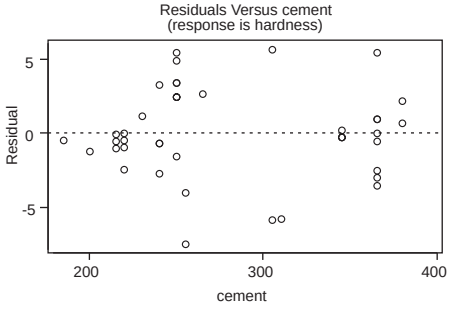
(d) $20 \times 0.186 = 3.72$.

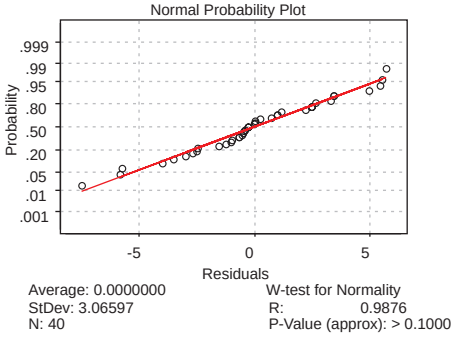
(e) A 95% confidence interval for the slope is $[0.170, 0.203]$. For every 1 g increase in cement, average hardness goes up by somewhere between 0.170 units and 0.203 units. [The interval also indicates that the slope is significantly greater than 0 as the interval is to the right of zero.]

(f) At $x = 275$, the 95% prediction interval is $[20.85, 33.58]$. We would expect the hardness of a new batch made with 275 g of cement to have a hardness somewhere between 20.85 and 33.58 units.

(g) Yes, but we can't be sure as 600 g is considerably larger than any of the cement levels that we have tested or observed (outside the range of the data). The relationship may be different out there.

(h) The following residual plot (below left) looks a little strange not only because of the apparent increased variability in the center of the plot but also because of replicated x-values.

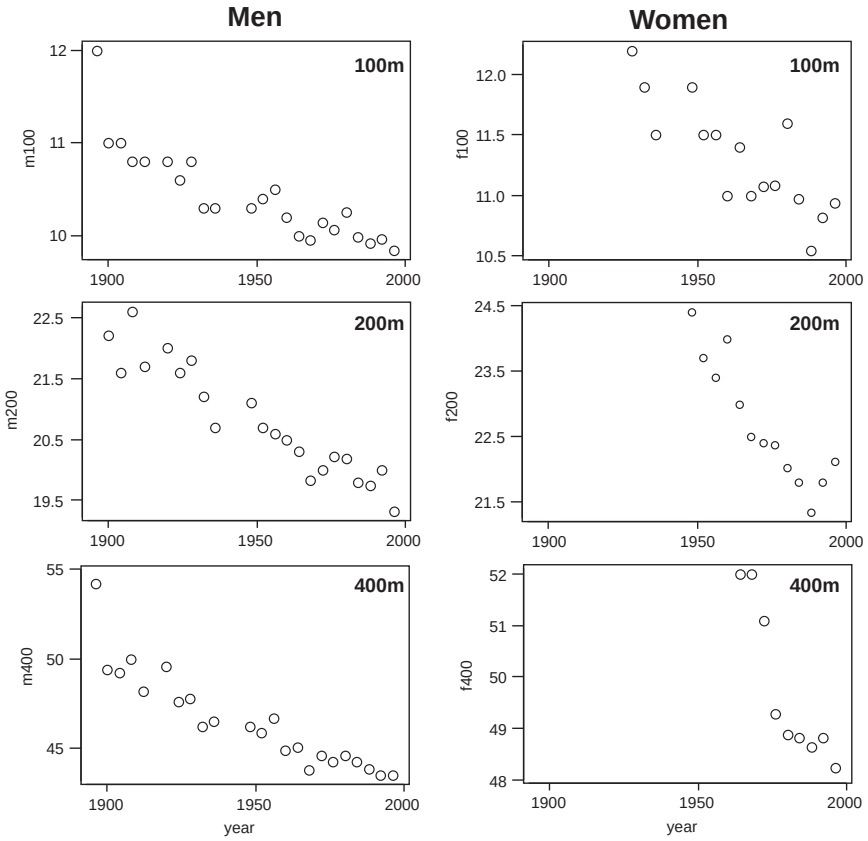




However, the plot is basically horizontal so that the linearity of the model does not appear to be in question. Perhaps, surprisingly, the Normal probability plot (above right) is closely linear, and the W-test for Normality is satisfactory.

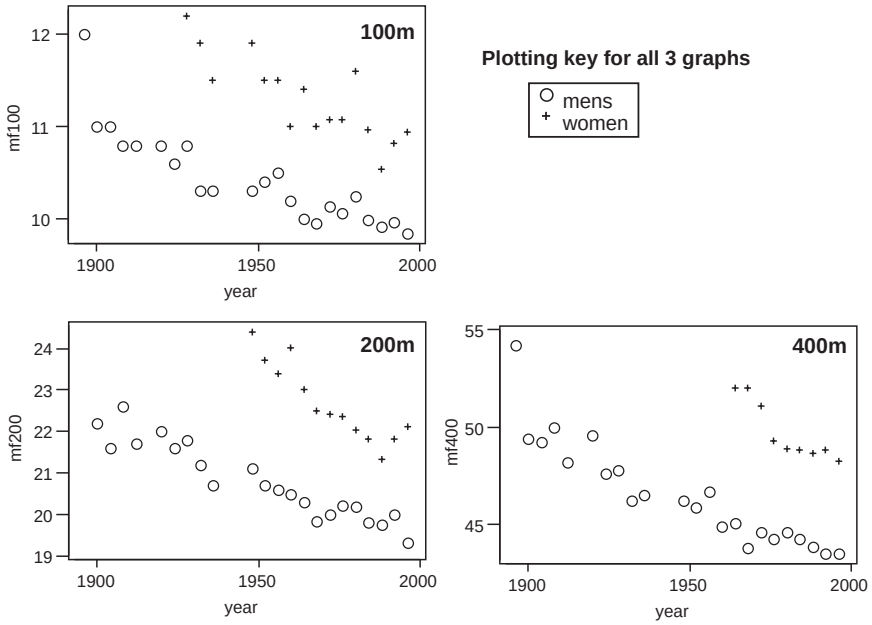
(i) The variability will depend on the degree of mixing of all the ingredients, namely, sand, chips, cement, and water. It will also depend on the quality, or characteristics, of each of the ingredients used and this might vary from batch to batch.

3. (a) All six plots are given below.



For the three men's races, the plots are surprisingly linear, apart from initial outliers for the 100-m and 400-m. There is no sign of leveling off for the 100-m and 200-m, though for the 400-m there appears to be a suggestion of some leveling off as the most recent points are above the general trend. The plots for the women's races are less well defined but, except for recent points, are still approximately linear for the 100-m and 200-m. In these last two, the trend is increasing for the last three points, indicating some leveling off. The plot for the 400-m is unusual, with a change in the slope of the downward trend, but the trend is still appears strongly down hill.

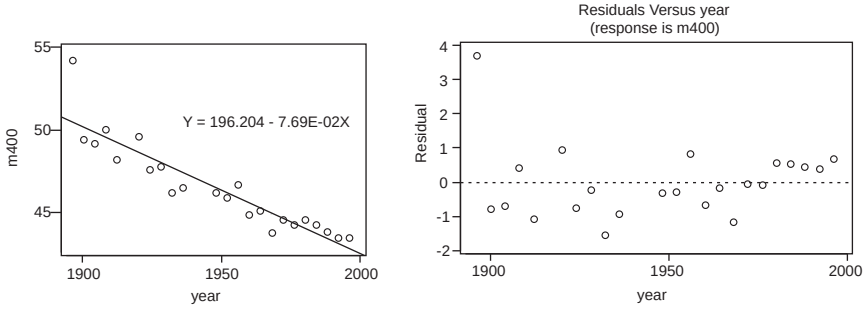
(b) To compare the men's and women's times we plot them on the same graphs.



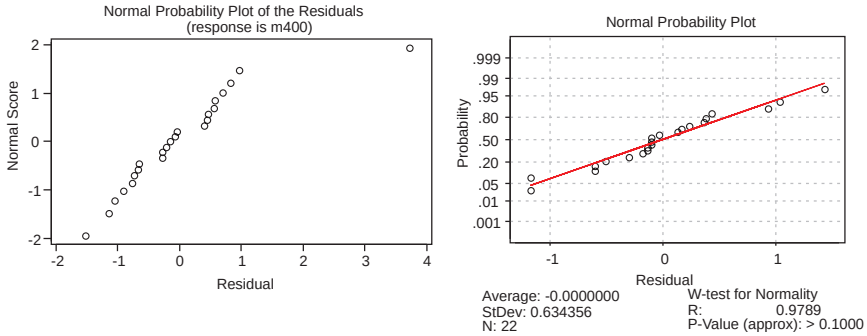
The women's times are slower, but have fallen more steeply than the men's times.

(c) A descending straight line must eventually cut the x-axis, giving a zero time and then it will go negative!

(d) **(i)** The plot follows (below left). The least-squares line is $\widehat{y} = 196 - 0.0769x$ and this is drawn on the plot.



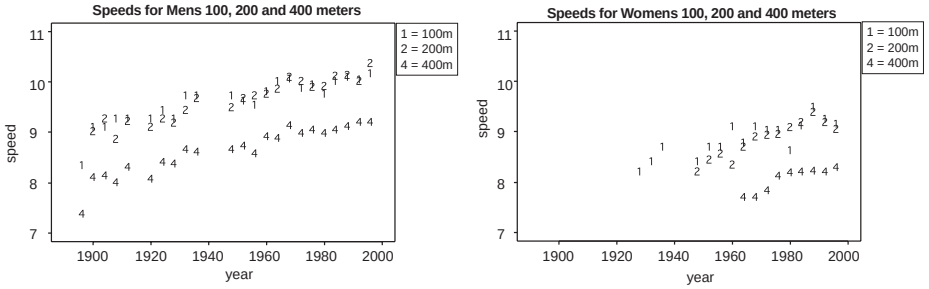
(ii) The residual plot (above right) shows that there are too many positive residuals at the right-hand end, thus suggesting that the linear model is breaking down at this end. Also, there is an outlier at 1900. The Normal probability plot (below left) shows up these problems.



(iii) Without the outlier, the line is now $\widehat{y} = 176 - 0.0666$. The Normal probability plot and the W-test (above right) do not indicate any non-Normality. The 95% prediction interval for the year 2000 is $[41.37, 44.31]$ which, because of its width, is not very informative. It is, however, reflective of the variability in winning times from olympiad to olympiad.

(e) For the reader.

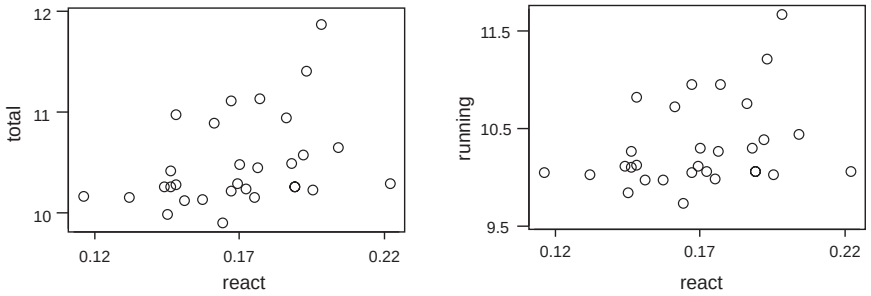
(f) For the men's races, the plots (below left) indicate that the speeds for the 100-m and 200-m are similar. The three plots are reasonably linear with less evidence of a leveling off.



(g) For women's races, the plots (above right) indicate that the speed for the 200-m in recent games has increased relative to that for the 100-m so that they are now similar. There has only been a slight increase in speed for the 400-m in the last few games.

If you are worried about the speed for the longer more tiring race being similar to that for the shorter race, recall that we are talking about average speeds here, not peak speeds. The races include a slower period at the start before the runners reach full pace.

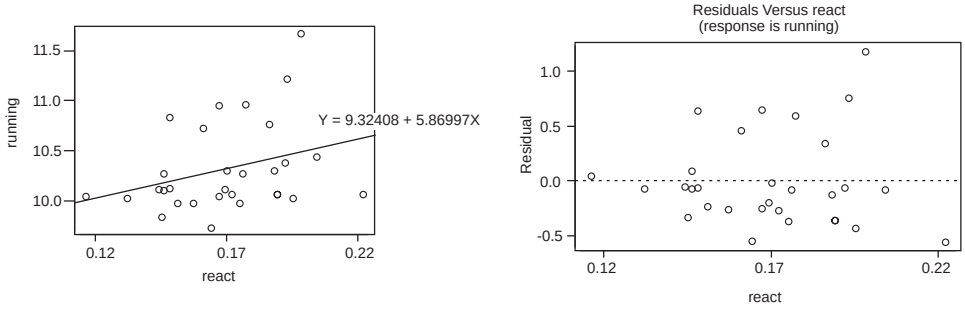
5. (a) The plot is given below left. The TOTAL time seems to increase with the REACTION time. The spread is increasing as REACTION time increases.



(b) The plot (above right) is almost the same as before. [We might perhaps have expected this as reaction time makes up such a small proportion of the total time.]

(c) **(i)** $r = 0.358$. **(ii)** $r = 0.312$. There is little difference between using TOTAL time and RUNNING time.

(d) The fitted line $\widehat{y} = 9.32 + 5.87x$ is given below left.



(e) Using either a test for zero ρ or an (equivalent) test for zero slope, we get P-value $= 0.094$ which indicates only weak evidence against the "no linear relationship" null hypothesis, i.e., we have only weak little evidence from this data of a relationship between REACTION time and RUNNING time.

(f) One interpretation of the scatter plot is that the physical qualities of athletes that lead to fast reactions and those leading to fast sprinting are closely related. It has also been suggested that there is more than one trend superimposed here (maybe three?). This could, for example, be due to the performance differences for the heats, semifinals, and the final, particularly with the better runners.

(g) We expect the residual plot to be fan shaped. It is given above right. We would not trust the least squares analysis, as the spread is not constant. We need to use a model that is more complicated than the simple linear model to analyze these data properly.

7. **(a)** Sometimes you may forget to record the data and fill it in later from memory. Such observations will be less reliable. The tank will not be completely empty each time. This does not matter much. More important is the variation in how close to full the tank is when "completely filled" because we will use the amount put into the tank as our measure of how much gas has been used since the last fill up. There is more than one month involved after some fill-ups, e.g., from the last few days of March through the first few days of April. Unless the car is always filled at the same time of the day, there is the problem of determining the number of days.

(b) DIST is the difference between the current odometer reading and the odometer reading recorded for the last fill up. LITERS is the amount put into the tank this time. KM.LITER is DIST divided by LITERS. DAYS was measured as the number of days since the last recorded date.

(c) A zero means that there is more than one fill-up on some days because of a long journey. There are several 0s and 1s close together in July and August (North American summer holiday period). It looks like the driver was on a travelling holiday at this time.

(d) The plot is perhaps curved. To us it looks more like 2 different lines joining at about month 7 in a "$\widehat{\ }$" shape. There are better consumption rates in the summer months. There is an outlier in month 4 with KM.LTR=11.61. It looks like a mistake has been made. If we halve this value we get a little under 6 which is similar to the other month 4 values (see beginning of the data). This and the large distance recorded (557) suggests that the previous fill up was not recorded.

(e) Such pairs of months might have similar average temperatures. Apart from apparent outliers, there seems to be an increasing trend from left to right, which appears to be related to temperature (and perhaps other weather conditions).

(f) We would want to see if there are any differences within the pairs of months that are linked together, and check overall for outliers. We would also look at a residual plot, as the trend does not look particularly linear. (We could see if an added quadratic term is significant.)

(g) The intercept of the straight line is the mean for the month of January ($x = 0$) and is estimated by 5.889. The 95% confidence interval is $[5.37, 6.41]$ telling us that the true average KM.LTR in January is somewhere between about 5.4 km/L and 6.4 km/L. [We are, however, a little suspicious of the fit of the linear model in January (i.e., at MO.JAN=0).]

(h) The slope of the straight line, estimated by 0.386, gives the increase in consumption as we move one month further away from January. The 95% confidence interval is $[0.243, 0.530]$, which says that the average consumption increases by between 0.24 and 0.53 km per liter for each month further from January.

(i) Yes, because the P-value for testing for zero slope is 0.000. Also because the confidence interval for the slope does not contain the value 0.

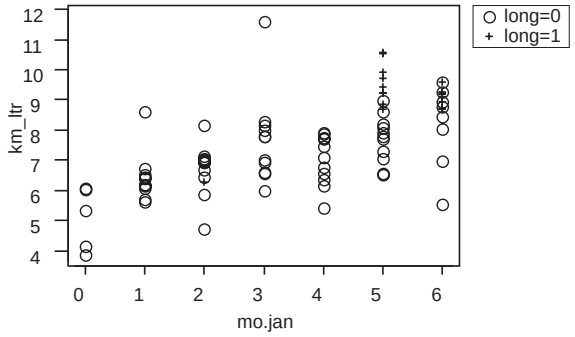
***(j)** For both June and August (MO.JAN=5) a 95% prediction interval is $[5.64, 10.00]$. This leads us to expect that the actual KM.LTR for a trip in June or August (rather than the average over many trips) will be somewhere between 5.6 km/L and 10 km/L.

(k) Yes, provided the standard deviations are not too different. The F-test indicates that the means are clearly not all equal. The null hypothesis for zero slope in the regression model is the same as the equal-means null hypothesis for the one-way ANOVA. However, the alternative hypothesis is more restrictive for the regression model in that it specifies that the means all lie on a line rather than just being not all equal, as in the case of the ANOVA model. [If the means really do lie on a line, the regression method will be more likely to detect that H_0 is false than will one-way ANOVA.]

(l) It means that there is no evidence against the null hypothesis that the quadratic coefficient is 0, i.e., we don't need a quadratic model, or looking at it another way,

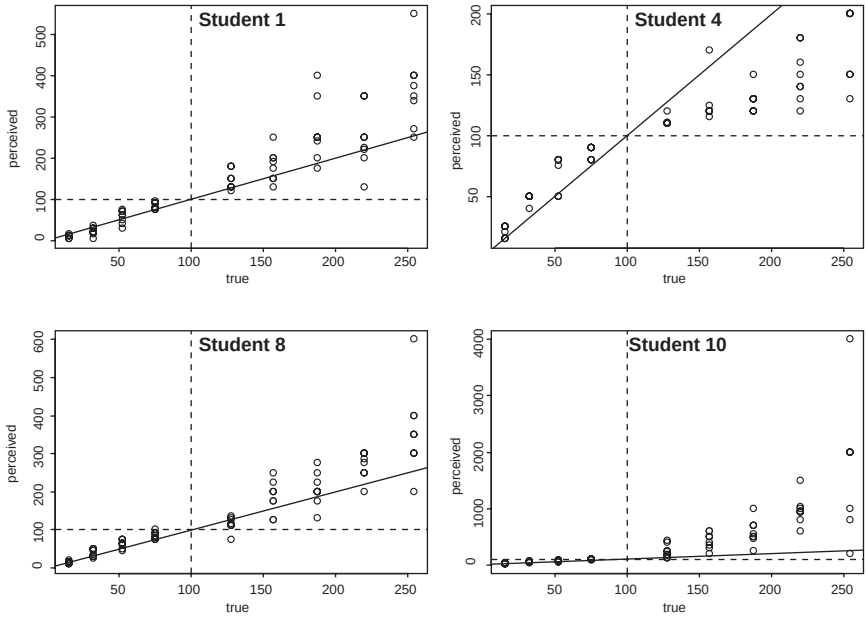
that any curvature we might think we see could just be produced by sampling variation.

(m) The plot with the added points is given below.



Yes, there appears to be a long-trip effect. There seems to be a decrease in consumption for long trips. Our impression of a long-trip effect does, however, seem to be coming entirely from those long trips occurring in August (there were none in June). Long trips in other months do not stand out.

9. (a) The four plots are given below. The line $y = x$ corresponds to the perceived size being the same as the true size.

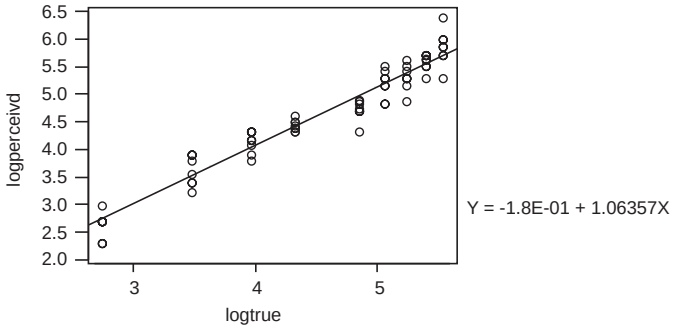


(b) The students vary a great deal in their perception of size. Students 1 and 8 are inclined to overestimate the sizes of the larger circles. Student 4 tended to overestimate the sizes of smaller circles and underestimate the sizes of larger circles. Student 10 strongly overestimated the sizes of larger circles.

(c) Using areas is not a good idea as there is substantial variation in people's perceptions of relative size. These people seem to make big errors in gauging relative

size and different people are making different sorts of errors.

(d) From the following plot the trend looks reasonably linear.



***(e)** Taking logs we get $\log(\textit{perceived size}) = \log a + b\log(\textit{area})$, which is a straight line with slope b.

(f) The fitted line is $\widehat{y} = -0.184 + 1.06x$. The line is superimposed above.

(g) A 95% confidence interval for the slope is $[1.000, 1.127]$. With 95% confidence, the power of area implicitly being used by this student is somewhere between 1.00 and 1.13.

***(h)** A 95% prediction interval for $\log y = \log(\textit{area})$ at $\log 300 = 5.704$ is given by $[5.388, 6.378]$. We can translate this into a prediction interval for $y = \textit{area}$ by taking $[\exp(5.388), \exp(6.378)]$, or $[218.70, 588.75]$.