Scientific Inference: Learning from data

Exercises for the reader

Simon Vaughan

Preface

These exercises are intended to complement the book *Scientific Inference: Learning from data* (Vaughan 2013; CUP). For each chapter I have included a few problems for the reader. Some are designed to test understanding of ideas in the text, others are there to enhance understanding e.g. by working through derivations or extensions not included in the book. Some of the data analysis problems require the use of a computer. The examples of R code given in the book should be enough to help you through them, but in principle they could be solved using any other programming language or good statistical software (e.g. Python, IDL, SPSS, Matlab, Maple, etc.)

Data files and additional material may be downloaded from the book web page at CUP: http://www.cambridge.org/9781107607590

Simon Vaughan, Leicester (September 19, 2013)

Document history

- 16/09/2013 – First complete version

1

Science and statistical data analysis

Exercises

1.1 For each of the five datasets discussed in Appendix B, decide

(a) whether the data are univariate, bivariate or multivariate,
(b) whether each variable is categorical, ordinal, discrete or continuous
(c) (where possible) whether each variable is explanatory or response.

1.2 Explain and give examples for the different types of inference: deduction, induction, abduction.

1.3 Several groups of students independently measure the voltage across a resistor in a circuit. Their data are shown below (in units of Volts).

- Group 1: 10.08 10.01 9.96 9.96 9.99
- Group 2: 9.522 9.526 9.520 9.538 9.516
- Group 3: 14.83 14.47 12.18 13.58 16.46
- Group 4: 8.7 13.0 2.6 8.5 16.0

Given that the experiment was designed so that the voltage was exactly 10 V, Which of the groups recorded the following?

(a) the most accurate measurements?
(b) the most precise measurements?
(c) the least accurate measurements?
(d) the least precise measurements?

1.4 What kind of variables are the following?

(a) the lengths of pieces of string
(b) the voltage across an ideal resistor carrying an ideal current
(c) the number of goals during a football match
(d) the height of a randomly selected person
(e) the number of heads from n flips of a coin

2

Statistical summaries of data

Exercises

2.1 The following numbers comprise a sample of data, the estimates of the metal content of 8 stars from the same star cluster ('cluster 1'). A value of 1 means the metal content is the same as the Sun.

```
1.060 1.020 0.873 0.781 0.929 0.949 0.852 1.070
```

Call these data x_i. Calculate the sample mean $\overline{x}$.

2.2 Calculate the sample variance and standard deviation of the 'cluster 1' data.

2.3 Calculate the standard error of the 'cluster 1' data. Present this as $\overline{x} \pm SE$.

2.4 Use the statistics computed so far to estimate how many stars you would need to observe to obtain a standard error as small as 0.01.

2.5 The following are measurements obtained by a colleague for the purpose of investigating how y varies with x

```
x: 9.42 11.7 13.0 12.7 11.5 10.6 6.23 8.91
y: 13.5 18.9 19.4 19.3 14.7 16.4 11.0 13.4
```

Use these to

(a) produce a scatter plot
(b) compute the sample means and variances for x and y
(c) compute the sample covariance s_{xy}
(d) compute the sample correlation coefficient r

2.6 You find one more data point to add to the table recorded by your colleague:

```
x: 4.0
y: 20.1
```

Recalculate the statistics of exercise 2.5 including this new data point. Your colleague previously identified this data point as an outlier. Is this reasonable?

2.7 Use the Hipparcos data (section B.4) to produce the following plots

(a) histograms for the variables `V.abs` and `BV`. Use the `hist()` function.
(b) smooth density curves for the variables `V.abs` and `BV`. Use the `density()` function.
(c) a matrix of scatter plots for the variables `dist`, `V.abs` and `BV`

2.8 Use the pion scattering data (section B.5) to produce the following plots

(a) cross-section as a function of energy (e.g. top panel of Fig 6.5, see R.Box B.11)
(b) repeat the plot but colour-code the data according to the target length (e.g. 10 cm target data blue, and 20 cm target data red)

Table 2.1 *Pulsar mass data show in figure 2.10. Masses are in unit of the Solar mass ($M_\odot$), and the error bars are shown in columns 3 and 4 (some error bars are asymmetric about the estimated value). The data are from Charles & Coe (2006, in* Compact stellar X-ray sources*, Eds: W. Lewin & M. van der Klis. Cambridge Astrophysics Series, No. 39, CUP). (A plain text data file is available from the book web page at CUP.)*

Name	Mass	σ_-	σ_+
4U 1700-37	2.44	0.27	0.27
Vela X-1	1.86	0.32	0.32
Her X-1	1.47	0.23	0.37
4U 1538-52	1.06	0.41	0.34
Cen X-3	1.21	0.21	0.21
LMC X-1	1.47	0.44	0.39
SMC X-1	1.17	0.36	0.32
J0045-7319	1.58	0.34	0.34
B1855+09	1.41	0.10	0.10
B1802-07	1.26	0.08	0.17
J1713+0747	1.34	0.20	0.20
J1012+5307	1.7	0.5	0.5
J1518+4904 (P)	1.56	0.13	0.44
J1518+4904 (C)	1.05	0.45	0.11
B2303+46 (P)	1.30	0.13	0.46
B2303+46 (C)	1.34	0.47	0.13
B2127+11 (P)	1.349	0.040	0.040
B2127+11 (C)	1.363	0.040	0.040
B1534+12 (P)	1.339	0.003	0.003
B1534+12 (C)	1.339	0.003	0.003
B1913+16 (P)	1.4411	0.00035	0.00035
B1913+16 (C)	1.3874	0.00035	0.00035

(c) Overlay a curve for the Breit-Wigner model (R.Box B.12) using various different parameter settings

2.9 Table 2.9 shows the pulsar mass data plotted in figure 2.10. Use these data to produce a dot chart (see R.Box 2.17).

3

Simple statistical inferences

Exercises

3.1 Use the t-statistic to assess whether the 'cluster 1' metalicity data (exercise 2.1) are consistent with Solar metalicity, i.e. $\mu_x = 1$.

3.2 The following numbers comprise a sample of data, estimates of the metal content of 8 stars from a different star cluster, 'cluster 2'. (Again, a value of 1 means the metal content is the same as the Sun.)

```
1.01 1.02 1.04 1.02 1.08 1.04 1.18 1.12
```

Call these data y_i. Use the two sample t-statistic to assess whether the 'cluster 1' and 'cluster 2' metalicity data are consistent with each other, i.e. same mean.

3.3 For each of the following model functions $y = f(x)$ identify the parameters of each model, and identify which models are linear in their parameters

(a) $f(x) = \alpha \log(x)$
(b) $f(x) = \alpha \log(x) + \beta$
(c) $f(x) = \alpha \log(\beta x)$
(d) $f(x) = A\mathrm{e}^{x/B}$
(e) $f(x) = Kx^S$
(f) $f(x) = A \sin(x) + B \cos(x)$
(g) $f(x) = C + Bx + Ax^2$
(h) $f(x) = (A^2/4)/(1 + x^2)$

3.4 Derive equation 3.11 by finding the minimum of SSE (equation 3.9). Then use substitution (equations 3.13) to derive the regression coefficients (equation 3.14)

3.5 Use the definitions of sample covariances and variances (chapter 2) to derive (equation 3.17) $b = s_{xy}/s_x^2$ for the slope of the linear regression model.

3.6 Derive equation 3.23 relating the correlation coefficient r to the SSM and SST terms of the linear regression model. (Note: r^2 is sometimes called the *coefficient of determination*.)

4

Probability theory

Exercises

4.1 Using the set notation given in chapter 4 write down the formulae for the events shaded in Figure 4.1.

4.2 What is the probability of the following events?

(a) rolling a six with one roll of a six-sided die? (assuming equal probabilities for each face)
(b) rolling any number other than six with one roll of a six-sided die?
(c) rolling a number less than 3 with one roll of a six-sided die?
(d) tossing heads, heads, then tails, from three flips of a fair coin?
(e) two heads and one tails (in any order), from three flips of a fair coin?
(f) at least two heads, from three flips of a fair coin?

4.3 Prove $\Pr(A) = odds(A)/[1 + odds(A)]$ using the definition of odds (Box 4.2).

4.4 Derive equation 4.42 using equations 4.39 and 4.41.

4.5 After a search of DNA records for 10^5 people a suspect has been identified for a recent crime. The probability of a DNA match from a random member of the population is estimated to be 10^{-6}. The DNA match is very reliable, so the probability of a match if the suspect is guilty can be assumed to be 1.

If G is the event 'suspect is guilty' (G^C the event 'suspect is not guilty'), then we can write $\Pr(M|G) = 1$ is the probability of a DNA match from the guilty person, while

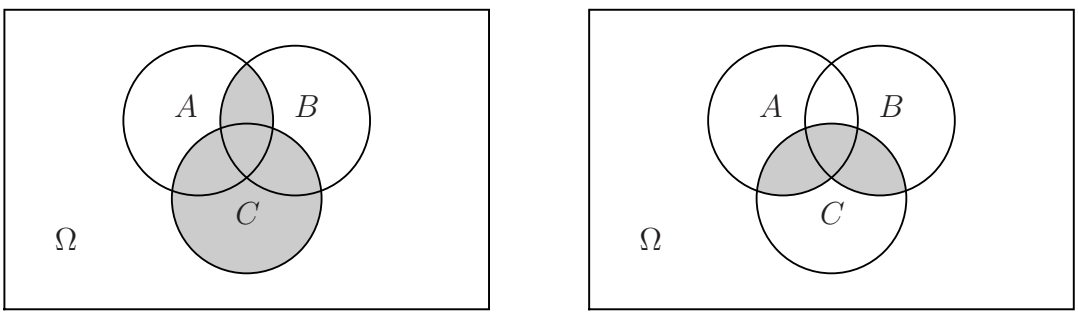

Figure 4.1 Venn diagrams for exercise 4.1.

$\Pr(M|G^C) = 10^{-6}$. Use Bayes theorem to deduce the probability of guilt given the DNA match $\Pr(G|M)$. (*Hint*: you will need to make a reasonable choice for $\Pr(G)$.)
Write out the assumptions involved in an analysis like this.

4.6 Imagine tossing a coin until it lands on heads, H. The number of tosses required until H occurs is a random variable X. The coin may not be perfectly fair, so we say the probability of H with each toss is θ (which may, or may not, be 0.5).
Write down a function for the probability mass function (pmf) for X, i.e. $p_X(x)$ the probability that H first occurs after x tosses, with $x = 1, 2, 3, \ldots$.

4.7 Using the pmf of exercise 4.6, derive the cumulative mass function (cmf) for X, i.e. $F(x) = \Pr(X \leq x)$.
Hint: it may be useful to know the following geometric series (for $r \neq 1$):

$$\sum_{i=1}^{n} ar^{i-1} = \sum_{i=0}^{n-1} ar^i = a\frac{1-r^n}{1-r}$$

4.8 Two friends happen to be in the same part of town. They each take a lunch break starting at 1.00 and finishing at 2.00. During this time they each spend exactly 15 (consecutive) minutes in the same a coffee shop. Calculate the probability of the two friends meeting in the coffee shop (i.e. being there at the same time).
Hint: assume the arrival time of each person is random and uniform ($U(a, b)$ for a suitable choice of a and b).

5

Random variables

Exercises

5.1 Use the definition of the expectation and variance for a continuous variable (equations 5.2 and 5.5) to prove $\mathrm{E}[aX + b] = a\mathrm{E}[X] + b$ and $\mathrm{V}[aX + b] = a^2\mathrm{V}[X]$

5.2 If we have a population of values $x = \{1, 2, 2, 3, 3, 3\}$ compute the following:

(a) the expectation (aka. population mean) of x.

(b) the population variance and standard deviation of x

5.3 Prove equation 5.11 (*Hint*: follow the argument of Box 5.2.)

5.4 Prove equation 5.19. (Note, this is can be a rather long derivation.)

5.5 Prove eqn 5.23, namely that

$$\mathrm{V}[\bar{X}] = \frac{\sigma^2}{n}$$

for independent random variables with equal variance (σ^2), using equation 5.19.

5.6 Prove that

$$x\binom{n}{x} = n\binom{n-1}{x-1}$$

using the definition of the binomial coefficient (Box 5.3, Appendix C).

5.7 Prove that if X is a random variable with a binomial pmf then $\mathrm{E}[X] = n\theta$.
Hint: The binomial pmf is

$$p(x|n, \theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

so its expectation is

$$E[X] = \sum_{i=0}^{n} x\binom{n}{x}\theta^x(1-\theta)^{n-x} = \sum_{i=1}^{n} x\binom{n}{x}\theta^x(1-\theta)^{n-x}$$

since the $x = 0$ term adds nothing. Now you should be able to complete the proof, making use of the result of exercise 5.6.

5.8 If we have a continuous variable X whose probability density function is $p(x) = \frac{1}{2}\mathrm{e}^{-x/2}$ (for $x \geq 0$)

(a) calculate the cumulative probability distribution $F(x)$

(b) Prove that the pdf integrates to 1 over the full range of x values.

(c) calculate the expectation $\mathrm{E}[X]$
(d) calculate the median, $x_{0.5}$
(e) sketch the pdf $p(x)$ and mark the mean and median values

(Note: this is a special case of the *chi-square* distribution, with $\nu = 2$ degrees of freedom.)

5.9 Using the binomial distribution compute the following probabilities:

(a) getting exactly 0, 1, 2, 3 and 4 heads in four tosses of a fair coin?
(b) getting at least two heads in four tosses of a fair coin?
(c) What is the probability that a family with three children has only same sex children? (Assuming probability of male birth is 0.5.)

5.10 Assuming cats are distributed randomly through town, if you expect to see three cats on your way home ($\lambda = 3$), use the Poisson distribution to compute the probability of observing:

(a) exactly four cats on a given day
(b) fewer than four cats
(c) at least four cats

5.11 Prove the variance of a Poisson variable with mean λ is $\mathrm{V}[X] = \lambda$.
Hint: Use the definition of variance $\mathrm{V}[X] = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$ (equation 5.6), and be prepared to use the Taylor series expansion of e^{λ}, and to use substitutions to simplify the formulae.

5.12 What is the distribution of the sum of $n = 20$ uniform random numbers? Write a routine to generate a sample of random numbers X_i, each one

$$X_i = \sum_{j=1}^{n} U_i - n/2.$$

5.13 The standard Normal distribution has the pdf given by eqn 5.31. Its expectation is given by

$$\mathrm{E}[z] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} z\mathrm{e}^{-z^2/2}\mathrm{d}z.$$

Evaluate this integral to find the expectation.

5.14 The integral of the Normal pdf cannot be evaluated using elementary methods. But using a transformation of variables it is possible to prove the standard Normal pdf integrates to 1 as expected. The integral we seek is

$$I = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \mathrm{e}^{-z^2/2}\mathrm{d}z$$

but instead we chose to solve

$$\begin{aligned} I^2 &= \frac{1}{2\pi} \left(\int_{-\infty}^{+\infty} \mathrm{e}^{-z^2/2}\mathrm{d}z \right) \left(\int_{-\infty}^{+\infty} \mathrm{e}^{-z^2/2}\mathrm{d}z \right) \\ &= \frac{1}{2\pi} \left(\int_{-\infty}^{+\infty} \mathrm{e}^{-z^2/2}\mathrm{d}z \right) \left(\int_{-\infty}^{+\infty} \mathrm{e}^{-y^2/2}\mathrm{d}y \right) \\ &= \frac{1}{2\pi} \left(\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \mathrm{e}^{-(z^2+y^2)/2}\mathrm{d}z\mathrm{d}y \right) \end{aligned}$$

Now make the transformation from (z, y) to polar coordinates (r, θ). Using $r^2 = z^2 + y^2$ we have

$$I^2 = \frac{1}{2\pi} \left(\int_0^{2\pi} \int_0^{+\infty} \mathrm{e}^{-r^2/2} r\mathrm{d}r\mathrm{d}\theta \right)$$

Evaluate this integral and use it to show the standard Normal pdf integrates to 1.

5.15 Using the definition of the chi-square pdf (equation 5.32)

(a) Write down the pdf for $\nu = 4$ degrees of freedom
(b) Find the mode of this pdf (*Hint*: find its maximum)
(c) Integrate the pdf to find the cdf $F(x)$
(d) Calculate the expectation $\mathrm{E}[X]$ (use equation 5.2)

5.16 Using the definition of Students' t distribution (eqn 5.35)

(a) Write down the pdf $p_t(x)$ for the t_1 distribution (i.e. with $\nu = 1$ degrees of freedom).
(b) Derive the corresponding cdf for t_1, $F_t(x)$.

(Note: this distribution is a special case of the *Cauchy distribution* and is closely related to the Breit-Wigner function used in section B.5.)

5.17 Derive the expectation and variance for the uniform distribution $U(a, b)$ of section 5.3.4.

5.18 Use equation 5.48 to compute the approximate variance of a new variable $z = x \log y$ (given estimates of x, y and their variances σ_x^2 and σ_y^2). Assume $y/\sigma_y \gg 1$ (so that $y = 0$ does not cause problems).

5.19 If $Y = \alpha + \beta X$ (and $\beta > 0$) prove that $\rho(X, Y) = 1$ using the definition of the correlation coefficient (eqn 5.14).

5.20 Given two random variables X and Y with zero means ($\mu_X = 0$ and $\mu_Y = 0$) show that

$$\mathrm{Cov}(X, Y) = \mathrm{E}[XY]$$

5.21 Prove the Cauchy-Schwarz inequality (eqn. 5.13), which can be written (using the result of exercise 5.20) as

$$|\mathrm{E}[XY]|^2 \leq \mathrm{E}[X^2]\mathrm{E}[Y^2]$$

Hint: one way to proceed is to use your knowledge of quadratic functions. It must be true that

$$\mathrm{E}[(\alpha X + Y)^2] \geq 0$$

The left side can be expanded as a quadratic in α, e.g. $a\alpha^2 + b\alpha + c$. The equality is true only for $Y = -\alpha X$, in which case the quadratic has one real root. In all other cases the quadratic has imaginary roots. This means the discriminant of the quadratic (the $b^2 - 4ac$ term) must zero or negative, i.e. $b^2 \leq 4ac$. This fact should be enough to complete the proof.

5.22 Use the Poisson distribution to model the following situation. Imagine running a hospital ward with $N_{bed} = 2$ beds. Per day each bed costs £100 to run, but each patient treated generates £500 income. Once the number of patients per day reaches N_{bed}, each additional patient incurs a fine of £100 (to cover transport, overtime work etc.) Assume the patients usually arrive at a rate of $\lambda = 2$ per day.

Compute the expected net gain (or loss) per day. Assume patients are discharged each night, so that day begins with N_{bed} empty beds. This problem can be solve analytically (e.g. using the first few terms of the Poisson distribution), or using computer simulation of Poisson random numbers.

Now repeat the exercise for different values of N_{bed} (e.g. 1, 3, 4). Which provides the most profit (or the smallest loss)?

5.23 If $X \sim U(0, 1)$ is a uniform random variable, i.e. $p_X = 1$ for $0 < x \leq 1$, find the pdf and cdf of the transformed variable $Y = -\log(X)$ (using the 'change of variables' rules of section 5.4). Note: we define X so as to avoid the $x = 0$ point. This transformation will be useful in Chapter 8.

6

Estimation and maximum likelihood

Exercises

6.1 Derive the maximum likelihood estimate (MLE) for the parameter θ for binomially distributed data x_i ($i = 1, 2, \ldots, N$), using the binomial distribution Binom(n, θ). *Hint*: follow the reasoning used for the Poisson distribution in section 6.2.

6.2 Derive the formula for the weighted mean of data y_i, if each data point has variance σ_i^2:

$$\hat{\theta} = \frac{\sum_{i=1}^{N} y_i/\sigma_i^2}{\sum_{i=1}^{N} 1/\sigma_i^2}$$

using the minimum X^2 criterion (eqn. 6.12).
Hint: assume a constant model $\mu_i = \theta$, insert this into the formula for the X^2 statistic, find the value of θ that minimises X^2.

6.3 Prove the sample variance s^2 (eqn 2.3) is unbiased for a random sample of data from an infinite population with finite mean μ and variance σ^2, i.e. that $\mathrm{E}[s^2] = \sigma^2$.
Hint: apply the expectation operator to s^2 and use $\mathrm{E}[(x_i - \mu)^2] = \sigma^2$ and $\mathrm{V}[\bar{x}] = \mathrm{E}[(\bar{x} - \mu)^2] = \sigma^2/n$ (exercise 5.5).

6.4 Find the relationship between the sample variance s^2 and the least-squares statistic X^2 when the data x_i are fitted with a constant model $\mu_i = \bar{x}$ (assume constant variance σ^2).

6.5 Fit the pion scattering data with the Breit-Wigner model to obtain the MLEs for parameters N, Γ_0 and E_0. (*Hint*: see R.Box 6.11)

6.6 Table 6.6 shows data on the timing of the binary pulsar PSR 1913 + 16. The data are from Taylor & Weisberg (1982; *Astrophysical Journal*, v235, pp908–920). They show the 'orbit phase residuals' from precise timing of the orbit of the system. If the orbit was (a) constant (at 7.76 hours) the residuals should be constant with time. If the orbit was (b) constant but its period was incorrectly determined the residuals should grow linearly with time. If the period of the system is constantly changing (c) there should be a parabolic change in the residual with time. A constantly increasing period (so quadratically decreasing phase residual) is what we would expect if gravitational waves are radiating energy from the system.
Use weighted least squares (or maximum likelihood assuming Normally distributed data) to fit the following models to the data:

(a) constant: $y = \alpha$
(b) linear: $y = \alpha + \beta x$

Table 6.1 *Data recorded by Taylor & Weisberg (1982). The observation times are shown in years, the time shift and its error are in units of seconds. (A plain text data file is available from the book web page at CUP.)*

observation (#)	time (year) (x)	time shift (y)	error (σ_y)
2	1974.93	-0.11412003	0.06048
3	1976.13	0.02017998	0.20736
4	1976.93	-0.28405002	0.05184
5	1977.58	-0.44069999	0.04320
6	1977.96	-0.46414001	0.02592
7	1978.23	-0.60691001	0.02592
8	1978.42	-0.56348002	0.06048
9	1978.82	-0.76762003	0.04320
10	1979.31	-0.96714002	0.02592
11	1980.10	-1.22854003	0.02592
12	1980.59	-1.45820999	0.07776
13	1980.59	-1.44239998	0.02592
14	1981.14	-1.75293997	0.01728

(c) quadratic: $y = \alpha + \beta x + \gamma x^2$

By plotting the data and models, or data-model residuals, decide which model or models give a reasonable match to the data.

7

Significance tests and confidence intervals

Exercises

7.1 Given the MLE for the θ parameter of a binomial distribution (exercise 6.1)

(a) derive the approximate variance of this MLE (following the reasoning in section 7.6.1).
(b) Express the formulae for the MLE in the form 'estimate±error'.
(c) Compare this to the variance of the sum of binomial variables (section 5.2.1).

7.2 Use the results from exercise 7.1 to quantify the uncertainty in the results of an opinion poll of $n = 1000$ people, where 53.7 per cent of those asked voted 'yes' (and 46.3 per cent voted 'no').

7.3 Apply a chi-square goodness-of-fit test to the model fit from exercise 6.5. Does the Breit-Wigner model provide a reasonable fit to the data? (*Hint*: see R.Box 7.1)

7.4 Express and interpret the results of the t-tests of exercises 3.1 and 3.2 using p-values (*Hint*: try the `t.test()` function.)

7.5 Apply a goodness-of-fit test to the model fits from exercise 6.6. Quantify which model(s) provide a reasonable fit to the data. (*Hint*: see R.Box 7.1)

7.6 Express the results of the above goodness-of-fit tests (the p-values) in terms of 'sigmas' (see R.Box 7.6).

7.7 Compute confidence intervals for the three parameters of the Breit-Wigner model from 6.5.

7.8 Compute confidence intervals for the parameters of the best fitting model of exercise 6.6

7.9 Use simple linear regression (R.Box 3.6) to fit the data from 2.5. Then

(a) use the SSE to estimate the 'errors' on the data (eqn 3.22)
(b) perform a chi-square goodness-of-fit test using these errors
(c) express the result in both p-value and 'sigma' terms
(d) explain why the result is not surprising
(e) recompute the goodness-of-fit including the outlier of exercise 2.6 (using the same 'error' estimate)

7.10 The data in table 7.11 are measurements of the 'flux' (brightness) of two stars (in arbitrary units) as a function of time (in days). Plot the data for 'star 1' and use a goodness-of-fit test and diagnostic plots to assess whether the star is variable or not.

7.11 The data for 'star 2' appear to show an oscillatory pattern, so we consider a model of the form:

$$y(t) = A\sin(2\pi Bt + C) + D$$

Table 7.1 *Star data. Time is in unit of days and the flux (and its error) are in arbitrary units. (A plain text data file is available from the book web page at CUP.)*

time	flux star 1	error	flux star 2	error
1.06	18.49	1.12	19.61	1.06
2.01	16.20	0.92	18.53	1.08
2.85	16.07	1.00	15.87	1.11
4.13	16.64	0.89	12.72	1.04
4.95	16.69	1.06	13.01	0.85
6.06	18.35	0.94	12.42	1.01
6.93	16.34	0.95	16.81	1.17
7.94	15.17	1.06	20.20	1.09
8.92	19.86	0.93	20.72	0.93
10.04	17.69	0.82	18.35	0.96
11.06	15.37	0.97	15.94	0.87
12.13	16.09	1.19	11.80	0.94
12.89	17.57	0.86	11.83	0.91
13.92	18.23	1.02	13.74	0.81
15.00	18.08	1.13	16.25	1.03
15.87	16.64	0.91	19.16	0.95
17.09	15.40	1.04	20.40	0.92
18.04	17.55	0.99	16.44	0.90
18.95	16.19	0.96	14.72	1.13
19.95	16.51	0.84	11.62	0.88

where y is the flux, t is the time. A, B, C and D are the various parameters of the model representing the amplitude, the frequency and the phase of the oscillation, and the mean level, respectively.

(a) Perform a goodness-of-fit test using a constant model ($y(t) = \mu$) to support the need for a different model
(b) Fit the sinusoid model to the data (assuming Normal errors)
(c) Use a goodness-of-fit test and diagnostic plots to test to assess whether the model is reasonable.
(d) Evaluate the confidence intervals on all four parameters

8

Monte Carlo methods

Exercises

8.1 Write a routine to simulate random numbers with a Normal distribution using numbers with a uniform distribution. The routine of exercise 5.12 is a very slow and approximate way to do this. A much better way is known as the Box-Muller method[1] and involves generating two random numbers, U_1 and U_2, from a $U(0,1)$ distribution, and applying the following transformations

$$X_1 = \sqrt{-2\log U_1}\cos(2\pi U_2)$$

$$X_2 = \sqrt{-2\log U_1}\sin(2\pi U_2)$$

The resulting numbers X_1 and X_2 should be drawn from a $N(0,1)$ distribution.

8.2 The following three exercises demonstrate the validity of the Box-Muller method to produce Normal random variables. Find the inverse transformations used in exercise 8.1 i.e. U_1 and U_2 as a functions of X_1 and X_2.

8.3 Find the Jacobian determinant of the inverse transformations from exercise 8.2

$$J = \left|\frac{\partial(U_1,U_2)}{\partial(X_1,X_2)}\right| = \begin{vmatrix} \frac{\partial U_1}{\partial X_1} & \frac{\partial U_1}{\partial X_2} \\ \frac{\partial U_2}{\partial X_1} & \frac{\partial U_2}{\partial X_2} \end{vmatrix} = \frac{\partial U_1}{\partial X_1}\frac{\partial U_2}{\partial X_2} - \frac{\partial U_1}{\partial X_2}\frac{\partial U_2}{\partial X_1}$$

8.4 Use the results from exercises 8.2 and 8.3 to write down the joint distribution $p_{X_1,X_2}(x_1,x_2)$ based on the joint distribution $p_{U_1,U_2}(u_1,u_2)$. Use the 'change of variables' rule (eqn 5.39) but in multi-variate form, making use of the Jacobian:

$$p(x_1,x_2) = p(u_1,u_2)\left|\frac{\partial(U_1,U_2)}{\partial(X_1,X_2)}\right|$$

Use the resulting joint distribution to show that X_1 and X_1 are independent, standard Normal variables.

8.5 Use the bootstrap method (R.Box 8.9) to compute 68.3 per cent confidence intervals on the difference between the means of the two samples from exercise 3.2. Compare this to the standard error on the difference of two means (see eqn 3.3).

[1] From Box & Muller (1958; *The Annals of Mathematical Statistics*, v29 (2), pp. 610-611)

8.6 Use the bootstrap method to compute 68.3 per cent confidence intervals on the coefficients (intercept and slope) of the linear regression from 7.9. How do these compare to the intervals returned by the standard linear regression theory (e.g. with the `lm()` function)?

8.7 Imagine you are planning to re-observe the star from exercise 7.11 and wish to estimate the amplitude of the oscillation with ≈ 1 per cent accuracy. Use Monte Carlo simulations to demonstrate how small the 'errors' of the fluxes need to be in order to achieve this accuracy on the amplitude parameter. Assume you can obtain $n = 20$ observations separated by exactly 1 night (i.e. $t = 1, 2, \ldots, n$)

Hint: use the result of exercise 7.11 as your best guess for the true oscillation. Simulate new data based on this model but with randomised errors (Normal distribution with standard deviation σ), and fit to determine the accuracy with which the parameters can be recovered. Then vary the magnitude of the flux errors systematically until the desired accuracy is obtained.

8.8 Now imagine you cannot decrease the error (σ) in exercise 8.7 but you can increase the number of observing nights (n). How large does n need to be in order to recover the amplitude of the oscillation with ≈ 1 per cent accuracy?