Solutions to Data Analysis Class Exercises

Have you attempted the exercises? If not, then go and do them first before looking here!

These model answers were prepared by Dr Chris Ferro to help show you how one should present statistical work correctly. Statistical work should be precise and concise. The data and methodology should be clearly explained so that in principle a reader could easily go away and reproduce the analysis (repeatability is essential for good science!). Statistical models and tests and their underlying assumptions should be clearly explained and you shouldn't be afraid to use proper mathematical symbols to write out statistical methods and models. The results should be presented clearly in well-labeled graphs (with meaningful captions and axis labels) and tables (with meaningful captions and column headings). The results should be interpreted in a careful way taking into account all sources of possible uncertainty such as sampling uncertainty caused by having only a finite sample of data. The evidence should be weighed up objectively rather than torturing the data with statistical methods to prove a point.

Exercise 1

Two histograms of daily mean temperatures, recorded in degrees Celsius (°C) at the Department field site during 2003, are shown in Figures 1 and 2. Figure 1 suggests that the temperatures have a bi-modal distribution, with two peaks around 9 and 16°C. This feature disappears when the interval width is increased in Figure 2. The second peak in Figure 1 is related to a record-breaking heat wave at the start of August 2003. The typical distribution of our field-site temperatures is better represented by the uni-modal distribution in Figure 2. Neither figure is wrong, but the data that are selected and the way in which they are plotted can strongly influence interpretation, so be careful!

The histograms show that the distribution of daily mean temperatures is roughly symmetric and there are no obvious outliers. The boxplot provides no additional insight for these data and so is omitted. Summary statistics are displayed in Table 1. Note the number of significant figures: since the quartiles and ranges correspond to particular observations they should have the same precision (one-tenth of a degree) as the data; the mean and standard deviation are given two decimal places but any more is unnecessary. For symmetrically distributed data, the mean and standard deviation provide a reasonable summary. For these data, the mean temperature is 10.96°C, with quite a wide spread given by the standard deviation of 5.84°C.

Mean	Std Dev	1st Quartile	Median	3rd Quartile	IQR	Range
10.96	5.84	6.8	10.4	15.6	5.2	26.8

Table 1. Statistics of field-site daily mean temperatures (°C) in 2003.

There are six missing values in the data record for 2003 and it is important to discover why this is so. For example, if missing values are recorded only when temperatures are very high (due perhaps to a fault in the thermometers) then our previous analysis of the data will be biased. Inspection of the data reveals that the missing values occur in three pairs that correspond to weekends, apparently randomly located through the year. It is unlikely, therefore, that the missing values are linked to the meteorological conditions, but are probably caused by technical faults that were not rectified until after the weekend.

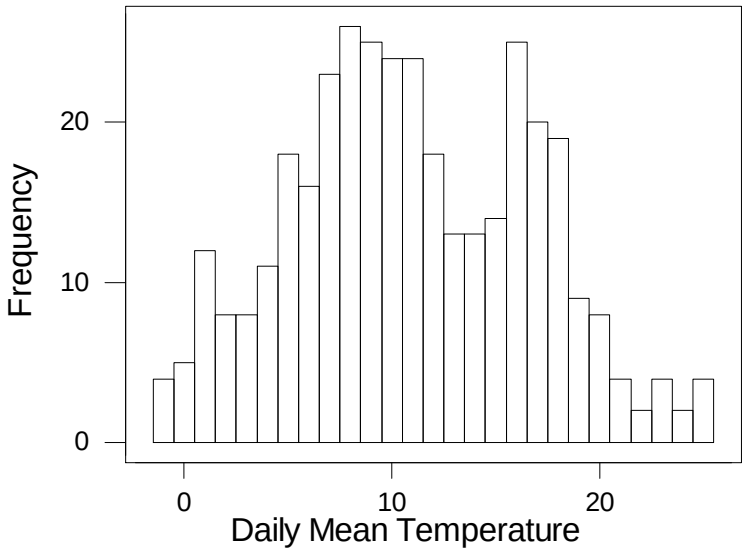


Figure 1. Histogram of field-site daily mean temperatures (°C) recorded in 2003.

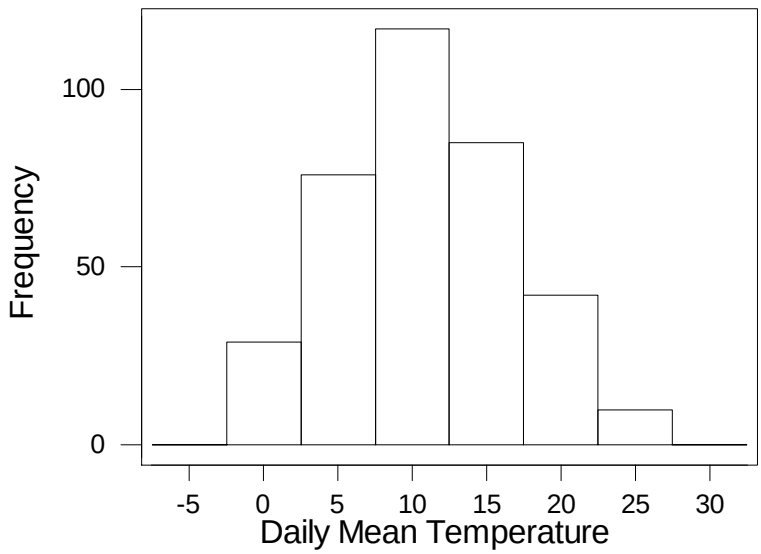


Figure 2. Histogram of field-site daily mean temperatures (°C) recorded in 2003.

Exercise 2

1. Temperatures

Note that this section is not written as a report, but only to point out some features of the data that you should have addressed. The wind-speed example in the next section gives you an idea of a report style.

The summary statistics for daily maximum temperatures are given in Table 1 and should immediately alert you to an error in the data: there is one value equal to 270°C. This occurs on day 167, immediately after two missing observations, so may be related to an instrumentation failure. Alternatively, the value may just have been stored incorrectly: perhaps the temperature has been recorded in degrees Kelvin, or maybe the decimal point is in the wrong position. The former is unlikely given the temperatures recorded on the preceding and succeeding days. In fact, I inserted the error myself (merely for demonstration purposes of course!) and the latter explanation is correct: it should be 27.0°C. The statistics after this correction has been made are shown in Table 2. Note that the mean, standard deviation and range change a lot, but the quartiles are more robust and are unaffected.

Mean	Std Dev	1st Quartile	Median	3rd Quartile	IQR	Range
16.28	15.12	10.3	15.2	20.5	10.2	272.6

Table 1. Statistics of daily maximum temperatures (°C) in 2003.

Mean	Std Dev	1st Quartile	Median	3rd Quartile	IQR	Range
15.62	6.93	10.3	15.2	20.5	10.2	36.1

Table 2. Statistics of corrected daily maximum temperatures (°C) in 2003.

Did you make any other checks on the data? One simple idea is to check for each day that the minimum temperature does not exceed the maximum temperature.

The summary statistics for daily minimum temperatures are shown in Table 3, and histograms of daily maximum and minimum temperatures, using the same axes in both plots, are given in Figures 1 and 2. Both distributions are roughly symmetric, with the mean of the minimum temperatures about 9°C lower than the mean of the maximum temperatures. The minimum temperatures also have a smaller spread (the standard deviation is about 1.5°C lower) than the maximum temperatures. Can you explain this phenomenon scientifically?

Mean	Std Dev	1st Quartile	Median	3rd Quartile	IQR	Range
6.55	5.42	2.3	6.6	10.7	8.4	26.5

Table 3. Statistics of daily minimum temperatures (°C) in 2003.

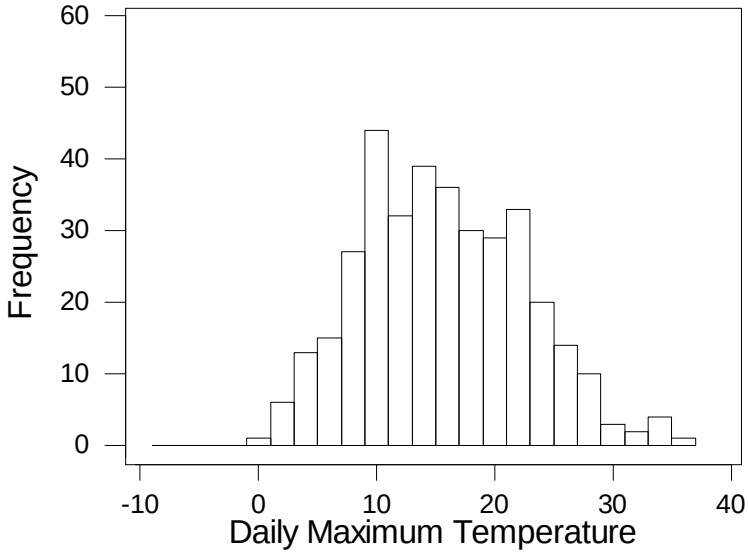


Figure 1. Histogram of daily maximum temperatures (°C) recorded in 2003.

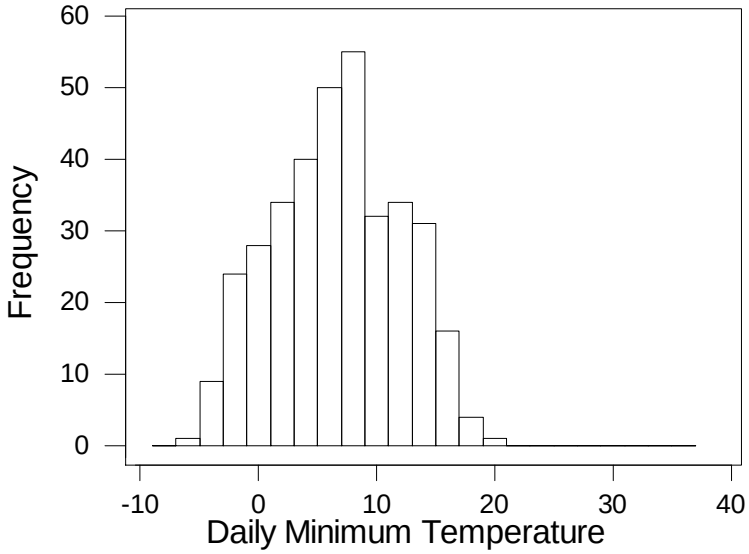


Figure 2. Histogram of daily minimum temperatures (°C) recorded in 2003.

2. Wind speeds

The 3-metre wind speeds recorded with a cup anemometer in metres per second (m/s) at the University of Reading meteorology station during 2003 will be examined. Two wind-speed measurements are available: daily means and maxima of five-minute averages. The study aims to compare the probability distributions of the wind-speeds and identify any unusual features. Observations are missing in both records for six days of the year and are omitted from the analysis. This is assumed to cause no bias in the results for reasons given in Exercise 1.

Mean wind speed equals zero for every day in the period 16 July – 29 October. It appears that the anemometer may have been broken. The wind speeds in this period, including the similarly small daily maxima, are omitted from the subsequent analysis for two reasons: the anemometer readings are probably unreliable, and the presence of a large number of zero or near-zero values can hide patterns in the remainder of the distribution.

Histograms of the remaining daily mean and maximum wind speeds are shown in Figures 3 and 4, and summary statistics are given in Table 4. Both distributions are slightly positively skewed. The mean of the daily mean wind speeds is approximately 7m/s lower that of the daily maxima, and the spread is also smaller. There is one outlying daily maximum observation of 30.1m/s that occurred on 30 October, immediately after the period of zero mean wind speeds. The mean wind speed on this day was 1.0m/s, so the maximum either corresponds to a short-lived gust or is a measurement error. The summary statistics are only slightly affected if this outlier is removed from the analysis, however, and the qualitative conclusions are unaffected.

	Mean	Std Dev	1st Quartile	Median	3rd Quartile
Mean	2.38	1.14	1.6	2.3	3.2
Maximum	9.69	3.61	7.6	9.2	11.5

Table 4. Statistics of daily mean and maximum wind speeds (m/s) in 2003.

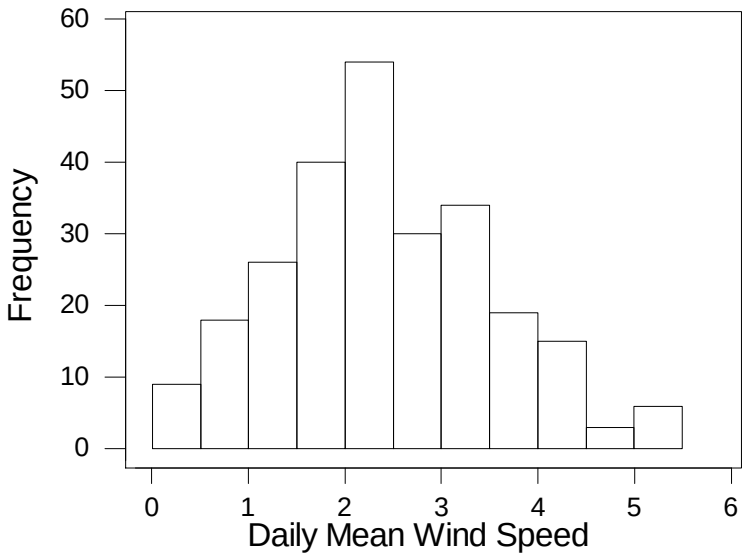


Figure 3. Histogram of daily mean wind speeds (m/s) recorded in 2003.

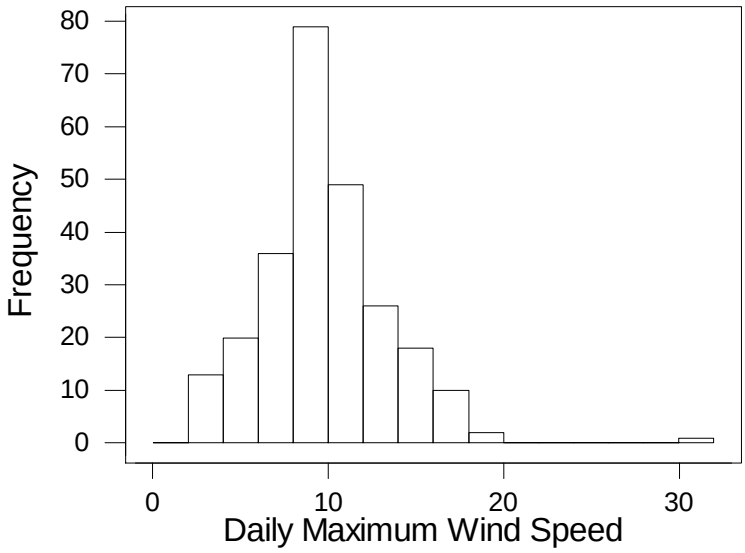
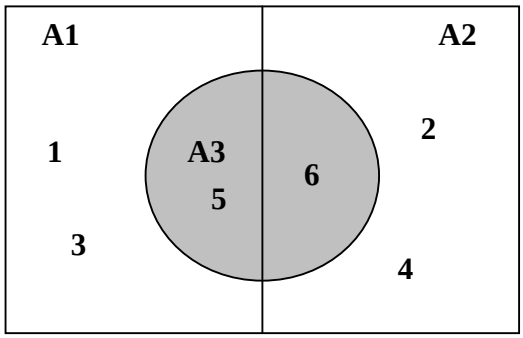


Figure 4. Histogram of daily maximum wind speeds (m/s) recorded in 2003.

Exercise 3

1. The sample space for X is {1, 2, 3, 4, 5, 6}; the Venn diagram is shown below.



   a. Pr(A1) = 1/2, Pr(A2) = 1/2, Pr(A3) = 1/3
   b. A1 and A2 are clearly mutually exclusive (a single roll cannot be even and odd) but it may not be obvious which events are independent. Pr(A1 and A2) = Pr({}) = 0, Pr(A1 and A3) = Pr({5}) = 1/6, Pr(A2 and A3) = Pr({6}) = 1/6. Since Pr(A1 and A2) = 0, A1 and A2 are indeed mutually exclusive. Since Pr(A1 and A3) = Pr(A1)Pr(A3) and Pr(A2 and A3) = Pr(A2)Pr(A3), both A1 and A2 are independent of A3.
   c. Pr(A1|A3) = Pr(A1 and A3)/Pr(A3) = 1/2 and Pr(A2|A3) = Pr(A2 and A3)/Pr(A3) = 1/2. So Pr(A1|A3) = Pr(A1) and Pr(A2|A3) = Pr(A2), as expected for independent events.
      Pr(A3|A1) = Pr(A1|A3)Pr(A3)/Pr(A1) = 1/3 and
      Pr(A3|A2) = Pr(A2|A3)Pr(A3)/Pr(A1) = 1/3.

2. Let R be the event ‘precipitation on Christmas day’, F the event ‘boundary layer below freezing’ and W the event ‘white Christmas’. We have Pr(R) = 1/2 and Pr(F) = 1/3. Assuming independence of R and F, Pr(W) = Pr(R and F) = Pr(R)Pr(F) = 1/6. If Pr(W) = 1/10 then Pr(R|F) = Pr(R and F)/Pr(F) = Pr(W)/Pr(F) = 3/10 < Pr(R). This appears to be inconsistent with independence of R and F, since in that case Pr(R|F) = Pr(R).

3. The results you get may differ because your random numbers will be different. For my five sample sizes I got proportions 0.4, 0.32, 0.358, 0.3306 and 0.33045. The Law of Large Numbers says that, if the data really are independent random numbers then the proportion should converge to Pr(A3) = 1/3 as the sample size increases. This is what we see here. Note that the convergence need not be monotonic: there is always some sampling variation.

4. The expected number of occurrences of each number is 100/6 = 16.7. In my sample I have the following counts.

	1	2	3	4	5	6
Observed	12	25	19	12	15	17

Again, there is some sampling variation but the observed counts are reasonably close to the expected values. Another way to test the random number generator is to count the number of times consecutive numbers are equal. The expected number is 99/6 = 16.5 since the probability that consecutive numbers are equal is 1/6. The observed number in my sample is 17.

5. $E(X) = 1(1/6) + 2(1/6) + 3(1/6) + 4(1/6) + 5(1/6) + 6(1/6) = 7/2$.
   $Var(X) = E(X^2) – E(X)^2 = 1(1/6) + 4(1/6) + 9(1/6) + 16(1/6) + 25(1/6) + 36(1/6) – (7/2)^2 = 91/6 – 49/4 = 35/12$.
   $E(X+Y) = E(X) + E(Y) = 2E(X) = 7$ since X and Y have the same expectation.
   $E(XY) = E(X)E(Y) = 49/4$ since X and Y are independent.
   $Cov(X, Y) = Cor(X, Y) = 0$ since X and Y are independent.
   $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y) = 2Var(X) = 35/6$.
   $Var(X – Y) = Var(X) + Var(Y) – 2Cov(X, Y) = 2Var(X) = 35/6$.

Exercise 4

1. Some example plots are shown in Figure 1.

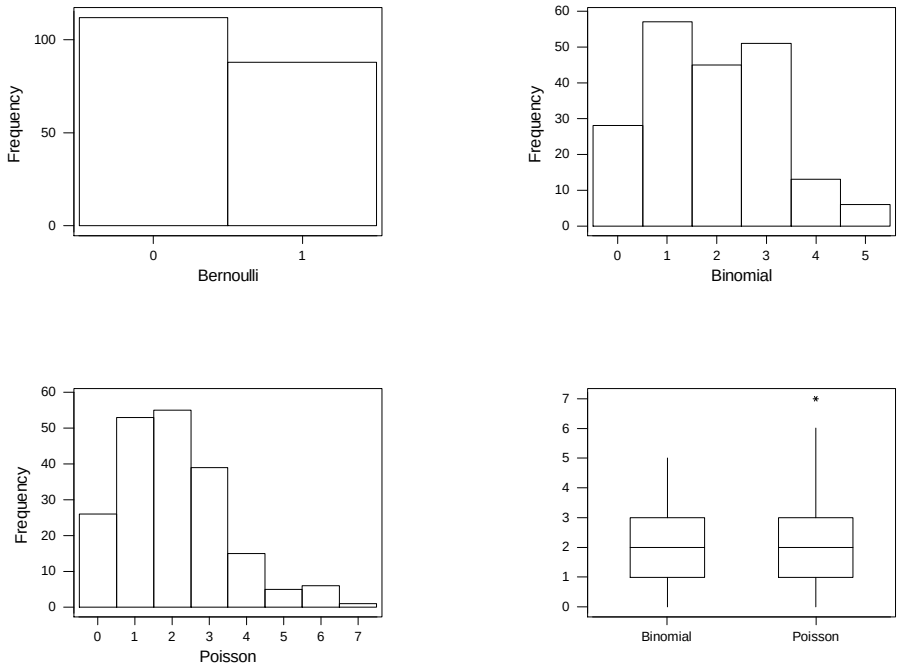


Figure 1. Histograms of samples of size 200 from three distributions: Bernoulli(0.4), Binomial(10, 0.2) and Poisson(2). Boxplots of the Binomial and Poisson data.

2. The Binomial distribution describes the number of successes in a fixed number of independent trials. Here, the number of 'trials' would be the number of days in a winter and a 'success' would be a stormy day. The Binomial distribution will be an unrealistic model, however, because days are unlikely to be approximately independent. The parameters of the model are $n = 120$ days and $\pi = 1/3$, with mean $n\pi = 40$ days and variance $n\pi(1 - \pi) = 80/3$. The probability of a winter having more than 40 stormy days is 0.4572. A plot of my 100 simulated winters is shown in Figure 2. The average number of stormy days in my sample is 40.5.

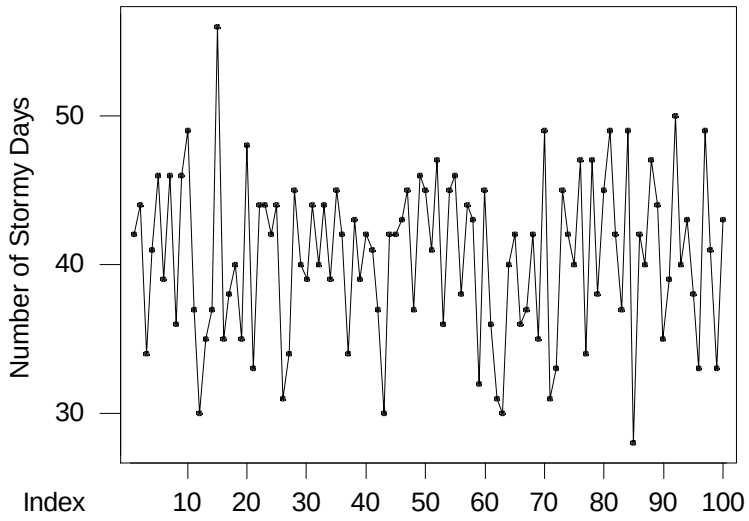


Figure 2. Time series of 100 simulated winters.

3. Some example plots are shown in Figure 3.

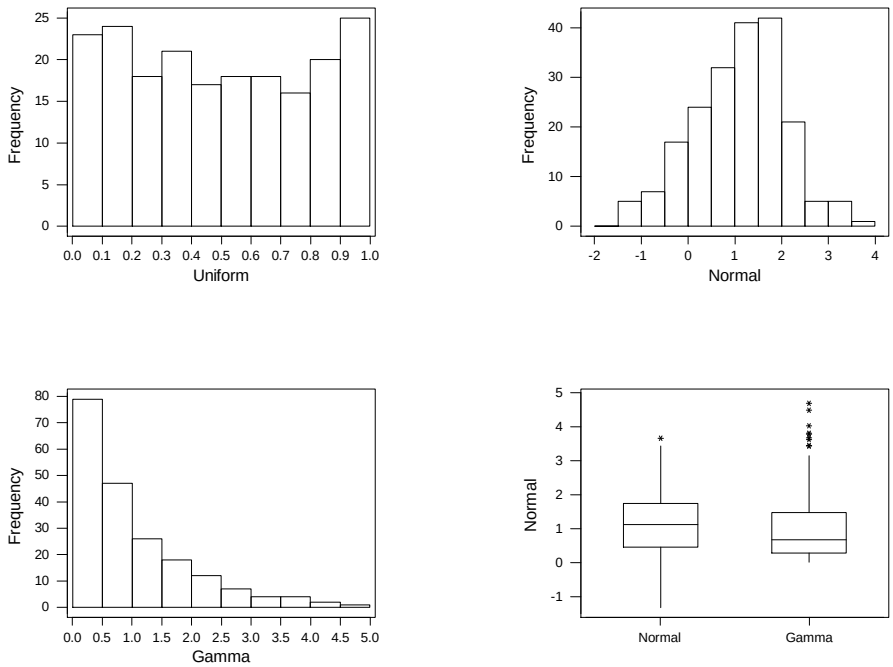


Figure 3. Histograms of samples of size 200 from three distributions: Uniform(0, 1), Normal(1, 1) and Gamma(1, 1). Boxplots of the Normal and Gamma data.

4. Histograms of the temperature data are shown in Figure 4. Normal densities with the same mean and variance as the data are superimposed on each plot. The daily mean temperatures appear reasonably Normal although there is an apparent discrepancy around 13°C. Daily minimum temperatures also appear reasonably Normal. Daily maximum temperatures, on the other hand, seem to have a longer upper tail and a shorter lower tail than the Normal distribution.

5. The probability that a Normal random variable with mean 10.964 and standard deviation 5.843 exceeds 15 is 0.245. This is lower than the sample proportion (0.29) because the Normal distribution is not a good fit to the data. Both values are likely to be poor estimates for 2004 because the temperature distribution in 2003 is unusual for Reading: a better estimate would be based on a longer temperature record.

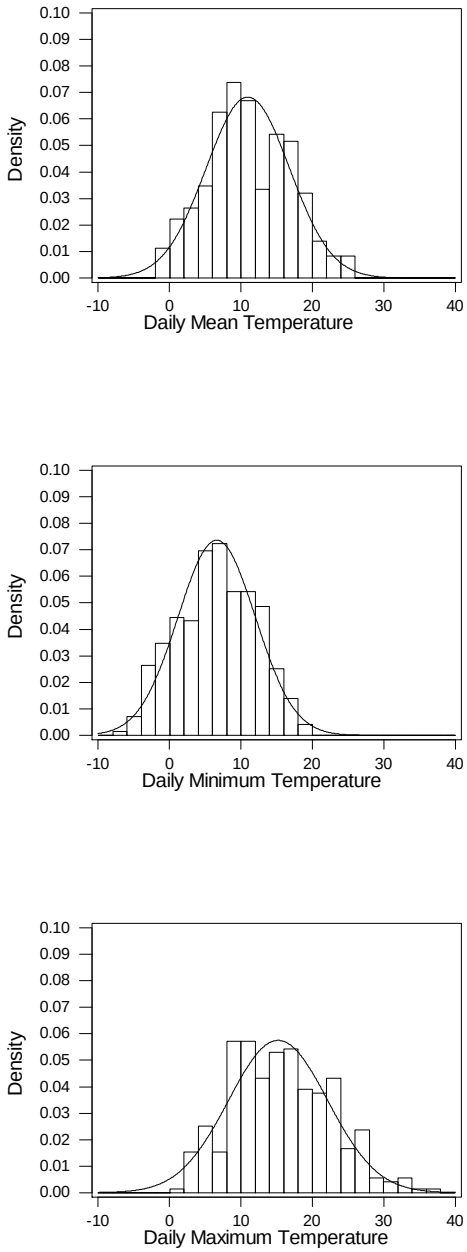


Figure 4. Daily mean, minimum and maximum temperatures (°C) with Normal densities superimposed.

Exercise 5

1. The mean of a large number of random variables should have a distribution that is approximately Normal according to the Central Limit Theorem. Each of the 100 means has expectation 4 and variance 4/100, so the approximating Normal distribution will have mean 4 and standard deviation 2/10. A histogram of the 100 means that I simulated is shown in Figure 1. The corresponding Normal density is superimposed and seems to be a good approximation.

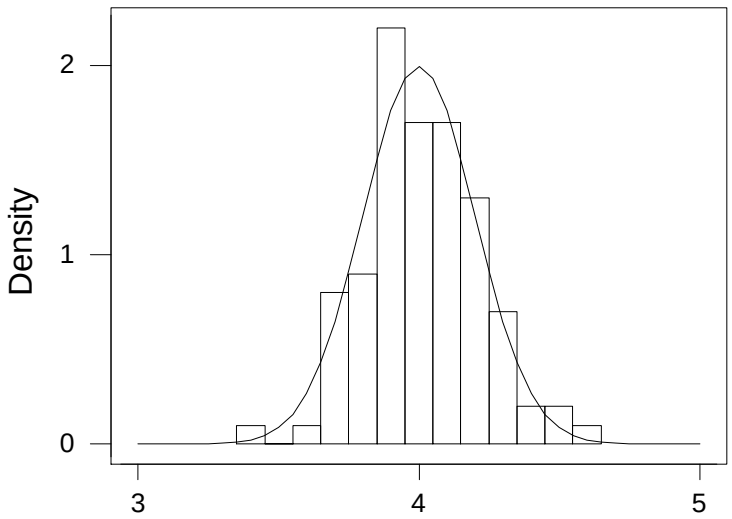


Figure 1. Histogram of 100 means and the approximating Normal density.

2. The usual point estimate is the sample mean, $\bar{x}$, which is 11.22°C for my data. If the true standard deviation, $\sigma = 5\,°\text{C}$, is known and the sample size is n then a $100(1-\alpha)$%-confidence interval is $(\bar{x} - c_\alpha \sigma/\sqrt{n}, \bar{x} + c_\alpha \sigma/\sqrt{n})$. For a 90%-confidence interval, $c_\alpha = 1.645$; for a 95%-confidence interval, $c_\alpha = 1.960$; for a 99%-confidence interval, $c_\alpha = 2.576$. The three confidence intervals for my data are therefore (10.40, 12.04), (10.24, 12.29) and (9.93, 12.51) degrees Celsius. In my case, only the 99%-confidence interval contains the true mean (10°C). The intervals are designed to contain the true mean $100(1-\alpha)$% of the time on average.

3. The standard deviation of my 100 sample means is 0.49. The theoretical standard error of the sample mean is $\sigma/\sqrt{n} = 1/2$. Of my 100 90%-confidence intervals,

90 of them contained the true mean (10°C), which is the proportion that we would expect in the long run.

4. The sampling distributions of my 100 sample means and medians are shown in Figure 2. The boxplots indicate that the two estimators are centred on the true mean but that the median has a greater spread and a skewed distribution. The greater spread is also reflected in the standard deviations: 0.49 and 0.57. Both estimators appear to be unbiased, but the sample mean is preferable because it has lower variance; the skewness has just arisen by chance in my samples. On the other hand, recall that the mean is more sensitive to outliers.

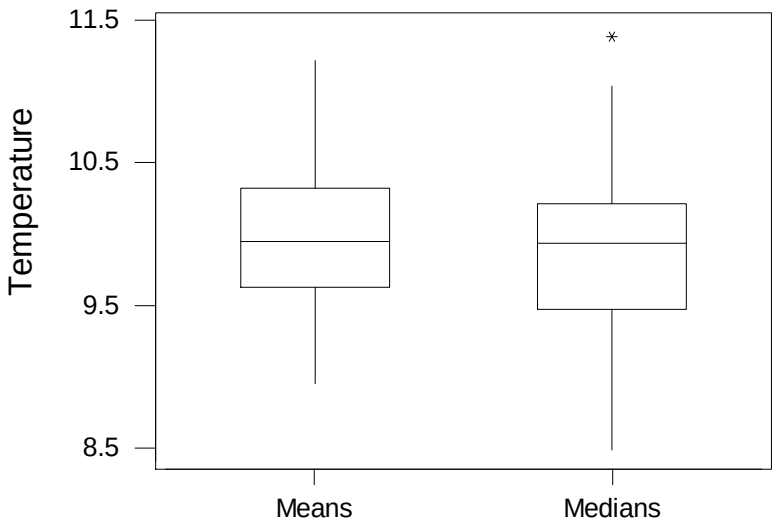


Figure 2. Boxplots of 100 sample means and medians.

Exercise 6

1. The boxplots in Figure 1 show that the rainfall amounts from both unseeded and seeded clouds are positively skewed and that there are some large, outlying observations. The median rainfall amounts for the unseeded and seeded clouds are 44 and 222 acre-feet, indicating that the seeded clouds tend to yield more rain. The boxplots show that the rainfall from the seeded clouds also has a greater spread than the unseeded clouds.

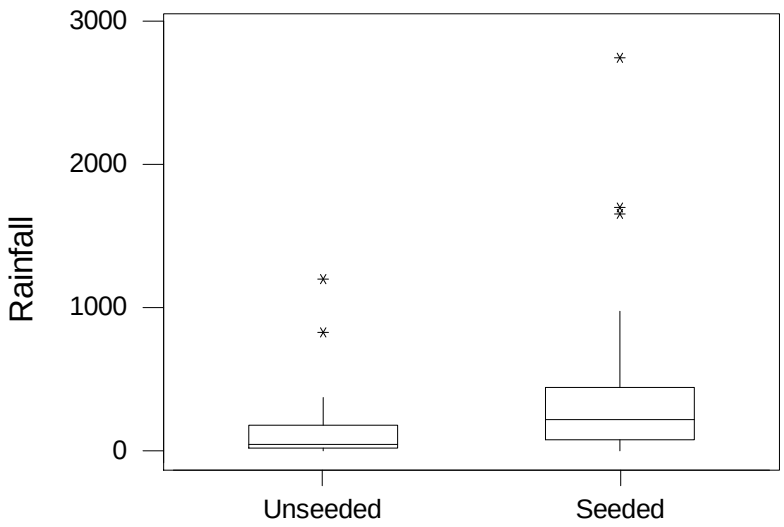


Figure 1. Rainfall amounts (acre-feet) for 26 unseeded and 26 seeded clouds.

2. A possible null hypothesis is $H_0 : \mu_1 = \mu_2$, where μ_1 is the mean rainfall from unseeded clouds and μ_2 is the mean rainfall from the seeded clouds. The general alternative hypothesis is $H_1 : \mu_1 \neq \mu_2$. (If the experimenters wanted to know if seeding a cloud *increased* rainfall then we could use $H_0 : \mu_1 \leq \mu_2$ and $H_1 : \mu_1 > \mu_2$, but we shall not consider this.) The two-sample t-test can be used to test such hypotheses. Since the twenty-six clouds in the two samples are not paired, the unpaired version of the test is appropriate. A reasonable significance level is anything between 1 and 10%, that is we conduct the test with the knowledge that we shall incorrectly reject the null hypothesis between 1 and 10% of the time on average.

3. The unpaired two-sample t-test assumes that each sample comprises independent Normal random variables with constant mean and variance, and that the two samples are independent with equal variances. The exploratory analysis in Question 1 shows that these assumptions are unreasonable for the data. Taking natural logarithms produces the data plotted in Figure 2, which are well approximated by Normal distributions. Furthermore, the large, outlying observations noted in Figure 1 are no longer evident and the standard deviations of the two samples, 1.64 and 1.60 for the unseeded and seeded clouds, are similar. The assumptions are acceptable for the transformed data, so applying the test to the logged data will give an accurate result.

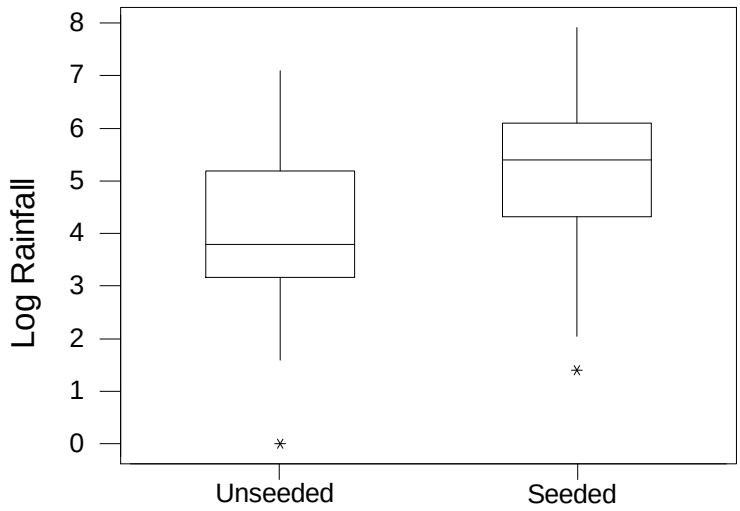


Figure 2. Log rainfall amounts for 26 unseeded and 26 seeded clouds.

4. The means of the transformed rainfall from the unseeded and seeded clouds are 3.99 and 5.13, with standard deviations 1.64 and 1.60. The difference in the means is 1.14 and the pooled standard deviation is 1.62. The t-statistic is $t = 1.14\sqrt{13}/1.62 = 2.54$, which is compared to the T-distribution with 50 degrees of freedom. For significance level 5% the critical values are ±2.01, which are shown below on the sketch of the density. The statistical tables also show that the two-sided p-value is between 0.01 and 0.02 since $2.40 < t < 2.68$. R yields the p-value 0.014. The null hypothesis is therefore rejected at the 5% level, but not at the 1% level for example. I conclude that there is quite strong evidence to suggest that seeding clouds affects the mean amount of rainfall produced by a cloud. The data indicate that the effect is to increase the amount of rainfall.

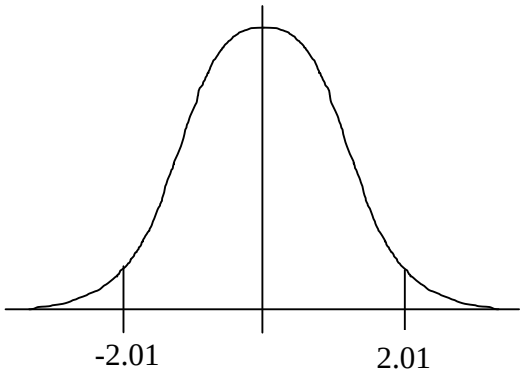


Figure 3. Density of the T_{50} distribution with critical values.

5. Without transforming the data the p-value is 0.051, and the null hypothesis would not be rejected at the 5% level.

6. Let X be the number of heads obtained in n = 250 spins of the coin, and let π be the probability that, when spun, the coin lands showing heads. We have no information to suggest that the spins are dependent, nor that the chance of heads changes during the experiment, so it is reasonable to assume that X has the Binomial distribution Bin(n, π). We wish to test the null hypothesis $H_0 : \pi = \pi_0 = 1/2$ that the coin is unbiased against the general alternative $H_1 : \pi \neq 1/2$ that the coin is biased.

   The experiment yielded x = 140 heads, so a point estimate for π is $\hat{\pi} = x/n = 14/25 = 0.56.$ Using the Normal approximation N(nπ, nπ (1 – π)) to the Binomial distribution, a 95% confidence interval for π is $\left(\hat{\pi} - 1.96\sqrt{\hat{\pi}(1-\hat{\pi})/n},\ \hat{\pi} + 1.96\sqrt{\hat{\pi}(1-\hat{\pi})/n}\right) \approx (0.498,\ 0.622)$. This interval contains π_0, so the null hypothesis is not rejected at the 5% level of significance.

   A hypothesis test based on the Normal approximation compares the z-statistic, $z = (\hat{\pi} - \pi_0)/\sqrt{\pi_0(1-\pi_0)/n} = 1.90$, under the null hypothesis to the standard Normal distribution. The probability of obtaining a z-statistic at least as large as this is 2Pr(Z > z) = 0.058. We conclude that there is only weak evidence that the coin used in the experiment is biased.

   A more accurate test that does not rely on the Normal approximation uses the Binomial distribution directly. The probability of obtaining a result as unlikely as 140 heads in 250 spins under the null hypothesis is Pr(X >= 140) + Pr(X <= 110) = 0.066.

   The claim that the Belgian euro is struck asymmetrically assumes that the result for this single coin holds for all Belgian euros. If this were the purpose of the investigation then a better experiment would be to spin n different coins instead of the same coin n times.

Exercise 7

1. The correlation is 0.82 (to two decimal places) for all four pairs, which suggests a strong, positive association in each case. The plots in Figure 1 show that the relationships are very different. The correlation is only meaningful for the first dataset, which has a roughly linear association. The second dataset has a strong, non-linear relationship that is not reflected in the correlation because it is a measure of only *linear* association. The third dataset appears to have an outlier; apart from that the relationship seems perfectly linear, but again this is not revealed by the correlation. The fourth dataset has only two distinct x-values, which makes it difficult to conclude anything about the relationship. Always plot the data if you are interested in the relationship between variables!

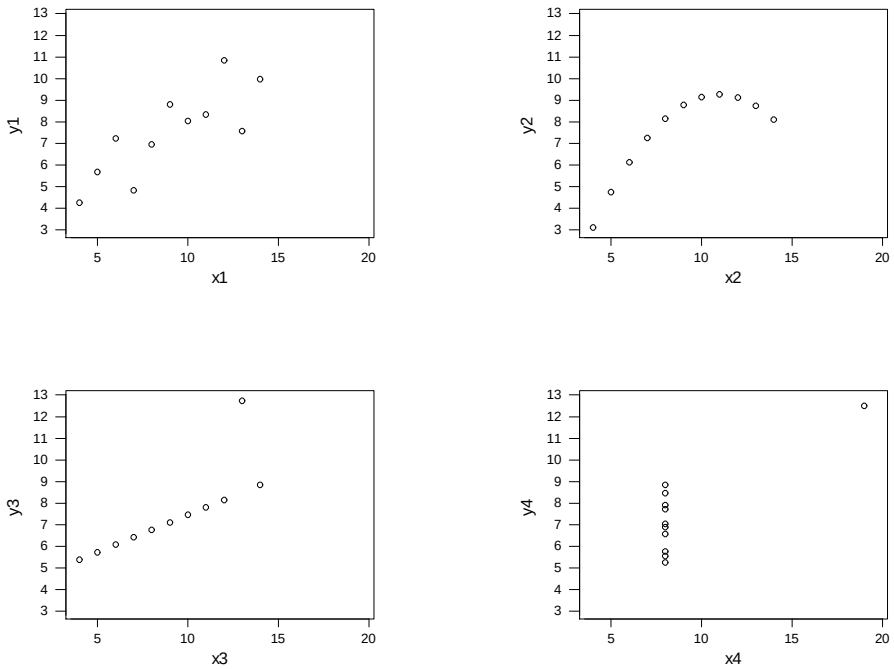


Figure 1. Scatter plots for four datasets.

2. The simple linear regression model is $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where Y_i is the response variable, x_i is the explanatory variable, and ε_i are independent $N(0, \sigma^2)$ random variables. For all four datasets the least-squares parameter estimates are $\hat{\beta}_0 = 3.0$ and $\hat{\beta}_1 = 0.5$.

3. The coefficients of determination are all $R^2 = 67\%$, and the p-values for testing $H_0 : \beta_1 = 0$ against a general alternative are all $p = 0.002$. These statistics indicate that the value of the x-variable is informative about the value of the y-

variable in each case. In particular, the x-variables explain about two-thirds of the variation in the y-variables.

4. Diagnostic plots reveal that the fitted model is acceptable for the first dataset only. The plots of residuals against fitted values in Figure 2 reveals the non-linear structure that is missed by the model for the second dataset, and highlights the influential outliers in the third and fourth datasets. Histograms and probability plots (not shown) reveal that the assumption of Normal residuals is also untenable for datasets 2 and 3.

   Only the first fitted linear model is useful; for the other datasets, the models will give misleading results. The second dataset needs a non-linear model, the linear model for the third dataset should be fitted in such a way that the outlier is accounted for, and the fourth dataset is not appropriate for modelling the bivariate relationship.

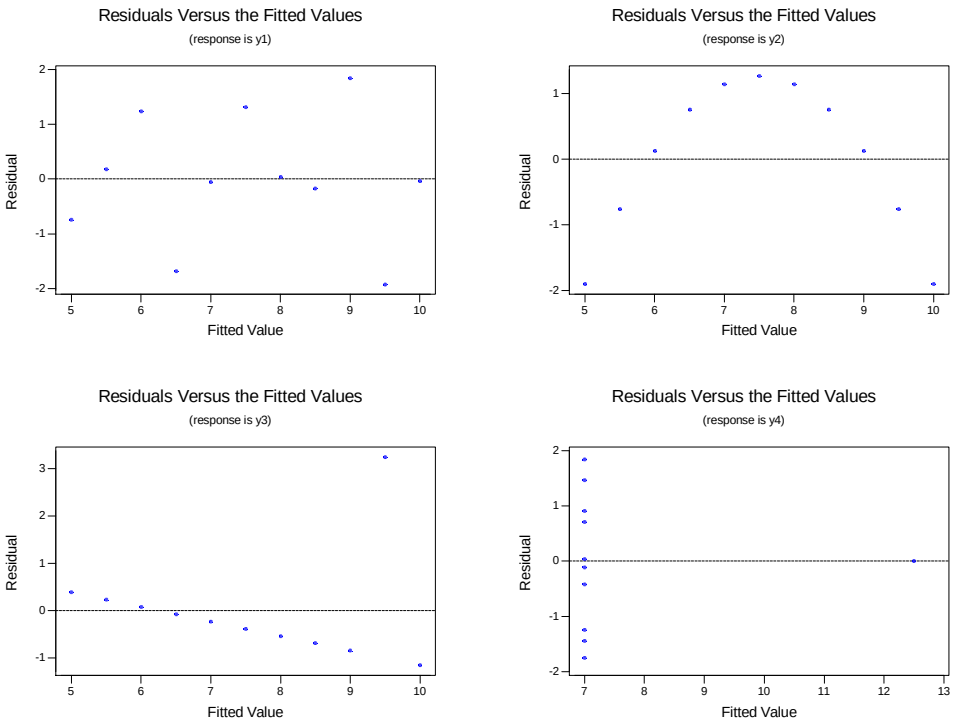


Figure 2. Diagnostic plots for the four linear regressions.

Exercise 8

This investigation aims to assess the evidence for time trends in wintertime Central England Temperature (CET). The data are mean winter CETs for the years 1866 to 1997 with winter defined to cover months December, January and February, the latter months identifying the year associated with a particular winter. Mean winter sea-level pressure (SLP) at four stations in Iceland, the Azores, Gibraltar and Lisbon for the same time period will also be used in the analysis.

The mean winter CETs are plotted against year in Figure 1. There appear to be some slow variations over time. The presence of an occasional very low temperature is reflected in the histogram of Figure 2, which is unimodal but has a longer cold tail. The mean temperature is 4.1°C and the standard deviation is 1.3°C.

A simple linear regression of temperature on year is fitted by ordinary least squares. The model is

$$T_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

where T_i is the temperature in year x_i and the ε_i are independent Normal random variables with zero mean and constant variance. The parameter estimates are $\hat{\beta}_0 = 1.1$ and $\hat{\beta}_1 = 0.0015$ with standard errors 5.7 and 0.003. The slope is not different from zero at any reasonable level of significance (p-value = 0.6) and the coefficient of determination is only 0.2%, suggesting that there is no discernable time trend in the wintertime CETs. Residual plots such as Figure 3 do not suggest any non-linear time trend that could be simply modelled.

A simple linear regression of temperature on mean SLP at Iceland is now estimated. The model is

$$T_i = \beta_0 + \beta_1 p_{Ii} + \varepsilon_i,$$

where p_{Ii} is the Icelandic SLP in year x_i. The parameter estimates are $\hat{\beta}_0 = 170$ and $\hat{\beta}_1 = -0.166$ with standard errors 13 and 0.013. The slope is highly significantly different from zero (p-value < 0.001) and the coefficient of determination is 55%. Icelandic SLP is a powerful explanatory variable for wintertime CET, explaining more than half of the variation in the temperatures. The CET increases by about 1°C for every decrease of 6hPa in Icelandic SLP. Residual plots indicate two, slight departures from the model assumptions: the distribution of the residuals has a long lower tail, and the residuals have a slight positive trend over time (Figure 4).

A linear time trend is added to the regression model:

$$T_i = \beta_0 + \beta_1 p_{Ii} + \beta_2 x_i + \varepsilon_i.$$

The parameter estimates are now $\hat{\beta}_0 = 165$, $\hat{\beta}_1 = -0.169$ and $\hat{\beta}_2 = 0.004$ with standard errors 13, 0.013 and 0.002. Although the time trend is slight, an increase in CET of about 0.4°C per century, it is now significantly different from zero at the 5% level (p-value =

0.043). The explanatory power of the regression model is only marginally improved, however: the coefficient of determination is 56%.

Perhaps the most illuminating model includes the Gibraltar SLP:

$$T_i = \beta_0 + \beta_1 p_{I\,i} + \beta_2 p_{Gi} + \beta_3 x_i + \varepsilon_i,$$

where $p_{G\,i}$ is the Gibraltar SLP in year x_i. The parameter estimates are $\hat{\beta}_0 = -104$, $\hat{\beta}_1 = -0.11$, $\hat{\beta}_2 = 0.21$ and $\hat{\beta}_3 = 0.004$ with standard errors 57, 0.02, 0.04 and 0.002. Both of the SLP terms are highly significantly different from zero (p-value < 0.001) and the time trend, of about 0.4°C per century remains significantly different from zero at the 5% level (p-value = 0.041). The opposite signs of the coefficients for the Iceland and Gibraltar SLP terms indicate that the (weighted) pressure difference between these two locations contributed most to the explanatory power of the model, which has coefficient of determination equal to 63%.

This analysis has shown that there is evidence, at the 5% level of significance, for an increasing time trend in CET of about 0.4°C per century. This trend is discernible only after accounting for the influence of atmospheric circulation patterns on CET, in particular the pressure differential between Iceland and Gibraltar.

One potential shortcoming of this analysis is the assumption that the residuals are independent: they might be expected to exhibit serial correlation from one year to the next. If such behaviour is not accounted for in the model then the results, such as the statistical significance of a time trend, can be misleading. Models that do account for correlated residuals are beyond the scope of this module, however, but fortunately the correlation is weak for the data examined in this exercise. Another potential shortcoming is the linear time trend. This is very common practice in climate science, but processes rarely exhibit linear trends: they are usually more complicated, but appropriate statistical models are also beyond the scope of this module.

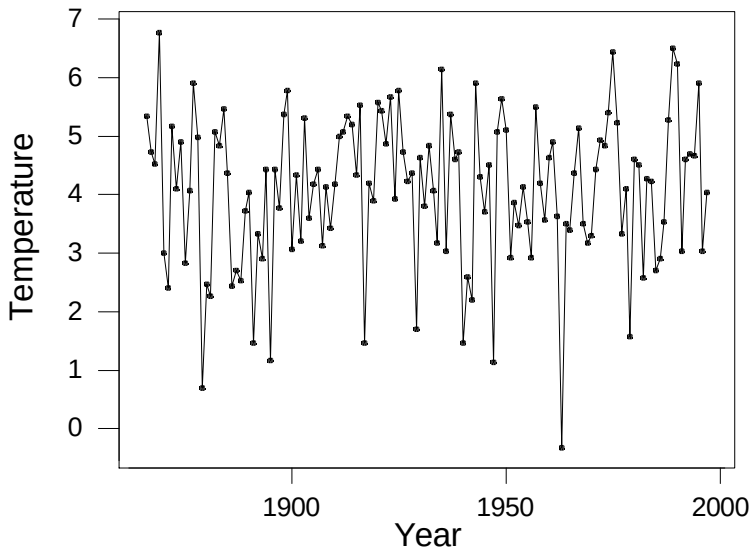


Figure 1. Mean winter CET (°C) against year for 1866 to 1997.

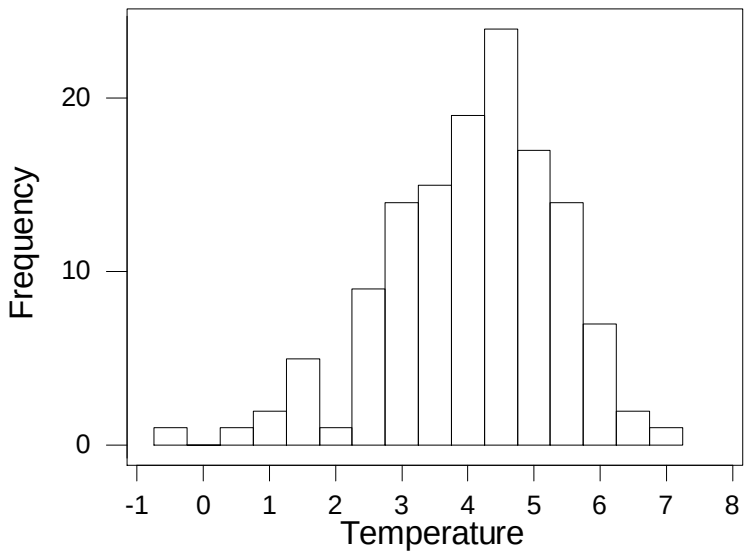


Figure 2. Histogram of mean winter CET (°C) from 1866 to 1997.

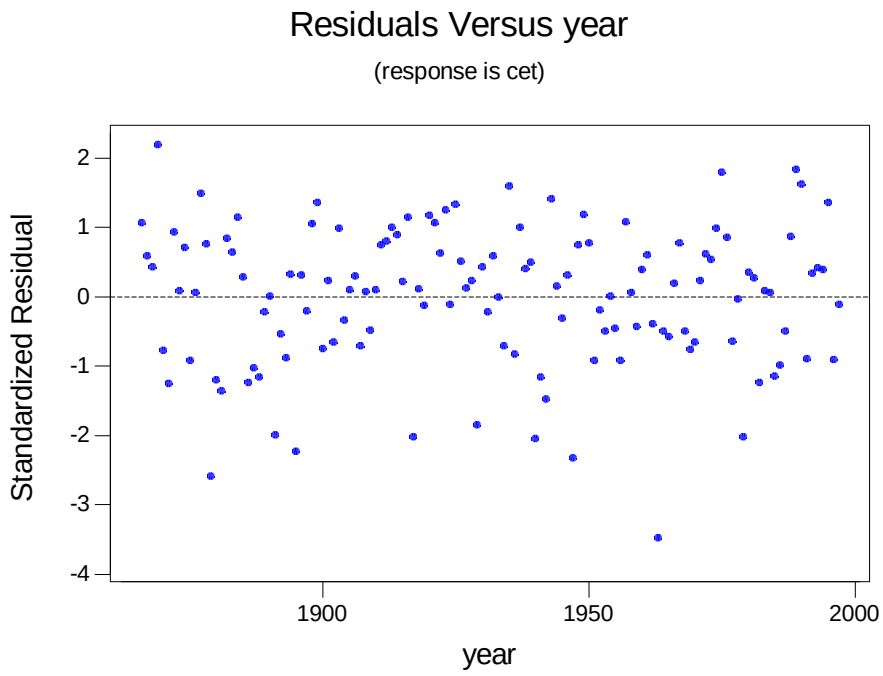


Figure 3. Residuals from the linear regression of CET (°C) on year.

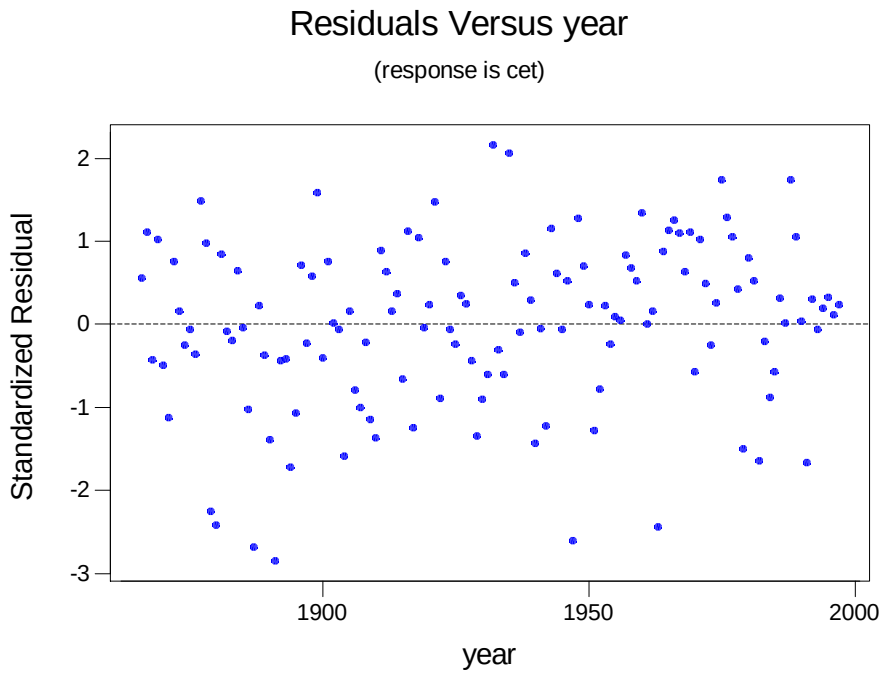


Figure 4. Residuals from the linear regression of CET (°C) on Iceland SLP.

Exercise 9

1. The Darwin SLP is plotted in Figure 1. There is a strong annual cycle that accounts for almost all of the variation in the data, which is spread evenly around the value 10; there is a slight increase over time.

2. For periodic data it is usual to take moving averages with lengths equal to multiples of the period. A moving average with length 60 months is shown in Figure 1. This highlights the main, low-frequency variation; shorter lengths have too much noise. The main pattern appears to be an overall increase with a dip in the early 1970s.

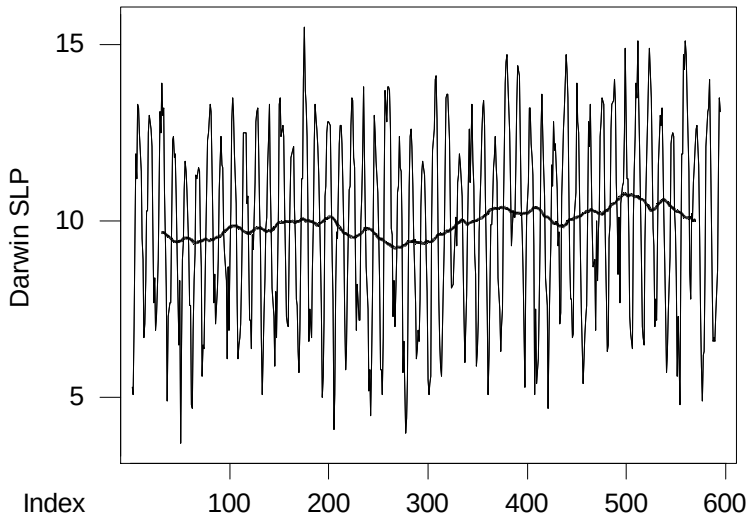


Figure 1. Monthly mean SLP (−1000hPa) at Darwin from January 1951 to July 2000. The centred moving average of length 60 months is superimposed.

3. The autocorrelation function is given in the lecture notes. The dependence at lag 12 months is illustrated in Figure 2.

4. The plot of the data after applying the backward difference filter of lag 12 is given in the lecture notes.

5. The series obtained after applying both lag 12 and lag 1 differences is shown in Figure 3. If this were white noise then the autocorrelation function would be zero for all lags. The actual autocorrelation function is shown in Figure 4 and has

large, negative correlations at lags 1 and 12. The differenced series is therefore not white noise, but has some remaining dependence structure.

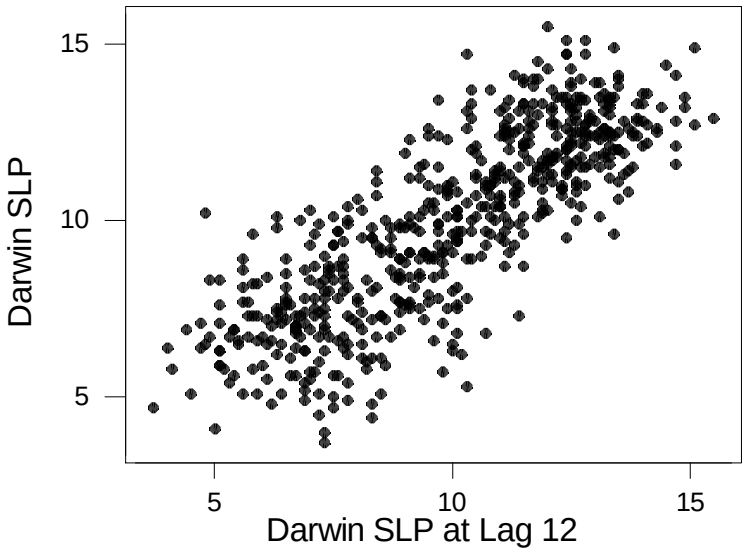


Figure 2. Darwin SLP plotted against itself at lag 12 months.

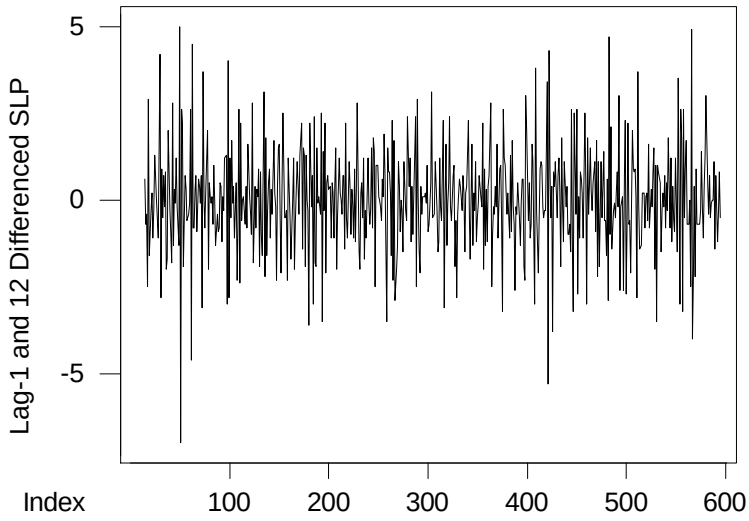


Figure 3. Darwin SLP after applying lag 1 and lag 12 backward difference filters.

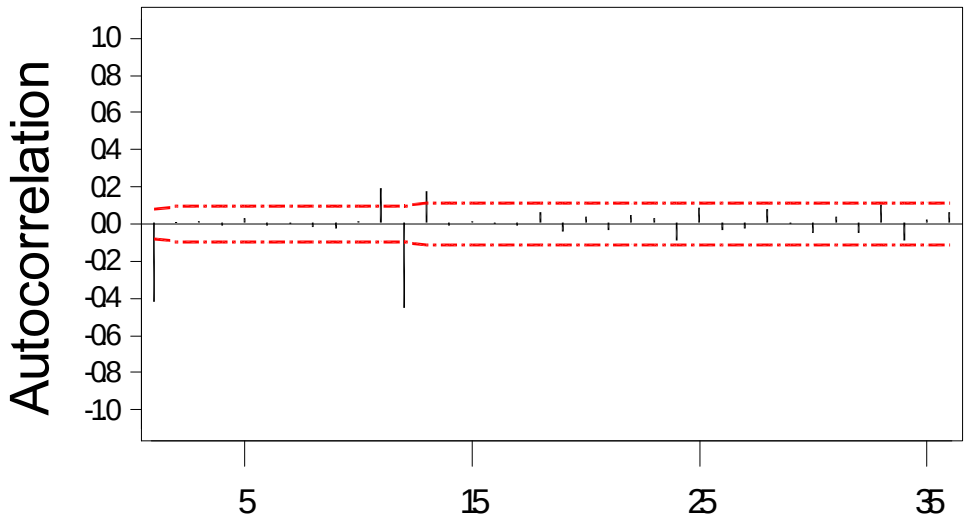


Figure 4. Autocorrelation for the series in Figure 3 with 95% confidence intervals.

6. The histogram and autocorrelation function from the seasonal ARIMA model are shown in Figure 5. The distribution of the residuals is approximately Normal (in fact the residuals have slightly heavier tails) and there is no significant autocorrelation at the 5% level, indicating that the residuals are well approximated by white noise. The fitted model has equation

$$Y_t = Z_t - 0.59Z_{t-1} - 0.96Z_{t-12} + 0.56Z_{t-13},$$

where Z_t is white noise and Y_t is the lag-1 and lag-12 differenced data, i.e.

$$X_t = X_{t-1} + (X_{t-12} - X_{t-13}) + Z_t - 0.59Z_{t-1} - 0.96Z_{t-12} + 0.56Z_{t-13}.$$

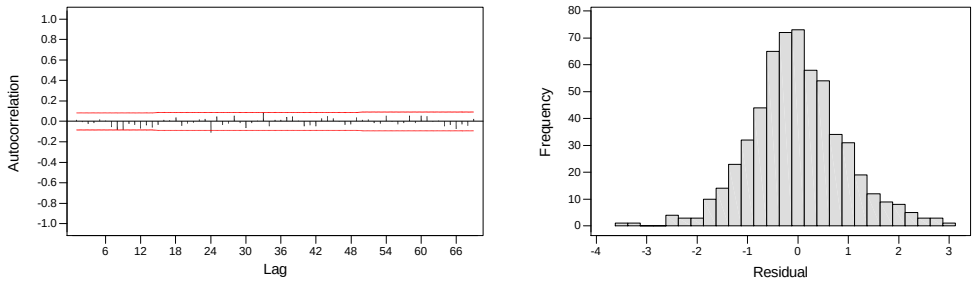


Figure 5. Autocorrelation function (with 95% confidence intervals) and histogram for the residuals from the seasonal ARIMA model.