Course Exercises

Data analysis for weather and climate research

These exercises are designed to help you to understand the concepts introduced in the lectures and to learn how to conduct, interpret and present basic statistical analyses. They are intended to be completed in order (later exercises assume knowledge acquired in earlier ones) and each exercise assumes that you have studied the corresponding chapter of the lecture notes. The exercises can be completed according to the following schedule:

Day	1	2	3	4	5
Exercise	1, 2	3, 4	5, 6	7, 8	9

Some `R` commands are provided to help you to complete the exercises, but you will need to spend some time reading the on-line documentation and help files in order to understand them. Example solutions are available on the course web-page.

Chris Ferro (Tutor) `c.a.t.ferro@reading.ac.uk`

Exercise 1. Introduction

1. Visit the course web-site via
   `www.met.rdg.ac.uk/cag/courses`
   and download the datasets `atldjf.txt` etc.

2. Open `R` and familiarise yourself with the help pages.
   ```
   > help.start()
   ```

3. Read the data in the file `temp.txt` into `R`.
   ```
   > data <- read.table(‘‘temp.txt’’, TRUE)
   > attach(data)
   ```

4. Produce a boxplot and a histogram of the daily mean temperatures.
   ```
   > boxplot(Tmean)
   > hist(Tmean)
   ```
   What do these plots tell you about the shape of the distribution? Are there any unusual values? If so, can you explain their occurrence? How will this affect your analysis?

5. Reproduce the histogram with the width of the intervals equal to 5 degrees.
   ```
   > hist(Tmean, breaks = seq(-5, 30, 5))
   ```
   Does this affect your view of the distribution? What implications does this have for plotting data?
   P.T.O.

6. Compute the mean and standard deviation of the daily mean temperatures.

   ```
   > mean(Tmean, na.rm = TRUE)
   > sd(Tmean, na.rm = TRUE)
   ```

   Compute the median, quartiles, range and inter-quartile range.

   ```
   > summary(Tmean)
   ```

   What do these statistics tell you about the data? There are some missing values in the data: why do you think they occurred? How does this affect your analysis?

7. Write a short summary of your analysis so far, including your answers to the earlier questions and any other conclusions you have drawn. To export figures from `R` see

   ```
   > ?Devices
   ```

8. Visit the `stats@met` web-site

   `www.met.rdg.ac.uk/cag/stats`

   and follow the link to the 'Rice Virtual Lab in Statistics'. Try out the demonstrations on 'Mean and Median', 'Histograms, Bin Widths, and Cross Validation' and 'Comparing distributions'. Browse some of the other links on the `stats@met` web-site.

   P.T.O.

9. Use a calculator, not `R`, to complete this question. For each of the age, height and weight datasets in the lectures notes (also available in the file `rdgmorph.txt`)

   (a) compute the mean and standard deviation;

   (b) rank the data in increasing order;

   (c) compute the median, quartiles, range and inter-quartile range; and

   (d) describe the shape of the distribution.

   Record your answers below.

	Mean	SDev	Q1	Q2	Q3	IQR	Range
Age							
Height							
Weight							

The age distribution...

The height distribution...

The weight distribution...

Exercise 2. Descriptive statistics

Produce a short report describing the daily *minimum* and *maximum* temperature data in `temp.txt` using something like the following approach, which can be taken as a template for exploratory data analysis more generally.

1. Make sure that you understand the data.
2. Decide what features of the data to investigate.
3. Decide what statistics and plots to use to do this.
4. Compute the statistics and plots you have chosen.
5. Interpret the results with regard to your chosen aims.
6. Do the results suggest further investigations? If so, repeat steps 2 to 5. If not, write up your findings.

Any data analysis report should contain the following information: an explanation of the data, a statement of the investigative aims, a description of the statistical methods used, a commentary on the results, and a summary of the conclusions. Remember that you are conducting a statistical analysis to enhance understanding, not to produce page after page of tables and graphs.

If you have time, produce a short report describing the wind-speed data in `wind.txt` as well.

Exercise 3. Probability concepts

1. Let X be the random variable representing the outcome of a single roll of a fair, six-sided die. Write down the sample space for X. Draw a Venn diagram showing the following three events: A_1, an odd number; A_2, an even number; and A_3, a number greater than four.

   (a) What are the values of the probabilities $\Pr(A_1)$, $\Pr(A_2)$ and $\Pr(A_3)$?

   $\Pr(A_1) =$
   $\Pr(A_2) =$
   $\Pr(A_3) =$

   (b) Which pairs of the events do you think are mutually independent? Which pairs do you think are mutually exclusive? Find $\Pr(A_1 \text{ and } A_2)$, $\Pr(A_1 \text{ and } A_3)$ and $\Pr(A_2 \text{ and } A_3)$, and check which pairs are independent and which are exclusive.

   $\Pr(A_1 \text{ and } A_2) =$
   $\Pr(A_1 \text{ and } A_3) =$
   $\Pr(A_2 \text{ and } A_3) =$

(c) Find the conditional probabilities $\Pr(A_1 \mid A_3)$ and $\Pr(A_2 \mid A_3)$. Are these what you expected? Compute the conditional probabilities $\Pr(A_3 \mid A_1)$ and $\Pr(A_3 \mid A_2)$ using Bayes's Theorem.

$\Pr(A_1 \mid A_3) =$
$\Pr(A_2 \mid A_3) =$
$\Pr(A_3 \mid A_1) =$
$\Pr(A_3 \mid A_2) =$

2. To get a white Christmas in London there needs to be precipitation and the boundary layer needs to be below freezing point. If the probability of precipitation on Christmas day is 1/2 and the probability that the boundary layer will be below freezing is 1/3 then calculate the probability of a white Christmas assuming independence of these two events. If observations show that a white Christmas in London happens, on average, once every 10 years then calculate the conditional probability of precipitation given that the boundary layer is below freezing and compare it with the unconditional probability of precipitation.

3. Generate a random sample of size 10 for the random variable X in question 1.

   ```
   > x <- sample(6, 10, TRUE)
   ```

   Write down the proportion of times event A_3 occurs.

   ```
   > mean(x > 4)
   ```

   Repeat for samples of sizes 100, 1000, 10 000 and 100 000. Are the results consistent with the Law of Large Numbers?

10	100	1000	10 000	100 000

4. Using your sample of size 100, compute the number of times each of the possible values of X occurs.

   ```
   > table(x)
   ```

   Compare these with the expected number for each value. Think of some other ways to check that `R` is able to generate seemingly random numbers, and try them out on your sample.

	1	2	3	4	5	6
Observed						
Expected						

5. What are the values of the expectation and variance of X?

   $E(X) =$
   $\mathrm{Var}(X) =$

   Let Y be the random variable representing the outcome of a second, *independent* roll of the same die. Write down the values of the following quantities without doing any calculations.

   $E(X+Y) =$ $\qquad$ $E(XY) =$
   $\mathrm{Cov}(X,Y) =$ $\qquad$ $\mathrm{Cor}(X,Y) =$
   $\mathrm{Var}(X+Y) =$ $\qquad$ $\mathrm{Var}(X-Y) =$

6. What would you guess is the chance of finding at least two people with the same birthday (not necessarily year) out of a sample of 30 people? By counting the number of possible birthdays for each person, calculate the probability that all of the 30 people have different birthdays. What is the probability that two or more of the people share a birthday? Think of reasons for any difference with your guess. An interactive demonstration of this 'Birthday Problem' is at
`www-stat.stanford.edu/~susan/surprise`
and an explanation is at
`www.mste.uiuc.edu/reese/birthday`.

Exercise 4. Probability distributions

1. Generate one sample of size 200 from each of the discrete Bernoulli, Binomial and Poisson distributions.
   ```
   > ?rbinom # see also rpois
   ```
   Compare the distributions of each sample with suitable plots. Change the parameters in each of the models to see how they influence the distributions.

2. The Binomial distribution is a possible probability model for the number of stormy days in a season. Do you think that this is a realistic model? If there are 120 days in a winter and the probability of a stormy day in winter is 1/3, write down the parameters, n and π, of the Binomial model. Write down the expectation and variance of the number of stormy days, and compute the probability of a winter having more than 40 stormy days. Generate 100 Binomial variables to represent a sequence of 100 winters and plot the simulated data. What is the average number of stormy days simulated?

   $n =$ Expectation =
   $\pi =$ Variance =

   Probability of more than 40 stormy days =
   Average number of stormy days simulated =

3. Repeat question 1 with continuous Uniform, Normal and Gamma distributions.

   ```
   > ?runif # see also rnorm and rgamma
   ```

4. Compare the three temperature measurements in the file `temp.txt` to Normal distributions with means and variance set equal to the sample statistics. Are there any noticeable departures from Normality?

   ```
   > qqnorm(Tmean)
   > abline(mean(Tmean), sd(Tmean))
   ```

5. Use the Normal distribution as a probabilistic model for daily mean temperatures at Reading, taking the mean and standard deviation of the model equal to the sample statistics. Find the probability that the daily mean temperature exceeds 15°C. Is this a useful estimate for the proportion of days on which mean temperature will exceed 15°C in 2004?

   ```
   > ?pnorm
   ```

6. Try the demonstration 'Normal Approximation to the Binomial Distribution' in the 'Rice Virtual Lab in Statistics'.

Exercise 5. Parameter estimation

1. Generate 100 samples of size 100 from a Poisson distribution with mean 4. Compute the 100 sample means. What distribution does the Central Limit Theorem say will approximate the distribution of these means? Assess this claim with a suitable plot.

```
> x <- matrix(rpois(10000, 4), 100, 100)
> m <- apply(x, 1, mean)
```

2. Suppose that the daily mean temperature at the Plato Cave weather station is precisely a sequence of independent Normal random variables with expectation 10°C and standard deviation 5°C. Generate a sequence of 100 daily mean temperatures. Now pretend that you do not know the true mean of the Normal distribution, and calculate a point estimate for it. By hand, also compute 90%, 95% and 99% confidence intervals. Do any of your intervals contain the true mean? What is the correct interpretation of the intervals?

   Point estimate =
   90% confidence interval =
   95% confidence interval =
   99% confidence interval =

3. Simulate 99 more samples of 100 temperatures and store the 100 sample means. Calculate the standard deviation of your sample means and compare it to the standard error that you would expect theoretically. Compute a 90% confidence interval from each sample, storing the lower and upper limits. What proportion of the intervals contain the true value, 10°C? What proportion would you expect?

   Sample standard deviation =
   Theoretical standard error =
   Proportion of intervals =
   Expected proportion =

4. Now store the 100 sample medians. Compare the sampling distributions of the sample means and medians. What do you conclude about properties of the two estimators? Which estimator do you prefer?

5. Carefully work through the demonstrations 'Sampling Distribution Simulation' and 'Confidence Intervals' in the 'Rice Virtual Lab in Statistics'.

Exercise 6. Hypothesis testing

1. The file `clouds.txt` contains data collected from a U.S. experiment in the early 1970s that dropped silver nitrate crystals from aircraft to 'seed' clouds and make rain. The data are the rainfall amounts (in acre-feet!) from a sample of 26 unseeded clouds and a sample of 26 seeded clouds. Explore the distributions of the two samples, noting any similarities and differences.

2. The experimenters want to know if seeding a cloud affects the amount of rainfall. One way to do this is to assess whether or not the two samples come from populations with different means. Write down the null and alternative hypotheses. Which statistical test can be used to test these hypotheses? Select an appropriate level of significance for the test.

3. What assumptions does your chosen test make about the data? Are these assumptions reasonable for the rainfall data? If not, transform the data (by taking square roots for example) so that the test will be appropriate.

4. Compute the test statistic for your test by hand. Write down the distribution of your test statistic and sketch its probability density. Determine critical regions for your test and add them to your sketch. Obtain an approximate p-value for your test statistic. Now perform the test using `R`, note the p-value and compare it with your chosen level of significance.

   ```
   > ?t.test
   ```

   Write down clearly what conclusion you draw from the test and what this tells the experimenters about cloud seeding.

5. Does the result change if you perform the test without transforming the data?

6. Shortly after the introduction of the euro coins in 2002, BBC on-line news published this article:

   'Meanwhile, two Polish statisticians have discovered something about euro coins that should gladden the hearts of confidence tricksters. The coin apparently favours heads. When Tomas Gliszczynski and Waclaw Zawadowski of the Podlaska Academy spun one Belgian euro coin 250 times, it came up with King Albert's head 140 times. "The euro is struck asymmetrically," Mr Gliszczynski told Germany's Die Welt newspaper. He said he hoped to experiment with German euro coins at a maths conference next month. "I know the phenomenon from other coins like the two zloty piece, which we have thrown more than 10,000 times," he said.'

   Write a short note describing a statistical test for the hypothesis that the Belgian euro is biased given these results, and comment on Mr Gliszczynski's claim.

Exercise 7. Linear regression

1. Read the data in the file `xy.txt` into `R`. Compute the sample correlation between x_1 and y_1 then repeat for the other three pairs.

   ```
   > ?cor
   ```

   Record the values in the table below. Also write down what these values tell you about the association between the four pairs of x and y variables. Plot y_1 against x_1 then repeat for the other three pairs.

   ```
   > ?plot
   ```

   Do these plots change your ideas about the associations? What implications does this have for the interpretation of correlations?

Pair	1	2	3	4
Correlation				

2. Write down a mathematical representation for the simple linear regression of each y-variable on the corresponding x-variable. Make sure that you know which is the response variable and which is the explanatory variable. Using a calculator instead of `R`, estimate the slope and intercept parameters in each of the four cases using the formulae in the lecture notes.

Dataset	$\hat{\beta}_0$	$\hat{\beta}_1$	R^2	p
1				
2				
3				
4				

3. Now perform the linear regressions with `R` and check that the parameter estimates agree with your calculations.

   ```
   > fit <- lm(y1 ~ x1)
   > summary(fit)
   ```

   What are the values of the coefficient of determination? What does this tell you about the linear models? What is the p-value for testing whether or not the slope is zero? What do you conclude about the explanatory power of the x variable in each case?

4. Compute the residuals from the fitted models and assess the model fits by making diagnostic plots.

   ```
   > z <- residuals(fit)
   ```

   Are any of the model assumptions inappropriate? Do these plots change your conclusion about the explanatory power of the x variable in each case?

Exercise 8. Multiple regression

In this exercise we look at the Central England Temperature (CET) dataset in the file `atldjf.txt`. The column 'cet' contains the mean winter central England temperature (°C) from 1866 to 1997. The next four columns contain mean sea-level pressure (SLP) measurements (hPa) at four locations. As you work through this exercise you should compile a short report of your analysis, including the regression equation for each model that you fit.

1. Plot CET against year and plot a histogram of the temperatures. What features do you notice in the series? What does the distribution look like?

2. One way to assess evidence for a time trend in CET is to fit a linear regression of temperature on year. Fit this model using `R` and interpret the results. What do you conclude about the changes in CET through time? Examine any diagnostic plots that you consider appropriate and comment on their consequences for your model.

3. Now regress CET on just the Iceland SLP and assess the model fit. What does the estimate of the slope parameter say about the relationship between CET and Iceland SLP? Can you explain this result scientifically?

4. The plot of residuals against observation order indicates a slight, increasing time trend. Fit the multiple regression of CET on Iceland SLP and year. Examine the significance of the two explanatory variables and assess the model fit. What conclusions do you draw now about any time trend in CET?

5. If you have time, experiment by including and excluding different SLP series from the regression and compare the fitted models. Also try adding a polynomial time trend by storing squared year in a new object.

Exercise 9. Time series analysis

1. The Darwin SLP data discussed in the lectures is in `darwin.txt`. Plot the data against time. What can you tell about any cycles or long-term time trends?
   ```
   > data <- read.table(‘‘darwin.txt’’, TRUE)
   > data <- c(t(data[, -1]))
   ```

2. Apply moving average filters of different lengths to obtain a clearer view of the long-term trends.
   ```
   > ma <- filter(data, rep(1 / 12, 12))
   ```

3. Compute the autocorrelation function up to lag 36 months.
   ```
   > acf(data, 36, na.action = na.pass)
   ```
   The cyclical behaviour indicates the seasonality of the data. If X_t is the SLP for month t, plot X_t against X_{t-12} to visualise this dependence.

4. Apply the backward difference filter with lag 12 to remove the seasonality, then plot the differenced series to obtain a clearer view of the inter-annual variations.
   ```
   > ?diff
   ```

5. (The final two questions are optional.) In addition, apply the backward difference filter with lag 1 to remove the month-to-month dependence and plot the resulting series, which we shall call Y_t. This looks like white noise. If this were so, what would you expect the autocorrelation function to look like? Compute

the autocorrelation function up to lag 36 and interpret the result.

6. The autocorrelation function for Y_t suggests that a particular seasonal ARIMA model, with moving average components at lags 1 and 12, will be a good description of the Darwin SLP series. Fit this model to the Darwin SLP series by selecting a seasonal model with period 12, one nonseasonal and one seasonal difference, one nonseasonal and one seasonal moving average term, and no constant term.

```
> arima(data, c(0, 1, 1),
+ list(order = c(0, 1, 1), period = 12))
```

Plot the autocorrelation function and histogram of the residuals. Do the residuals look like white noise?