Exercises for the course on:

Statistical concepts in environmental science

Dr. David B. Stephenson

www.met.reading.ac.uk/cag/courses

1. Introduction

1. Browse and bookmark the web sites on the course web site (see above). Learn how to find statistics information in online textbooks and glossaries.

2. Become familiar with the MINITAB package by starting it up on the PCs (see Annex B). Get used to the options and the on-line help.

3. Type in a column of 10 numbers guessed at random and then try to analyse them using the **Stats** options. (this will be covered in detail in the next lecture).

4. Read chapter 1 of the lecture notes and go and look up some of the reference books in the library.

2. Descriptive sample statistics

1. Read chapter 2 of the lecture notes. Think of questions to ask the lecturer!

2. Try to reproduce the analyses presented in this lecture by using MINITAB on the data set **rdgmorph.txt**. Use the **Stats** and **Graph** menus.

3. Prepare a new data set based on a sample of your colleagues and repeat the analyses. Comment on any notable differences between the statistics of your sample and the sample presented in the lecture notes.

4. By inspecting the histogram, estimate the mode (most likely value) for both samples and compare with the sample means and medians.

5. Try comparing the two data samples by doing a stacked box plot (i.e. two boxplots side by side) – this can be found in **Stats – Anova** in MINITAB.

6. Remove the largest and the smallest heights from your sample and repeat the analyses to find out how sensitive the various statistics are to removal of extremes.

7. Add a "troll" with height 0.5 metres and weight of 150kg to your data set and repeat the descriptive statistics. Note which statistics change the most and hence are the least resistant to having a troll (i.e. outlier) in the sample.

8. Go through the demonstration on means and medians in the Rice Virtual Laboratory of statistics (see course web page). Read all the instructions and information and complete all the online exercises.

3. Basic probability concepts

1. Read chapter 3 of the lecture notes. Think of questions to ask the lecturer!

2. The Star Wars version of Trivial Pursuit does not contain a die, but instead a "pseudo-random" number generator, in the shape of R2-D2. Matt recorded all of the values produced by the generator. These are in **random.txt** on the course website.

   What is the probability of getting a 6 on a single roll of a fair die?

   What is the probability of getting a 6 in the "R2-D2" data? (Hint: try using **stat->tables->tally** in MINITAB)

   Generate your own sample of random numbers using the **calc->random data->integers**… option in MINITAB and compare this with the "R2-D2" data

   Are they consistent?

   Now look at the conditional probability of getting a 6 followed by a 6 for your random sample (hint: you may need to use the **calc->calculator->lag** option)? Compare these with the R2-D2 number generator; is the generator a good approximation to the die?

3. 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 of the boundary layer to 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 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.

4. Three horses only are going to race in a horse race. A bookmaker gives odds of 3:1, 2:1, and 1:1 for each of the horses to win the race. Use these odds to calculate the corresponding probabilities for each horse to win. By adding up all the probabilities and comparing to one, estimate the overall bias of the bookmaker.

5. Have a look at the probability demos and course notes on www-stat.stanford.edu/~susan/surprise.

4. Probability distributions

1. Read chapter 4 of the lecture notes. Think of questions to ask the lecturer!

2. You can generate samples of random variables from different probability models by using **Calc->Random data->"Probability model"** in MINITAB. Use this to generate samples of 200 discrete random variables from Bernoulli, Binomial, and Poisson distributions and then compare these sample distributions using **Graph->Histogram**. Change the parameters in each of the models to see how they influence the distributions.

3. If the probability of a day being stormy in winter is 1/3, use the binomial distribution to calculate the distribution of number of stormy days in the winter season having 120 days. Write down approximations for the mean and standard deviation of the number of stormy days.

4. Repeat Question 2 but now generate continuous random variables from the Uniform, Normal, and Gamma distributions. In addition to histograms, look at boxplots of the samples of data to see what happens to the quartiles and outliers.

5. Theoretical (as opposed to empirical) probabilities can be calculated in MINITAB for different probability distributions using the **Calc->Probability Distributions->"Probability model"** option . Generate a sequence of values from –10 to 10 in steps of 0.1 using **Calc->Make Patterned Data->Simple set of numbers** and then calculate theoretical probabilities for these values assuming a normal distribution. Use **Graph->Plot** to make plots of the probability density function (p.d.f.) and the cumulative distribution function (c.d.f.).

6. By considering the distribution of numbers of legs per person, explain why you have more than the average number of legs. Draw a probability distribution of leg number to illustrate the situation.

7. What would you guess the chance is of finding at least two people who have their birthdays on the same day in the year out of a sample of 30 people? Write down the maximum probability you would expect.

8. By counting the number of possible ways for each person, calculate the probability Pr(X=0) of no people having their birthdays on the same day of the year out of a sample of 30 people (ignore leap years). Then use Pr(X>0)=1-Pr(X=0), to calculate the probability of one or more people out of 30 having their birthdays on the same day of the year. How does this compare with your guessed answer to question 7? Give a possible reason for any large discrepancies. A nice interactive demonstration of this can be found at www-stat.stanford.edu/~susan/surprise and a clear explanation is provided at www.mste.uiuc.edu/reese/birthday/

5. Parameter estimation

1. Read chapter 5 of the lecture notes. It is essential that you fully understand the concepts of parameter estimation and sampling uncertainty in order to do good science. Think of questions to ask the lecturer!

2. Simulate 10 different samples of 30 heights by using **Calc->Random Data->Normal** with mean 175.0 and standard deviation 8.0. Calculate the sample means of these 10 columns/samples of data by doing **Stat->Basic Statistics->Store Descriptive Statistics** with mean only selected in the **Statistics** options. This will calculate a sample mean for each of the samples and will then store them in columns 11-20.

   Stack these sample means in a new single column using the **Manip->Stack->Stack Columns** option. Now explore the *sampling distribution* of these sample means by making histograms, boxplots, etc.

   Calculate the standard error of your sample means and compare it to what you would expect theoretically for means of 30 normally distributed variables. How many of your sample means lie within the 90% confidence interval expected from theory?

3. Carefully work through the 2 demonstrations on Sampling distributions and Confidence Intervals in the Rice Virtual Laboratory of statistics (see course web page). Read all the instructions and information and complete all the online exercises.

6. Statistical hypothesis testing

1. Hypothesis testing is an area of statistics that many scientists misunderstand – make sure you are not one of them by reading carefully chapter 6 of the lecture notes!

2. The data set **cloud.txt** on the course web site contains data collected from a U.S. experiment in the early 1970s aimed at dropping silver nitrate crystals from aircraft in order to "seed" clouds to make rain. The data set contains the resulting rainfall (in acre-feet!) from a sample of 26 unseeded clouds and a sample of 26 seeded clouds.

   Explore the distributions of the two data samples by plotting histograms and also compare the two samples by plotting unstacked boxplots: **Stats->Anova->One-way (unstacked).** Do their distributions look very different?

   Which statistical test can be used to test the hypothesis that the two samples come from populations with different means? Write down clearly the null and alternative hypotheses. Choose a reasonable level of significance.

   One of the usual assumptions in t-tests is that the data are normally distributed. Do you think this is the case for the rainfall for each of these samples? If not, use a normalizing transformation such as logarithm to transform the rainfall to something more normally distributed (Use **Calc->Calculator** to generate two new columns of data).

   Now perform a test on the means of the two samples using the appropriate test in **Stat->Basic Statistics.** Write down the resulting test statistic and its p-value and compare the p-value with your previously chosen level of significance. What do you conclude from this hypothesis test?

3. Shortly after the introduction of the new euro coins in 2002, BBC online news on 4 Jan 2002 published this article:

   Heads up

   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.

   By using a one-sample z test for normally distributed proportions (see page 40 of the notes) test whether the Belgian euro is biased given the Polish statisticians results for 250 coin tosses. What test statistic do you get and what is its associated p-value? What can you conclude?

4. Read through the hyperstat online help to make sure you fully understand what's going on in hypothesis testing (http://davidmlane.com/hyperstat/logic_hypothesis.html).

7. Basic linear regression

1. Read chapter 7 of the lecture notes. Think of questions to ask the lecturer!

2. Is there a linear time trend in the wintertime mean Central England Temperature?

   a) Read the **atldjf.txt** data into MINITAB and then perform a regression analysis of CET (column 6) on year (column 1) in order to test this hypothesis.

   b) Find the estimated rate of change per year, the coefficient of determination, and the confidence. Does the trend explain a lot of the temperature variance and how significant is the fit?

   c) Carefully examine your linear fit to the Central England Temperature series by making diagnostics of the residuals. Do you think that the CET trend is well described by the linear fit?

   d) Based on your best estimates of the linear fit, what is the predicted value of CET for 2050? Give uncertainty estimates on this prediction of future temperature.

3. Read through the StatSoft online help on regression which nicely illustrates such as things as influential values using dynamic graphics.

8. Multiple and nonlinear regression

1. Read chapter 8 of the lecture notes. Think of questions to ask the lecturer!
2. Extend your linear modelling of CET in data set **atldjf.txt** by performing a multiple regression that uses the mean wintertime sea-level pressures as explanatory variables.

a) Does including Iceland SLP as well as time improve the fit?

b) Use the sea-level pressures in data set **atldjf.txt** to explain the CET response by performing a multiple regression excluding time. Which pressure observations are most important in explaining the CET?

c) Test the sensitivity of the above multivariate regression to the inclusion of time as an explanatory factor. Are pressures alone sufficient for explaining the long-term trend in CET since 1950?

3. Make a new column containing year squared and then do a nonlinear multiple regression for CET that includes both time and time-squared as explanatory variables. Hence determine whether the trend in time is nonlinear.

9. Introduction to time series

1. Read chapter 9 of the lecture notes. Think of questions to ask the lecturer!

2. Read into MINITAB the montly mean time series of Darwin sea-level pressures in file **darwinraw.txt** and reproduce the analyses presented in this chapter of the notes.

3. Using the ARIMA option in MINITAB, fit both an AR(1) and AR(2) time series model to the series and compare the fits.

4. Gaussian white noise is unpredictable. However, backward differences of Gaussian white noise are well correlated with preceeding values. Calculate both theoretically and with some randomly generated numbers the lag-1 autocorrelation for such a

series. Explain why an AR(1) fit to such a series could not be used as a forecasting scheme to predict white noise.

Annex A: Datasets used in the notes and exercises

Dataset 1: rdgmorph.txt

Small biometric survey of nearby colleagues in the meteorology department at the University of Reading taken on 10 Aug 2000.

PERSON AGE in years WEIGHT (cm) HEIGHT (kg)

Dataset 3: atldjf.dat

Some wintertime means of historical time series from around the North Atlantic for the period from 1866 to 1997(Note: 1866 refers to mean of Dec 1865 to Feb 1866).

Columns 2-6 are the mean sea-level pressures observed at certain stations frequently used to construct the North Atlantic Oscillation index. CET is the Central England
Temperature observed at several central stations in England. The pressures can be used to explain the variations in the temperatures. In other words, the central england temperature is a response to changes in explanatory factors related to the atmospheric circulation.

Original time series data can be found on http://www.met.rdg.ac.uk/cag/NAO/ under Indices.

YEAR	ICELAND	AZORES	GIBRALTAR	LISBON	CET
1866	995.1	1021.57	1022.63	1022.37	5.33

Dataset 3: darwin.txt and darwinraw.txt

Monthly mean Sea-level pressure measured at Darwin in northern Australia
from Jan 1951 to Jul 2000 freely obtained from
http://www.cpc.noaa.gov/data/indices/
Note that 1000mb has been subtracted from all values.

YEAR JAN FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC

Darwinraw.txt consists of the 600 consecutive monthly mean values of the pressure at Darwin starting in January 1951. Note that the annual cycle is present.

Annex B: Using Minitab – some essential features

Note: Minitab 13 is used in the ITS Labs, but an earlier version is used on Met PCs. Differences are mentioned as they arise.

To open Minitab: Go to Start→Statistics→MINITAB 13 for windows→MINITAB

To download the files for the practical: Go to the course website at:

http://www.met.rdg.ac.uk/cag/courses/Stats

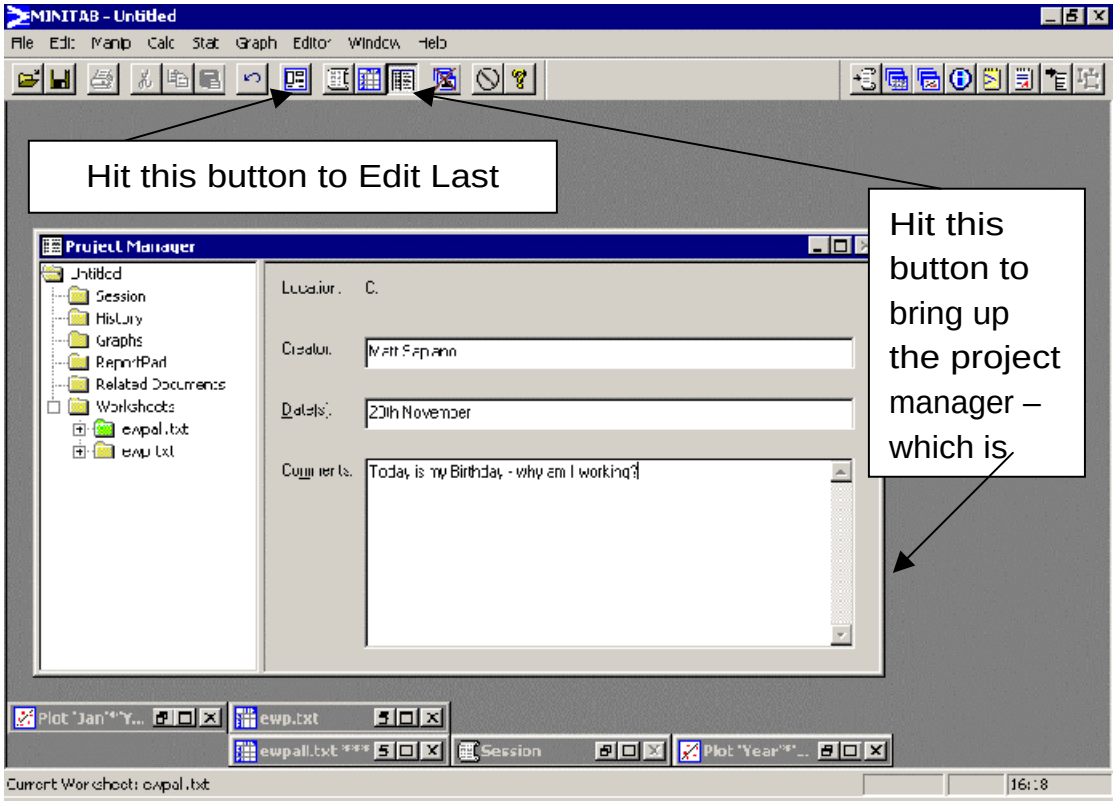
here is a complete list of datasets under the heading of *Practicals*, right click on the link to the dataset, a menu appears, choose **Save target as …** from the menu.

Select the location to download to and save the file

To open the file in Minitab: Choose File→Open worksheet… from Minitab. Change the **Files of type** box to **all**, and select the file. The file can then be saved as a worksheet (*.mtw – Minitab worksheet) or as a project (*.mpj – Minitab projects take up a lot of memory, but store *all* of your work).

The Minitab Project

This is the workspace within Minitab, and all worksheets and graphs are elements of this project. In Minitab 13 (ITS Lab), these elements can be seen with the project manager.

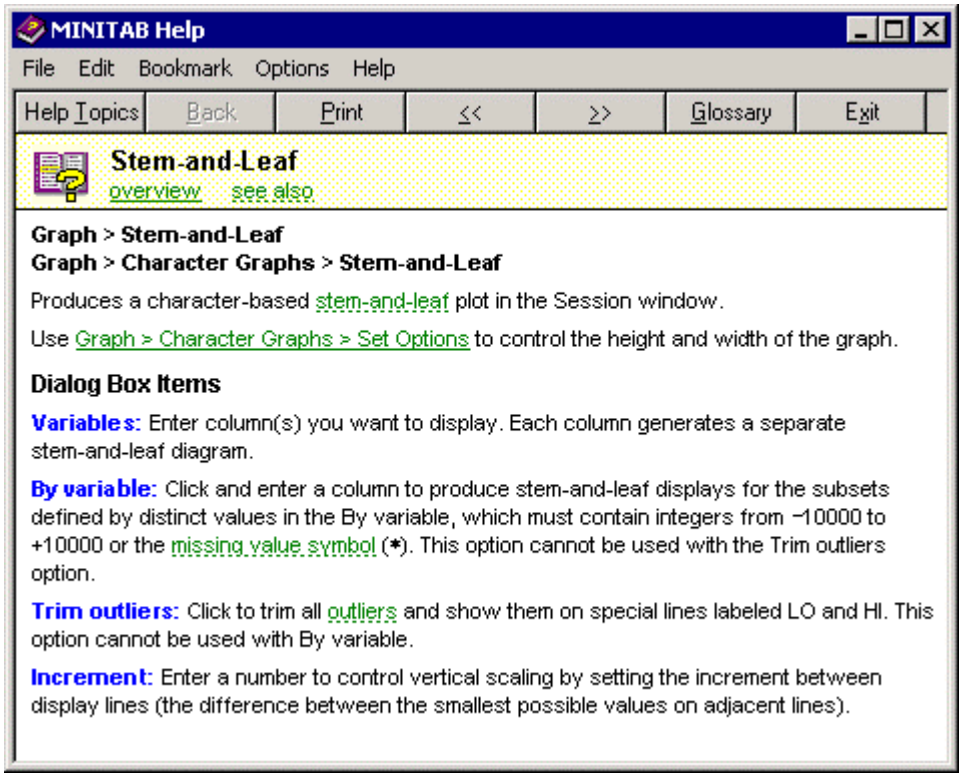


After you have produced some plots and summaries, use the program manager to investigate your project – ask for help if you are unsure of anything.

Using Minitab help

Minitab has the usual windows based help, with a search function. It is most useful within a dialogue box, where there will be a help button at the bottom left of the dialogue box. This gives the help for that box only, with links to relevant explanations.

The help screen below is for the stem and leaf plot dialogue:



It explains all possible functions, with links to other useful parts of help. The overview is about graphs in Minitab, see also gives related subjects etc. When using a dialogue, take a minute to familiarise yourself with the possibilities – you may learn something useful, particularly when producing graphs.

Where now?

The beauty of Minitab is that it is very easy to analyse data very quickly. In fact it is almost too easy – use the help menu to help you understand what you are doing, and check that the analysis makes sense.

Don't be afraid to experiment with Minitab: as with any computer package, the only way to learn how to make full use of it is to mess around with it.

"Many shall run to and fro, and knowledge shall increase"