

Contents

Contents	1
I Background notes	8
1 Equations needed for this module	9
1.1 Basic statistics and histograms	9
1.1.1 Central tendency	9
1.1.2 Variability	10
1.1.3 z-score	10
1.1.4 Symmetry	10
1.1.5 Accuracy and precision	11
1.1.6 Central Limit Theorem	11
1.2 Confidence intervals	12
1.3 Hypothesis testing	14
1.4 Inferences from two samples	15
1.5 Correlation and regression	17
1.6 Propagating errors	18
2 Cumulative probability tables	20
2.1 The standard normal distribution	20
2.2 t-distribution	23
3 Introduction to Excel	26
3.1 Creating a an Excel Worksheet	27
3.2 Getting help	29
3.3 Undoing your work	29
3.4 Saving an Excel File	29
3.5 Doing calculations in Excel	29
3.6 Printing an Excel Worksheet	30
3.7 Closing files and closing Excel	30
3.8 Installing Add-Ins	30
3.9 Cell referencing	31
3.9.1 Relative cell references	31

<i>CONTENTS</i>	2
3.9.2 Absolute cell references	31
3.9.3 Mixed cell references	32
4 Using Excel for different graphs	33
4.1 Using Excel to construct a frequency distribution	33
4.2 Using Excel for Histograms	33
4.3 Using Excel for frequency polygons	35
5 Introduction to Matlab	36
5.1 Overview	36
5.2 The command window	37
5.3 Numbers and arithmetic operations in MATLAB	38
5.4 Vectors, Matrices / arrays	39
5.5 Saving and reloading your work	41
5.6 Help	42
5.7 Demos	43
5.8 Interrupting a running program	43
5.9 Long command lines	44
5.10 Plotting and histograms	44
5.11 Reading data into MATLAB	44
5.12 MATLAB resources on the Internet	44
6 Table of equations and functions	46
References	50
II Statistics exercises	51
1 Week 3 practical	1
1 Week 7 practical	1
1.1 Neat stuff in MATLAB	2
2 Week 8 practical	1
2.1 Excel Method	1
2.2 MATLAB method	0
3 Week 9 practical	1
3.1 Excel procedure for generating 100 simulated samples	2
3.2 MATLAB procedure for generating 100 simulated samples	2
3.3 Fancy things in MATLAB: Function m-file	0

<i>CONTENTS</i>	3
4 Week 10 practical	1
4.1 Overview	1
4.2 Questions asked and statistics	1
4.3 Excel	2
4.4 MATLAB	3
4.5 Fancy stuff in MATLAB	6
5 Week 11 practical	1
5.1 Fitting a power law	1
5.1.1 Excel method	1
5.1.2 MATLAB method	2
5.2 Using your fit to tell you something about Mars	4
5.3 More fancy stuff in MATLAB	5

Preamble

General information

This course on data analysis will enable you to plan and analyse data collected as part of a scientific investigation and arrive at the correct conclusions. Some of the notes have are based on the book by Triola (2009); however some people may prefer the text book by Kirkup (2002). Triola's book however is more relevant to this particular course. However, through experience I have found that Excel can be quite limited and sometimes requires a lot of effort (and patience), therefore I have tried to include information in how to use MATLAB to do the analysis in place of Excel; an excellent reference for those getting started in MATLAB is Hanselman and Littlefield (2001). By all means feel free to look at these books; however, don't expect to find the solutions to the Excel projects: they do not contain them. In order to solve these problems you will have to work at them yourself and may find it necessary to ask myself or the demonstrators for pointers. Also, don't be afraid to speak to each other if you are having problems understanding certain aspects of what we are doing.

A course on data analysis and statistics is seen as a necessity for environmental and earth scientists these days. If it is your desire to published peer reviewed articles on your work then they are subject to criticism if the methods used to summarise your data or arrive at you conclusions are not statistically robust.

I'll mention the text book in class, but my personal opinion is that the notes you receive in lectures are adequate to pass the course; however, if you want to learn more about statistical methods, with lots and lots of examples that are relevant to the environmental sciences, including ANalysis Of VAriance (ANOVA) or statistical process control (which will allow you to say whether global warming is out of statistical control), then by all means get the book.

I will not follow the book exactly, there is far too much information to cover in a lecture, since statistics is quite a large subject. Consequently it is expected that you will read them in your own time. Whether you want to read before the lecture or after is up to you entirely: I wont lecture you on the best way for you to learn! You are all different and all learn in different ways; however, I think you'll agree that it is impossible for me to pour the knowledge into your mind, and so *you* still have to do the work.

In this course there will be no end of semester exam, but you will be tested as you go through the material (see the schedule on the handouts to be sure you don't miss any assignments). The aim of the practicals are for you consolidate your knowledge by 'learning by doing'. If there is one thing I hope you take from this course, it is not to memorise a bank of statistical procedures and formulae. Rather, it is that I hope you'll remember when you come to plan an investigation or analyse data, that there are certain rules that should be adhered to and that you'll be able to look them up and use them throughout your careers. Hence, while I won't lecture

you on note keeping (you're all adults!). I will say that it may be a good idea to save the spreadsheets that you work on in the practicals so that you can come back to them should you need to.

Where will I use statistical methods?

Here are some examples where you may use statistical methods:

- In your project next semester (EART20170)
- In your final year dissertation: projects that use robust statistical methods are generally the ones that attain the highest marks.
- If you go on to get a job in: hazards, research into environmental health, toxicology, forecasting, consultancy, etc.
- Some BIOL modules that environmental science student take have used ANOVA techniques; while I won't be covering it here, the methods developed in this course will make using ANOVA that much easier. I can offer notes and guidance to those interested.

Why use statistical methods?

Basically, because if you can not show that your results are significant or meaningful then why should anyone take any notice or what you are saying?

- You will get better project marks if you are competent at data analysis.
- If you want to publish any results you will have a hard time unless you show the results are significant.

The example I like to use is say you have a bucket full of different coins: 1, 2, 5, 10, 50 pence pieces, say. You are asked to take out 5 coins, one by one at random, and by chance you manage to take out the coins in ascending order (e.g. you take out a 1p, a 2p,...,50p). You could wrongly conclude that the value of the coin you take out depends on the order you take them out. Or course if you were to repeat this experiment again you would have little chance of taking the coins out in order of monetary value.

MATLAB vs Excel

Its no secret that I am a MATLAB advocate, and now python. Python is excellent, but MATLAB is probably the easier language to pick up if this is the first time you've done any programming. I would rarely use Excel for any type of analysis. Excel is very good for keeping track of spreadsheets and accounts—which is what it was designed for. Throughout the course I will give examples in both MATLAB and Excel in the hope that you'll see how powerful MATLAB is. However, it is *mostly* up to you which on you choose to make use of in the practicals. I am pretty sure I wont convert the most stalwart Excel warriors to MATLAB.

Syllabus

EART20170: Data analysis

Semester 1: Lecture Monday 14:00-14:50 weeks 3-5 and 7-10, G.54
 Semester 1: Practical Tuesday 10:00-12:00, Simon 6.06 in weeks 3-4 and 7-10
 Practical Tuesday 11:00-13:00, Simon 6.06 in week 5
 BSc Environmental Science
 Geology
 MEarthSci
 Geochem
 ERG
 BSc Geology)
 Geology with Planetary Science
 Semester 1 is worth 5 credits

There is no class or practical in week 6 (reading week), but there will be a Blackboard assessment.

In semester 2 geology students take part in a mineral deposits project for the other 5 credits, whereas environmental science will complete an environmental project.

You are required to achieve a pass mark in both 1st and 2nd semester parts of the module in order to pass this course.

Instructor for semester 1

Dr. Paul Connolly
 Centre for Atmospheric Science
 School of Earth, Atmospheric and Environmental Sciences
 Simon Building, Room 3.08 (dial 63921 to be admitted)
 paul.connolly@manchester.ac.uk
<http://www.manchester.ac.uk/research/paul.connolly/>
 Phone: +44 (0)161 306 3921

Useful texts

See references at the back.

Assessment

Semester 1 counts for 5 credits. In semester 1, the assessment for the module is as follows:

Assessment	Weighting
Weekly practicals on Blackboard	60%
Week 6 mid-term assessment on Blackboard	40%

Feedback will be given in the practicals and during on-line assessments.

Weekly assessments on Blackboard

Log in to your student portal and then into Blackboard. You should then click on the Blackboard page for this course (EART20170). These assessments will be usually be available for 1 week or so, so important you do them. They will be graded after your first try and you can complete them as often as you like.

Mid-term assessments

These delivered in a similar manner to the weekly assessments. You are not permitted to work in groups for this assessment, but you may use your notes. *The Mid-Term assessment will be available to you from 26th October 13:00 and the deadline for completing it the 8th November at 17:00.*

Nature of assessment

The nature of the assessment is as follows:

- During the practical you will be set a data analysis project. You will be expected to work through the project and submit it to Blackboard—either at the end of the session or during the week. The feedback you will get for these projects will be from the demonstrators and myself in the practical class. Therefore it is important that you come to these classes.

Attendance

Attendance in lectures is compulsory and will be monitored by the undergraduate office. Those not attending will be contacted about their lack of attendance.

Feedback

You will get feedback in several ways:

- By talking to us in classes about your work.
- By taking the online weekly assessments—feedback is given electronically at the end of every assessment.

Schedule

See the web / Blackboard (<http://www.seaes.manchester.ac.uk/study/undergraduate/undergraduate-courses/environmental-science-bsc/course-unit-spec/?unitcode=EART20170>)

Part I

Background notes

Data analysis 1

Equations needed for this module

1.1 Basic statistics and histograms

1.1.1 Central tendency

Revise calculating the mode, median and mean. Formula for the mean:

$$\bar{x} = \frac{\sum_{i=1}^N x_i}{N} \quad (1.1)$$

If you have data tabulated in a frequency table you can calculate the mean by multiplying the frequencies by the mid-points of the bins, adding them up and dividing by the sum of all frequencies. For example:

Example 1.1 *The UK Met.Office (Met.Office) publish climatologies for the number of days of sunshine in the UK. Their data state that, for the month of January, the 30 year average hours of sunshine between 1971 and 2000, had a distribution over the UK and was binned into a histogram with the mid points:*

<i>Mid-points (hrs)</i>	<i>Frequency</i>
20.0	9610
22.5	27343
27.5	14380
32.5	11265
37.5	31694
42.5	4704
47.5	3793
52.5	3623
57.5	2525

What was the mean hours of sunshine for January over this period?

Calculate the sum of the product of bin mid points and frequencies and divide this by the sum of frequencies:

$$\begin{aligned}
 \text{Sum of bin mid points} \times \text{freqs.} &= (20 \times 9610 + \dots + 57.5 \times 2525) = 3.47 \times 10^6 \\
 \text{Sum of frequencies} &= (9610 + 27343 + \dots + 2525) = 108937 \\
 \therefore \text{Mean} &= \frac{347 \times 10^6}{108937} = 31.88 \text{ hours.}
 \end{aligned}$$

1.1.2 Variability

Revise range, interquartile range and standard deviation. Calculating the range:

$$range = x_{largest} - x_{lowest} \quad (1.2)$$

Calculating the standard deviation for a sample of data:

$$s = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N - 1}} \quad (1.3)$$

Calculating the standard deviation for the whole population of data:

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N}} \quad (1.4)$$

Example 1.2 Calculate the standard deviation of the numbers 1,2 and 5.

- in Excel enter in a cell = STDEV(1, 2, 5)
- in MATLAB enter std([1, 2, 5])

The variance is just the standard deviation squared.

$$s^2 = \frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N - 1} \quad (1.5)$$

Example 1.3 Calculate the variance of the numbers 1,2 and 5.

- in Excel enter = VAR(1, 2, 5) or = STDEV(1, 2, 5) ^ 2
- in MATLAB enter var([1, 2, 5]) or std([1, 2, 5]) ^ 2

1.1.3 z-score

It is often useful to know how many standard deviations a value is away from the mean. The formula that gives you this is:

$$z = \frac{x - \mu}{s} \quad (1.6)$$

where x is the value or measurement, μ is the mean and s is the standard deviation.

1.1.4 Symmetry

The symmetry of a dataset can be described using skewness

$$g_1 = \frac{\sqrt{N} \sum_{i=1}^N (x_i - \bar{x})^3}{\left(\sum_{i=1}^N (x_i - \bar{x})^2 \right)^{3/2}} \quad (1.7)$$

If more data is less than the mean than greater than the mean then we say the data is *positively skewed* (the value g_1 is positive) and vice-versa.

Example 1.4 Calculate the skewness of the numbers 1,2 and 5.

- in Excel enter = SKEW(1, 2, 5)
- in MATLAB enter skewness([1, 2, 5], 0)

Another measure of symmetry is kurtosis:

$$g_2 = \frac{N \sum_{i=1}^N (x_i - \bar{x})^4}{\left(\sum_{i=1}^N (x_i - \bar{x})^2\right)^2} - 3 \quad (1.8)$$

Example 1.5 Calculate the kurtosis of the numbers 1,2, 5 and 6.

- in Excel enter = KURT(1, 2, 5, 6) - 3
- in MATLAB enter kurtosis([1, 2, 5, 6], 0)

1.1.5 Accuracy and precision

The *accuracy* of an instrument is related to *systematic error*. It is given by the magnitude of the difference between the actual value of the property being measured and the mean of measured values.

$$accuracy = |\bar{x} - x_{true}| \quad (1.9)$$

The *precision* of an instrument is related to *random error*. In other words you if take a measurement of the same thing with an imprecise instrument you will get a different answer every time. It can be quantified by the standard error of the mean (also called the standard deviation of the sample means), SE . In other words we can improve the precision of a measured property by taking lots of measurements and calculating the mean of those measurements. The standard error of the mean tells us how confident we can be in our measurement.

$$SE_{\bar{x}} = \frac{s}{\sqrt{N}} \quad (1.10)$$

here N is the number of measurements taken and s is the standard deviation. So this means that precision of your measurements can be improved by making more measurements and averaging them together (i.e. increasing the sample size N).

1.1.6 Central Limit Theorem

The standard error of the mean finds application in the *Central Limit Theorem*. The Central Limit Theorem states that if you take a sample of data and calculate the mean and then repeat this lots of time, to create a list of sample means, then those sample means will be normally distributed with a standard deviation equal to the standard error of the mean, $SE_{\bar{x}}$ (which is defined above).

Example 1.6 *The freezing temperatures of a population of cloud drops are normally distributed with a mean of $\mu = -21.33^\circ\text{C}$ and standard deviation of $\sigma = 0.93^\circ\text{C}$. If you select 1 of those drops what is the probability of the freezing temperature being less than -22°C ?. What is the probability of a sample of 10 drops having a mean freezing temperature less than -22°C ?*

1. *Sample from the normal distribution. For the single drop: $\text{normcdf}(-22, -21.33, 0.93)$ in MATLAB or $= \text{NORMDIST}(-22, -21.33, 0.93, 1)$ in Excel. The answer is 0.24 so there is a 24% chance.*
2. *For the sample of 10 drops, the standard deviation is equal to $0.93/\sqrt{10} = 0.29$. Therefore there will be less chance of the sample of 10 drops having a mean freezing temperature less than -22°C : $\text{normcdf}(-22, -21.33, 0.29)$ in MATLAB or $= \text{NORMDIST}(-22, -21.33, 0.29, 1)$ in Excel. The answer is 0.01 so there is only 1% chance.*

Also, because of this concept any error bars that you put on your sample data should have width equal to, $SE_{\bar{x}}$.

1.2 Confidence intervals

Confidence interval on population parameters from sample statistics.

$$\hat{p} - E < p < \hat{p} + E \quad (1.11)$$

where \hat{p} is the sample proportion of successes (e.g. the proportion of people who say yes in a survey) and p is the population proportion (i.e. the proportion of the population who would say yes if they took the survey). E is the margin of error on the estimate.

It can also be written:

$$\hat{p} \pm E \quad (1.12)$$

where

$$E = z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{N}} \quad (1.13)$$

Note that $\hat{q} = 1 - \hat{p}$.

People doing surveys as part of a third year project take note... of this important fact...

You can estimate the size of your sample to have a certain level of confidence in your answer by calculating:

$$N = \frac{[z_{\alpha/2}]^2 \times 0.25}{E^2} \quad (1.14)$$

where, $z_{\alpha/2}$ is the confidence interval at a significance level of $\alpha/2$ and E is the desired margin of error on your answer. This is very important as it tells you how many people you need to ask to be able to infer something about a population.

Example 1.7 How many people should you ask in a survey whether they believe human induced climate change is true, if you want to be 90% confident that you can state the confidence interval on the whole population's opinion to within a margin of error of 20%.

- calculate $z_{\alpha/2}$ for a two tailed distribution at 90% confidence (e.g. $\text{norminv}(0.1/2, 0, 1)$).
- calculate N using Equation 1.14 with $E = 0.20$
- You should have $N \cong 17$, so you need to ask at least 17 people.

Also, if you know the confidence interval and want to get the best estimate of the population proportion and the margin or error you can do this by using the fact that the best estimate is in the middle of the confidence interval and the confidence interval has a width of twice the margin of error.

$$\hat{p} = \frac{\text{upper confidence interval limit} + \text{lower confidence interval limit}}{2} \quad (1.15)$$

$$E = \frac{\text{upper confidence interval limit} - \text{lower confidence interval limit}}{2} \quad (1.16)$$

The confidence interval on population mean from sample mean is calculated in a similar way to that for proportions:

$$\bar{x} - E < \mu < \bar{x} + E \quad (1.17)$$

where \bar{x} is the sample mean and μ is the population mean. E is the margin of error. This is equivalent to:

$$\bar{x} \pm E \quad (1.18)$$

where

$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{N}} \quad (1.19)$$

You can estimate the size of the sample you need to have a certain level of confidence in your answer:

$$N = \left[\frac{z_{\alpha/2} \times \sigma}{E} \right]^2 \quad (1.20)$$

so for instance, say you had a requirement to estimate a population mean to a certain level of accuracy. You would have take N measurements to be able to strictly state this. Note that if you do not know what the population standard deviation, σ , is (and you often wont), then use the sample standard deviation, s , in place of it, and instead of using the normal distribution to calculate $z_{\alpha/2}$, use the t-distribution.

If you know the confidence interval and want to get the best estimate of the population mean and the margin or error you can do this by using the fact that the best estimate is in the middle of the confidence interval and the confidence interval has a width of twice the margin of error.

$$\mu = \frac{\text{upper confidence interval limit} + \text{lower confidence interval limit}}{2} \quad (1.21)$$

$$E = \frac{\text{upper confidence interval limit} - \text{lower confidence interval limit}}{2} \quad (1.22)$$

1.3 Hypothesis testing

Test statistic for proportion:

$$z = \frac{\hat{p} - p}{\sqrt{pq/N}} \quad (1.23)$$

Example 1.8 A sample of 20 people surveyed are asked whether they think it rains all the time in Manchester. The proportion who said yes was 0.7. Test the hypothesis that in the population, the proportion of people who think it rains all the time is larger than 0.5 i.e. more people think it rains more than those who dont.

- the null hypothesis is that the sample proportion is close enough to 0.5 to be considered the same.
- Calculate the z statistic above $\frac{0.7-0.5}{\sqrt{0.5 \times 0.5/20}} = 1.79$.
- Assign a confidence level (how confident you want to be), e.g. 90%, so significance level is $100 - 90 = 10\%$.
- Calculate the z_α for a one-tailed test (e.g. $\text{norminv}(0.1, 0, 1)$), which should equal -1.28 —ignore the sign.
- The z -statistic is greater than the critical value, so we can reject the null hypothesis. Hence, we accept the alternate hypothesis which is more people think it rains all the time than do not.

Test statistic for mean:

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{N}} \quad (1.24)$$

or if you do not know the population standard deviation (often the case):

$$t = \frac{\bar{x} - \mu}{s / \sqrt{N}} \quad (1.25)$$

with $N - 1$ degrees of freedom.

Example 1.9 An EU directive states that PM10 loadings (particulate matter with size less than $10\ \mu\text{m}$) must not exceed an mean over a 24 hour period of $0.05\ \text{mg}\ \text{m}^{-3}$ (milli grams of PM10 per cubic metre of air). 20 spot measurements are made over a 24 hour period which have a mean of 0.055 and a sample standard deviation of 0.04. Has the PM10 exceeded the directive?

- We can choose the null hypothesis to be that the sample mean is close enough to be considered the same. The alternate hypothesis is that they are different.
- Calculate the t statistic above $\frac{0.055-0.05}{0.04/\sqrt{20}} = 0.56$.
- Assign a confidence level (how confident you want to be), e.g. 90%, so significance level is $100 - 90 = 10\%$.
- Calculate the t_α for a one tailed test (e.g. `tinvt(0.1, 19)` in MATLAB or `= tinvt(0.1 * 2, 19)` in Excel), which should equal -1.33 —ignore the sign.
- The t -statistic is less than the critical value, so we can accept the null hypothesis, which is the sample mean is not significantly different to 0.05. Hence we cannot say that there has been an exceedance.
- To the untrained scientist, the fact that the sample mean exceeds 0.05 might seem like evidence there has been an exceedance, but the standard deviation is quite high, so there is not enough evidence to be conclusive.

1.4 Inferences from two samples

Note, those doing surveys might want to compare different groups—e.g. Men vs Women, both asked the same yes/no question.

Usually you will state the null hypothesis: e.g. H_0 : men have different opinions than women.

$$z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}}} \quad (1.26)$$

$$\hat{p}_1 = \frac{x_1}{n_1} \quad (1.27)$$

$$\hat{p}_2 = \frac{x_2}{n_2} \quad (1.28)$$

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} \quad (1.29)$$

$$\bar{q} = 1 - \bar{p} \quad (1.30)$$

where x_1 is the number of successes in the 1st sample (i.e. the number of people saying yes to a question) and n_1 is the size of the 1st sample group and x_2 and n_2 are the same for the 2nd sample group.

Example 1.10 A survey is conducted in which a question is "yes/no would you buy an electric car?". The data can be stratified into both male and female responses. 10 females were asked and 5 males. The proportion of females that said "yes" was 60% (i.e. 6 out of the 10 asked), while the proportion of males who said yes was 20% (i.e. 1 out of the 5 asked). Is there a difference of opinion between males and females?

- The null hypothesis is that there is no difference of opinions between groups.
- $\bar{p} = \frac{6+1}{15} = 0.4667$ and $\bar{q} = \frac{4+4}{15} = 0.5333$
- Calculate the z -value above: $z = \frac{0.6-0.2}{\sqrt{\frac{0.4667 \times 0.5333}{10} + \frac{0.4667 \times 0.5333}{5}}} = 1.4638$
- Assign a significance level, say 0.05 (i.e. we want to be 95% confident in our claim), and calculate $z_{\alpha/2}$ (i.e. $\text{norminv}(0.05/2, 0, 1)$), which should equal -1.96 (ignore the sign).
- Since the z -statistic is less than the critical value of z we accept the null hypothesis. There is not enough data to say the two groups have different opinions.
- If the question we asked was is the proportion of women who would buy an electric car larger than men, we would have to conduct a one-tailed test. Make sure you know what the differences would be.

If you want to test if collected data sets have the same population mean:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} \quad (1.31)$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)} \quad (1.32)$$

with $n_1 + n_2 - 2$ degrees of freedom.

These statistics are then either compared to a normal or t-distribution at some significance level, depending, and if the statistic is large compared to the distribution we reject the null hypothesis. If it is less than the value from the distribution we accept the null hypothesis.

Example 1.11 You observe the freezing of two samples of 100 cloud drops. One consists of pure water only, and the other contains 10 g L^{-1} of volcanic ash. The sample with pure water freezes with a mean freezing temperature of $\bar{x}_1 = -21.33^\circ\text{C}$ and a standard deviation of $s_1 = 0.93$, whereas the sample with volcanic ash freezes with a mean freezing temperature of $\bar{x}_2 = -18.20^\circ\text{C}$ and standard deviation of $s_2 = 0.88$. Does volcanic ash affect the freezing temperature of cloud drops?

- The null hypothesis is that there is no volcanic ash does not affect the freezing temperature.
- Calculate the t -value above. You will first have to calculate S_p^2 and then t , the answer should be $S_p^2 = 0.82$ and $t = -24.45$.
- Assign a significance level, say 0.01 (i.e. we want to be 99% confident in our claim), and calculate $t_{\alpha/2}$ (i.e. `tinvs(0.01/2, 198)` in MATLAB or `tinvs(0.01, 198)` in Excel), which should equal -2.60 (ignore the sign).
- Note because we are interested if the two are different it is a two-tailed test.
- Since the t -statistic much larger (in magnitude) than the critical value of t we reject the null hypothesis. Volcanic ash clearly affects the freezing of cloud drops.
- If we asked the question is the mean freezing temperature of the pure drops lower than the volcanic ash laden drops we would have to conduct a one-tailed test. Make sure you understand what the difference in the procedure would be.

1.5 Correlation and regression

In order to test whether the population correlation coefficient is equal to zero (no correlation).

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \quad (1.33)$$

with $n - 2$ degrees of freedom. Here r is the correlation coefficient.

In order to fit a straight line of the form $y = b_0 + b_1 \times x$:

$$b_1 = r \frac{s_y}{s_x} \quad (1.34)$$

$$b_0 = \bar{y} - b_1 \bar{x} \quad (1.35)$$

Note that not all equations you fit will be linear, you should think about what form of equation should best fit the data.

For instance, if you were looking at the freezing of cloud drops to form ice crystals, then the theory suggests that $N_{drop}(t) = N \exp(-Jt)$, so if you were to fit a curve you should fit a linear-exponential curve. The same method can be used if you fit $N_{drop}(t)$ vs $\exp(t)$.

Also, some data might fit a power law, $y = ax^b$. For example, the terminal velocity of water drops in air vs their size or similarly sediments in free-fall in a fluid. Again, linear regression can be used in the following way:

1. Take logs of both sides:

$$\begin{aligned}\ln y &= \ln(ax^b) \\ \ln y &= \ln a + b \ln x\end{aligned}$$

2. This is just the same as fitting a straight line if we fit the natural logarithm of y against the natural logarithm of x .
3. The gradient will be b and the intercept will be $\ln a$. Therefore a can be found by raising e to the power of the intercept.
4. This seems hard but it is quite easy if you get your head round it.

1.6 Propagating errors

This is not assessed. If you make two measurements x and y , which have independent errors σ_x and σ_y , and use those measurements to calculate a new number, z , the corresponding error in z , σ_z is shown in the table below.

Relationship	Error propagation
$z = x + y$	$\sigma_z^2 = \sigma_x^2 + \sigma_y^2$
$z = x - y$	$\sigma_z^2 = \sigma_x^2 + \sigma_y^2$
$z = x \times y$	$\left(\frac{\sigma_z}{z}\right)^2 = \left(\frac{\sigma_x}{x}\right)^2 + \left(\frac{\sigma_y}{y}\right)^2$
$z = \frac{x}{y}$	$\left(\frac{\sigma_z}{z}\right)^2 = \left(\frac{\sigma_x}{x}\right)^2 + \left(\frac{\sigma_y}{y}\right)^2$
$z = k \times x$	$\sigma_z = k \times \sigma_x$
$z = x^n$	$\frac{\sigma_z}{z} = n \frac{\sigma_x}{x}$
$z = \ln(x)$	$\sigma_z = \frac{\sigma_x}{x}$
$z = \exp(x)$	$\frac{\sigma_z}{z} = \sigma_x$

Example 1.12 You make a measurement of the length of along strike of a mineralised seam. Your tape measure is 10 m long, which is just shorter than the seam so you measure the length in two steps. The two readings are 1000 ± 1 cm and 403 ± 1 cm. As the precision of the tape measure is $\sigma \cong \pm 1$ cm, propagate the error in your measurement.

- Use the first formula in the table—i.e. for $z = x + y$; $\sigma_z^2 = \sigma_x^2 + \sigma_y^2$.
- This means that the error in z can be calculated as: $\sigma_z = \sqrt{1^2 + 1^2} \cong 1.41$ cm.

Example 1.13 You want to estimate the mass loading of aerosol between 2.5 and 10 μm . You have two separate instruments that measure either PM2.5 or PM10 (i.e. the total number less than 2.5 μm and the total number less than 10 μm). Both instruments have quoted precisions of $\sigma = 0.01 \text{ mg m}^{-3}$. Your measurements are PM2.5 = $0.02 \pm 0.01 \text{ mg m}^{-3}$ and PM10 = $0.05 \pm 0.01 \text{ mg m}^{-3}$. What is the mass loading between 2.5 and 10 μm ?

- Use the second formula in the table—i.e. for $z = x - y$; $\sigma_z^2 = \sigma_x^2 + \sigma_y^2$.
- This means that the error in z can be calculated as: $\sigma_z = \sqrt{0.01^2 + 0.01^2} \cong 0.0141$.
- So the value you quote should be $0.03 \pm 0.014 \text{ mg m}^{-3}$.

Example 1.14 You want to estimate the density of a rock. You do this by independently measuring the mass of the rock and also its volume. In order to measure the rock's mass you place on the scales and get $M = 30000 \pm 10 \text{ g}$ (10 is the quoted precision). In order to measure the rock's volume, you place the rock in a rectangular tub of water, that has an area of $2500 \pm 10 \text{ cm}^2$, and note the increase in height level of the water. Before the rock is placed in the water the height of the water in the tub is $30 \pm 1 \text{ cm}$ and after it is $33 \pm 1 \text{ cm}$.

- This calculation has several points where the errors need to be considered and propagated.
- Firstly, consider the change in height level of the water $l = 33 \pm 1 - (30 \pm 1)$. To propagate this error use the second formula in the table: $\sigma_l = \sqrt{1^2 + 1^2} = 1.41 \text{ cm}$.
- Secondly consider the volume of the rock $v = l \times A = 7500 \text{ cm}^3$. To propagate this error use the 3rd formula in the table: $\left(\frac{\sigma_z}{z}\right)^2 = \left(\frac{\sigma_x}{x}\right)^2 + \left(\frac{\sigma_y}{y}\right)^2$, so $\sigma_v = v \sqrt{\left(\frac{\sigma_l}{l}\right)^2 + \left(\frac{\sigma_A}{A}\right)^2}$. So $\sigma_v = 7500 \sqrt{\left(\frac{1.41}{3}\right)^2 + \left(\frac{10}{2500}\right)^2} \cong 3.5 \times 10^3 \text{ cm}^3$.
- Finally, the density of the rock is calculated by dividing the mass by the volume, $\rho = \frac{M}{v}$. To propagate the error use the fourth formula in the table, which is the same form as above $\sigma_\rho = \rho \sqrt{\left(\frac{\sigma_M}{M}\right)^2 + \left(\frac{\sigma_v}{v}\right)^2}$. So $\sigma_\rho = 4 \sqrt{\left(\frac{10}{30000}\right)^2 + \left(\frac{3.5 \times 10^3}{7500}\right)^2} \cong 1.87 \text{ g cm}^{-3}$.
- So the answer you should quote is $\rho = \frac{30000}{7500} \cong 4 \pm 1.87 \text{ g cm}^{-3}$.

Example 1.15 You wish to estimate the volume of a cuboid with side $l = 10 \pm 1 \text{ cm}$.

- Use the sixth formula in the table: i.e. for $z = x^n$ use $\frac{\sigma_z}{z} = n \frac{\sigma_x}{x}$.
- Therefore $z = 10^3 = 1000$ and $\sigma_z = z \times n \frac{\sigma_x}{x} \cong 300 \text{ cm}^3$.
- Quote your answer as $z = 1000 \pm 300 \text{ cm}^3$.

Data analysis 2

Cumulative probability tables

2.1 The standard normal distribution

The tables below show the cumulative probability for a given number of standard deviations away from the mean, z . Read the tables as follows: if you want to know the cumulative probability for a value of z (to up to 2 dps), find the nearest value of z in the column for z , then count across the columns for the second decimal place. The value in the table that corresponds to this row and column is the cumulative probability.

For example: what is the cumulative probability for $z = -2.05$? *Answer: go down the 1st column until you get to a z -value of -2.0 then go along the rows until the top column reads 0.05: this means $z = -2.05$. Then read of the probability, which is 0.0202.*

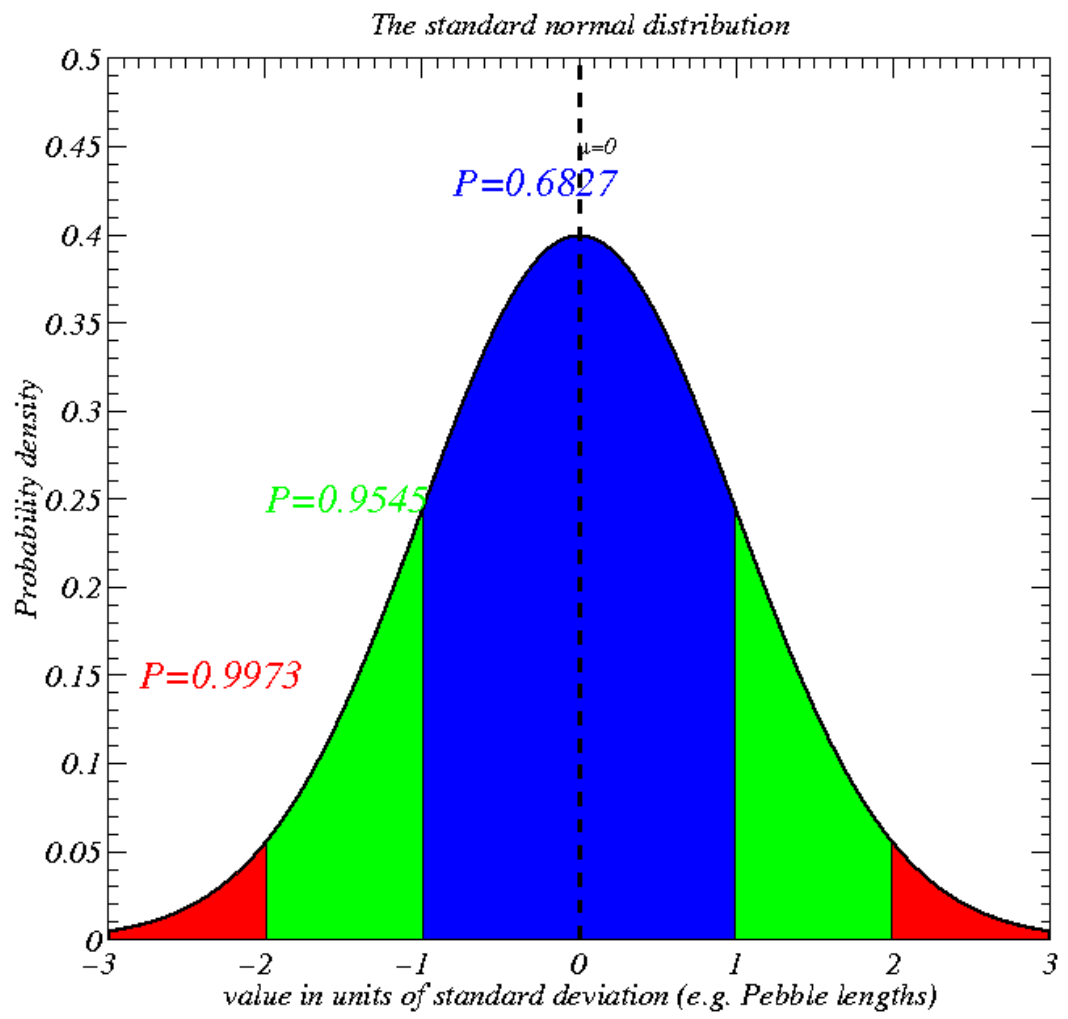
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

The same concept applies for the right side of the distribution. For example: what is the cumulative probability for a value of $z = 2.05$? *Answer: go down the 1st column until you get to a z-value of 2.0 then go along the rows until the top column reads 0.05: this means $z = 2.05$. Then read of the probability, which is 0.0202.*

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998

IMPORTANT: you can use these tables for other course, reports, etc; however, for this course I'd like you to either use Excel, or Matlab to calculate the cumulative probabilities. The functions are:

- Excel: =normdist($z, \mu, \sigma, 1$)—calculates the cumulative probability of the normal distribution for a z-score, z .
- MATLAB: normcdf(z, μ, σ)—calculates the cumulative probability of the normal distribution for a z-score, z .
- Excel: =norminv(P, μ, σ)—calculates the z-score for a cumulative probability, P , for the normal distribution.
- MATLAB: norminv(P, μ, σ)—calculates the z-score for a cumulative probability, P , for the normal distribution.

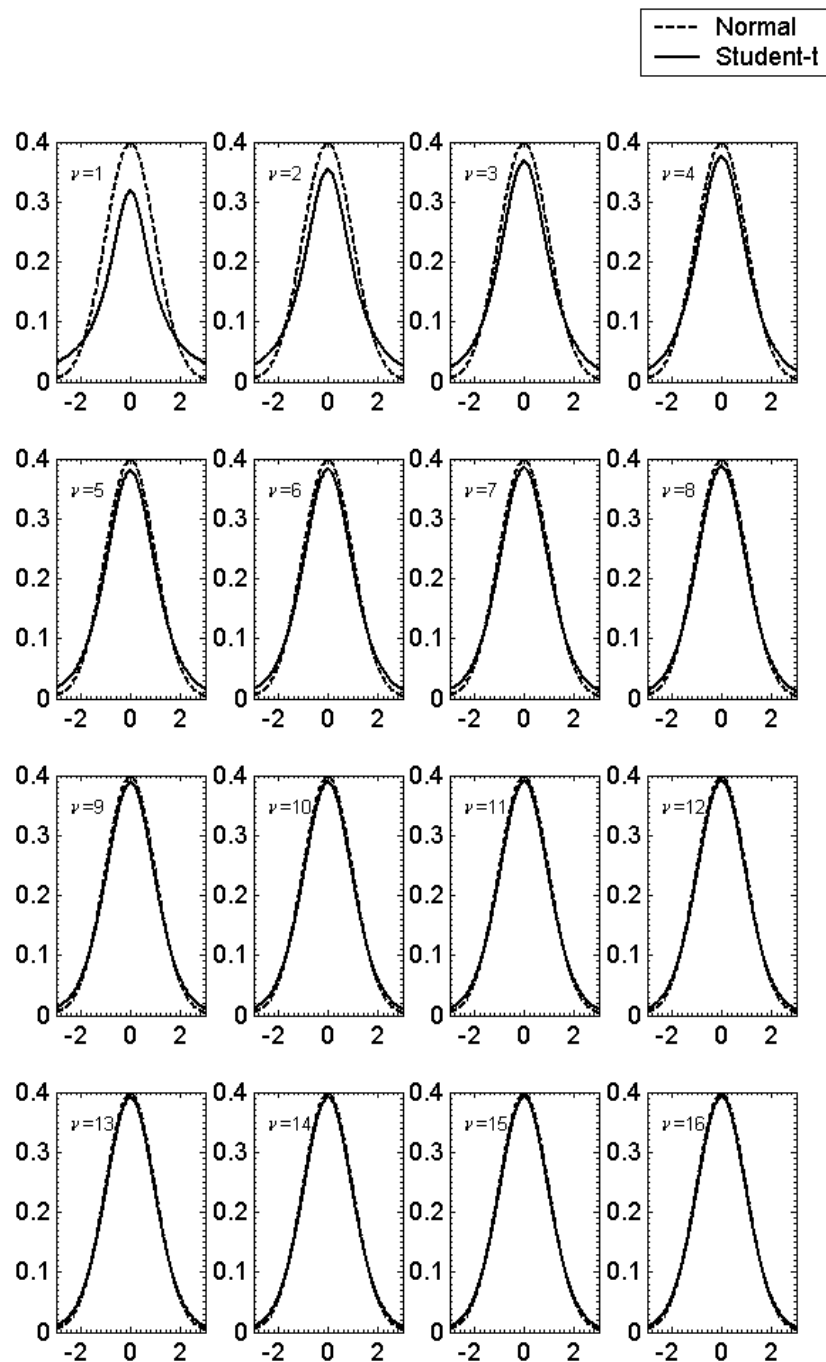


2.2 t-distribution

The table of the t-distribution is presented in a slightly different way to the normal distribution. For a given probability and degrees of freedom it gives the t-score (i.e. distance from the mean). For example: if degrees of freedom is = 10, what is the t-score where the cumulative probability (i.e. area to the left of t-score) is 0.025? *This is a one tailed example, so find the column where the "area in one tail" is 0.025 and the row for df = 10. This gives t = 2.228.*

Degrees of freedom	Area in one tail				
	0.005	0.01	0.025	0.05	0.10
	Area in two tails				
	0.01	0.02	0.05	0.10	0.20
1	63.657	31.821	12.706	6.314	3.078
2	9.925	6.965	4.303	2.920	1.886
3	5.841	4.541	3.182	2.353	1.638
4	4.604	3.747	2.776	2.132	1.533
5	4.032	3.365	2.571	2.015	1.476
6	3.707	3.143	2.447	1.943	1.440
7	3.499	2.998	2.365	1.895	1.415
8	3.355	2.896	2.306	1.860	1.397
9	3.250	2.821	2.262	1.833	1.383
10	3.169	2.764	2.228	1.812	1.372
11	3.106	2.718	2.201	1.796	1.363
12	3.055	2.681	2.179	1.782	1.356
13	3.012	2.650	2.160	1.771	1.350
14	2.977	2.624	2.145	1.761	1.345
15	2.947	2.602	2.131	1.753	1.341
16	2.921	2.583	2.120	1.746	1.337
17	2.898	2.567	2.110	1.740	1.333
18	2.878	2.552	2.101	1.734	1.330
19	2.861	2.539	2.093	1.729	1.328
20	2.845	2.528	2.086	1.725	1.325
21	2.831	2.518	2.080	1.721	1.323
22	2.819	2.508	2.074	1.717	1.321
23	2.807	2.500	2.069	1.714	1.319
24	2.797	2.492	2.064	1.711	1.318
25	2.787	2.485	2.060	1.708	1.316
26	2.779	2.479	2.056	1.706	1.315
27	2.771	2.473	2.052	1.703	1.314
28	2.763	2.467	2.048	1.701	1.313
29	2.756	2.462	2.045	1.699	1.311
30	2.750	2.457	2.042	1.697	1.310
31	2.744	2.453	2.040	1.696	1.309
32	2.738	2.449	2.037	1.694	1.309
33	2.733	2.445	2.035	1.692	1.308
34	2.728	2.441	2.032	1.691	1.307
35	2.724	2.438	2.030	1.690	1.306
36	2.719	2.434	2.028	1.688	1.306
37	2.715	2.431	2.026	1.687	1.305
38	2.712	2.429	2.024	1.686	1.304
39	2.708	2.426	2.023	1.685	1.304
40	2.704	2.423	2.021	1.684	1.303
45	2.690	2.412	2.014	1.679	1.301
50	2.678	2.403	2.009	1.676	1.299
60	2.660	2.390	2.000	1.671	1.296
70	2.648	2.381	1.994	1.667	1.294
80	2.639	2.374	1.990	1.664	1.292
90	2.632	2.368	1.987	1.662	1.291
100	2.626	2.364	1.984	1.660	1.290
200	2.601	2.345	1.972	1.653	1.286
300	2.592	2.339	1.968	1.650	1.284
400	2.588	2.336	1.966	1.649	1.284
500	2.586	2.334	1.965	1.648	1.283
1000	2.581	2.330	1.962	1.646	1.282
2000	2.578	2.328	1.961	1.646	1.282
∞	2.576	2.326	1.960	1.645	1.282

Note that students t-distribution is practically the same as a normal distribution for large degrees of freedom, ν .



IMPORTANT: you can use this table for other course, reports, etc; however, for this course I'd like you to either use Excel, or Matlab to calculate the t-scores. The

functions are:

- Excel: `=tdist(t, df, 1)`—calculates the cumulative probability of the t distribution for a t-score, t .
- MATLAB: `tcdf(t, df)`—calculates the cumulative probability of the t distribution for a t-score, t .
- Excel: `=tinv(P*2, df)`—calculates the t-score for a cumulative probability, P , for the t distribution. To calculate the t-score for two-tails do `=tinv(P, df)`
- MATLAB: `tinv(P, df)`—calculates the t-score for a cumulative probability, P , for the t distribution. To calculate the t-score for two-tails do `tinv(P/2, df)`

Data analysis 3

Introduction to Excel

Microsoft Excel is a widely used program commonly found throughout business and industry. Job applicants with a knowledge of Excel enjoy an advantage over those without that knowledge. Excel is a spreadsheet program that includes procedures for calculations and graphs; a *spreadsheet* is a popular way to organize data.

A spreadsheet is a collection of data organized in an array of cells arranged in rows and columns, and it is used to summarize, analyze, and perform calculations with the data.

OFFICE BUTTON: Click on this button for Excel options, including opening a file, saving a file, and printing a worksheet.

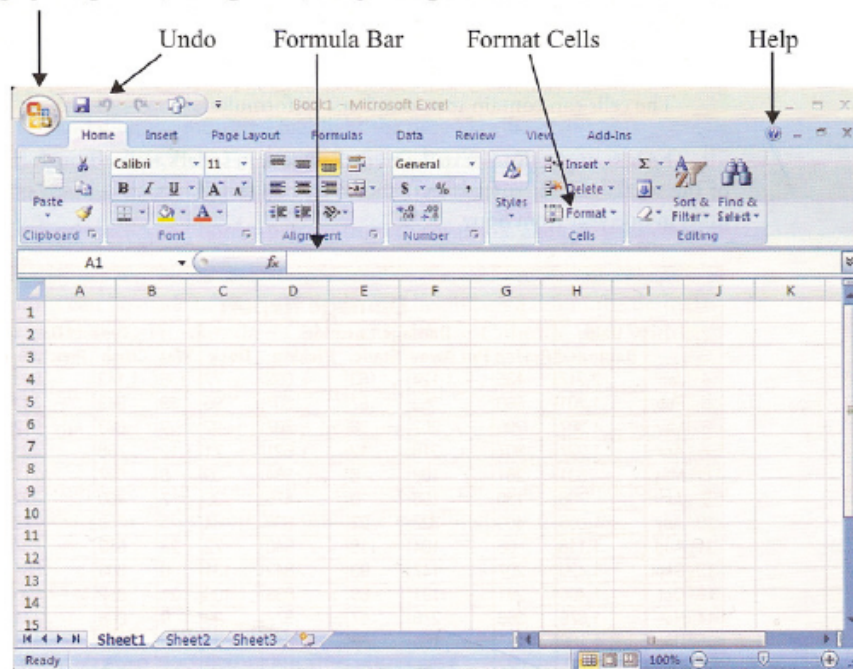


Figure 3.1: Spreadsheet in Excel 2007

Figures 3.1 and 3.2 show the basic format of an Excel spreadsheet. The body of the spreadsheet consists of cells in columns labeled A, B, C, ... , and rows labeled 1, 2, 3, ... An individual spreadsheet is part of a larger Excel structure, described as follows.

Excel file = Workbook = 3 worksheets

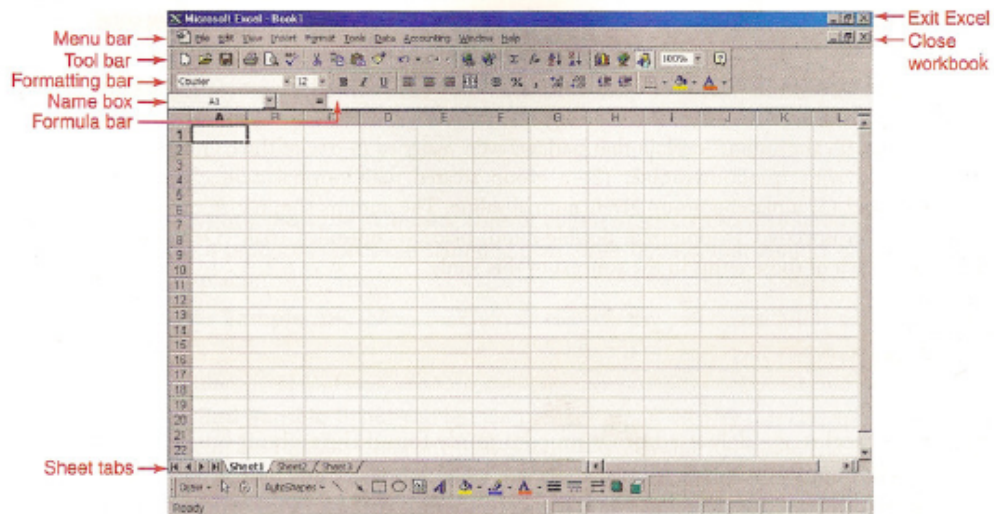


Figure 3.2: Spreadsheet in Excel 2003

(see Figure 3.1 or Figure 3.2)

Although Excel is actually a spreadsheet program, we typically refer to worksheets instead of spreadsheets. (The number of worksheets is 3 by default, but you can insert additional worksheets or delete existing worksheets.)

An Excel worksheet is one page or sheet (or spreadsheet), consisting of cells arranged in an array of rows and columns (as in Figure 3.1 or Figure 3.2). The cells can contain text, numbers, or formulas.

The best way to learn Excel is to actually use it, so let's walk through the creation of the worksheet (or spreadsheet) shown in Figure 3.3, which is based on damaged units in one year from Sony.

3.1 Creating a an Excel Worksheet

1. Launch Excel to get a blank worksheet. This typically requires that you double-click on the Excel icon. Shown below are the Windows desktop icons for Excel 2007 and Excel 2003.



(Excel should be installed on your computer. If it is not installed, follow the directions provided with the software.)

	A	B	C	D	E	F	G	H	I	J	K
1	Damage Report										
2		Units	Damage Location				Type of Damage				
3		Damaged	Rcving	Put Away	Static	Picking	Dock	Wet	Crush	Puncture	Scuffing
4	Jan	2,217	135	1,124	183	698	77	136	1,142	903	36
5	Feb	1,701	560	352	87	607	95	39	943	442	277
6	Mar	1,387	295	257	36	707	92	2	863	512	10
7	Apr	1,227	307	216	12	621	71	12	768	428	19
8	May	831	281	109	8	394	39	8	597	211	15
9	June	798	130	158	0	475	35	7	460	322	9
10	July	1,506	405	327	22	668	84	5	758	743	0
11	Aug	1,118	96	184	118	648	72	23	539	487	69
12	Sept	1,420	207	147	93	842	131	0	806	567	47
13	Oct	1,495	291	191	69	835	109	21	852	590	32
14	Nov	1,676	365	278	57	877	99	0	866	774	36
15	Dec	1,186	263	155	38	643	87	0	481	673	32
16	Total	16,562	3,335	3,498	723	8,015	991	253	9,075	6,652	582

Figure 3.3: Excel Spreadsheet of a Damage Report

2. Click on the cell in the upper left corner and type "Damage Report" and , then press the Enter key. To modify the title so that it appears as shown in Figure 3.3, follow these steps.
 - (a) Click on the inside of the cell and, while holding the mouse button down, drag the mouse to the right so that columns A through K in row 1 are all highlighted. Release the mouse button.
 - (b) Click on the menu item of Format, select Format Cells (or Cells), select Alignment, then proceed to select the Horizontal option of Center. Also, click on the box to the left of Merge cells. Click OK.
 - (c) Click on the menu item of Format Cells (or Format), select Format Cells (or Cells), select Font and proceed to select the Font style of Bold and the Size of 14. Click OK.
3. Click on the fourth cell in the first column and enter "Jan." Proceed to enter the other data shown in Figure 3.3. The titles of "Damage Location" and "Type of Damage" can be modified by using the same procedures described in parts a, b, and c of Step 2. The bold borders can be created as follows:
 - (a) Click and drag the mouse to highlight the desired cells.
 - (b) Click on Format, then Format Cells (or Cells), then Border.
 - (c) Click on the desired line Style.
 - (d) Click on the desired border pattern from the available choices, then click OK.
4. Columns can be made narrower or wider by clicking on a cell in the desired column, selecting Format, then Column Width (or Column, then Width). Enter a value for the desired width, such as 15. You may have to experiment with

different values to find your desired column width. After creating the Excel worksheet shown in Figure 3.3, we could save the worksheet, print it, edit it, perform calculations, do statistical analyses, and construct graphs. Some of these important functions will be discussed in the chapters that follow. For now, we consider a few of the more important general functions of Excel.

3.2 Getting help

To get help while you are running Excel: In Excel 2007, click on the question mark icon located on the top menu bar. In Excel 2003, click on Help located to the right of the menu bar at the top, then click on the question mark icon.

3.3 Undoing your work

When using Excel, it often happens that you do something terribly wrong and you wish that you hadn't done it. Great news: You can undo your last action. Simply click on the counterclockwise arrow located on the menu bar. It looks like this:



3.4 Saving an Excel File

1. In Excel 2007, click on the large round Office Button in the upper left corner, as shown in Figure 3.1; in Excel 2003, click on File. Select the option of Save As.
2. You should now see a dialog box labeled "Save As."
 - (a) In the "Save in" box at the top, select the directory and folder in which you want your file saved. To change the drive that -is shown, click on the dropdown button and then click on the desired location, such as (C:). (If you select C, you will get a list of folders available in directory C, and you should click on the desired folder, such as "My Documents.")
 - (b) Enter a name (such as "Damages") in the File name box.
 - (c) In the "Save as type" box, select Excel Workbook (or Microsoft Excel Workbook) .
 - (d) Click on Save.

3.5 Doing calculations in Excel

By entering an equal sign followed by an arithmetic expression in the formula bar, you can perform calculations. Because the standard keyboard does not have special keys for operations such as division, exponents, and square roots, we must use the keys or commands designated for those purposes. Here are a few of the more common operations, with their Excel expressions:

Operation	Excel Expression	Excel Example	Result
Addition	+	=2+3	5
Subtraction	-	=6-4	2
Multiplication	*	=3*4	12
Division	/	=12/3	4
Exponentiation	^	=2^3	8
Square root	SQRT()	=SQRT(49)	7
Absolute value	ABS()	=ABS(-5)	5

Using operations such as those listed, we can use Excel to evaluate arithmetic expressions as follows.

1. Click on an empty cell or the formula bar.
2. Enter an equal sign (=) followed by the desired arithmetic expression.
3. Press the Enter key.

Whether using calculators or computer software, it often happens that the result of an arithmetic expression is given in a form of scientific notation, and we should be able to express such results in a standard format. See the following two examples.

Math Expression	Excel Expression	Excel Result	Result in Standard Format
0.2^{10}	=0.2^10	1.024E-07	0.0000001024
3^{25}	=3^25	8.47289E+11	847,289,000,000

3.6 Printing an Excel Worksheet

To print all of the cells in the current worksheet: In Excel 2007, click on the Office Button, then select Print; in Excel 2003, click on File, then select Print.

3.7 Closing files and closing Excel

It is the end of the day, your work is done, you have saved your Excel file, and you now want to close your Excel file so that you can proceed to shut down the whole Excel program, shut down the computer, turn out the lights, and go home to feed your pet iguana.

- Closing an Excel file: In Excel 2007, click on the Office Button in the upper left corner, then select Close; in Excel 2003, click on File, then click on Close.
- Closing the entire Excel program: Simply click on the X box located in the upper right corner of the screen.

3.8 Installing Add-Ins

In the following lectures, we will describe the use of Excel for a variety of different statistical functions. Those functions typically come from two sources:

1. Click on *fx* found on the tool bar. You will have access to many built-in functions.
2. Use the Data Analysis set of statistical procedures. In Excel 2007, click on Data, then click on Data Analysis; in Excel 2003, click on Tools, then click on Data Analysis.

Instructions for installing Excel add-ins are available through Excel's Help feature. In Excel 2007, go to the Help feature by clicking on the question mark icon (?) located near the upper-right corner; in Excel 2003, click on Help. In the Search box, enter "add-in" and press the Enter key. Follow Excel's instructions for installing add-in programs. Select **Analysis ToolPak**.

3.9 Cell referencing

You can reference the values or text within a Excel cell by using what is referred to as either *absolute, relative or mixed cell references*, these can save a lot of time and are best illustrated with examples.

3.9.1 Relative cell references

As a first example enter the number 0 in cell A1 of an Excel spreadsheet. Now in cell A2 enter =A1+1, and you should find that the number displayed in cell A2 is 1. Click on cell A2 and then copy it's contents (ctrl-c), then go down to cell A10 and click on that cell with the mouse; hold the keys ctrl + shift and press the up arrow key; you should find that cells A2 to A10 are now highlighted. Now paste the contents of what you copied by pressing ctrl-v. You should find that the numbers 0-9 are in cells A1 to A10. The reason being is that you told each cell below A1 to be equal to the number in the cell above plus 1.

Excel 'knew' to add 1 to the cell above it because we used relative cell references; that is we referenced the cells using only the letter corresponding to the column and the number corresponding to the row.

Relative cell references are useful when you want to multiply (or divide or add or subtract) two lists of numbers to make a third list.

3.9.2 Absolute cell references

The other main type of cell reference is an absolute cell reference, again best illustrated with an example.

As an example enter the number 0 in cell A1 of an Excel spreadsheet. Now in cell A2 enter =\$A\$1+1, and you should find that the number displayed in cell A2 is 1. Click on cell A2 and then copy it's contents (ctrl-c), then go down to cell A10 and click on that cell with the mouse; hold the keys ctrl + shift and press the up arrow key; you should find that cells A2 to A10 are now highlighted. Now paste the contents of what you copied by pressing ctrl=v. You should find that the number 0 is in cell A1 and the number 1 is in cells A2 to A10. The reason being is that you told each cell below A1 to be equal to the number in the cell A1 plus 1, which is always 1.

Excel ‘knew’ to add 1 to always reference the cell A1 because we used absolute cell references; that is we referenced the cells a dollar symbol before the letter corresponding to the column and a dollar symbol before the number corresponding to the row.

Absolute cell references are useful when you want to multiply (or perform some mathematical operation) a list of numbers by the same number each time.

3.9.3 Mixed cell references

Mixed cell references can be useful in certain circumstances, such as computing 2-dimensional tables of quantities and costs and 2-dimensional surface functions (plus countless other things that are hard to explain in words).

Consequently I’ve put a video on YouTube showing examples of where you might find relative, absolute and mixed cell references useful—see <http://www.youtube.com/watch?v=y3xb1uhWY6k>

Data analysis 4

Using Excel for different graphs

Excel sucks for histogram / frequency distribution generation. Instead of reading the pages that follow take a look at my YouTube vid <http://www.youtube.com/watch?v=mdiXhAW0>.

4.1 Using Excel to construct a frequency distribution

You can use Excel to construct a frequency distribution from a list of sample data. In Excel, the process of constructing a frequency distribution is called binning the data, because each category acts like a separate bin into which we can pour some of the individual data values. Correct interpretation of an Excel frequency distribution requires that you know this important principle:

Excel's bins (classes) are based on upper class limits.

The following Excel procedure for constructing a frequency distribution uses the menu item of "Histogram," which is a type of graph. However, the same dialog box is used to generate a frequency distribution or a histogram graph.

1. Enter your sample data in a column of the Excel worksheet, or open an existing data set. To manually enter data, type the first value, then press the Enter key. Type the second value, then press the Enter key, and so on.
2. If using Excel 2007, click on Data, then click on Data Analysis; if using Excel 2003, click on Tools, then click on Data Analysis.
3. Click on Histogram, then click OK.
4. You should now see a Histogram dialog box
 - (a) First enter the Input Range. For example, if the data are listed in cells 1 through 40 of column A, enter the input range of A1:A40.
 - (b) For the Bin Range, you have two options: (1) Leave the bin range blank and let Excel decide how to construct the frequency table; (2) if you want specific class limits, specify a range of cells that you have previously filled in with the values of the upper class limits.
 - (c) Select Output Range, then enter a cell location for the frequency table, such as E2. Finally, click OK.
5. You can also use Excel to generate a frequency table with cumulative percentages. Follow the same five steps listed above, but in part c of Step 4, click on the box labeled "Cumulative Percentage" before clicking OK.

4.2 Using Excel for Histograms

CAUTION: Be very careful when reading and interpreting an Excel-generated histogram, because the values shown on the horizontal scale might appear to be class midpoints, but they are actually upper class limits.

The following procedure is quite complicated, but uses techniques that will also apply to many other Excel graphs, so the effort and time spent now will help later.

1. To generate a histogram, follow the procedure described in Section 4.1 for constructing a frequency distribution, but make the following addition. In part c of Step 4, click on the box labeled Chart output before clicking OK. Excel's default histogram requires considerable effort to make it suitable, so proceed patiently with the following steps.
2. Enlarge the Graph The original histogram will be quite small, so enlarge it. Click any area just inside of the border. Click and drag the black boxes on the outer edges so that the graph is enlarged.
3. Delete the Frequency Box To delete the "Frequency" box located to the right of the histogram, right-click on that box and then select either Delete or Clear.
4. Delete the Spaces Between Bars If using Excel 2007, right click on one of the bars, select Format Data Point, set the "Gap Width" to 0, then click on Close. If using Excel 2003, double click on one of the bars, click on the Options tab, change the "Gap width" to 0, then click OK.
5. Delete the Bin Label Remove the label of "Bin" for the horizontal axis by right-clicking on it, then selecting either Delete or Clear.
6. Delete the Bin Values on the Horizontal Scale The default values on the horizontal scale are bin values, so delete them as follows: Use the mouse to right-click on one of those values, then click on the option of Delete or Clear.
7. Delete the Background Color If using Excel 2003, delete the gray background color by right-clicking on any portion of that background, selecting Format Plot Area, clicking on the box colored white, then clicking OK.
8. Insert Class Boundary Values: Using Excel 2007 to insert class boundaries: Click on the Insert tab, then click on Text Box and release the mouse button. Move the cursor to the upper left portion of the area that will contain the text for the class boundary values (or class midpoint values), then hold the mouse button and slide the mouse to get a box large enough to contain the text for all of the class boundaries (or class midpoints). (Hint: You can make more room for the text box by clicking on an empty area just above a bar of the histogram, then click and hold the mouse on the lower right corner of the box that contains the bars, then slide the cursor upward.) Enter the class boundaries (or class midpoints) in the text box and use spacing to position them correctly. To change the size or style of the text, use the formatting options available by clicking on Home and selecting Font.
Using Excel 2003 to insert class boundaries: Double-click on the Text Box icon at the bottom. (The Text Box icon looks like a page of a book with a large "A" in the upper left corner.) Now move the mouse cursor to the area just below the graph. While holding the mouse button down, slide the mouse to enclose a region that will be suitable for entering the class boundaries. (If

your text box isn't suitable, change it by clicking and dragging one of the small squares on its border.) Type the values of the class boundaries and use spacing to position them correctly. Hint 1: To change the size or style of the text, use the mouse to click and drag over all of the text so that it is highlighted, then select Format to change the text as desired. Hint 2: If there is not enough room for the text box at the bottom, click on the inner border that is attached to the histogram, then use the mouse to click and drag (upward) the black box in the middle of the bottom portion of that border. Hint 3: To delete a border around the text box itself, right-click on that border, select Format Text Box, choose Colors and Lines, and in the "Line" section of options, click on the down arrow to the right of "Color" and click on No Line.

9. Insert a Label for the Horizontal Scale The histogram should have a label for the values on the horizontal scale, so enter it by using the procedure in Step 8. That is, create a text box below the values of the class boundaries and enter the label within that box. Remember, the size and style of the font can be edited as desired.

Excel's "Chart Wizard" is very flexible and allows you to generate a wide variety of different graphs. You can usually edit graphs by right-clicking or double-clicking on the element that you want to edit.

4.3 Using Excel for frequency polygons

The following Excel procedure for generating a frequency polygon is much easier than the procedure for generating a histogram.

1. Start with a clear worksheet. In column A, enter 0 in the first cell, followed by the class frequency, followed by a 0 at the end. For example, if you want to generate a frequency polygon based on a table (with frequencies of 12, 14, 11, 1, 1, 0, 1), enter these frequencies in column A: 0, 12, 14, 11, 1, 1, 0, 1, 0.
2. Click on Insert, then select Chart and, among the various charts, select Line.
3. Select the line graph in the upper left corner.
4. The frequency polygon will be displayed without the correct class midpoint values listed along the horizontal axis, so that must be fixed along with any other modifications that would improve the graph. The displayed values along the horizontal axis can be deleted, and the correct values can be inserted as a text box by using the procedure given in Step 8 for creating a histogram.

Data analysis 5

Introduction to Matlab

Topics discussed here include the Command Window, numbers and arithmetic operations, saving and reloading a work, using help, MATLAB demos, interrupting a running program, long command lines, and MATLAB resources on the Internet.

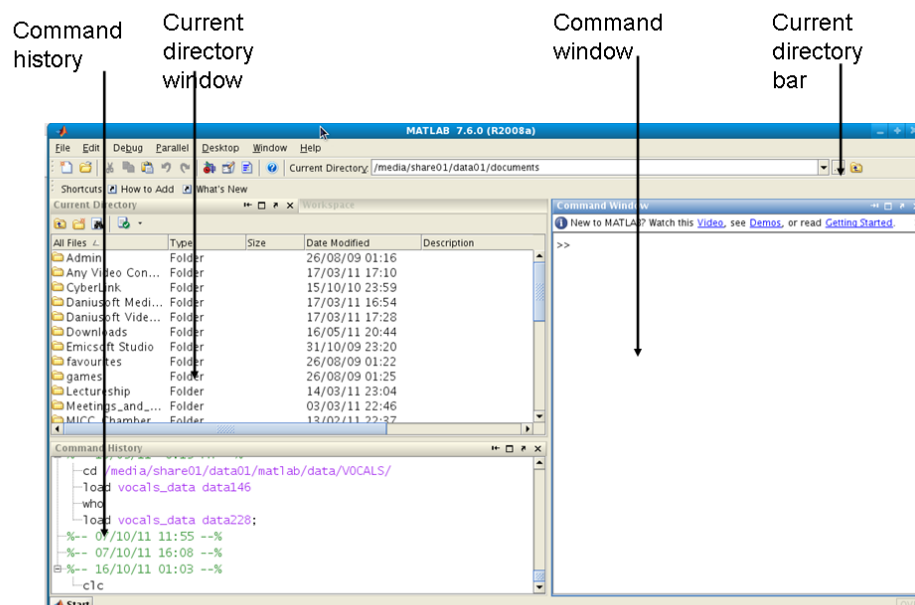
There are some statistics functions that do not ship with the standard version of MATLAB. You may need to download these for the course (unless you have the statistics toolbox). Go to Blackboard, 'Course content→ Semester→ Matlab resources' and download the *stats_toolbox.zip* file to your hard drive and unzip it. It will need to be on your MATLAB path so that you can use it.

5.1 Overview

You can start MATLAB by double clicking on the MATLAB icon that should be on the desktop of your computer. This brings up the MATLAB program and contained within the whole program (amongst other things) are:

- the Command Window.
- the Command History window
- the Current Directory Window
- the Current Directory bar.

as shown below.



The Command History prints a list of anything entered into the Command Window. The Current Directory shows where you are within the System File Structure.

The Command Window is perhaps the most important window. It is where commands can be entered and is now described.

5.2 The command window

This window allows a user to enter simple commands. To clear the Command Window type `clc` and next press the Enter or Return key. To perform a simple computations type a command and next press the Enter or Return key. For instance,

```
1 s = 1 + 2
2 s =
3     3
4 fun = sin(pi/4)
5 fun =
6     0.7071
```

In the second example the trigonometric function sine and the constant π are used. In MATLAB they are named `sin` and `pi`, respectively.

Note that the results of these computations are saved in variables whose names are chosen by the user. If they will be needed during your current MATLAB session, then you can obtain their values typing their names and pressing the Enter or Return key. For instance,

```
1 s
2 s =
3     3
```

Variable name begins with a letter, followed by letters, numbers or underscores. MATLAB recognizes only the first 31 characters of a variable name.

To change a format of numbers displayed in the Command Window you can use one of the several formats that are available in MATLAB. The default format is called short (four digits after the decimal point.) In order to display more digits click on File, select Preferences..., and next select a format you wish to use. They are listed below the Numeric Format. Next click on Apply and OK and close the current window. You can also select a new format from within the Command Window. For instance, the following command

```
1 format long
```

changes a current format to the format long. To display more digits of the variable `fun` type

```
1 fun
2 fun =
3     0.70710678118655
4
5 % To change a current format to the default one type
6 format short
7 fun
8 fun =
9     0.7071
```

To close MATLAB type exit in the Command Window and next press Enter or Return key. A second way to close your current MATLAB session is to select File in the MATLAB's toolbar and next click on Exit MATLAB option. All unsaved information residing in the MATLAB Workspace will be lost.

5.3 Numbers and arithmetic operations in MATLAB

There are three kinds of numbers used in MATLAB:

1. integers
2. real numbers
3. complex numbers

Integers are entered without the decimal point

```
1 xi = 10
2 xi =
3      10
```

However, the following number

```
1 xr = 10.01
2 xr =
3      10.0100
```

is saved as the real number.

Complex numbers in MATLAB are represented in rectangular form. The imaginary unit $\sqrt{-1}$ is denoted either by i or j

```
1 i
2 ans =
3      0 + 1.0000i
```

However, we will not make use of complex numbers in this course.

In addition to classes of numbers mentioned above, MATLAB has three variables representing the nonnumbers:

- -Inf
- Inf
- NaN

The Inf and Inf are the IEEE representations for the negative and positive infinity, respectively. Infinity is generated by overflow or by the operation of dividing by zero. The NaN stands for the not-a-number and is obtained as a result of the mathematically undefined operations such as 0.0/0.0 or ∞ or $-\infty$.

The list of basic arithmetic operations in MATLAB include six operations

Operation	Symbol
addition	+
subtraction	-
multiplication	*
division	/ or \
exponentiation (to the power)	^

MATLAB has two division operators /—the right division and \—the left division. They do not produce the same results

```

1 rd = 47/3
2 rd =
3     15.6667
4 ld = 47\3
5 ld =
6     0.0638

```

as you can see, the first returns the value of $47 \div 3$ and the second $3 \div 47$.

5.4 Vectors, Matrices / arrays

MATLAB is excellent for dealing with vectors and matrices (it stands for MATrix LABoratory after all). Let us call a vector a 1-d list of numbers, you can type some in using the commandline as follows:

```

1 vector=[1; 2; 1; 3; 7 ; 10]
2
3 % and matlab prints the following
4 vector =
5
6     1
7     2
8     1
9     3
10    7
11   10

```

The ‘;’ tells MATLAB to start a new row. You can use ‘,’ instead to start a new column:

```

1 rowvector=[1, 2, 1, 3, 7, 10]
2
3 % and matlab prints the following
4 rowvector =
5
6     1     2     1     3     7    10

```

You can also append two column vectors or row vectors together:

```

1 columnvector1=[1; 2; 1; 3; 7; 10];
2
3 columnvector2=[2; 3; 7; 10];
4
5
6 new=[columnvector1;columnvector2]
7
8 % and MATLAB will print
9 new =
10
11     1
12     2
13     1
14     3
15     7
16    10
17     2
18     3
19     7
20    10

```

If you want to append two row vectors together the syntax is slightly different:

```

1 rowvector1=[1, 2, 1, 3, 7, 10];
2
3 rowvector2=[2, 3, 7, 10];
4
5
6 new=[rowvector1 ,rowvector2]
7
8 % and MATLAB will print
9 new =
10
11      1      2      1      3      7      10      2      3      7      10

```

If two vectors have the same dimensions you can add / subtract them:

```

1 columnvector1=[1; 2; 1; 3];
2
3 columnvector2=[2; 3; 7; 10];
4
5
6 new=columnvector1+columnvector2
7
8 % and MATLAB will print
9 new =
10
11      3
12      5
13      8
14     13

```

or you can multiply or divide each column (or row). When multiplying or dividing vectors in this way you must use the `.*` or `./` syntax (the `.'` is important it tells MATLAB to multiply or divide each column (or row)):

```

1 columnvector1=[1; 2; 1; 3];
2
3 columnvector2=[2; 3; 7; 10];
4
5
6 new=columnvector1.*columnvector2
7
8 % and MATLAB will print
9 new =
10
11      2
12      6
13      7
14     30

```

Finally here we append two matrices together:

```

1 matrix1=[1, 2, 4; 5, 7, 10 ]; % 2 by 3 matrix
2
3 matrix2=[2, 3, 7; 6, 4, 1; 0, 1, 4; 3, 4, 3]; % 4 by 3 matrix
4
5
6 new=[matrix1 ;matrix2 ]
7
8 % and MATLAB will print
9 new =
10
11      1      2      4
12      5      7      10
13      2      3      7

```



```

14 |     6     4     1
15 |     0     1     4
16 |     3     4     3

```

There are ways you can multiply and divide matrices and vectors that are different to element wise multiplication above. These will not be covered in this course, but they are very useful.

There is an interesting point about arrays in computer languages, which allows us to gain insight into how computer memory works. Try the two following pieces of code, which both do the same thing:

```

1 a=zeros(10000,10000);
2
3 tic
4 for i=1:10000
5     for j=1:10000
6         a(i,j)=a(i,j)+1;
7     end
8 end
9 toc

```

on my computer the output is: Elapsed time is 1.024994 seconds.

```

1 a=zeros(10000,10000);
2
3 tic
4 for j=1:10000
5     for i=1:10000
6         a(i,j)=a(i,j)+1;
7     end
8 end
9 toc

```

on my computer the output is: Elapsed time is 0.498409 seconds.

So we see that the way in which you access data (memory) can make a difference to the speed of the computations. Very important!

Matlab treats its arrays so that the first dimension (rows) are arranged sequentially in memory. If you access columns sequentially the computer has to work over time to go forward and backward to access the data.

5.5 Saving and reloading your work

All variables used in the current MATLAB session are saved in the Workspace. You can view the content of the Workspace by clicking on File in the toolbar and next selecting Show Workspace from the pull-down menu. You can also check contents of the Workspace typing whos in the Command Window. For instance,

```

1 whos
2 Name          Size          Bytes    Class
3 ans           1x1           16      double array (complex)
4 fun           1x1           8       double array
5 ld            1x1           8       double array
6 rd            1x1           8       double array
7 s             1x1           8       double array
8 xi            1x1           8       double array
9 xr            1x1           8       double array
10 Grand total is 7 elements using 64 bytes

```

shows all variables used in current session. You can also use command `who` to generate a list of variables used in current session

```
1 who
2
3 Your variables are:
4 ans      ld      s
5 fun      rd      xi
```

To save your current workspace select Save Workspace as... from the File menu. Chose a name for your file, e.g. filename.mat and next click on Save. Remember that the file you just created must be located in MATLAB's search path. Another way of saving your workspace is to type

```
1 save <filename>
```

where <filename> is the name of the file you want to save. in the Command Window. The following command

```
1 save <filename> s
```

where <filename> is the name of the file, saves only the variable `s`.

Another way to save your workspace is to type the command:

```
1 diary <filename>
```

where <filename> is the name of the file, in the Command Window. All commands and variables created from now will be saved in your file. The following command:

```
1 diary off
```

will close the file and save it as the text file. You can open this file in a text editor, by double clicking on the name of your file, and modify its contents if you wish to do so.

To load contents of the file named filename into MATLAB's workspace type

```
1 load <filename>
```

in the Command Window.

More advanced computations often require execution of several lines of computer code. Rather than typing those commands in the Command Window you should create an m-file, which is just a text file with commands to be executed one after another. The m-file is saved with a '.m' file extension. Each time you will need to repeat computations just invoke your m-file by typing the name in the MATLAB command window (without the '.m' extension). Another advantage of using m-files is the ease to modify its contents.

5.6 Help

One of the nice features of MATLAB is its help system. To learn more about a function you are to use, say `normpdf`, type in the Command Window

```

1 help normpdf
2
3 NORMPDF Normal probability density function (pdf).
4 Y = NORMPDF(X,MU,SIGMA) returns the pdf of the normal distribution with
5 mean MU and standard deviation SIGMA, evaluated at the values in X.
6 The size of Y is the common size of the input arguments. A scalar
7 input functions as a constant matrix of the same size as the other
8 inputs.
9
10 Default values for MU and SIGMA are 0 and 1 respectively.
11
12 See also NORMCDF, NORMFIT, NORMINV, NORMLIKE, NORMRND, NORMSTAT.

```

The `helpwin` command, invoked without arguments, opens a new window on the screen. To find an information you need double click on the name of the sub-directory and next double click on a function to see the help text for that function. You can go directly to the help text of your function invoking `helpwin` command followed by an argument. For instance, executing the following command

```

1 helpwin zeros
2 ZEROS Zeros array.
3 ZEROS(N) is an N-by-N matrix of zeros.
4 ZEROS(M,N) or ZEROS([M,N]) is an M-by-N matrix of zeros.
5 ZEROS(M,N,P,...) or ZEROS([M N P ...]) is an M-by-N-by-P-by-...
6 array of zeros.
7 ZEROS(SIZE(A)) is the same size as A and all zeros.
8 See also ONES.
9 generates an information about MATLAB's function zeros.

```

5.7 Demos

To learn more about MATLAB capabilities you can execute the demo command in the Command Window:

```

1 demo

```

or click on Help and next select Demos from the pull-down menu. You can then use the explorer bar on the left to learn about different aspects of MATLAB and its toolboxes. Some of the MATLAB demos use both the Command and the Figure windows.

To learn about plots and graphics in MATLAB open the demo window using one of the methods described above. In the left pane select Graphics and in the right pane select 2-D Plots. The information will be displayed in the help pane. You can run a demo by clicking the link "Run in the Command Window", you may have to return to the command window to press return—and follow instructions.

I recommend trying some of these demos and looking at what MATLAB commands are needed to do them. There is enough information to 'learn by doing' for those interested (and motivated to do so).

5.8 Interrupting a running program

To interrupt a running program press simultaneously the Ctrl-c keys. Sometimes you have to repeat pressing these keys a couple of times to halt execution of your

program. This is not a recommended way to exit a program, however, in certain circumstances it is a necessity. For instance, a poorly written computer code can put MATLAB in the infinite loop and this would be the only option you will have left.

5.9 Long command lines

To enter a statement that is too long to be typed in one line, use three periods, ... , followed by Enter or Return. For instance,

```
1 x = sin(1) - sin(2) + sin(3) - sin(4) + sin(5) - ...
2     sin(6) + sin(7) - sin(8) + sin(9) - sin(10)
3 x =
4     0.7744
```

You can suppress output to the screen by adding a semicolon after the statement

```
1 u = 2 + 3;
```

5.10 Plotting and histograms

To plot a graph of $f(x) = x^2$ on the interval $0 \leq x \leq 10$ you may type:

```
1 X=0:0.1:10 % create a vector x ranging from zero to 10 in steps of 0.1
2 plot(X,X.^2) % plot out the function f(x)=x^2
3 xlabel('x_variable');
4 ylabel('f(x)=x^2')
```

To plot a histogram of some random data:

```
1 X=rand(1000,1); % create a vector of 1000 random numbers ranging from 0 to 1
2 hist(X,[0:0.2:1]) % create a histogram of those values binned into intervals of 0.2
3 xlabel('x')
4 ylabel('Frequency')
```

There are many other ways of visualising data in MATLAB. Take a look at the Graphics and 3-D visualisation demos for some examples.

5.11 Reading data into MATLAB

For this course you can use `csvread` to read in a csv (comma separated variable) file:

```
1 dat=csvread('filename') % 'filename' is the name of the file to be read into the variable 'dat'
```

You may also be able to `xlsread` to read in an Excel spreadsheet (try 'help xlsread') to learn about this function:

```
1 [dat ,text]=xlsread('filename') % 'filename' is the name of the file to be read into the variables 'dat'
```

5.12 MATLAB resources on the Internet

- The MathWorks Web site: <http://www.mathworks.com/> The MathWorks, the makers of MATLAB, maintains an important Web site. Here you can find information about new products, MATLAB related books, user supplied files and much more.

- Especially this link: http://www.mathworks.co.uk/academia/student_center/tutorials/launchpad.html, which gives plenty of resources and interactive tutorials.
- The Mastering Matlab Web site: <http://www.eece.maine.edu/mm> Link to the books: Mastering Matlab. A Comprehensive Tutorial and Reference, which are also available from the John Rylands library.

Table of equations and functions

Here I give a quick reference guide to common things you may want to do and the corresponding functions to look up in either Microsoft Excel or Matlab—see table opposite.

What you want to do	Equation	Excel functions	Matlab functions
Import data	N/A	N/A	dat=csvread; [dat, text]=xlsread
Create a frequency distribution of your data	N/A	frequency	hist
Create a cumulative frequency distribution of your data	N/A	use frequency and then in another column calculate the cumulative sum using absolute and relative cell references.	use [N, X]=hist(dat, bins) and then N2=cumsum(N) to calculate the cumulative frequency
Create a relative frequency distribution of your data	N/A	use frequency then in another column reference the frequency distribution and divide by sum() of all values in the frequency distribution (using absolute cell references)	use [N, X]=hist(dat, bins) and then divide by sum(N)
Create a bar chart of your data	N/A	Insert chart	bar
Create a pie chart of your data	N/A	Insert chart	pie
Create an x-y scatter plot of your data	N/A	Insert chart	plot(x, y)
Calculate the arithmetic mean	$\bar{x} = \frac{\sum x}{N}$	average	mean(x)
Calculate the median	N/A	median	median(x)
Calculate the mode	N/A	mode	mode(x)

Continued on next page

What you want to do	Equation	Excel functions	Matlab functions
Calculate the range	N/A	max() -min(), and pass cell references	range(x)
Calculate the sample standard deviation	$s = \sqrt{\frac{\sum(x-\bar{x})^2}{N-1}}$	stdev	std(x,0)
Calculate the population standard deviation	$\sigma = \sqrt{\frac{\sum(x-\bar{x})^2}{N}}$	stdevp	std(x,1)
Calculate a percentile	N/A	percentile(arr,k)	prcentile
Calculate an interquartile range	$Q_3 - Q_1$	percentile(arr,75)-percentile(arr,25)	iqr
Calculate a z-score	$z = \frac{x-\bar{x}}{s}$		[Z, MU, SIGMA] = zscore(X)
Calculate probability density for the normal distribution	$y = \frac{\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)}{\sigma\sqrt{2\pi}}$	normdist(x,mean,s,0)	normpdf(x,mean,s)
Calculate cumulative probability for the normal distribution	$\frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\right]$	normdist(x,mean,s,1)	normcdf(x,mean,s)
Sample from the normal distribution	N/A	norminv(P,mean,s)	norminv(P,mean,s)
Calculate cumulative probability for the student t-distribution	nasty	tdist(P,degf,tails) note that tails can be either 1 or 2 and refers to whether the area, P, is on both sides or one	tcdf this is the one sided cumulative dist.

Continued on next page

What you want to do	Equation	Excel functions	Matlab functions
Sample from the t-distribution	N/A	<code>tinv(P, degF)</code> note this returns the two-tailed distribution. Multiply P by two to return the one-tailed	<code>tinv(P, degF)</code> note this returns the one-tailed distribution. Divide P by two to return the two-tailed.
Calculate the correlation coefficient	see sheet of formulas	<code>CORREL(x,y)</code>	<code>corrcoef(x,y)</code>
Perform linear regression	N/A	Use regression in Analysis toolpak	<code>P=polyfit(x,y,1)</code> P will contain the gradient and the intercept
Calculate linear regression line	N/A	N/A	<code>y=polyval(P,x)</code> Where x is a list of numbers spanning the lower and upper values (e.g. <code>x=-10:0.01:10</code>) and P contains the output of <code>polyfit</code>

References

- Hanselman, D., and B. Littlefield, 2001: *Mastering Matlab 6: A comprehensive tutorial and reference*. Pearson.
- Kirkup, L., 2002: *Data analysis with Excel: An introduction for physical scientists*. 1st ed., Cambridge University Press.
- Triola, M. F., 2009: *Elementary statistics using Excel*. 4th ed., Pearson education.

Part II
Statistics exercises

Data analysis 1

Week 3 practical

There are two exercises to do this week:

1. the '*Week 3 project*' practical in the computing cluster. It is designed to help you learn about graphing in Excel and MATLAB. In order to do this practical you should be familiar with the bullet points below.
2. the assessment '*Introduction to statistics*' on Blackboard. You should be able to do this in your own time, but may get to finish it in class.
 - How to do an x-y scatter plot in Excel 2007—see <http://www.youtube.com/watch?v=6drKbhJA2EM>.
 - How to do an x-y scatter plot in Matlab—see <http://www.youtube.com/watch?v=KiASzo8Yukc>.
 - How to do a histogram in Excel 2007—see <http://www.youtube.com/watch?v=mdixXI5hAW0>.
 - How to do a histogram in Matlab—see http://www.youtube.com/watch?v=OT_ZzHQkc98.
 - Where to download the Excel spreadsheets from on Blackboard: Go to Course content→Spreadsheets and data.

In this practical you will be required to input your answers into Blackboard. You may wish to do the exercise in the spreadsheet or in MATLAB first though before opening the test.

First download the ELECTRIC.xls and FHEALTH.xls and MHEALTH.xls spreadsheets. Also if you haven't done so, download the notes for the course to your p-drive.

1. In the spreadsheet ELECTRIC.xls, four columns of data are available (energy consumption in kWh; cost of electricity; degree days; and average temperature in Fahrenheit) (Course content→Spreadsheets and data) use the 22 average daily temperatures and the corresponding 22 amounts of energy consumption (kWh) to construct a scatterplot in Excel or MATLAB. Based on the result, is there a relationship between the average daily temperatures and the amounts of energy consumed? Try to identify at least one reason why there is or isn't a relationship.
2. Construct a histogram of energy consumptions (kWh) with bin mid-points for every 500 kWh. Label your axes.
3. Manually constructed graphs are generally unsuitable for publications and presentations. Computer-generated graphs are much better for such purposes. From the Blackboard website for this course download the xls files FHEALTH.xls and MHEALTH.xls. These contain pulse rates (labelled PULSE) of females and males, and they are available as files that can be opened by Excel or imported into MATLAB. Using either Excel or MATLAB, analyse the FHEALTH and MHEALTH xls files, generate three histograms:
 - (a) a histogram of the pulse rates of females with bin spacing of 5 BPM;
 - (b) a histogram of the pulse rates of males with bin spacing of 5 BPM;

(c) a histogram of the combined list of pulse rates of males and females.

Compare the histograms. Does it appear that the pulse rates of males and females have similar characteristics? (Later in this course, we will present more formal methods for making such comparisons.)

Save your spreadsheet or MATLAB workspace (e.g. type: `save <filename>`) somewhere for later reference.

Data analysis 1

Week 7 practical

There is one exercise to do this week:

1. the '*Week 7 project*' in the computing cluster. It is an exercise looking at what determines whether a data point is an *outlier*.
2. MATLABers can learn how to plot out the landgrav data using the code on the page opposite. Either enter each line into the command window, or put the text into a text file and call it 'landgrav.m', save it on your path and type 'landgrav' into the command window.

This will be done in the practical session on Monday. You may wish to do the exercise in a spreadsheet or MATLAB before opening the test.

In order to do this practical you should be familiar with the following:

- How to calculate various descriptive statistics and z-scores using Excel—see <http://youtu.be/Q7RuUnLpXjM>.
 - If you want to use MATLAB then see how to calculate descriptive statistics and z-scores using MATLAB—<http://youtu.be/c4yPS1kgcko>.
 - To read in the landgrav_all.csv file in MATLAB type `dat=csvread('landgrav_all.csv',2,0)`; as described on lines 1-6 on the opposite page.
1. Calculate the mean and the standard deviation of the duration times of Eruptions of the Old Faithful geyser (based on the OLDFAITH.xls dataset—see Course content→Spreadsheets and data).
 - (a) What is the difference between a duration time of 110 and the mean of the column of data?
 - (b) How many standard deviations is that (the difference found in part (a)?);
 - (c) Convert the duration of 110 sec to a z-score;
 - (d) If we consider '*usual*' duration times to be those that convert to z-scores between -2 and 2, is the duration time 110 usual or unusual?
 2. Calculate the mean and the standard deviation of the Bouguer anomaly using the data in the csv file landgrav_all.csv. Convert the following Bouguer anomalies to z-scores (2 d.p.s) and say whether they are usual or unusual
 - (a) 40 mGal;
 - (b) -45 mGal.
 3. Refer to the VOLTAGE.xls dataset, which includes measured voltage levels from a US home, a generator, and an uninterruptible power supply. Generate histograms for all three columns, with a range of 122 to 125.2V (bin width = 0.2V), and calculate statistics to compare the three sets of data. Are there any outliers? (hint: calculate the z-scores). Do all three power sources appear to provide electricity with properties that are basically the same?

Save your spreadsheet or MATLAB workspace (e.g. type: `save <filename>`).

1.1 Neat stuff in MATLAB

You can plot out the `landgrav_all.csv` data set using MATLAB. The function to plot out 3-d scattered data is `scatter(x,y,siz,z,'filled')` (where `siz` is the size of the dots on the image), which plots the x and y . locations of the measurements and colours them by the measurement value, z . Try typing `help scatter` for more information.

The code excerpt shows how to do use `scatter` to plot out the `landgrav` data.


Note, if you get an error that MATLAB cannot find the `coast` variable or file you will need to download it. It is available on Blackboard: Course content->Semester 1->Matlab resources->`coast.mat`. Download it and put in your working directory. You could also just comment lines 21, 22, 34 and 46.

```

1 % Note there are two headerlines in the file
2 % (hence the 2 for the second argument)
3 % The columns are: latitude , longitude , elevation (m),
4 % observed gravity -980000 (mGal), Bouguer anomaly (mGal),
5 % Bouguer density
6 dat=csvread('landgrav_all.csv',2,0);
7
8 % Locate values that equal 999999 and set to 'NaN'
9 dat(find(dat(:)==999999))=NaN;
10
11 % Open a figure to write the plot to, using the 'openGL' renderer for speed
12 figure('name','Observed_gravity','renderer','openGL')
13 scatter(dat(:,2),dat(:,1),2,dat(:,4),'filled');
14 h=colorbar;
15 ylabel(h,'Observed_gravity_-980000_(mGal)')
16 hold on;
17
18 % 'coast' is a low resolution coastline dataset that is part of matlab.
19 % it contains the arrays 'long' and 'lat', which are the locations of the
20 % coastlines.
21 load coast
22 plot(long(:,1),lat(:,1));
23 axis([-11 2 49 60])
24 xlabel('longitude_(^\ circ)');
25 ylabel('latitude_(^\ circ)');
26 grid on;
27
28 figure('name','Bouguer_anomaly','renderer','openGL')
29 scatter(dat(:,2),dat(:,1),2,dat(:,5),'filled');
30 h=colorbar;
31 ylabel(h,'Bouguer_anomaly_(mGal)')
32 hold on;
33
34 plot(long(:,1),lat(:,1));
35 axis([-11 2 49 60])
36 xlabel('longitude_(^\ circ)');
37 ylabel('latitude_(^\ circ)');
38 grid on;
39
40 figure('name','Elevation','renderer','openGL')
41 scatter(dat(:,2),dat(:,1),2,dat(:,3),'filled');
42 h=colorbar;
43 ylabel(h,'Elevation_(m)')
44 hold on;
45
46 plot(long(:,1),lat(:,1));
47 axis([-11 2 49 60])
48 xlabel('longitude_(^\ circ)');

```

```
49 ylabel('latitude_{\circ}');  
50 grid on;
```



Data analysis 2

Week 8 practical

In this practical you will be required to input your answers into Blackboard. You should be familiar with the following:

- How to use the NORMDIST function in Excel for calculating probabilities—see <http://youtu.be/JZX231rTlg>.
- How to use the NORMINV function in Excel for calculating confidence intervals—see <http://youtu.be/5aA1Tk-C21c>.
- How to use the normcdf function in MATLAB for calculating probabilities—see <http://youtu.be/Lo7tkbWIKco>.
- How to use the norminv function in MATLAB for calculating confidence intervals—see <http://youtu.be/aIi79zNJPTS>.

Now start the week 8 Blackboard practical and work through questions 1 to 7.

For questions 8-12 on the *Blackboard practical* you will need to generate some random data by sampling from a distribution. You should be familiar with:

- How to sample randomly from a normal distribution function in Excel—see <http://youtu.be/108MXJs0uz4>.
- How to sample randomly from a normal distribution function in MATLAB—see <http://youtu.be/p70ScjpZTWA>.

Now read and complete the rest of this sheet before attempting questions 8-12

Our example is as follows: *when two consecutive years of 149 measurements of PM10 (pollution aerosols) around Manchester were compared, the reduction in PM loading was $1.0 \mu\text{g m}^{-3}$, with a standard deviation of $2.2 \mu\text{g m}^{-3}$.*

The following Excel and MATLAB procedures show how to generate normally distributed data and then shift the mean and standard deviation to desired values. This can be useful in a number of situations, but here we just compare the data to a theoretical normal distribution.

We will sample data from a distribution with mean equal to 0 and standard deviation equal to 1 and then shift the mean and standard deviation to desired values. If we multiply every sample value by a constant the standard deviation is multiplied by that constant, and if we then add a constant to every sample value the same constant is added to the mean but the standard deviation does not change.

See Section 2.1 for the Excel method and Section 2.2 for the MATLAB method.

2.1 Excel Method

1. Generate a list of sample values from a population with a mean of 0 and a standard deviation of 1. If using Excel 2007, click on Data, then Data Analysis; if using Excel 2003, click on Tools, then Data Analysis. Select Random Number Generation, then click OK. In the dialog box, enter 1 for the number of variables, enter 149 (or any desired sample size) for the number of random values, select a Normal distribution, and use the default parameter values of Mean = 0 and Standard Deviation = 1. Select the default output option of New Worksheet Ply so that the sample data will be put in column A. Click OK.

2. Now find the existing standard deviation of the generated values. Click on any empty cell, say D1 then type =STDEV() and enter the range of cells containing the generated sample data, such as A1:A149 for 149 values in rows 1 through 149 of column A.
3. Transform the sample data so that the standard deviation is changed from the existing standard deviation found in Step 2 to the desired standard deviation (what is the desired standard deviation and why?). To accomplish this, multiply each generated sample value by the desired standard deviation, and divide each result by the existing standard deviation (calculated above). e.g.:
 - (a) In cell B1, enter this expression:
$$=A1 * 2.2 / \$D\$1$$
where the 2.2 is the desired standard deviation.
 - (b) Press the Enter key, then click and drag the lower right corner of cell B1 down to the same number of rows as in column A.
4. Column B now should contain data with the desired standard deviation. Now transform column B so that the mean is changed to the desired mean. e.g.:
 - (a) Click on any empty cell, say D2, and find the mean of the values in column B by using =average(C) and enter the range of values for column B, such as B1:B149. Record the existing mean.
 - (b) Click on cell C1 and enter this expression:
$$=B1 + 1.0 - (\$D\$2)$$
where, the 1.0 above is the desired mean and \$D\$2 is the actual mean found above.
 - (c) Press the Enter key, then click and drag the lower right corner of cell C1 down to the same number of rows as in column A.
5. Column C should now contain the desired number of sample values having the desired mean and the desired standard deviation. Verify this by using the Descriptive Statistics module found in the Data Analysis menu.

To check whether the data are indeed normally distributed, complete the following (you may want to watch <http://youtu.be/108MXJs0uz4> again):

1. Create a histogram of the data, divide the counts in each bin by the bin width and divide again by the sum of all bins.
2. On the same plot, use NORMDIST with cumulative flag=0 to plot a normal distribution with a mean of 1.0 and standard deviation equal to 2.2.
3. Do a similar comparison for the cumulative distribution (i.e. compare the cumulative frequency to the cumulative probability using NORMDIST with cumulative flag=1).

2.2 MATLAB method

1. Generate 149 normally distributed random numbers with a mean of 0 and a standard deviation of 1:

```
1 % Note rand(149,1) generates a column of 149 random numbers between 0 and 1
2 % norminv is therefore sampled with a random probability
3 dat=norminv(rand(149,1),0,1);
```

2. Transform these variables so that they have a standard deviation of 2.2

```
1 % Multiply by the desired standard deviation and divide by the actual sd.
2 dat2=dat.*2.2./std(dat);
```

Note the `.` before the `*` and `/`. The reason for this is it allows *all values* in the table (or array) to be multiplied or divided. `dat2`, now contains the 149 variables with a standard deviation of 2.2.

3. Transform `dat2` so that it has a mean of 1.0

```
1 % Subtract the actual mean and add the desired mean
2 dat3=dat2+1.0-mean(dat2);
```

4. `dat3` will now have 149 values that have a mean of 1.0 and standard deviation of 2.2. Check this by calculating summary statistics.

To check whether the data are indeed normally distributed, complete the following:

1. Do a normalised histogram:

```
1 % N is frequency, X is bin centre
2 [N,X]=hist(dat3);
3 % Note: 'diff' calculates the difference between values in a list.
4 bin_spacing=diff(X);
5 % This is a plot of the normalized frequency density
6 % i.e. N divided by sum(N) and divided by the bin width
7 plot(X,N./(sum(N)./bin_spacing(1)), 'k')
```

2. Compare to the theoretical normal distribution:

```
1 % hold on; holds the plot so that we can plot another graph on top
2 hold on;
3 % Plot the theoretical normal prob. density.
4 plot(X, normpdf(X,1.0,2.2), 'r');
```

3. Compare the cumulative distribution to the theoretical cumulative distribution:

```
1 figure;
2 % Cumulative frequency: cumsum creates a cumulative sum of data
3 plot(X,cumsum(N)./sum(N), 'k'); hold on;
4 % theoretical cumulative distribution:
5 plot(X, normcdf(X,1.0,2.2), 'r');
```

Save your spreadsheet or MATLAB workspace (e.g. type: `save <filename>`).

Data analysis 3

Week 9 practical

In this practical you will be required to input your answers into Blackboard. You may wish to do the exercise in the spreadsheet first though before opening the test. *It is designed to give you an appreciation why the hypothesis testing method works.*

In this practical you will use Excel or MATLAB for hypothesis testing. You should be familiar with the following:

- how to do a hypothesis test about a mean in Excel—<http://youtu.be/wb1q4FW0wTY>
- and the same in MATLAB—<http://youtu.be/o7yPcDnrXws>
- how to conduct the sampling procedure (Section 3.1) in Excel—<http://youtu.be/4Vcdn9gMWWA>
- and the same (Section 3.2) in MATLAB—<http://youtu.be/mwDi8tTkrW0>

This practical considers two ways to test a hypothesis, through theory and sampling.

1. Open the file RUBBISH.xls (in either Excel or Matlab—e.g. using `xlsread`), which contains weights in kg of household rubbish discarded each week. Use the sample weights of discarded plastic (PLAS or column 4) to *test the claim that the mean weight of plastic discarded in a week is less than 1.0 kg*. Use a 0.05 significance level ($\alpha = 0.05$).

Is the mean weight of plastic than 1.0 kg? if so is it significantly below 1.0 kg?

2. Estimate how often we would calculate a sample mean less than 0.95 kg, if we were to sample from a population with $\mu = 1.0$ kg and $\sigma = 0.4843$ kg. To do this do the following:

- Calculate the value of t for $\mu = 1.0$, $\sigma = 0.4843$ and $\bar{x} = 0.95$:

$$t = \frac{\bar{x} - \mu}{s / \sqrt{N}}$$

- Use `tdist` in Excel or `tcdf` in MATLAB, with $N-1$ degrees of freedom and the above value of t to calculate the p-value (probability of sampling a value of t , less than the value of t being considered.):

Excel: `=TDIST(ABS((0.95-1)/(0.4843/SQRT(62))), 61, 1)` (note we have used '1' for the third input, why?)

MATLAB: `tcdf((0.95-1)./(0.4843./sqrt(62)), 61)`

- Your answer should be about 0.2 or so. Make sure you understand how these functions work as they will be used in later practicals.

3. Now, lets look into why we actually use these methods. We will generate some data by *sampling* from a normal distribution with $\mu = 1.0$ kg and $\sigma = 0.4843$.

- Using the procedure described, in Section 3.1 (Excel users) or 3.2 (MATLAB users), below generate 100 samples of 62 weights from a normally distributed population having the assumed mean of 1.0 kg (as in the null hypothesis, i.e. $H_0 : \mu = 1.0 \text{ kg}$). Use a standard deviation, s , of 0.4843, like the in sample data.
4. Testing the hypothesis: Based on the sample means that are found from the simulated samples, determine whether a sample mean of 0.95, is ‘unusual’ or can easily occur by chance (i.e. count them and calculate a fraction that are less than 0.95)
- How many of the sample means are less than or equal to 0.95 kg? What fraction is less than or equal to 0.95 kg?
 - What do the simulated results suggest about the claim that the mean weight of discarded plastic is less than 1.0 kg? How does your answer compare to the p-value you calculated above?

Save your spreadsheet or MATLAB workspace (e.g. type: save <filename>).

3.1 Excel procedure for generating 100 simulated samples

- If using Excel 2007, click on Data, then select Data Analysis; if using Excel 2003, click on Tools, then select Data Analysis.
- In the Data Analysis window, select Random Number Generation, click OK.
- In the dialog box, enter 100 for the number of variables (or samples), enter a sample size of 62 (as in the sample of weights of plastic), select the distribution option of Normal, enter a mean of 1 (as assumed) and enter a standard deviation of 0.4843 kg, then click OK.
- Find the sample mean of the first generated sample by going to cell A63 and using the Excel command =AVERAGE(A1:A62). Click and drag the lower right corner of cell A63 to the right so that all of the 100 sample means are found. Inspect the 100 sample means to determine how many are less than 0.95 kg. Return back to point 4 above to complete the practical.

3.2 MATLAB procedure for generating 100 simulated samples

- Use the rand() function to sample from a normal distribution using norminv.

```

1 % generate a table of 100 by 62 random numbers.
2 dat_rand=rand(100,62);
3
4 % generate 100 lots of 62 normally distributed random samples.
5 dat=norminv(dat_rand,1,0.4843);

```

- Find the sample means of all 100 samples

```

1 sample_means=mean(dat,2);

```

which takes the mean over the 2nd dimension.

- Inspect the 100 sample means to determine how many are less than 0.95 kg:

```
1 number_less_than=length(find(sample_means < 0.95))
```

which finds the length of all rows that have a sample mean less than 0.95.

- Return to point 4 to complete the practical.

3.3 Fancy things in MATLAB: Function m-file

You can save commands in a text file so that you can run common tasks easily. In this example we will save a *function m-file* that allows us to test the hypothesis of whether a sample is drawn from a distribution with a specified population mean.

First, type `edit` on the MATLAB command window. In the editor that opens, type the following:

```
1 function [accept , pvalue]=hypothesis01 (data , alpha , pop_mean , tails )
2 % Test whether a sample is drawn from a distribution with a population mean
3 % equal to pop_mean at a significance , alpha. Tails is either 1 or 2 depending
4 % on whether a one or two tailed test is needed.
5 % It returns which hypothesis is accepted in 'accept' and the 'pvalue'
6
7 % Tails must be 1 or 2
8 if(tails ~= 1 & tails ~=2)
9     disp('tails must be 1 or 2');
10    return;
11 end
12 xbar=mean(data); % calculate the mean
13 std1=std(data); % calculate the standard deviation
14 N=length(data); % sample size
15 t=abs((xbar-pop_mean)./(std1./sqrt(N))); % t-statistic - absolute value
16 % adjust the probability depending on how many tails.
17 if(tails==1)
18     alpha1=alpha;
19 elseif(tails==2)
20     alpha1=alpha./2;
21 end
22 tcrit=abs(tinv(alpha1,N-1)); % critical value of t
23 pvalue=tcdf(-t,N-1); % p-value
24
25 % If t is less than the critical value , accept the null , etc.
26 if(t<=tcrit)
27     accept='Null';
28 else
29     accept='Alternate';
30 end
```

Save it as `hypothesis01.m` on your file path and run it for a 1 tailed test at a significance level = 0.05 by typing:

```
1 [accept , pvalue]=hypothesis01 (data ,0.05 ,1 ,1)
```

on the MATLAB commandline, where `data` contains your data. The function will tell you whether to accept the null or alternate hypothesis and also give a p-value.

Data analysis 4

Week 10 practical

Is most of the planet covered with clouds, on average, or not?

Is the liquid cloud coverage greater (or smaller), on average, than the ice cloud coverage?

Do months within an El-Nino year have a higher cloud coverage, on average, than those that are not in an El-Nino?

These are the questions to be addressed in this practical.

First download the `MODIS_satellite_data.zip` file from *Spreadsheets and data* and unzip it on your hard drive. MATLAB users will also want to download the function m-files in this weeks folder (`additional_matlab_files.zip`) and put them in the same folder as the satellite data.

4.1 Overview

In this practical you will test hypotheses about a sample mean and also about two sample means.

- The data used are real data of *cloud fraction* sampled using the MODIS satellite sensor.
- The MODIS sensor is on a satellite platform that is *polar orbiting* (see http://www.youtube.com/watch?v=MLrPyhT_YNg) and so it samples a small fraction of the globe at any one time.
- Nevertheless, eventually the sensor will sample data from ever point on the globe, many times during a month.
- *Cloud fraction* is the fraction of times a point on the globe is sampled to be cloudy to the times the same point is sampled.
- In this practical you are given 12 one-monthly datasets of global cloud fraction and asked to infer questions about the population mean cloud fraction from within and out of El-Nino years.
- You therefore need to calculate a global average for each month and use the 12 global averages as your sample data to test hypotheses against.

At the end, save your spreadsheet or MATLAB workspace (e.g. type: `save <filename>`).

4.2 Questions asked and statistics

1. Is the global total cloud fraction (liquid plus ice) from months within El-Nino years significantly greater than 0.5 (in other words is most of the planet covered with clouds)?

Here we are testing a hypothesis about a mean (as covered in the last practical). The statistic to be tested is therefore:

$$t = \frac{\bar{x} - \mu}{s / \sqrt{N}}$$

with $N - 1$ degrees of freedom.

2. Is the global liquid cloud fraction from months within an El-Nino year greater (or smaller) than the ice cloud fraction?

Here we are testing a hypothesis about two means (as covered in the Monday lecture). The statistic to be tested is therefore:

$$S_p^2 = \frac{(n_1 - 1) s_1^2 + (n_2 - 1) s_2^2}{n_1 + n_2 - 2}$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S_p^2/n_1 + S_p^2/n_2}}$$

with $n_1 + n_2 - 2$ degrees of freedom.

3. Is the global total cloud fraction (liquid plus ice) from months within El-Nino years significantly greater than those not in El-Nino years?

Here we are testing a hypothesis about two means (as covered in the Monday lecture). The statistic to be tested is therefore as in question 2.

For further information see the sections on either Excel or MATLAB below.

4.3 Excel

If you click on the files they will open up in Excel. You'll notice the data start in cell A1 and end in row 181 and column MW. The first row is the value of *longitude* that the data are taken at and the first column is the value of *latitude*. The rest of the data are *cloud fraction* values at those *longitude* and *latitude*.

1. For the first question:
 - You can take the global mean of a particular dataset by typing: =average(B2:MW181) in an empty cell.
 - You can also reference a different work book to where you calculate the average by opening up another workbook and calculating the means by using the mouse to select the numbers you want.
 - Calculate the means for all csv files that begin with the letter A and end in _all.csv; there should be 12 means in total. These are the total cloud fractions for the El-Nino year.
 - You should have all you need to test the hypothesis now.
2. For the second question:

- Repeat the above process for the files starting with A and ending with `_liquid.csv` and `_ice.csv`; there should be two lots of 12 means (24 values in total).
- You should have everything you need to test the hypothesis that the liquid cloud fraction is different to the ice cloud fraction.

3. For the third question:

- Repeat the above process for the files starting with C (this is the non El-Nino year) and ending with `_all.csv`; there should be 12 additional means to those you generated in question 1 (24 values in total).
- You should have everything you need to test the hypothesis that the cloud fraction is different in El-Nino years is different to the cloud fraction in non El-Nino years.

4.4 MATLAB

Question 1: To test whether more of the earth is covered by clouds than is not, take a look at the following code.

```

1  function [means]=process_cloud_fraction2
2  % Read in the cloud fraction data and calculate a global mean for each
3  % month. Is the global mean >50%?
4
5  d=dir('A*_all.csv');
6
7  N=length(d); % sample size
8  means=zeros(size(d));
9  for i=1:length(d)
10     dat=csvread(d(i).name);
11     cloud_fraction=dat(2:end,2:end);
12     means(i)=nanmean(cloud_fraction(:));
13
14     % % problem:
15     % [longs,lats]=meshgrid(dat(1,2:end),dat(2:end,1));
16     % weighted=cloud_fraction.*cos(lats.*pi./180);
17     % means(i)=nanmean(weighted(:));
18 end
19
20 % Is mean cloud cover >0.5? test at the 5% significance level
21 t_statistic=abs((mean(means)-0.5)/(std(means)./sqrt(N)))
22
23 t_critical=abs(tinv(0.05,N-1))
24
25 if (t_critical>t_statistic)
26     disp(['Accept the null hypothesis:', ...
27         'there is no difference between 0.5 and the sample data']);
28 else
29     disp(['Reject the null hypothesis:', ...
30         'there is a difference between 0.5 and the sample data']);
31 end

```

This code is downloadable from Blackboard in the folder for this weeks practical. If you put it on your hard disk in the same directory as the csv files you will be able to run it from the MATLAB commandline by typing:

```
1 [means]=process_cloud_fraction2;
```

and `means` will contain the 12 global means.

Question 2: To test whether more of the earth is covered by liquid clouds than ice clouds, take a look at the following code. *You may need to alter the significance level for your test.*

```

1  function [means1, means2]=process_cloud_fraction3
2  % Read in the cloud fraction data and calculate a global mean for each
3  % month. Is the global mean fraction of liquid clouds greater than ice
4  % clouds?
5
6  d1=dir('A*_liquid.csv');
7  d2=dir('A*_ice.csv');
8
9  N=length(d1); % sample size
10 means1=zeros(size(d1));
11 means2=zeros(size(d1));
12 for i=1:length(d1)
13     dat1=csvread(d1(i).name);
14     cloud_frac_liquid=dat1(2:end,2:end);
15     means1(i)=nanmean(cloud_frac_liquid(:));
16
17     dat2=csvread(d2(i).name);
18     cloud_frac_ice=dat2(2:end,2:end);
19     means2(i)=nanmean(cloud_frac_ice(:));
20
21 end
22
23 % Is mean liquid cloud cover > mean ice cloud cover?
24 % test at the 5% significance level
25 Sp2=(N-1).*std(means1).^2+(N-1).*std(means2).^2;
26 Sp2=Sp2./((N-1)+(N-1));
27 t_statistic=abs((mean(means1)-mean(means2))./(sqrt(Sp2./(N+Sp2./N)) )
28
29 t_critical=abs(tinvt(0.05,N+N-2))
30
31 if(t_critical>t_statistic)
32     disp(['Accept the null hypothesis:', ...
33          'there is no between the fraction of liquid and ice clouds']);
34 else
35     disp(['Reject the null hypothesis:', ...
36          'there is a difference between the fraction of', ...
37          'liquid and ice clouds']);
38 end

```

This code is downloadable from Blackboard in the folder for this weeks practical. If you put it on your hard disk in the same directory as the csv files you will be able to run it from the MATLAB command line by typing:

```
1 [means1, means2]=process_cloud_fraction3;
```

means1 and means2 will contain the 12 liquid and ice fraction global means respectively.

Question 3: 2006-7 was an El-Nino year. To test whether the cloud fraction is different in months of an El-Nino year vs months not in an El-Nino year look at the following code. *You may need to alter the significance level for your test.*

```

1 function [means1, means2]=process_cloud_fraction4
2 % Read in the cloud fraction data and calculate a global mean for each
3 % month. Is the global mean fraction of liquid clouds greater than ice
4 % clouds?
5
6 d1=dir('A*_all.csv');
7 d2=dir('C*_all.csv');
8
9 N1=length(d1); % sample size
10 N2=length(d2); % sample size
11 means1=zeros(size(d1));
12 means2=zeros(size(d2));
13 for i=1:length(d1)
14     dat1=csvread(d1(i).name);
15     cloud_frac_2006=dat1(2:end,2:end);
16     means1(i)=nanmean(cloud_frac_2006(:));
17 end
18 for i=1:length(d2)
19     dat2=csvread(d2(i).name);
20     cloud_frac_2011=dat2(2:end,2:end);
21     means2(i)=nanmean(cloud_frac_2011(:));
22 end
23
24 % Is mean liquid cloud cover > mean ice cloud cover?
25 % test at the 10% significance level
26 Sp2=(N1-1).*std(means1).^2+(N2-1).*std(means2).^2;
27 Sp2=Sp2./((N1-1)+(N2-1));
28 t_statistic=abs((mean(means1)-mean(means2))./(sqrt(Sp2./N1+Sp2./N2)))
29
30 t_critical=abs(tinv(0.10,N1+N2-2))
31
32 if(t_critical>t_statistic)
33     disp(['Accept the null hypothesis:', ...
34         ' there is no between the first year and the second year']);
35 else
36     disp(['Reject the null hypothesis:', ...
37         ' there one year is cloudier than the other']);
38 end

```

This code is downloadable from Blackboard in the folder for this weeks practical. If you put it on your hard disk in the same directory as the csv files you will be able to run it from the MATLAB command line by typing:

```
1 [means1, means2]=process_cloud_fraction4;
```

means1 and means2 will contain the 12 El-Nino means and non El-Nino global means respectively.

4.5 Fancy stuff in MATLAB

Display a pseudo colour plot of the cloud fraction from September 2012:

```

1 % Display the cloud fraction for September 2012
2 dat=csvread('C201209_all.csv');
3
4 % parse 'dat' into long, lat and cloud fraction.
5 long1=dat(1,2:end); % First row of data
6 lat1=dat(2:end,1); % First column of data
7 cloud_fraction=dat(2:end,2:end); % rest of the data
8
9 figure('name','Cloud_fraction')
10 pcolor(long1,lat1,cloud_fraction); shading flat;
11 colormap gray;
12 xlabel('longitude');
13 ylabel('latitude');
14 title('Cloud_fraction_for_September_2012')
15 colorbar
16
17 % Display coast line data
18 load coast
19 hold on;
20 plot(long,lat,'color',[1 1 1]); % plot in white [1 1 1]
21 axis equal; axis tight;
22 % Print the figure as a png
23 print -dpng pcolor_cf.png

```

Display a contour plot of the cloud fraction from September 2012:

```

1 % Display the cloud fraction for September 2012
2 dat=csvread('C201209_all.csv');
3
4 % parse 'dat' into long, lat and cloud fraction.
5 long1=dat(1,2:end); % First row of data
6 lat1=dat(2:end,1); % First column of data
7 cloud_fraction=dat(2:end,2:end); % rest of the data
8
9 figure('name','Cloud_fraction')
10 [c,h]=contourf(long1,lat1,smoothn(cloud_fraction,[7 7]));
11 clabel(c,h);
12 colormap gray;
13 xlabel('longitude');
14 ylabel('latitude');
15 title('Cloud_fraction_for_September_2012')
16 colorbar
17
18 % Display coast line data
19 load coast
20 hold on;
21 plot(long,lat,'color',[1 1 1]); % plot in white [1 1 1]
22 axis equal; axis tight;
23
24 % Print the figure as a png
25 print -dpng contour_cf.png

```

Data analysis 5

Week 11 practical

This practical assesses how to perform correlation and linear regression using Excel or MATLAB.

You may open the Blackboard assessment and work through questions 1-7. Following which you may wish to look at the explanations below (either Excel or MATLAB).

5.1 Fitting a power law

5.1.1 Excel method

When fitting a curve to data in Excel the trickiest thing can be pre-processing your data so you remove the bad values. In this example you will fit a power law of the form $y = ax^b$, where a and b are constants.

1. Download the `AustralianCirrusClouds.csv` file from Blackboard and save somewhere on your computer. Open it up in Excel by clicking on it.
2. You'll see that the file contains 3 columns of data: time in decimal days; Extinction from the lidar; and Ice water content from the *in-situ* probe.
3. Next Insert a column between columns A and B and call it time. Excel time format is the number of decimal days from the year 1900, whereas the spreadsheet has number of decimal days from 0 A.D. Set column B equal to the first row minus 693960 (which is the number of days between 0 AD and the year 1900). Repeat this for all rows.
4. Set the formatting to date and time format—highlight all cells in column B and right click on them, and select `Format cells`. Select 'custom' then in the box on the right scroll down to select `dd/mm/yyyy hh:mm:ss`. Press OK.
5. Plot both datasets on the same time vs y scatter graph (i.e. time on the x-axis and both extinction and ice water content on the y-axis) to inspect for any possible correlations.
6. As discussed we will be fitting a power law to this data. This means a straight line of $\log Ext$ (log of extinction) against $\log IWC$ (log of the ice water content). In two adjacent columns calculate the log of the *Extinction* and the *Ice water content*. You'll notice that there are some errors as you will have taken the log of zero, which is undefined. These will need to be removed from the data, before fitting a curve.
7. Remove the bad data by setting every value that is *not-a-number* to something that can be filtered following the procedure below:
 - (a) In an adjacent column, e.g. column I, type:

$$=IF(NOT(ISNUMBER(F2)), "", F2)$$

can you see what this does?. In this example the cell in F2 should be the log of the Extinction and the cell in G2 should be the log of the Ice water content. Copy this down for all Extinction and Ice water contents (using relative references).

- (b) It is desirable to get rid of the blank gaps in the list of numbers. Do this by first selecting both of the two columns just created right clicking on the cells and selecting copy and then right click on another cell and select paste special->values. Secondly select the new columns of data (which also contain blanks) and press F5. A pop-up box will open; click on Special then select the constants radiobutton from the list and un-check all of the check boxes below the radio button, except Text. Press OK, the blanks should be selected. Right click on the selection and delete all of the selected cells. The blank cells are now deleted!
8. Calculate the correlation coefficient, r , using `=CORREL(array1,array2)` where array1 is the x-values and array2 is the y-values.
 9. Now we wish to fit a straight line to the two columns of data. You can either do this using the LINEST function or the Regression tool in the *data analysis toolbox*. The latter is straight forward—just go to data analysis and select regression, while the former is a little trickier as it is an array function.
 - (a) To use LINEST select two empty adjacent cells (in the horizontal direction) and type `=LINEST(Array1,Array2)`, where Array1 are your y-values in the list and Array2 are your x-values in the list.
 - (b) Press `ctrl+shift+enter` and you should get two numbers in the cells, which are the fit coefficients for a straight line—e.g. $y = b_1x + b_0$.
 - (c) You are fitting $y = ax^b$ or rather $\log y = \log a + b \log x$. Therefore in order to calculate a you should take the anti-log (i.e. 10^{b_0}) of the b_0 fit coefficient (do you see why?). b is just equal to b_1 in this example.
 10. Once you have the fit coefficients try plotting the *actual* Extinction against Ice water content and then plot Extinction against *parameterised* ice water content (i.e. against $aExt^b$). If you've got this correct you should find that the parameterised form agrees well with the data!

5.1.2 MATLAB method

When fitting a curve in MATLAB the trickiest thing is probably getting your data into MATLAB.

1. Download the `AustralianCirrusClouds.csv` file from Blackboard and save somewhere on your computer. Read into MATLAB by changing the MATLAB directory to the location of the file on your computer and reading in using:

```
1 dat=csvread('AustralianCirrusClouds.csv',1,0);
```

The 1 means to skip the text in the first row and the 0 means do not skip any columns. You'll notice that `dat` is a table / matrix of size 1257 by 3 (e.g. `size(dat);`)

- Calculate the log of the two columns (Extinction and Ice Water content) by doing:

```
1 dat2=log10(dat(:,2:3));
```

- Remove the bad values (because you've taken log of a zero) by searching for the indices of the good values and referencing the `dat2` matrix with those indices. Do this as follows:

```
1 index=find(~isinf(dat2(:,1)) & ~isinf(dat2(:,2)) ...
2           & (~isnan(dat2(:,1)) & ~isnan(dat2(:,2))));
```

The fragment of code above means find all of the elements in `dat2(:,1)` and `dat2(:,2)` that are not infinity or *not-a-number* (i.e. all the real numbers!). The MATLAB code `~` means *not*. If you were to exclude it, `index` would contain all of the values that are either infinity or *not-a-number*. The `&` means *and* in the expression above.

- Secondly, once you've found the indices that you want to keep, create a new array / table by using `index` to reference `dat2`:

```
1 dat3=dat2(index,:);
```

`dat3` should now contain all the data, with blanks removed.

- Now calculate the correlation coefficient using:

```
1 corrcoeff(dat3(:,1),dat3(:,2))
```

The elements that aren't unity are the correlation coefficient. You could also use:

```
1 corr2(dat3(:,1),dat3(:,2))
```

- Now fit a straight line of the form $y = b_0x + b_1$ by using:

```
1 P=polyfit(dat3(:,1),dat3(:,2),1);
```

The output, `P` contains both the fit coefficients. You may need to type `format long` to print the required significant figures.

- In order to calculate a and b in $y = ax^b$ take the anti-log of `P(2)` (ie. $10^{P(2)}$).
- Try plotting actual extinction vs ice water content as a scatter plot:

```
1 plot(dat(:,2),dat(:,3),'.')

```

then hold this plot and plot the extinction vs parameterised ice water content and set the axes to display on log scale:

```
1 hold on;
2 plot(dat(:,2),10^P(2).*dat(:,2).^P(1),'k-');
3 set(gca,'yscale','log','xscale','log')
```

5.2 Using your fit to tell you something about Mars

Download the `martian_clouds.csv` file from Blackboard. Either click on it to open it in Excel or type:

```
1 datm=csvread('martian_clouds.csv',2,0);
```

to read into MATLAB.

1. It contains an extinction vertical profile from Mars for both dust and clouds (look at the headers: height; dust extinction; cloud extinction). Use the cloud extinction and the parameterised equation to derive the vertical profile of ice water content.
2. What is the peak ice water content? (i.e. calculate another column using your power law).
3. What is the average ice water content between the levels of *height* = 2700 and *height* = 3900 metres inclusive? E.g. calculate the average ice water content between these heights:

```
1 ind=find(datm(:,1)>=2700 & datm(:,1)<=3900);  
2 xbar=mean(iwc(ind));
```

where `iwc` is the *ice water content* calculated from your power law.

At the end, save your spreadsheet or MATLAB workspace (e.g. type: `save <filename>`).

5.3 More fancy stuff in MATLAB

MATLAB can differentiate and integrate functions for you. For example lets say you wanted to know the derivative of:

1. $x^2 + \frac{1}{x}$
2. $ax^3 \exp(-bx)$
3. $\frac{1}{\sqrt{ax^2+bx+c}}$

just type the following:

```
1 diff('x^2+1/x','x') % differentiate wrt x and display result
2
3 diff('a*x^3*exp(-b*x)','x') % differentiate wrt x and display result
4
5 diff('1/sqrt(a*x^2+b*x+c)','x') % differentiate wrt x and display result
```

MATLAB will display:

```
1 ans =
2
3 2*x - 1/x^2
4
5 ans =
6
7 3*a*x^2*exp(-b*x) - a*x^3*b*exp(-b*x)
8
9 ans =
10
11 -1/2/(a*x^2+b*x+c)^(3/2)*(2*a*x+b)
```

thus doing the differentiation for you.

Say you wanted to know the indefinite integral:

1. $\int x^2 + \frac{1}{x} dx$
2. $\int ax^3 \exp(-bx) dx$
3. $\int \frac{1}{\sqrt{ax^2+bx+c}} dx$

just type the following:

```
1 int('x^2+1/x','x') % integrate wrt x and display result
2
3 int('a*x^3*exp(-b*x)','x') % integrate wrt x and display result
4
5 int('1/sqrt(a*x^2+b*x+c)','x') % integrate wrt x and display result
```

MATLAB will display:

```
1 ans =
2
3 1/3*x^3+log(x)
4
5 ans =
6
7 -(6+6*b*x+3*x^2*b^2+x^3*b^3)*a*exp(-b*x)/b^4
8
9 ans =
10
11 log((1/2*b+a*x)/a^(1/2)+(a*x^2+b*x+c)^(1/2))/a^(1/2)
```

thus doing the integrations for you.