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

Data analysis methods in weather and climate research

Dr. David B. Stephenson
D.B.Stephenson@reading.ac.uk
Department of Meteorology
University of Reading

July 20, 2005

Course outline

1. Introduction
2. Descriptive statistics
3. Basic probability concepts
4. Probability distributions
5. Parameter estimation
6. Statistical hypothesis testing
7. Basic linear regression
8. Multiple and nonlinear regression
9. Introduction to time series

Course Aim: To introduce the basic statistical concepts relevant to environmental science and to provide experience in the correct use and interpretation of the various statistical methods currently used in the analysis of weather/climate observed and model simulated data.

Practical Exercises Each topic covered in the lectures will be followed by exercises analyzing real data in practical computing classes using R statistical software.

Prerequisites Minimal statistical knowledge but some basic mathematics and computer skills will be assumed.

"Some people hate the very name of statistics but I find them full of beauty and interest. Whenever they are not brutalized, but delicately handled by the higher methods, and are warily interpreted, their power of dealing with complicated phenomena is extraordinary."

- Sir Francis Galton [1]

[1] the person who invented the concept of *correlation* and the word *anticyclone*!

Acknowledgements

The development of this course has benefited from many stimulating discussions with colleagues and students over the past few years. In particular, I wish to thank Nils Gunnar Kvamsto, Rasmus Benestad, and students at the University of Bergen for their helpful comments on the first version of this course.

In addition, I would like to thank my kind colleagues in the Climate Analysis Group at the University of Reading, who took the time to read the whole of these lecture notes and provided invaluable comments (Chris Ferro, Abdel Hannachi, Sergio Pezzulli, Barbara Casati, and Matt Sapiano).

Despite all this feedback, there is a non-zero probability that these notes might still contain some mistakes - if you find any mistakes or have any suggestions for improvements then please let me know by sending an email to D.B.Stephenson@reading.ac.uk.

Contents

Chapter 1

Introduction

Aim: The aim of this chapter is to explain the purpose and structure of this course and to present a brief introduction to the philosophy of statistics and available statistical software.

1.1 Purpose of this course

This course will help you to develop:

- statistical expertise necessary for atmospheric and climate research
- an ability to choose appropriate analysis methods
- the practical skills needed to apply statistical methods
- the ability to critically interpret the results of analyses
- a deeper appreciation of statistical science

Many important topics will be covered but not at very great depth due to time limitations. Instead, it is hoped that this course will provide you with the basic understanding and skills necessary for applying and interpreting statistics in climate and environmental research. The underlying concepts and pitfalls of various methods will be explained in order to avoid you treating statistical analyses as magical "black box" numerical recipes.

1.2 Brief history of statistics

The Oxford English etymological dictionary defines statistics as follows:

> **statistics** - first applied to the political science concerned with the facts of a state or community XVIII; all derived immediately from German *statistisch* adj., *statistik* sb.; whence *statistician* XIX.

Statistics is concerned with exploring, summarising, and making inferences about the *state* of complex systems, for example, the state of a nation (*official statistics*), the state of peoples' health (*medical and health statistics*), the state of the environment (*environmental statistics*), etc.

Table 1.1 gives a brief summary of some of the earlier key developments in statistics and probability in Europe over the last five centuries. The initial development of statistics in the 16th and 17th centuries was motivated by the need to make sense of the large amount of data collected by population surveys in the emerging European nation states. Then, in the 18th century, the mathematical foundations were improved significantly by breakthroughs in the theory of probability inspired by games of chance (gambling). In the 19th century, statistics started to be used to make sense of the wealth of new scientific data. Finally, in the 20th century, modern statistics has emerged and has continued to progress rapidly throughout the whole century. The development of electronic computers in the 1950s and ever increasing amounts of available data have both played key roles in driving statistics forwards. For a more complete historical review of statistics refer to the books by David (1962), Johnson and Kotz (1998) and Kotz and Johnson (1993).

Year	Event	Person
1532	First weekly data on deaths in London	Sir W. Petty
1539	Start of data collection on baptisms, marriages, and deaths in France	
1608	Beginning of parish registry in Sweden	
1654	Correspondence on gambling with dice	P. de Fermat B. Pascal
1662	First published demographic study based on bills of mortality	J. Graunt
1693	Publ. of *An estimate of the degrees of mortality of mankind drawn from curious tables of the births and funerals at the city of Breslaw with an attempt to ascertain the price of annuities upon lives*	E. Halley
1713	Publ. of *Ars Conjectandi*	J. Bernoulli
1714	Publ. of *Libellus de Ratiocinus in Ludo Aleae*	C. Huygens
1714	Publ. of *The Doctrine of Chances*	A. De Moivre
1763	Publ. of *An essay towards solving a problem in the Doctrine of Chances*	Rev. Bayes
1809	Publ. of *Theoria Motus Corporum Coelestium*	C.F. Gauss
1812	Publ. of *Théorie analytique des probabilités*	P.S. Laplace
1834	Establishment of the Statistical Society of London	
1839	Establishment of the American Statistical Association (Boston)	
1889	Publ. of *Natural Inheritance*	F. Galton
1900	Development of the χ^2 test	K. Pearson
1901	Publ. of the first issue of *Biometrika*	F. Galton et al.
1903	Development of Principal Component Analysis	K. Pearson
1908	Publ. of *The probable error of a mean*	"Student"
1910	Publ. of *An introduction to the theory of statistics*	G.U. Yule
1933	Publ. of *On the empirical determination of a distribution*	A.N. Kolmogorov
1935	Publ. of *The Design of Experiments*	R.A. Fisher
1936	Publ. of *Relations between two sets of variables*	H. Hotelling

Table 1.1: Summary of some the earlier key events in the development of statistics in Europe. For more historical details, refer to Johnson and Kotz (1998).

1.3 What exactly is statistics ?

The purpose of statistics is to develop and apply methodology for extracting useful knowledge from both experiments and data. In addition to its fundamental role in data analysis, statistical reasoning is also extremely useful in data collection (design of experiments and surveys) and also in guiding proper

scientific inference (Fisher, 1990).

Major activities in statistics include:

- design of experiments and surveys to collect data (e.g. meteorological observation networks)
- exploration and visualization of sample data (e.g. graphs, maps, tables)
- summary description of sample data (e.g. climatology)
- modelling relationships between variables (e.g. predictability studies)
- stochastic modelling of uncertainty (e.g. stochastic parameterisations)
- forecasting based on suitable models (e.g. seasonal forecasting)
- hypothesis testing and statistical inference (e.g. detection and attribution of climate change)

Statistics is neither really a science nor a branch of mathematics. It is perhaps best considered as a meta-science (or language) for dealing with data collection, analysis, and interpretation. As such its scope is enormous and it provides much guiding insight in many branches of science, business, etc. Critical statistical reasoning can be extremely useful for making sense of the ever increasing amount of information becoming available (e.g. via the web). Knowledge of statistics is a very useful transferable skill!

1.4 Some fundamental concepts

Statistical data analysis can be subdivided into **descriptive statistics** and **inferential statistics**. Descriptive statistics is concerned with exploring and describing a **sample** of data, whereas inferential statistics uses statistics from a sample of data to make general statements about the whole **population**. Note that the word "data" is plural and a single element of data is called a "datum", so avoid saying things like "the data has been ...".

Descriptive statistics is concerned with exploring, visualising, and summarizing data sampled from a population but without fitting any probability models to the data. This kind of **Exploratory Data Analysis (EDA)** is used to explore sample data in the initial stages of data analysis. Since no probability models are involved, it can not be used to test hypotheses or to make testable out-of-sample predictions about the whole population. Nevertheless, it is a very important preliminary part of analysis that can reveal many interesting features in the sample data.

Inferential statistics is the next stage in data analysis and involves the **identification** of a suitable probability model. The model is then fitted to the data to obtain an optimal **estimation** of the model's **parameters**. The model then undergoes **evaluation** by testing either **predictions** or **hypotheses** of the model. Models based on a unique sample of data can be used to infer generalities about features of the whole population.

Much of climate analysis is still at the descriptive stage, and this often misleads climate researchers into thinking that statistical results are not as testable or as useful as physical ideas. This is not the case and statistical thinking and model-based inference can be exploited to much greater benefit to make sense of the complex climate system.

1.5 Statistical software

The development of computer technology since the 1950s has led to the creation of many very useful statistical software packages for analysing data. Off-the-shelf statistical software now makes it possible to perform analyses on a personal computer that would have been completely impossible in the pre-computer era. For this reason, **computational statistics** is now a large and rapidly advancing branch of modern statistics. Many diverse statistical software packages are currently available that offer a wide variety of capabilities. They can be broadly classified into three main categories:

1. **Powerful language-based packages**
   For example, S-PLUS, R, and SAS, which are packages that allow the user to develop their own statistical macros and functions in addition to the comprehensive range of statistical routines available. These powerful language-based packages are used by many practising statisticians. They are not particularly user-friendly but once mastered can be extremely powerful tools. The R software is freely available to run on many different computer platforms from www.r-project.org.

2. **Interactive packages**
   For example, MINITAB and SPSS, which are packages that allow the user to perform many standard statistical operations at the click of a mouse. These are quick and easy to use and are useful for applying standard methods but not ideally suited for developing new functions. A big danger with such packages is that the user can easily perform operations that they do not understand. This can create a "black box" view of statistical methods that often leads to poor interpretations.

3. **Packages with statistical libraries**
   For example, MATLAB and PV-Wave/IDL, which are primarily data analysis and visualization programs/languages that also include libraries including statistical functions. These packages can be useful in climate analysis since they can cope with the large gridded data sets quite easily and can also be used to quickly visualise spatial data. A problem with these packages is that the libraries often contain only a subset of standard statistical functions, and do not benefit from input from professional statisticians. This is particularly the case with certain spreadsheet packages such as EXCEL that contain rather idiosyncratic and poorly developed statistical libraries.

4. **Home made subroutines**
   Many climate researchers have a bad habit of doing statistics using subroutines in Fortran that they have either written by themselves, obtained from a friend, or copied from numerical recipes. This Do-It-Yourself cookbook approach has several disadvantages that include time being wasted reinventing the wheel programming routines rather than time

being spent thinking about the appropriate choice of method etc., and lack of any input or contact with professional statisticians. The lack of statistical input can lead to ignorance about the range of possible methods available, and the problems associated with the different methods. Just as good surgeons make use of the best professional instruments for their work rather than using just a few home made tools, so one should expect scientists to use the best data analysis software at their disposal rather than something they just hacked together. Good analysis requires the expert use of good tools.

1.6 Further reading for this course

These notes have been written to give you a readable introduction to basic statistics. The lectures will cover most (but not all) of the material in these notes, so please make an effort to read all the chapters.

There are many good books available covering all aspects of statistics. A cheap and readable introduction to univariate statistics can be found in the Schaum outline series book by Spiegel (1992). This book is well written and contains many good examples, but is poor on explaining the underlying concepts in statistics. The introductory text books by DeGroot and Schervish (2002), Rice (1995) and Wackerley et al. (1996) provide a clear and much deeper coverage of basic statistics. In addition, to these books, there are several recent books on statistics and data analysis written specifically for either meteorologists or oceanographers, for example, Wilks (1995) and von Storch and Zwiers (1999), and Emery and Thomson (1997). An interesting review of the history of probability and risk can be found in the popular book by David (1962)

In addition to books, there is also a large amount of useful statistical help and information available online - some of the most useful links are listed on the web page for this course:

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

By reading the help on these pages, you will be able to deepen your knowledge of statistics and learn from statisticians how best to do things. Remember that it is most likely that you are not the first person to have used a particular method and that there is a wide range of statistical information and advice available in text books or via the web. There is no excuse for not knowing about the method you have used to analyse your data - good scientists always know how to use their tools properly !

Chapter 2

Descriptive statistics

Aim: The aim of this chapter is to present common methods used to summarise and describe simple data sets. This first stage in data exploration should be approached with an inquisitive and open mind following the motto "let the data speak for themselves". The aim is not to torture the data into confirming some prior belief!

2.1 Tabulation and the data matrix

Small samples of data are best presented in the form of a table [1]. For example, Table 2.1 below presents the age, height, and weight of a sample of some colleagues at the Department of Meteorology in the University of Reading. There are $n = 11$ **objects** (or **individuals** or **units**) in the **sample** with $p = 3$ observed **variables** age, height, and weight. Note that for good clarity and typesetting, published tables should not include ANY vertical lines or shading even if certain word processors allow such features.

The table of numbers can be considered to be a rectangular **data matrix** $\mathbf{X}$ having $n = 11$ rows and $p = 3$ columns. The data matrix $\mathbf{X}$ has dimension $(n \times p)$ and elements x_{ij} where the first subscript $i = 1, 2, \ldots, n$ is the object

[1] this is good practice when publishing scientifically since it allows others to examine your data and reproduce your results !

Person	Age (years)	Height (cm)	Weight (kgs)
1	30.9	180	76
2	26.9	164	64
3	33.2	176	87
4	28.5	172	75
5	32.3	176	75
6	37.0	180	86
7	38.3	171	65
8	31.5	172	76
9	32.8	161	75
10	37.7	175	85
11	29.1	190	83

Table 2.1: Some bodily characteristics of a small sample of (unnamed !) meteorologists in the Department of Meteorology at the University of Reading.

index and the second subscript $j = 1, 2, \ldots, p$ is the variable index. Note: it is conventional to denote variables by columns and sample objects by rows.

The rest of this lecture will focus on the special case of descriptive methods for **univariate** data $\{x_i : i = 1, 2, \ldots, n\}$ having only one variable ($p = 1$). Many descriptive methods have also been developed for exploring **multivariate** data having more than one ($p > 1$) variable and some of these will be covered in later lectures.

2.2 Descriptive statistics for univariate data

2.2.1 Key attributes of sample data

Numerical summaries of univariate data, such as the heights in the previous table, are useful for making quantitative comparisons with other samples. The following quantities are of paramount interest:

- **Sample size** is the number of objects making up the sample. It is also the number of rows n in the data matrix. It strongly determines the power of inferences made from the sample about the original population from which the sample was taken. For example, sample statistics based on a sample with only 11 people are not likely to be very representative of statistics for the whole population of meteorologists at the University of Reading.

- **Central Location** is the typical average value about which the sampled values are located. In other words, a typical size for the variable based on the sample. It can be measured in many different ways, but one of the most obvious and simplest is the arithmetic **sample mean**:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.1}$$

  For the example of height in the previous table, the sample mean is equal to 174.3cm which gives an idea of the typical height of meteorologists in Reading.

- **Scale** is a measure of the spread of the sampled values about the central location. The simplest measure of the spread is the **range**, $r = \max\{x_1, \ldots, x_n\} - \min\{x_1, \ldots, x_n\}$, equal to the difference between the largest value and the smallest value in the sample. This quantity, however, is based on only the two most extreme objects in the sample and ignores information from the other $n-2$ objects in the sample. A more democratic measure of the spread is given by the **standard deviation**

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{2.2}$$

  which is the square root of the **sample variance**. [2]

- **Shape** of the sample distribution can be summarized by calculating higher moments about the mean such as

$$b_1 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s}\right)^3 \tag{2.3}$$

$$b_2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s}\right)^4 \tag{2.4}$$

  b_1 is called the moment measure of **skewness** and measures the **asymmetry** of the distribution. $b_2 - 3$ is the moment measure of **kurtosis** and measures the flatness of the distribution.

2.2.2 Resistant statistics

Observations often contain rogue **outlier** values that lie far away from the bulk of the data. These can be caused by measurement or recording errors or can be due to genuine freak events. Especially when dealing with small samples, outliers can bias the previous summary statistics away from values representative for majority of the sample.

This problem can be avoided either by eliminating or downweighting the outlier values in the sample (**quality control**), or by using statistics that are **resistant** to the presence of outliers. Note that the word **robust** should not be used to signify resistant since it is used in statistics to refer to insensitivity to choice of probability model or estimator rather than data value. Because the range is based on the extreme minimum and maximum values in the sample, it is a good example of a statistic that is not at all resistant to the presence of an outlier (and so should be interpreted very carefully !).

[2] a denominator of $n-1$ is often used rather than n in order to ensure that the sample variance gives an unbiased estimate of the population variance.

2.2.3 Empirical quantiles

One way of obtaining resistant statistics is to use the **empirical quantiles** (percentiles/fractiles). The quantile (this term was first used by Kendall, 1940) of a distribution is the number x_p such that a proportion p of the values are less than or equal to x_p. For example, the 0.25 quantile $x_{0.25}$ (also referred to as the 25th percentile or lower **quartile**) is the value such that 25% of all the values fall below that value.

Empirical quantiles can be most easily constructed by sorting (ranking) the data into ascending order to obtain a sequence of **order statistics** $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ as shown in Figure 2.1b. The p'th quantile x_p is then obtained by taking the **rank** $r = (n+1)p$'th order statistic $x_{((n+1)p)}$ (or an average of neigbouring values if $(n+1)p$ is not integer):

$$x_p = \begin{cases} x_{((n+1)p)} & \text{if } (n+1)p \text{ is integer} \\ 0.5 * (x_{([(n+1)p])} + x_{([(n+1)p]+1)}) & \text{otherwise} \end{cases} \quad (2.5)$$

where p is the probability $\Pr(X \leq x_p) = r/(n+1)$ and $[a]$ is the greatest integer not exceeding a. Note that the empirical probability $p = r/(n+1)$ is only defined at discrete values - quantiles for other values of p can be obtained either by interpolation ($1 \leq p(n+1) \leq n$) or by extrapolation ($p < 1/(n+1)$ or $p > n/(n+1)$). The use of $(n+1)$ rather than n in the denominator of $p = r/(n+1)$ prevents issuing probabilities that are either zero or one (i.e. perfect certainty) based on only a finite sample of data. As an example, the quartiles of the height example are given by $x_{0.25} = x_{(3)} = 171$ (**lower quartile**), $x_{0.5} = x_{(6)} = 175$ (**median**), and $x_{0.75} = x_{(9)} = 180$ (**upper quartile**).

Unlike the arithmetic mean, the **median** $x_{0.5}$ is not at all influenced by the exact value of the largest objects and so provides a resistant measure of the central location. Likewise, a resistant measure of the scale can be obtained using the **Inter-Quartile Range (IQR)** given by the difference between the upper and lower quartiles $x_{0.75} - x_{0.25}$. In the asymptotic limit of large sample size ($n \to \infty$), for normally (Gaussian) distributed variables (see Chapter 4), the sample median tends to the sample mean and the sample IQR tends to

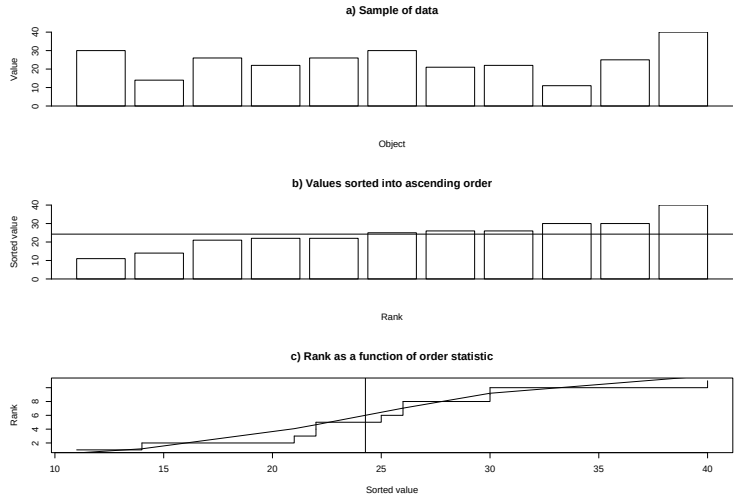


Figure 2.1: Diagram showing how the empirical distribution is obtained for the heights given in Table 2.1. All heights are relative to a reference height of 150cm in order to make the differences more apparent.

1.34 times the sample standard deviation. Resistant measures of skewness and kurtosis also exist such as the dimensionless **Yule-Kendall skewness statistic**

$$\gamma_{YK} = \frac{x_{0.25} - 2x_{0.5} + x_{0.75}}{x_{0.75} - x_{0.25}} \quad (2.6)$$

and **Moors kurtosis statistic**

$$\tau_M = \frac{(x_{0.875} - x_{0.625}) + (x_{0.375} - x_{0.125})}{x_{0.75} - x_{0.25}}$$

There also exist other resistant measures based on all the quantiles such as L-moments, but these are beyond the scope of this course - refer to Wilks (1995) and von Storch and Zwiers (1999) for more discussion.

2.2.4 Example: Summary statistics for the height data

Statistic	Symbol	Value	Comment
sample size	n	11	number of objects/individuals
mean	$\overline{x}$	174.3cm	non-resistant measure of location
standard deviation	s_x	7.9cm	non-resistant measure of scale
range	$x_{max} - x_{min}$	29cm	VERY non-resistant measure of scale!
skewness	b_1	0.17	non-resistant measure of skewness
kurtosis	$b_2 - 3$	0.003	non-resistant measure of kurtosis
median	$x_{0.5}$	175cm	resistant measure of location
interquartile range	$x_{0.75} - x_{0.25}$	9cm	resistant measure of scale
Yule-Kendall	γ_{YK}	0.11	resistant measure of skewness

Table 2.2: Summary statistics for the sample of height data in Table 2.1

.

2.3 Graphical representation

The human eye has evolved to be extremely good at analysing and recognising patterns. This most useful tool can be exploited for data analysis by using it to critically examine plots of the data. Statisticians have developed numerous methods for visualizing samples of both univariate and multivariate data. Some of the standard univariate plots will be illustrated below using the height data:

Boxplot (box-and-whiskers plot) The boxplot is a useful way of plotting the 5 quantiles x_0, $x_{0.25}$, $x_{0.5}$, $x_{0.75}$ and x_1 of the data. The ends of the whiskers show the position of the minimum and maximum of the data whereas the edges and line in centre of the box show the upper and lower quartiles and the median. Sometimes shorter whiskers that extend 1.5 IQR above and below the median are drawn instead of ones that cover the whole range (see the software help for details). The whiskers show at a glance the behaviour of the extreme outliers, whereas the box edges and mid-line summarize the sample in a resistant manner. For symmetrically distributed data the mid-line (median)

is half way between the upper and lower edges of the box (the upper and lower quartiles). The Yule-Kendall skewness statistic in Eqn. (2.6) is a standardised measure of how far the median is from the middle of the box.

Boxplots are particularly useful for comparing multiple samples of data from, say, different experiments. The boxplots for each sample can be *stacked* side-by-side to allow easy visual comparison of the between and within sample spreads of the different samples.

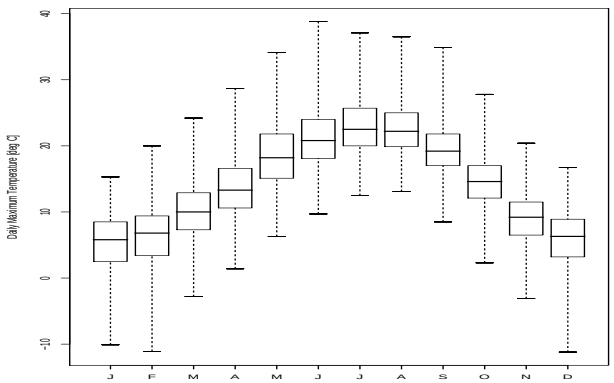


Figure 2.2: Boxplots of 20th century daily maximum temperatures recorded at Uccle, Belgium, split by month.

Histogram The range of values is divided up into a finite set of **class intervals** (**bins**). The number of objects in each bin is then counted and divided by the sample size to obtain the **frequency of occurrence** and then these are plotted as vertical bars of varying height. It is also possible to divide the frequencies by the bin width to obtain **frequency densities** that can then be compared to probability densities from theoretical distributions. For example, a suitably scaled normal **probability density function** has been

superimposed on the frequency histogram in Figure 2.3.

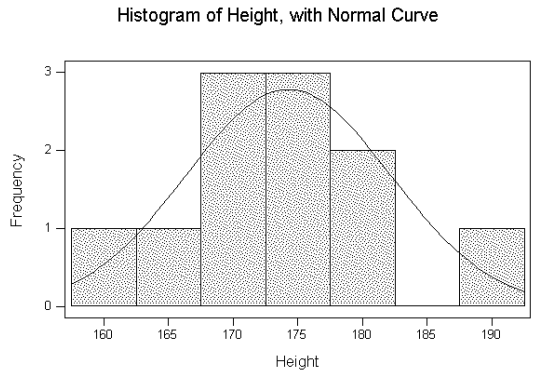


Figure 2.3: Histogram of sample heights showing frequency in each bin with suitably scaled normal density curve superimposed.

The histogram quickly reveals the location, spread, and shape of the distribution. The shape of the distribution can be **unimodal** (one hump), **multimodal** (many humps), **skewed** (fatter tail to left or right), or more-peaked and fatter tails (**leptokurtic**), or less-peaked and thinner tails (**platykurtic**) than a normal (Gaussian) distribution.

Quantile-quantile plot An alternative way to compare the frequency distribution of a sample to a theoretical distribution such as the normal distribution is by plotting the order statistics $x_{(1)}, \ldots, x_{(n)}$ against the corresponding $1/(n+1), \ldots, n/(n+1)$ quantiles of the theoretical distribution. If the sample follows the theoretical distribution then the points will fall approximately along a straight line. This plot is particularly good for revealing discrepancies in the lower and upper tails of the distributions. It can also be used to compare two samples by plotting the two sets of order statistics against one another.

Figure 2.4 shows that the July daily maximum temperatures at Uccle have shorter lower and upper tails than a normal distribution.

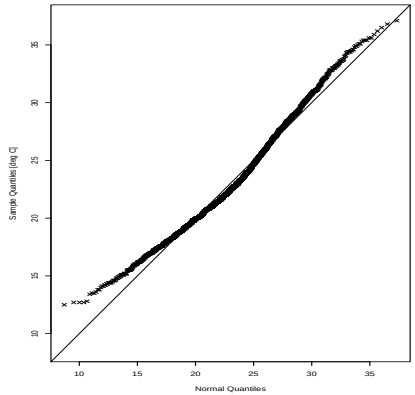


Figure 2.4: Normal-quantile plot of 20th century daily maximum temperatures recorded at Uccle, Belgium, in July.

It can also be useful to plot the **empirical distribution function (e.d.f)** and the theoretically-derived **cumulative distribution function (c.d.f)**. The **e.d.f** (or **ogive**) is a bar plot of the accumulated frequencies in the histogram and the **c.d.f** is the integral of the density function - e.g. the staircase and smooth curve respectively shown in the lower panel of Fig. 2.1. These cumulative distribution functions give directly empirical probabilities p as a function of quantile value x_p. Mathematical definitions of these quantities will be given later in Chapter 4.

2.4 Transformation of data

Transformations are widely used in statistics to make patterns easier to see, or to reduce data to standard forms. Some common methods of re-expressing

data are as follows:

- **Centering** The sample mean (column mean) $\overline{x}$ is subtracted from the data values x_i in order to obtain **centered** "anomalies" $x_i - \overline{x}$ having zero mean. All information about mean location is lost.

- **Standardizing** The data values are centered and then divided by their standard deviations to obtain "normalised anomalies" (meteorological notation) having zero mean and unit variance. All knowledge of location and scale is lost and so statistics based on standardised anomalies are unaffected by any shifts or rescaling of the original data. Standardizing makes the data dimensionless and so is useful for defining standard indices. Certain statistics are unaffected by any linear transformations such as standardization (e.g. correlation, see Chapter 3).

- **Normalizing**

  Normalizing transformations are non-linear transformations often used by statisticians to make data more normal (Gaussian). This can reduce bias caused by outliers, and can also transform data to satisfy normality assumptions that are assumed by many statistical techniques. Note that the transformation just puts the data on a different scale; it needn't change the information content. Note also that meteorologists (and even some statisticians) often confusingly say "normalizing" when what they really mean is "standardizing"!

  A much used class of transformations is the Box-Cox power law transformation $y = (x^\lambda - 1)/\lambda$, where λ can be optimally tuned. In the limit as $\lambda \to 0$, one obtains the $y = \log x$ transformation much used to make postively skewed quantities such as stockmarket prices more normal. The $\lambda = 0.5$ square root transformation is often a good compromise for postively skewed variables such as rainfall amounts (Stephenson et al. 1999).

2.5 Further reading

Many of the modern techniques used in Exploratory Data Analysis (EDA) such as box plots were introduced by Tukey (1977) who emphasized the great importance of using novel descriptive statistics. Chapter 3 of Wilks (1995) covers standard descriptive techniques in more detail.

Chapter 3

Basic probability concepts

Aim: The aim of this chapter is to present a brief introductory overview into the fascinating concept of probability. A good grasp of probability is essential for understanding statistical analysis and for making correct statistical inferences.

3.1 Motivation

In environmental science and many other subjects there is certainly no shortage of uncertainty, for example, in our knowledge about whether it will rain or not tomorrow. Uncertainty about such events arises naturally from errors and gaps in measurements, incomplete and incorrect knowledge of the underlying mechanisms, and also from the overall complexity of all the possible interactions in real-world systems. We try to describe this uncertainty qualitatively by using words such as "likely", "probably", "chance", etc.. However, to make progress scientifically it is necessary to use a more quantitative definition of uncertainty. In 1812, the French mathematician Pierre Simon Laplace defined the word "probability" to mean a number lying between 0 and 1 that measures the amount of certainty for an event to occur. A probability of 1 means the event is completely certain to occur, whereas a probability of 0 means that the event will certainly never occur. Probability is essential for understanding

how samples of data can be drawn from the underlying population, and for making inferences about this population based on sample statistics.

3.2 Events and event space

Probability is the chance of a random event happening. An **event** A is a set (or group) of possible outcomes of an uncertain process e.g. {Heads in a single coin toss}, {rain}, {no rain}, $\{T > 20°C\}$, $\{10° \leq T < 20°C\}$. Events can be **elementary** (indivisible) or **compound** e.g. {Heads in a single coin toss} (elementary), {One head and one tail in 2 coin tosses} (compound). The set of ALL possible elementary events defines **event space** (**sample space**), which sometimes can be represented visually by using an Euler (or Venn) diagram. Figure 3.1 shows the event space for two events $\{A_1, A_2\}$. As an example, $\{A_1\}$ could be the event "precipitating cloud" and $\{A_2\}$ could be the event "boundary layer below freezing". In order to get snow falling on the ground, it is necessary that both events occur at the same time (the intersection region).

3.3 Random variables

A **random variable**, X, is a label allocated to a random event A (e.g. $X = 1$ if a tornado occurs and $X = 0$ otherwise). In statistical literature, random variables are often abbreviated by "r.v." and are denoted by upper case letters (e.g. X). Actual observations (samples) of a particular random variable are denoted by the corresponding lower case letter (e.g. x). In other words, x is a possible value that random variable X can take. Data $x_1, \ldots, x_n$ making up a sample can often be thought of as repeated observations of the same random variable X.

Random variables can either be categorical, discrete numbers (i.e. integers), or continuous numbers (i.e. real numbers). Categorical variables can either be nominal (no ordering) e.g. {sun}, {rain}, {snow}, or cardinal (ordered) e.g. $\{T \leq 0°C\}$, $\{T > 0°C\}$. Discrete random variables can be binary

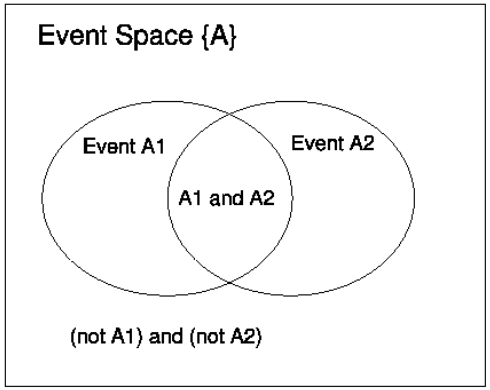


Figure 3.1: Euler diagram showing event space for 2 events.

(e.g. $X = 0$ or $X = 1$) or can be count variables (e.g. $X = 0, 1, 2, \ldots$) representing the number of events (e.g. number of hurricanes). The probability $\Pr(X = x_i)$ of a random variable X taking different observable values, $\{x_i\}$ defines the probability distribution discussed in the next chapter.

3.4 How is probability defined?

The probability of an event $\Pr(A)$ is a measure between 0 and 1 of whether the event is likely to happen. The word "probability" is derived from the word "probable" that comes from the Latin word probābilis meaning *provable*, which is rather paradoxical since only when the the probability is exactly one or zero can anything be definitely proved! The different ways that can be used to define the concept of probability are briefly described below:

3.4.1 Definition 1: Number of symmetric ways

If an event A can happen in w_A ways out of a total of w *equally likely* possible ways, then the probability of A is given by $\Pr(A) = w_A/w$. For example, the probability of getting an odd number when throwing a 6-sided die is given by 3/6 since there are 3 ways to get an odd number (i.e. numbers {1,3,5}) out of a total of 6 equally likely outcomes (i.e. numbers {1,2,3,4,5,6}).

3.4.2 Definition 2: Relative frequency of repeated event

For repeated events, probability can be estimated by the "long-run" relative frequency of an event out of a set of many trials. If an event occurs m times in n trials then the relative frequency m/n provides an unbiased estimate of the probability of the event. In the asymptotic limit as the number of trials n tends to infinity, the relative frequency converges to the true probability of the event ("Law of large numbers"). This interpretation involving repeated trials is known as the "frequentist" approach to probability.

3.4.3 Definition 3: Non-frequentist subjective approach

The frequentist approach has a number of disadvantages. Firstly, it can not be used to provide probability estimates for events that occur once only or rarely (e.g. climate change). Secondly, the frequentist estimates are based ENTIRELY on the sample and so can not take into account any prior belief (common sense) about the probability. For example, an unbiased coin could easily produce 2 heads only when tossed 10 times and this would lead to a frequentist probability estimate of 0.2 for heads. However, our belief in the rarity of biased coins would lead us to suspect this estimate as being too low. In other words, the frequentist estimate does not really reflect our true beliefs in this case.

In such cases a more subjective approach to probability must be adopted that takes into account ALL the available information. The subjective probability of an event A can be defined as the price you would pay for a *fair* bet

on the event divided by the amount you would win if the event happens. Fair means that neither you or the bookmaker would be expected to make any net profit. To make a fair bet all the prior information must be taken into account - e.g. the biasedness of coins, the previous form of a horse in a horse race, etc. This can be done most conveniently by making use of Bayes' theorem (covered later in section 3.5 of this chapter). The Bayesian approach takes a more relativist view of probability and instead uses data to update **prior probability** esimates to give improved **posterior probability** estimates.

3.4.4 Definition 4: The axiomatic approach

None of the above definitions is entirely satisfactory or applicable for all situations. The Russian scientist Kolmogorov, therefore, proposed that probability should be defined axiomatically by stating three necessary and sufficient axioms (assumptions/properties):

1. All probabilities are greater than or equal to zero: $\Pr(A_i) \geq 0$ for all events $\{A_i\}$ (i.e. no event is more unlikely than a zero probability event).
2. The probabilities of all events in event space always sum up to one (i.e. something must happen !).
3. The probability of either one or other *mutually exclusive* events (i.e. events that cannot happen at the same time) is equal to the sum of the probabilities of each event alone. In other words, $\Pr(A_1 \text{ or } A_2)=\Pr(A_1)+\Pr(A_2)$ for all mutually exclusive events A_1 and A_2.

All the previous definitions satisfy these axioms and so provide valid and complementary **interpretations** of probability.

3.5 Joint and conditional probabilities

We are often interested in the case when two events happen at the same time. For example, to get snow falling on the ground, it is necessary that two events,

$\{A_1$="precipitating cloud"$\}$ and $\{A_2$="boundary layer below freezing"$\}$ occur at the same time. The probability of two events happening at the same time, $\Pr(A_1 \text{ and } A_2)$, is known as the **joint probability** of events A_1 and A_2. For mutually exclusive events that never occur at the same time, the joint probability is zero.

It is also useful to define the probability of an event GIVEN that another event has happened. This approach is very powerful and is known as **conditioning**. The conditional probability of an event A_1 given A_2 (i.e. conditioned on A_2) is defined as

$$\Pr(A_1|A_2) = \frac{\Pr(A_1 \text{ and } A_2)}{\Pr(A_2)} \tag{3.1}$$

For example, to estimate the probability of rain during El Niño episodes we could use a conditional probability conditioned on El Niño events (rather than all events).

For **independent** events, $\Pr(A_1 \text{ and } A_2) = \Pr(A_1)\Pr(A_2)$ and so the conditional probability $\Pr(A_1|A_2) = \Pr(A_1)$ - in other words, conditioning on independent events does not change the probability of the event. This is the definition of **independence**.

By equating $\Pr(A_1 \text{ and } A_2) = \Pr(A_1|A_2)\Pr(A_2)$ and $\Pr(A_1 \text{ and } A_2) = \Pr(A_2|A_1)\Pr(A_1)$, one can derive the following useful identity

$$\Pr(A_1|A_2) = \frac{\Pr(A_2|A_1)\Pr(A_1)}{\Pr(A_2)} \tag{3.2}$$

This is known as Bayes' theorem and provides a useful way of getting from the unconditioned **prior probability** $\Pr(A_1)$ to the **posterior probability** $\Pr(A_1|A_2)$ conditioned on event A_2. Event A_2 is invariably taken to be the occurence of the sample of available data. In other words, by conditioning on the available data, it is possible to obtain revised estimates of the probability of event A_1.

3.6 Odds

A common way of expressing probability is in the form of the **odds** of an event. The odds of an event is defined as the ratio of the probability of the event occurring to the probability of it not occurring i.e. $\Pr(A)/\Pr(notA)$. So an event with probability 0.001 has odds of 1/999 (or 999:1 **against** in gambling jargon). Odds can range from zero to infinity and are equal to one for events whose occurrence and non-occurence are equally likely (known as **evens** by gamblers). Odds can be used to assess the total risk of a set of independent events by simply multiplying together the odds of the individual events.

3.7 Expectation, (co-)variance, and correlation

If probabilities are known for all events in event space, it is possible to calculate the **expectation** (**population mean**) of a random variable X

$$E(X) = \sum_i x_i \Pr(A_i) = \mu_X \quad (3.3)$$

where x_i is the value taken by the random variable X for event A_i; i.e. $A_i = \{X = x_i\}$. As an example, if there is a one in a thousand chance of winning a lottery prize of £1500 and each lottery ticket costs £2 then the expectation (expected long term profit) is -£0.50= $0.001 \times$£(1500-2)) + $0.999 \times$ (-£2). A useful property of expectation is that the expectation of any linear combination of two random variables is simply the linear combination of their respective expecations

$$E(aX + bY) = aE(X) + bE(Y) \quad (3.4)$$

where a and b are (non-random) constants. Note also that $E(XY) = E(X)E(Y)$ if X and Y are independent random variables.

The expectation can also be used to define the **population variance**

$$Var(X) = E((X - E(X))^2) = E(X^2) - (E(X))^2 = \sigma_X^2 \quad (3.5)$$

which provides a very useful measure of the overall uncertainty in the random variable. The variance of a linear combination of two random variables is given by

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2abCov(X, Y) \quad (3.6)$$

where a and b are (non-random) constants. The quantity $Cov(X, Y) = E((X - E(X))(Y - E(Y)))$ is known as the **covariance** of X and Y and is equal to zero for independent variables. The covariance can be expressed as

$$Cov(X, Y) = Cor(X, Y)\sqrt{Var(X)Var(Y)} \quad (3.7)$$

where $Cor(X, Y)$ is a dimensionless number lying between -1 and 1 known as the **correlation** between X and Y. Correlation is widely used to measure the amount of **linear association** between two variables.

Note that the quantities $E(.)$ and $Var(.)$ refer specifically to **population parameters** and NOT **sample** means and variances. To avoid confusion the sample mean of an observed variable x is denoted by $\overline{x}$ and the sample variance is denoted by s_x^2. Sample covariance is denoted s_{xy} and sample correlation is denoted by r_{xy}. These provide estimates of the population quantities but should never be confused with them !

3.8 Summary of statistical notation

Mathematical notation in statistics can often be a source of confusion. Here is a brief summary of some commonly used conventions:

- Upper case Roman letters are used to denote random variables (e.g. X, Y, Z, etc.) whereas lower case Roman letters are used to denote their specific values (e.g. x, y, and z). For example, the probability of a (generic) random variable X exceeding a (specific) value x is given by $\Pr(X > x)$. Therefore, sample data such as measurements are denoted by lower case Roman letters (e.g. a data sample with values $\{x_1, x_2, \ldots, x_n\}$).

- Unobservable quantities such as model parameters and noise are denoted using lower case Greek letters. For example, the linear regression model that explains random variable Y in terms of X is given by $Y = X\beta+\alpha+\epsilon$.

- The hat symbol is used to denote estimated and predicted values. For example, $\hat{\beta}$ is an estimate of the model parameter β, and $\hat{Y}$ is a prediction of the random variable Y.

- Bold face upper case Roman letters are used to denote data matrices containing multiple variables. For example, the $(n \times p)$ data matrix $\mathbf{X}$ contains elements x_{ij} where row index $i = 1, 2, \ldots, n$ refers to the data items/objects, and column index $j = 1, 2, \ldots, p$ refers to the variables.

- The symbol n is often used to denote the sample size (the number of data objects in the sample).

- The population mean μ_X is written in terms of the expectation operator as $\mu_X = E(X)$ whereas the sample mean is denoted using an overline (e.g. $\overline{x}$). Sometimes climate studies using the quantum mechanics bra-ket notation $< . >$ to denote expectation but this non-standard practice should be avoided.

3.9 Further reading

Chapter 6 of Spiegel (1992) provides a clear yet elementary introduction to the basic concepts of probability together with many good examples. Chapter 2 of Wilks (1995) also presents a good introduction to probability with examples taken from meteorological situations. An interesting review of the history of probability and risk can be found in the popular book by David (1962)

Chapter 4

Probability distributions

Aim: The aim of this chapter is to define the concept of theoretical probability distributions useful for modelling both discrete and continuous random variables. Examples of several commonly used distributions will also be discussed.

4.1 Motivation

Statisticians have identified several classes of function suitable for explaining the probability distribution of both discrete and continuous variables. These classes of functions can be considered to be **probability models** for explaining observed data, and provide the necessary link for inferring population properties from sample data. The theoretical (population) probability distribution of a random variable is determined by a small set of population parameters that can be estimated using the sample data. This chapter will present definitions and examples of probability distributions for both discrete and continuous variables.

4.2 Distributions of discrete variables

4.2.1 Definition

The **probability distribution** of a discrete variable X is the set of probabilities $p(x) = \Pr(X = x)$ for all the possible values x of X in the event space. So for a discrete random variable that can take k distinct values in the set $\{x_1, x_2, \ldots, x_k\}$, the probability distribution is defined by the k probabilities $\{\Pr(X = x_1), \Pr(X = x_2), \ldots, \Pr(X = x_k)\}$. The probability distribution contains complete information about ALL the statistical properties of X; for example, once the probability distribution is known, the expectation of any function of X can be calculated.

4.2.2 Empirical estimates

A simple **empirical** estimate of the probability distribution is obtained from a sample of data by calculating the **relative frequency** $f_i = n_i/n$ of occurrence of each event $\{X = x_i\}$. In the limit of large sample sizes, these empirical estimates $\hat{p}(x_i) = f_i$ of the distribution provide increasingly accurate unbiased estimates of the population probabilty distribution. The empirical distribution can be displayed graphically by plotting bars of height f_i for the different values x_i.

4.2.3 Theoretical discrete distributions

The probability distribution $p(x)$ is completely defined by specifying the k values $\{p(x_1), p(x_2), \ldots, p(x_k)\}$. However, in many cases, it is possible to obtain a very good approximation to the distribution by assuming a simple functional form $p(x) = f(x; \theta_1, \ldots, \theta_m)$ determined by a smaller number ($m < k$) of more meaningful **population parameters** $\{\theta_1, \theta_2, \ldots, \theta_m\}$. Over the years, statisticians have identified several **theoretical probability distributions** $p(x) = f(x; \theta_1, \ldots, \theta_m)$ that are very useful for **modelling** the probability distributions of observed data. These distributions are known as **parametric**

probability models since they are completely determined by a few important parameters $\{\theta_1, \theta_2, \ldots, \theta_m\}$. The following subsections will briefly describe some of the functions that are used most frequently to model the probability distribution of discrete random variables.

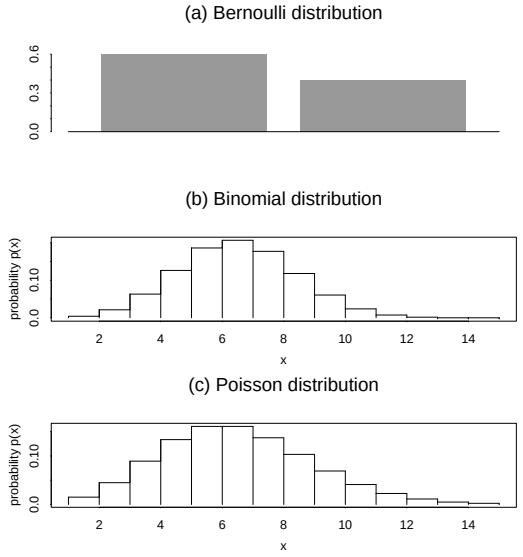


Figure 4.1: Examples of discrete distributions: (a) Bernoulli $\pi = 0.4$, (b) Binomial with $n = 15$ and $\pi = 0.4$, and (c) Poisson with $\mu = 6$.

Example 1: Bernoulli distribution

A Bernoulli (binary) variable is a random variable that can take only the value of either 1 (**success**) or 0 (**failure**). Bernoulli variables are commonly

used for describing binary processes such as coin tossing, rain/no rain, yes/no decisions etc.. The Bernoulli distribution uses one population parameter π to define the probability of success $\Pr(X = 1) = \pi$ and the probability of failure $\Pr(X = 0) = 1 - \pi$. This can be written more succinctly as

$$\Pr(X = x) \quad = \quad \pi^x (1 - \pi)^{1-x} = f(x; \pi) \tag{4.1}$$

where x takes the value of either 0 or 1. The parameter π completely determines the population distribution and all possible statistics based on X, for example, the population mean is given by $E(X) = \pi.1 + (1 - \pi).0 = \pi$ and the population variance is given by $Var(X) = E(X^2) - E(X)^2 = \pi.1^2 + (1 - \pi).0^2 - \pi^2 = \pi(1 - \pi)$. A random variable X distributed with a Bernoulli distribution is described as $X \sim \text{Be}(\pi)$ by statisticians (the $\sim$ symbol means "distributed as").

Example 2: Binomial distribution

Suppose we are interested in counting the number of times X a Bernoulli event with probability π happens in a fixed number n of independent trials. For example, we might be interested in counting the total number of times hurricanes hit Florida out of a specified number of hurricane events. The probability distribution of such a count variable is given by the Binomial distribution $X \sim Bin(n, \pi)$ defined as

$$\Pr(X = m) \quad = \quad \frac{n!}{(n - m)!m!} \pi^m (1 - \pi)^{n-m} = f(m; n, \pi) \tag{4.2}$$

for $m = 0, 1, \ldots, n$ where $n! = n(n - 1)(n - 2) \ldots 1$ and $0! = 1$. The fraction containing factorials on the left hand side is the number of possible ways m successes can happen out of n events, and this can often be surprisingly large. A binomially distributed variable has expectation $E(X) = n\pi$ and variance $Var(X) = n\pi(1 - \pi)$. In the limit of large n, the binomial distribution is well approximated by a normal distribution with mean $n\pi$ and variance $n\pi(1 - \pi)$. So for example, if the probability of a hurricane hitting Florida is $\pi = 0.1$, then out of 200 hurricanes, one would expect a mean of $200 \times 0.1 = 20$ hurricanes

to hit Florida with a standard deviation of $\sqrt{200 \times 0.1 \times 0.9} = 4.2$ hurricanes. The binomial distribution is useful for estimating the fraction of binary events X/n such as the fraction of wet days, or the fraction of people voting for a political party.

Example 3: Poisson distribution

Often we do not know the total number of trials, but we just know that events occur independently and not simultaneously at a mean rate of μ in a certain region of space or in an interval time. For example, we might know that there are a mean number of 20 hurricanes in the Atlantic region per year. In such cases, the number of events X that occur in a fixed region or time interval is given by the Poisson distribution $X \sim Poisson(\mu)$ defined by

$$\Pr(X = m) \quad = \quad \frac{\mathrm{e}^{-\mu}\mu^m}{m!} = f(m;\mu) \tag{4.3}$$

for $m = 0, 1, \ldots$. A Poisson distributed count variable has expectation $E(X) = \mu$ and variance $Var(X) = \mu$. The Poisson distribution approximates the Binomial distribution in the limit of large n and finite $\mu = n\pi$. The sum of two independent Poisson distributed variables is also Poisson distributed $X_1 + X_2 \sim Poisson(\mu_1 + \mu_2)$. Meteorological events such as storms often satisfy the independence and non-simultaneity criteria necessary for a **Poisson process** and so the number of such events in a specified region or time interval can be satisfactorily modelled using the Poisson distribution.

Example 4: Uniform distribution

A random variable has a discrete uniform distribution over a set of values $\{x_1, \ldots, x_k\}$ if each value is equally likely, that is $\Pr(X = x_i) = 1/k$ for each $i = 1, \ldots, k$. A simple example is the outcome from a roll of a fair, standard die.

4.3 Distributions of continuous variables

4.3.1 Definition

Because there is an infinite continuum of possible values x for a continuous random variable X, the probability of X being exactly equal to a particular value is zero $\Pr(X = x) = 0$! Therefore, the approach used to define the probability distribution of discrete random variables can not be used to describe the distribution of continuous random variables. Instead, the **probability distribution** of a continuous variable is defined by the probability of a random variable being less than or equal to a particular value

$$\Pr(X \leq x) \quad = \quad F(x) \tag{4.4}$$

The probability distribution function, $F(x)$, is close to zero for large negative values of x and increases towards one for large positive values of x.

The probability distribution function is sometimes referred to more specifically as the **cumulative distribution function** (c.d.f). The probability of a continuous random variable X being in a small interval $(a, a + \delta x]$ is given by

$$\Pr(a < X \leq a + \delta x) \quad = \quad F(a + \delta x) - F(a) \approx \left[\frac{dF}{dx}\right]_{x=a} \delta x \tag{4.5}$$

The derivative of the probability distribution, $f(x) = \frac{dF}{dx}$, is known as the **probability density function** (p.d.f.) and can be integrated with respect to x to find the probability of X being in any interval

$$Pr(a < X \leq b) \quad = \quad \int_a^b f(x)dx = F(b) - F(a) \tag{4.6}$$

In other words, the probability of X being in a certain interval is simply given by the integrated area under the probability density function curve.

4.3.2 Empirical estimates

An **empirical** estimate of the population probability distribution can be obtained from a sample of data by calculating the **cumulative frequencies** of

objects in the sample having values less than x. A cumulative frequency curve can be calculated either by accumulating up the frequencies in a histogram, or by sorting all the data into ascending order to get estimates of the empirical quantiles $x_q = \hat{F}^{-1}(q)$. In the limit of large sample sizes, these empirical estimates $\hat{F}(x)$ provide increasingly accurate unbiased estimates of the population probability distribution.

4.3.3 Theoretical continuous distributions

Because the estimates are based on a finite number of sample values, the empirical (cumulative) distribution function (e.d.f.) goes up in small discrete steps rather than being a truly smooth function defined for all values. To obtain a continuous differentiable estimate of the c.d.f., the probability distribution can be smoothed using either moving average filters or smoothing splines or kernels. This is known as a **non parametric** approach since it does not depend on estimating any set of parameters.

An alternative approach is to approximate the empirical distribution function by using an appropriate class of smooth analytic function. For a particular class of function (probability model), the location, spread, and/or shape of the probability density function $f(x; \theta_1, \theta_2, \ldots, \theta_m)$ is controlled by the values of a small number of population parameters $\theta_1, \theta_2, \ldots, \theta_m$. This is known as a **parametric** approach since it depends on estimating a set of parameters.

The following sections will describe briefly some (but not all) of the most commonly used theoretical probability distributions.

Example 1: Uniform distribution

A random variable is uniformly distributed $X \sim U(a, b)$ when $f(x) = 1/(b-a)$ for $a \leq x \leq b$ and zero otherwise. In other words, the random variable is equally likely to take any value in the interval $[a, b]$. Standard (pseudo)random number generators on computers and pocket calculators generate random numbers from 0 to 1 that are distributed as $X \sim U(0, 1)$.

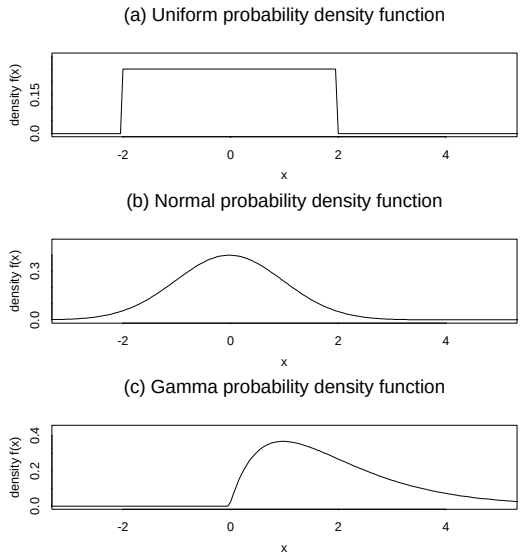


Figure 4.2: Examples of continuous probability density functions: (a) Uniform with $a = -2$ and $b = 2$, (b) Normal with $\mu = 0$ and $\sigma = 1$, and (c) Gamma with $\alpha = 2$ and $\beta = 1$.

Example 2: Exponential distribution

A positive random variable is exponentially distributed $X \sim Expon(\beta)$ when $f(x) = \beta \exp(-\beta x)$ for $x > 0$ and $\beta > 0$. In other words, the random variable is more likely to take small rather than large positive values. The single parameter, β, fully determines the exponential distribution and all its moments, for example, $E(X) = 1/\beta$ and $Var(X) = 1/\beta^2$.

Example 3: Normal (Gaussian) distribution

A random variable is **normally** distributed $X \sim N(\mu, \sigma^2)$ when

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (4.7)$$

where x is any real number and $\sigma > 0$. This is the famous symmetric "bell-shaped" curve that provides a remarkably good description of many observed variables. Although often referred to the **Gaussian distribution** in recognition of the work by K.F. Gauss in 1809, it was actually discovered earlier by De Moivre in 1714 to be a good approximation to the binomial distribution: $X \sim Bin(n, \pi) \approx N(n\pi, n\pi(1-\pi))$ for large n. Rather than refer to it as the "Gaussian" (or "Demoivrian" !) distribution, it is better to simply refer to it as the "normal" distribution.

The reason why the normal distribution is so effective at explaining many measured variables is explained by the **Central Limit Theorem**, which roughly states that the distribution of the mean of many independent variables generally tends to the normal distribution in the limit as the number of variables increases. In other words, the normal distribution is the unique invariant **fixed point** distribution for means. Measurement errors are often the sum of many uncontrollable random effects, and so can be well-described by the normal distribution.

The **standard normal** distribution with zero mean and unit variance $X \sim N(0, 1)$ is widely used in statistics. The area under the standard normal curve, $F(x)$, is sometimes referred to as the **error function** and given its own special symbol $\Phi(x)$, which can be evaluated numerically on a computer to find probabilities. For example, the probability of a normally distributed variable X with mean $\mu = 10$ and $\sigma = 2$ being less than or equal to $x = 14$ is given by $\Pr(X \leq x)$, which is equal to $\Phi((x-\mu)/\sigma) = \Phi((14-10)/2) = \Phi(2) = 0.977$.

Example 4: Gamma distribution

A positive random variable is **gamma** distributed $X \sim Gamma(\alpha, \beta)$ when

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \mathrm{e}^{-\beta x} \quad (4.8)$$

where $\alpha, \beta > 0$. The parameter α is known as the **shape parameter** and determines the shape (skewness) of the distribution, whereas parameter β is known as the **inverse scale parameter** and determines the scale/width of the distribution. The population mean $E(X) = \alpha/\beta$ and the population variance $var(X) = \alpha/\beta^2$. The coefficient of variation, $\sigma/\mu = 1/\sqrt{\alpha}$, provides a simple (moment) method estimate of the shape parameter. The mode of the distribution is less than the mean and located at $(\alpha - 1)/\beta$ when $\alpha > 1$. For $\alpha \leq 1$, the Gamma density is inverse J-shaped with the mode at $x = 0$.

The gamma distribution is useful for describing positively skewed positive variables such as rainfall totals. A nice additive property of gamma distributed variables is that if X_1 and X_2 are independent with $X_1 \sim Gamma(\alpha_1, \beta)$ and $X_2 \sim Gamma(\alpha_1, \beta)$, then $X_1 + X_2 \sim Gamma(\alpha_1 + \alpha_2, \beta)$. For example, the sum S of n independent rainfall totals distributed as $X \sim Gamma(\alpha, \beta)$ will also be Gamma distributed as $S \sim Gamma(n\alpha, \beta)$.

Several commonly used distributions are special cases of the gamma distributions. The exponential distribution $X \sim Expon(\beta)$ is the special case of the Gamma distribution when $\alpha = 1$ i.e. $X \sim Gamma(1, \beta)$. The special case $X \sim Gamma(n/2, 1/2)$ is also known as the **chi-squared distribution** $X \sim \chi^2_n$ with n degrees of freedom. The chi-squared distribution describes the distribution of the sum of squares of n independent standard normal variables, and so for example, the sample variance of n independent normal variates is distributed as $s^2/\sigma^2 \sim \chi^2_{n-1}$ (there are $n-1$ degrees of freedom rather than n since one is lost in estimating the sample mean).

4.4 Further reading

Probability distributions are the main building bricks used by statisticians to model data, and so are covered in most basic statistics books. They are also well described on many online glossarys, which often include instructive graphical demonstrations (e.g. StatSoft's online textbook). Other important examples not discussed in this chapter are Student's t distribution, the beta distribution and extreme-value distributions.

Chapter 5

Parameter estimation

Aim: The aim of this chapter is to explain how **sample statistics** can be used to obtain accurate estimates of **population parameters**. Due to the finite size of samples, all estimates are uncertain but the amount of sampling uncertainty can be estimated using sampling theory.

5.1 Motivation

Imagine that we assume a certain random variable to be distributed according to some distribution $X \sim f(\theta)$ and that we wish to use a sample of data to estimate the **population parameter** θ. For example, we may be interested in estimating either the mean μ or the variance σ^2 (or both) of a variable that is thought to be normally distributed $X \sim N(\mu, \sigma^2)$. A single value **point estimate** $\hat{\theta}$ may be obtained by choosing a suitable sample statistic $\hat{\theta} = t(x)$, for example, the sample mean $t(x) = \overline{x}$ provides a simple (yet far from unique) way of estimating the population mean μ. However, because sample sizes are finite, the sample estimate is only an approximation to the true population value - another sample from the same population would give a different value for the same sample statistic. Therefore, rather than give single value point estimates, it is better to use the information in the sample to provide a range of possible values for θ known as an **interval estimate**. To take account of

the **sampling uncertainty** caused by finite sample size, it is necessary to consider the probability distribution of sample statistics in more detail.

5.2 Sampling distributions

The probability distribution of a sample statistic such as the sample mean is known as a **sampling distribution** (and should not be confused with the probability distribution of the underlying random variable). For example, it can be shown that the sample mean of n independent normally distributed variables $X \sim N(\mu, \sigma^2)$ has a sampling distribution given by $\overline{X} \sim N(\mu, \sigma^2/n)$. In other words, the sample mean of n normal variables is also normally distributed with the same mean but with a reduced variance σ^2/n that becomes smaller for larger samples. Rather amazingly, the sample mean of any variables no matter how distributed has a sampling distribution often tends to normal $\overline{X} \sim N(\mu, \sigma^2/n)$ for sufficiently large sample size. This famous result is known as the **Central Limit Theorem** and accounts for why we encounter the normal distribution so often for observed quantities such as measurement errors etc.

The sampling distribution $f_T(.)$ of a sample statistic $T(X)$ depends on:

- The underlying probability distribution $X \sim f(.)$ determined by its population parameters $(.) = (\mu, \sigma^2$, etc).
- The choice of the particular sample statistic. Simple analytic forms for the sampling distribution can only be derived for a few particularly simple sample statistics such as the sample mean and sample variance. In other cases, it is necessary to resort to computer based simulation methods such as resampling.
- The sample size n. The larger the sample size, the less the uncertainty (spread) in the sampling distribution. For example, the mean heights of two different samples of meteorologists could be quite different for small samples with $n = 11$, whereas the mean heights are unlikely to be very different for large samples with $n = 11000$.

5.3 Sampling errors

For infinitely large sample sizes, the spread of the sampling distribution (sampling uncertainity) tends to zero. However, for finite samples, there is always uncertainty in the estimate due to the finite spread of the sampling distribution (except in the unlikely event that the sample is the whole finite population).

A traditional physicist approach to providing an estimate of this sampling uncertainty is to quote the **standard error**, which is defined as the standard deviation s_t of a sample statistic t (i.e. the spread of the sampling distribution). For example, the heights of meteorologists in Table 2.1 have a sample mean of 174.3cm and a sample standard deviation of 7.9cm, and therefore an estimate of the population mean would be 174.3cm with a standard error of 2.4cm ($s_{\overline{x}} = s/\sqrt{n}$). Physicists write this succintly as $t \pm s_t$ e.g. 174.3 ± 2.4cm. The interval $[t - s_t, t + s_t]$ is known as an **error bar** and it is often stated that a "measurement without an error bar is meaningless". In other words, to interpret a estimate meaningfully you need to have an idea of how uncertain the estimate may be due to sampling.

Sampling errors of linear combinations of independent random variables can easily be estimated by summing sampling variances. If random variable Z is a linear combination $aX + bY$ of two independent and normally distributed variables $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$, then Z is also normally distributed $Z \sim N(\mu_Z, \sigma_Z^2)$ with mean $\mu_Z = a\mu_X + b\mu_Y$ and variance $\sigma_Z^2 = a^2\sigma_X^2 + b^2\sigma_Y^2$. Therefore, the standard error s_Z of $Z = aX + bY$ is $\sqrt{a^2 s_X^2 + b^2 s_Y^2}$, and so, for example, the standard error of the difference of two sample statistics $Z = X - Y$ is simply $\sqrt{s_X^2 + s_Y^2}$ - the quadrature sum of the standard errors of X and Y.

5.4 Confidence intervals

An error bar is a simple example of what statisticians refer to as an **interval estimate**. Instead of estimating a point value for a parameter, the sample data is used to estimate a range of estimates that are likely to cover the true

population parameter with a prespecified probability known as the **confidence level**.

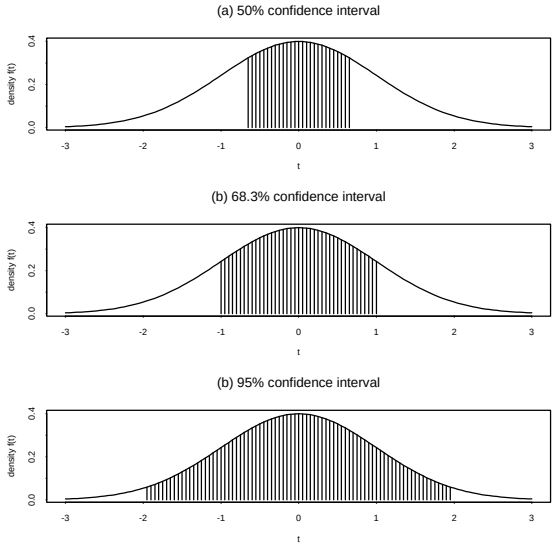


Figure 5.1: Sampling density function of statistic t showing the t_L and t_U lower and upper confidence limits.

A $100(1-\alpha)\%$ **confidence interval** (C.I.) contains the true value of the population parameter θ with probability $1 - \alpha$ (the **confidence level**). The interval $[t_L, t_U]$ is defined by lower and upper **confidence limits** t_L and t_U, which are functions of the data. In other words, if C.I.s were calculated for many different samples drawn from the full population then a $(1 - \alpha)$ fraction of the C.I.s would cover the true population value. These intervals are shown schematically in Fig. 5.1. To be precise, if $F_T(t \mid \theta) = \Pr(T(X) \leq$

t) is the sampling cumulative distribution function that depends on θ then $\Pr\{F_T^{-1}(\alpha/2 \mid \theta) \leq T(X) \leq F_T^{-1}(1-\alpha/2 \mid \theta)\} = 1-\alpha$ and the two inequalities can be rearranged to give $\Pr(t_L \leq \theta \leq t_U) = 1-\alpha$ for some t_L and t_U. In classical (but not Bayesian) statistics, the true population parameter is considered to be a fixed constant and not a random variable, hence it is the C.I.s that randomly overlap the population parameter rather than the population parameter that falls randomly in the C.I.

Statisticians most often quote 95% confidence intervals, which should cover the true value in all but 5% of repeated samples. For normally distributed sample statistics, the 95% confidence interval is about twice as wide as the ± 1 error bar used by physicists (see example below). The ± 1 error bar corresponds to the 68.3% confidence interval for normally distributed sample statistics. In addition to its more precise probabilistic definition, another advantage of the C.I. over the error bar is that it is easily extended to skewed statistics such as sample variance.

5.4.1 Example 1: Confidence Interval for population mean

The sampling distribution for the sample mean tends in the limit of large n to $\overline{X} \sim N(\mu, \sigma^2/n)$. Therefore, the **pivotal quantity** or **test statistic** $Z = (\overline{X} - \mu)/(\sigma/\sqrt{n})$ is distributed as $N(0,1)$. The $(1-\alpha)100\%$ confidence interval for μ can be written

$$\overline{x} - Z_c \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{x} + Z_c \frac{\sigma}{\sqrt{n}} \tag{5.1}$$

where $Z_c(\alpha) = -\Phi^{-1}(\alpha/2)$ is the the half-width of the $(1-\alpha)100\%$ confidence interval measured in standard errors. Z_c is sometimes referred to as a **critical value**. Table 4.1 gives some critical values for various confidence limits:

α	$1-\alpha$	Z_c	Description
0.50	0.50	0.68	50% C.I. ± one "probable error"
0.32	0.68	1.00	68% C.I. ± one "standard error"
0.10	0.90	1.65	90% C.I.
0.05	0.95	1.96	95% C.I. about ± two standard errors
0.001	0.999	3.29	99.9% C.I. about ± three standard errors

Table 5.1: Critical values for various common confidence levels

5.4.2 Example 2: Confidence Interval for sample proportion

Quantities such as the fraction of dry days etc. can be estimated by using the sample proportion. The sample proportion of a binary event is given by the sample mean $\overline{X}$ of a Bernoulli distributed variable $X \sim Be(p)$. The sampling distribution of the number of cases $X = 1$ is given by $n\overline{X} \sim Bin(n,p)$. For large sample size ($n \geq 30$), the binomial distribution $Bin(n,p)$ approximates the normal distribution $N(np, np(1-p))$, and hence the sampling distribution becomes $\overline{X} \sim N(p, p(1-p)/n)$. Therefore, for large enough samples, the proportion is estimated by $\hat{p} = \overline{X}$ with a standard error of $s_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})/n}$. Note the standard errors shrink when $\hat{p}$ gets close to either zero or one. For small samples, the normal approximation can not be used and the C.I.'s are asymmetric due to the skewed nature of the binomial distribution.

5.4.3 Example 3: Confidence Interval for sample variance

For independent and identically distributed (i.i.d) variables $X \sim N(\mu, \sigma^2)$, the sample variance s^2 is the sum of squared normal variates, and is therefore distributed as $s^2/\sigma^2 \sim \chi^2_{n-1}$. The C.I. for sample variance is therefore determined by the $\alpha/2$ and $1-\alpha/2$ quantiles of the chi-squared distribution with $n-1$ degrees of freedom. Because the chi-squared distribution is positively skewed,

the C.I. is asymmetric with a bigger interval between the upper limit and the sample estimate than between the sample estimate and the lower limit.

5.5 Choice of estimator

Generally, a population parameter can be estimated in a variety of different ways by using several different sample statistics. For example, the population mean can be estimated using estimators such as the sample mean, the sample median, and even more exotic sample statistics such as trimmed means etc.. This raises the question of which method to use to choose the best estimator. The three most frequently used estimation approaches are:

1. **Moment method** - the sample moments, $\overline{x}$, $\overline{x^2}$, $\overline{x^3}$, etc., are used to provide simple estimates of the location, scale, and shape parameters of the population distribution. Although these are the most intuitive choices for estimator, they have the disadvantage of giving biased estimates for non-normal distributions (non-robust), and can also be unduly influenced by the presence of outliers in the sample (non-resistant).

2. **Robust estimation** - instead of using moments, robust estimation methods generally use statistics based on quantiles e.g. median, interquartile range, L-moments, etc.. These measures are more robust and resistant but have the disadvantage of giving estimators that have larger sampling errors (i.e. less efficient estimators - see next section).

3. **Maximum Likelihood Estimation (MLE)** - These are most widely used estimators because of their many desirable properties. MLE estimates are parameter values chosen so as to maximise the likelihood of obtaining the data sample. In simple cases such as normally distributed data, the MLE procedure leads to moment estimates of the mean and variance. For more complicated cases, the MLE approach gives a clear and unambiguous approach for choosing the best estimator.

5.6 Accuracy and bias of estimators

The **accuracy** of an estimator $\hat{\theta}$ can be evaluated by its **Mean Squared Error** $E((\hat{\theta}-\theta)^2)$. The MSE can be decomposed into the sum of two parts:

$$E((\hat{\theta}-\theta)^2) = (E(\hat{\theta})-\theta)^2 + var(\hat{\theta}) \tag{5.2}$$

The first term is the square of the mean **bias** $E(\hat{\theta})-\theta$ and measures the difference the mean of ALL sample estimates and the true population parameter. The second term is the variance of the sample estimate caused by sampling uncertainty due to finite sample size. Bias can sometimes be reduced by choosing a different estimator but often at the expense of increased variance.

Estimators with smaller MSE are called more **efficient** estimators, and ones with the smallest MSE are called **Least Squared Error (LSE)** estimators. To obtain the smallest MSE, it is necessary to have small or even no bias (unbiased) and low variance.

5.6.1 Example 1: The sample mean

The expectation of the sample mean is calculated as follows:

$$\begin{aligned} E(\overline{X}) &= E(\frac{1}{n}\sum_{i=1}^{n} X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} E(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu \\ &= \mu \end{aligned} \tag{5.3}$$

Hence, the bias of the sample mean $E(\overline{X})-\mu$ is zero and the sample mean is an "unbiased" estimate of the population mean. As discussed earlier, the variance of the sample mean is given by σ^2/n and, therefore, the MSE of the sample mean estimate is simply given by σ^2/n. As the sample size increases, the MSE tends to zero and the sample mean estimate converges on the true population value. This smooth unbiased convergence is what allows us to use sample means to estimate population means.

5.6.2 Example 2: The sample variance

The sample variance $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$ underestimates the population variance σ^2. Using the same approach as in the previous example (try it !), it is possible to show that $E(s^2) = \sigma^2(n-1)/n$, and therefore the bias $E(s^2) - \sigma^2 = -\sigma^2/n$. This underestimate of the true population variance is greatest when the sample size is very small, for example, the mean sample variance is only 2/3 of the true population variance when $n = 3$. To obtain an unbiased variance estimate, the sample variance is sometimes defined with $n-1$ in the denominator instead of n i.e. $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$. However, it should be noted that this larger estimator also has larger variance than s, and is therefore a less efficient estimator. It is also worth noting that although this estimator gives by design an unbiased estimate of the population *variance*, $\hat{s}$ still remains a biased (over)estimate of the population *standard deviation*.

5.7 Further reading

Chapters 8 and 9 of Spiegel (1992) explain estimation clearly albeit in a rather old fashioned approach. They also introduce briefly the important idea of re-sampling which has not been covered in this chapter. Chapter 3 of Emery and Thomson (1997) give a clear and more in-depth discussion of estimation as used in oceanography. Von Storch and Zwiers (1999) explains the L-moment estimation method based on the published articles of Hosking and collaborators.

Chapter 6

Statistical hypothesis testing

Aim: The aim of this chapter is to explain how sample statistics can be used to make binary true/false inferences (decisions) about the population distribution. How unlikely does the value of a test statistic have to be before we can reject the idea that it might have just happened by chance sampling ?

6.1 Motivation

6.1.1 The basic approach

The previous chapter on estimation showed how it is possible to use sample statistics to make estimates of population parameters that include estimates of sampling uncertainty. However, sometimes we would like to go further and use sample statistics to test the binary true/false validity of certain hypotheses (assumptions) about population parameters. In other words, we want to make true/false decisions about specified hypotheses based on the evidence provided by the sample of data. This "Sherlock Holmes" detective approach is obviously more risky than simply estimating parameters, but is often used to clarify conclusions from scientific work. In fact, the radical idea underlying the whole of natural science is that hypotheses and theories are not only judged by their intrinsic beauty but can also be tested for whether or not they explain observed

data. This is exemplified by the Royal Society's [1] revolutionary famous motto "nullis in verba", which is taken from a poem by the roman poet Horace and means do not (unquestioningly) accept the words (or theories) of anyone !

Suppose, for example, we suspect there might be a non-zero correlation between two variables (e.g. sunspot numbers and annual rainfall totals in Reading). In other words, our scientific hypothesis H_1 is that the true (population) correlation between these two variables is non-zero. To test this hypothesis, we examine some data and find a sample correlation with, let's imagine, a quite large value. Now did this non-zero sample correlation arise because the hypothesis H_1 is really true, or did it just happen by chance sampling ? The hypothesis that the large sample value happened just by chance is known as the **null hypothesis** H_0. In order to claim that the **alternative hypothesis** H_1 is true, we must first show that the large value would be very unlikely to happen by pure chance. In other words, we must use the data to **reject** the null hypothesis H_0 in favour of the more scientifically interesting alternative hypothesis H_1. By using the data to reject H_0, we are able to make a binary decision about which of the two hypotheses is least inconsistent with the data.

6.1.2 A legal example

So the approach is not to use the data to **accept** the alternative hypothesis we are interested in proving, but instead to use the data to **reject** the null hypothesis that our peculiar sample of data might just have happened by chance. In other words, we try to **falsify** the pure chance hypothesis. To help understand why this somewhat perverse approach actually makes sense, consider the legal case of trying to prove whether a suspect is guilty of murder. Based on the available evidence, a decision must be made between the alternative hypothesis that "the suspect is guilty" and the null hypothesis that "the suspect is innocent". [2] If we assume the alternative "guilty" hypothesis, then to avoid conviction we must find evidence of innocence e.g. no sign of a

[1] one of the world's oldest scientific academies founded in 1660

[2] in this particular example, the alternative hypothesis is the negation of the null hypothesis as is often the case in statistical hypothesis testing.

murder weapon with the suspect's fingerprints. However, no sign of a murder weapon (or any other evidence of innocence) does not prove that the suspect is innocent since it could just be that the suspect is guilty but the murder weapon has not yet been found. A different sample of evidence a few years later may contain the murder weapon and invalidate the earlier evidence of innocence. Now consider what happens if we start by considering that the null "innocent" hypothesis is true, and then look for evidence inconsistent with this hypothesis (e.g. the murder weapon). If we find the murder weapon with the suspect's fingerprints, we can clearly reject the null hypothesis that the suspect is innocent, and thereby deduce that the suspect is guilty of murder. One bit of data inconsistent is enough to falsify a hypothesis, but no amount of consistent data can verify a non-trivial hypothesis! Look at what happened to Newton's laws of motion - the theory was consistent with all observed data over several centuries, until finally measurements of the speed of light in the 20th century showed that the whole theory was fundamentally wrong. Therefore, in statistical inference as in science, the correct approach is to use data to falsify rather than verify hypotheses.

6.1.3 Getting rid of straw men

So to test the alternative hypothesis, we set up a "straw man" null hypothesis that we then try to knock down by using the data - this is a fairground analogy where we imagine throwing balls (data) at a straw man (hypothesis) in order to knock it over. If the data are found to be inconsistent with the null hypothesis, we can reject the null hypothesis in favour of the alternative hypothesis - something signficantly different from sampling has happened. If the data are found to be consistent with the null hypothesis, this does not imply that the null hypothesis is necessarily true but only that "data are not inconsistent with the null hypothesis". In other words, we are not allowed in this case to make a big fuss about how exciting the result is - it could easily have happened by chance. This may seem a depressing kind of result but in fact non-rejections can often be as informative and as useful as rejections and so deserve to be published. For example, Sir Gilbert Walker successfully isolated

the Southern Oscillation as a leading global climate pattern by using statistical testing to eliminate all the other confusing world-wide correlations that were due to chance.

6.2 Decision procedure

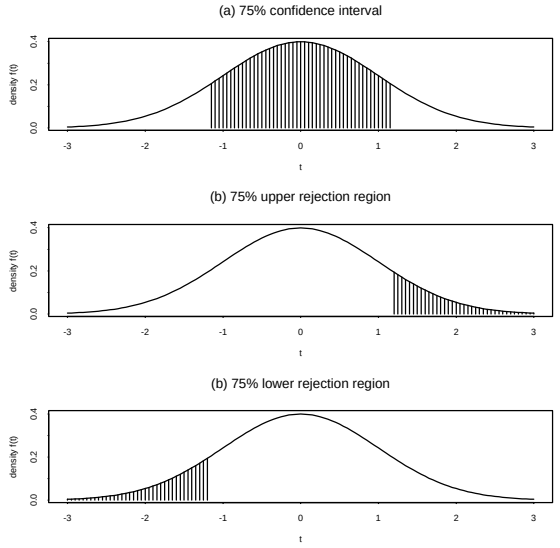


Figure 6.1: Schematic showing the sampling density function of a test statistic assuming a certain null hypothesis. The critical levels are shown for an (unusually large) level of significance $\alpha = 0.25$ for visually clarity.

Statistical testing generally uses a suitable **test statistic** $T(X)$ that can

be calculated for the sample. Under the null hypothesis, the test statistic is assumed to have a sampling distribution that tails off to zero for large positive and negative values of t. Fig. 6.1 shows the sampling distribution of a typical test statistic such as the Z-score $Z = (\overline{X} - \mu)/\sigma$, which is used to test hypotheses about the mean. The decision-making procedure in classical statistical inference proceeds by the following well-defined steps:

1. Set up the most reasonable null hypothesis $H_0 : X \sim f(.)$
2. Specify the **level of significance** α you are prepared to accept. This is the probability that the null hypothesis will be rejected even if it really is true (e.g. the probability of convicting innocent people) and so is generally small (e.g. 5%).
3. Use H_0 to calculate the sampling distribution $T \sim f_T(.)$ of your desired **test statistic** $T(X)$
4. Calculate the **p-value** $p = Pr\{\|T\| \geq t\}$ of your observed sample value t of the test statistic. The p-value gives the probability of finding samples of data even less consistent with the null hypothesis than your particular sample assuming H_0 is true i.e. the area in the tails of the probability density beyond the observed value of the test statistic. So if you observe a particularly large value for your test statistic, then the p-value will be very small since it will be rare to find data that gave even larger values for the test statistic.
5. Reject the null hypothesis if the p-value is less than the level of significance $p < \alpha$ on the grounds that the data are inconsistent with the null hypothesis at the α level of significance. Otherwise, do not reject the null hypothesis since the "data are not inconsistent" with it.

The level of significance defines a **rejection region** (**critical region**) in the tails of the sampling distribution of the test statistic. If the observed value of the test statistic lies in the rejection region, the p-value is less than α and the null hypothesis is rejected. If the observed value of the test statistic lies

closer to the centre of the distribution, then the p-value is greater than or equal to α and the null hypothesis can not be rejected. All values of t that have p-values greater than or equal to α define the $(1 - \alpha)100\%$ confidence interval.

Example: Are meteorologists taller or shorter than other people ?

Let us try and test the hypothesis that Reading meteorologists have different mean heights to other people in Reading based on the small sample of data presented in Table 2.1. Assume that we know that the population of all people in Reading have heights that are normally distributed with a population mean of μ_0 =170cm and a population standard deviation of σ_0 =30cm. So the null hypothesis is that our sample of meteorologists have come from this population, and the alternative hypothesis is that they come from a population with a different mean height. Mathematically the hypotheses can be stated as:

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu \neq \mu_0 \end{aligned} \tag{6.1}$$

where $\mu_0 = 170$cm. Let's choose a typical level of significance equal to 0.05. Under the null hypothesis, the sampling distribution of sample means should follow $\overline{X} \sim N(\mu_0, \sigma_0^2/n)$, and hence the test statistic $Z = (\overline{X} - \mu_0)/(\sigma_0/\sqrt{n}) \sim N(0, 1)$. Based on our sample of data presented in Table 2.1, the mean height is $\overline{X} = 174.3cm$ and so the test statistic z is equal to 0.48, in other words, the mean of our sample is only 0.48 standard errors greater than the population mean. The p-value, i.e. the area in the tails of the density curve beyond this value, is given by $2(1 - \Phi(|z|)$ and so for a z of 0.48 the p-value is 0.63, which is to say that there is a high probability of finding data less consistent with the null hypothesis than our sample. The p-value is clearly much larger than the significance level and so we can not reject the null hypothesis in this case - at the 0.05 level of significance, our data is not inconsistent with coming from the population in Reading. Based on this small sample data, we can not say that meteorologists have different mean heights to other people in Reading.

6.3 Alternative hypotheses

So far we have been considering simple situations in which the alternative hypothesis is just the complement of the null hypothesis (e.g. $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$). The rejection region in such cases includes the tails on both *sides* of $t = 0$, and so the tests are known as **two-sided tests** or **two-tailed tests**. However, it is possible to have more sophisticated alternative hypotheses where the alternative hypothesis is not the complement of the null hypothesis. For example, if we wanted to test whether means were not just different but were larger than the population mean, we would use instead these hypotheses $H_0 : \mu = \mu_0$ and $H_1 : \mu > \mu_0$. The null hypothesis would be rejected in favour of the alternative hypothesis only if the sample mean was significantly greater than the population mean. In other words, the null hypothesis would only be rejected if the test statistic fell in the rejection region to the right of the origin. This kind of test is known as a **one-sided test** or **one-tailed test**. One-sided tests take into account more prior knowledge about how the null hypothesis may fail.

	H_0 true	H_1 true
$p > \alpha$	Correct non rejection	Missed rejection (Type II error)
don't reject H_0	probability $1 - \alpha$	probability β
$p \leq \alpha$ reject	False rejection (Type I error)	Correct rejection
reject H_0	probability α	probability $1 - \beta$

Table 6.1: The four possible situations that can arise in hypothesis testing

Table 6.1 shows the four possible situations that can arise in hypothesis testing. There are two ways of making a correct decision and two ways of making a wrong decision. The false rejection of a true null hypothesis is known as a **Type I error** and occurs with a probability exactly equal to the level of significance α for a null hypothesis that is true. In the legal example, this kind of error corresponds to the conviction of an innocent suspect. This kind of error is made less frequent by choosing α to be a small number typically 0.05, 0.01, or 0.001. The missed rejection of a false null hypothesis is known

as a **Type II error** and corresponds to failing to convict a guilty suspect in the legal example. For a true alternative hypothesis, type II errors occur with an unspecified probability β determined by the sample size, the level of significance, and the choice of null hypothesis and test statistic. The probability $1 - \beta$ is known as the **power** of the test and this should ideally be as large as possible for the test to avoid missing any rejections. There is invariably a trade off between Type I and Type II errors, since choosing a smaller α leads to fewer overall rejections, and so fewer type I errors but more type II errors. To reduce the number of type II errors it is a good idea to choose the null hypothesis to be the simplest one possible for explaining the population ("principle of parsimony").

6.4 Examples of bad practice

The atmospheric sciences literature is full of bad practice concerning statistical hypothesis testing. Hopefully, after this course, you will not contribute to the continuation of these bad habits ! Here are some of the classic mistakes that are often made in the literature:

- **"The results ... are statistically significant"**
  Complete failure to state clearly which hypotheses are being tested and the level of significance. In addition, the way this is stated treats statistical inference as just a way of rubber stamp approving results that the author found interesting. This is not what inference is about.

- **"... and are 95% significant"**
  What the author is trying to say is that the null hypothesis can be rejected at the 0.05 level of significance. In statistics, levels of significance are kept small (e.g. $\alpha = 0.05$), whereas levels of confidence are generally large (e.g. $1 - \alpha = 0.95$). This abuse of convention is particularly bad in the atmospheric science literature (as also noted by von Storch and Zwiers 1999).

- **"the results are not significant at the 0.05 level but are significant at the 0.10 level"**
  The idea of hypothesis testing is that you FIX the level of significance BEFORE looking at the data. Choosing it after you have seen what the p-values are so as to reject the null hypothesis will lead to too many rejections. If the p-values are quite large (i.e. greater than 0.01) then it is good practice to quote the values and then let the reader make the decision.

- **"however, the results are significant in certain regions"**
  This is stating the obvious since the more data samples (variables) you look at, the more chance you will have of being able to reject the null hypothesis. With large gridded data sets, there is a real danger of "data mining" and fooling yourself (and others) into thinking you have found something statistically significant. The total number of data samples or variables examined needs to be taken into account when doing many tests at the same time. See Wilks (1995) for a discussion of these kind of multiplicity and dimensionality problems.

- **"the null hypothesis can not be rejected and so must be true"**
  Wrong. Either the null hypothesis is true OR your data sample was just not the right one to be able to reject the null hypothesis. If the null hypothesis can not be rejected all you can say is that the "data are not inconsistent with the null hypothesis" (remember this useful phrase !).

6.5 One sample tests in environmental science

This and the following section will briefly list some of the hypothesis tests most frequently used in the atmospheric and environmental sciences. Some basic one sample tests will be described in this section, and then some tests for comparing two samples of data (e.g. control and perturbed experiments) will be presented in the following section.

One sample tests are used to test whether a particular sample could have been drawn from a population with known parameters. The tests compare an observed sample statistic with a given population parameter.

6.5.1 Z-test on a mean with known variance

Does a sample come from a population with mean μ_0 and variance σ_0 ?

$$\begin{aligned} H_0 : \mu &= \mu_0 \\ \sigma &= \sigma_0 \\ H_1 : \mu &\neq \mu_0 \end{aligned} \tag{6.2}$$

Test using a Z-score test statistic with the sampling distribution

$$Z = \frac{\overline{X} - \mu_0}{\sigma_0/\sqrt{n}} \sim N(0,1)$$

6.5.2 T-test on a mean with unknown variance

Does a sample come from a population with mean μ_0 ?

$$\begin{aligned} H_0 : \mu &= \mu_0 \\ H_1 : \mu &\neq \mu_0 \end{aligned} \tag{6.3}$$

Since the population variance is no longer known we must estimate it using the sample variance s^2. This increases the uncertainty and modifies the sampling distribution of the test statistic slightly for small sample sizes $n < 30$. Instead of being distributed normally, the test statistic is distributed as **Student's t distribution** $T \sim t_\nu$ with $\nu = n - 1$ degrees of freedom. Student's t distribution has a density $f(t) \propto (1 + t^2)^{-n/2}$, which resembles the normal density except that it has slightly fatter tails (leptokurtic) and so provides more chance of having values far from the mean. This test is often referred to as a one-sample t-test on the mean. Test using a T-score test statistic with the sampling distribution

$$T = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

6.5.3 Z-test for non-zero correlation

Are two variables significantly correlated with a population correlation ρ different from zero ?

$$\begin{aligned} H_0 : \rho &= 0 \\ H_1 : \rho &\neq 0 \end{aligned} \tag{6.4}$$

Two-sided test using a t statistic with Student's t sampling distribution

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$

6.6 Two sample tests

Two sample tests are used to test whether two samples of data could have come from the same population. For example, we might be interested in comparing results from an experiment with those from a control experiment. Two sample tests compare the values of the sample statistic observed in the two different samples. When more than two samples need to be tested (e.g. 3 experiments), instead of performing two sample tests between all pairs of possible samples, it is better to use a more integrated multiple testing approach such as ANalysis Of VAriance (ANOVA). See Von Storch and Zwiers (1999) for more discussion.

6.6.1 T-test on unpaired means with unknown variance

Do two samples come from populations with the same mean ?

$$\begin{aligned} H_0 : \mu_1 - \mu_2 &= 0 \\ H_1 : \mu_1 - \mu_2 &\neq 0 \end{aligned} \tag{6.5}$$

Two-sided test using a T test statistic based on the difference in sample means that has a Student's t distribution with $n_1 + n_2 - 2$ degrees of freedom

$$T = \frac{\overline{X_1} - \overline{X_2}}{s_p/\sqrt{n}} \sim t_{n_1+n_2-2}$$

where $\frac{1}{n} = \frac{1}{n_1} + \frac{1}{n_2}$ and s_p^2 is the **pooled** estimate of variance

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1+n_2-2)} \tag{6.6}$$

6.6.2 T-test on paired means with unknown variance

Do two *paired* samples come from populations with the same mean μ_0 ?

Sometimes two samples are either generated or gathered as **pairs** of values $\{(X_1, X_2)\}$ rather than as two separate samples $\{X_1\}$ and $\{X_2\}$, e.g. heights of twins. In this case, the two-sample test on means described above is inappropriate and a **paired** test has to be used. The paired test is based on testing the mean difference of all pairs $D = X_1 - X_2$ for zero mean.

$$\begin{aligned} H_0 : \mu_D &= 0 \\ H_1 : \mu_D &\neq 0 \end{aligned} \tag{6.7}$$

Test using a T-score test statistic with the sampling distribution

$$T = \frac{\overline{D} - \mu_0}{s_D/\sqrt{n}} \sim t_{n-1}$$

6.6.3 F-test for equal variances

Do two samples come from populations with the same variance ?

$$\begin{aligned} H_0 : \sigma_1 &= \sigma_2 \\ H_1 : \sigma_1 &\neq \sigma_2 \end{aligned} \tag{6.8}$$

Two-sided test using an F test statistic distributed as an F distribution with n_1 and n_2 degrees of freedom . The F distribution is named after the famous statistician Sir R. **F**isher who discovered it while inventing the widely used **ANalysis Of VAriance** techniques.

$$F = \frac{s_1^2}{s_2^2} \sim F(n_1 - 1, n_2 - 1)$$

6.6.4 Z-test for unpaired equal correlations

Do two samples come from populations with the same correlation ?

$$\begin{aligned} H_0 : \rho_1 &= \rho_2 \\ H_1 : \rho_1 &\neq \rho_2 \end{aligned} \tag{6.9}$$

The trick here is to transform correlations into variables that are approximately normally distributed by using Fisher's Z transformation

$$Z = \frac{1}{2}\log_e\left(\frac{1+r}{1-r}\right) \tag{6.10}$$

The variance of Z is independent of r and is given by $s_Z^2 = 1/(n-3)$. The hypotheses can now be tested easily by using a 2-sample Z-test on unpaired means of normally distributed variables Z_1 and Z_2.

$$Z = \frac{\overline{Z_1} - \overline{Z_2}}{s_p} \sim N(0,1)$$

where the **pooled** estimate of variance is given by

$$s_p^2 = s_1^2 + s_2^2 = \frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}$$

6.7 Further reading

Because of its obvious importance, inferential statistics is covered in many statistics books. Chapter 10 of Spiegel (1992) provides a clear introduction to statistical hypothesis testing with many examples. A more in-depth discussion is given in Chapter 5 of Wilks (1995). Chapter 3 of Emery and Thomson (1997) has a concise summary of hypothesis testing.

The whole of this chapter is based on *classical* statistical inference, which assumes that the population parameters are fixed constants with no uncertainty. Bayesian statisticians relax this assumption and treat both variables and population parameters as random variables. The extra uncertainty in the

population parameters can sometimes lead to different Bayesian inferences to those found using classical inference. A clear non-mathematical introduction to Bayesian statistics can be found in Berry (1996).

Many wise and amusing words on the difficult art of inference and its potential pitfalls can be found in the stories of Sherlock Holmes by Sir Arthur Conan Doyle.

Chapter 7

Basic Linear Regression

Aim: The aim of this chapter is to introduce the concept of a linear regression model and show how it can be used to model the response of a variable to changes in an explanatory variable.

7.1 A few words on modelling strategy ...

Models provide compact ways of summarising observed relationships and are essential for making predictions and inferences. Models can be considered to be maps (representations) of reality and like any map can not possibly describe everything in reality (nor do they have to !). This is nicely summarised in the famous quotation:

> "All models are wrong, but some are useful" - G.E.P. Box

Good modellers are aware of their models' strengths and weaknesses and use the models appropriately. The process of choosing the most appropriate models is very complex and involves the following stages:

1. **Identification**
   By analysing the data critically using descriptive techniques and thinking about underlying processes, a class of potentially suitable models can be

   identified for further investigation. It is wise in this stage to consider first the simplest possible models (e.g. linear with few parameters) that may be appropriate before progressing to more complex models (e.g. neural networks).

2. **Estimation**
   By fitting the model to the sample data, the model parameters and their confidence intervals are estimated by usually using either least-squares or maximum likelihood methods. Estimation is generally easier and more precise for **parsimonius models** that have the least number of parameters.

3. **Evaluation**
   The model fit is critically assessed by carefully analysing the **residuals** (errors) of the fit and other diagnostics. This is sometimes referred to as validation and/or verification by the atmospheric science community.

4. **Prediction**
   The model is used to make predictions in new situations (i.e. independent data to that used in making the fit). The model predictions are then verified to test whether the model has any real skill. Predictive skill is the ultimate test of any model. There is no guarantee that a model which provides a good fit, will also produce good predictions. For example, non-parsimonious models having many parameters that provide excellent fits to the original data often fail to give good predictions when applied to new data (over-fitting).

By iterating at any stage in this process, it is possible with much skill and patience to find the most appropriate models.

7.2 Linear regression

It is a reasonable hypothesis to expect that body height may be an important factor in determining the body weight of a Reading meteorologist. This depen-

dence is apparent in the **scatter plot** below showing the paired weight versus height data (x_i, y_i) for the sample of meteorologists at Reading. Scatter plots are useful ways of seeing if there is any relationship between multiple variables and should always be performed before quoting summary measures of linear **association** such as correlation.

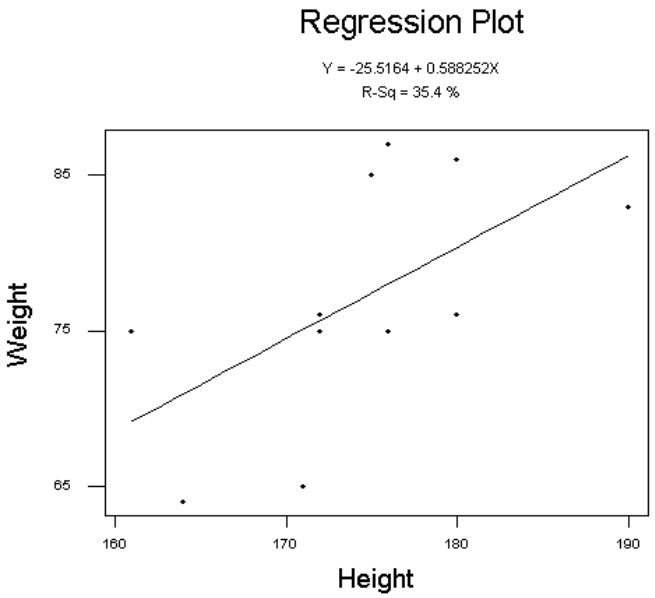


Figure 7.1: Scatter plot of body weight versus height for the sample of meteorologists at Reading. Best least squares fit regression line is superimposed.

The **response variable** (weight) is plotted along the y-axis while the **explanatory variable** (height) is plotted along the x-axis. Deciding which variables are responses and which variables are explanatory factors is not always easy in interacting systems such as the climate. However, it is an important first step in formulating the problem in a testable (model-based) manner. The explanatory variables are assumed to be error-free and so ideally should be control variables that are determined to high precision.

The cloud of points in a scatter plot can often (but not always!) be imagined to lie inside an ellipse oriented at a certain angle to the x-axis. Mathematically, the simplest description of the points is provided by the additive linear regression model

$$y_i \;=\; \beta_0 + \beta_1 x_i + \epsilon_i \tag{7.1}$$

where $\{y_i\}$ are the values of the response variable, $\{x_i\}$ are the values of the explanatory variable, and $\{\epsilon_i\}$ are the left-over noisy **residuals** caused by **random effects** not explainable by the explanatory variable. It is normally assumed that the residuals $\{\epsilon_i\}$ are uncorrelated Gaussian noise, or to be more precise, a sample of independent and identically distributed (i.i.d.) normal variates.

Equation 7.1 can be equivalently expressed as the following probability model:

$$Y \sim N(\beta_0 + \beta_1 X, \sigma_\epsilon^2) \tag{7.2}$$

In other words, for a given value of X, the Y values are normally distributed about a mean that is linearly related to X i.e. $\mu_{Y|X} = \beta_0 + \beta_1 X$. This makes the probability distribution clearly apparent and reveals ways of how to extend regression to more complicated situations.

The **model parameters** β_0 and β_1 are the y-intercept and the slope of the linear fit, and σ_ϵ is the standard deviation of the noise. These three parameters can be **estimated** using **least squares** by minimising the sum of squared residuals

$$SS = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \quad (7.3)$$

By solving the two simultaneous equations

$$\frac{\partial SS}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \quad (7.4)$$

$$\frac{\partial SS}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \quad (7.5)$$

it is possible to obtain the following least squares estimates of the two model parameters:

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} \quad (7.6)$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \quad (7.7)$$

$$(7.8)$$

Since the simultaneous equations involve only first and second moments of the variables, least squares linear regression is based solely on knowledge of means and (co)variances. It gives no information about higher moments of the distribution such as skewness or the presence of extremes.

7.3 ANalysis Of VAriance (ANOVA) table

When MINITAB is used to perform the linear regression of weight on height it gives the following results:

```
The regression equation is
Weight = - 25.5 + 0.588 Height
```



```
Predictor       Coef    StDev       T       P
Constant       -25.52   46.19   -0.55   0.594
Height         0.5883  0.2648    2.22   0.053

S = 6.606       R-Sq = 35.4%     R-Sq(adj) = 28.2%

Analysis of Variance

Source          DF      SS      MS      F       P
Regression       1      215.30  215.30  4.93    0.053
Residual Error   9      392.70   43.63
Total           10      608.00
```

The **regression equation** $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ is the equation of the straight line that "best" fits the data. The hat symbol ˆ is used to denote "predicted (or estimated) value". Note that regression is not symmetric: a regression of x on y does not generally give the same relationship to that obtained from regression of y on x.

The **Coef** column gives the best estimates of the model parameters associated with the explanatory variables and the **StDev** column gives an estimate of the standard errors in these estimates. The standard error on the slope is given by

$$s_{\beta_1} = \frac{1 - r^2}{\sqrt{n}} \frac{s_y}{s_x} \quad (7.9)$$

where r is the correlation between x and y and s_x and s_y are the standard deviations of x and y respectively.

The other two columns can be used to assess the statistical significance of the parameter estimates. The **T** column gives the ratio of the parameter estimate and its standard error whereas the **P** column gives a p-value (**p**robability **value**) for **rejection** of the null hypothesis that the parameter is zero (i.e. not a significant linear factor). For example, a p-value of 0.05 means that there is

5% chance of finding data less consistent with the null hypothesis (zero slope parameter) than the fitted data. Small p-values mean that it is unlikely that the slope was non-zero purely by chance.

The overall goodness of fit can be summarised by calculating the fraction of total variance explained by the fit

$$R^2 = \frac{var(\hat{y})}{var(y)} = \frac{var(\hat{\beta}_0 + \hat{\beta}_1 x)}{var(y)} \tag{7.10}$$

which is also known as the **coefficient of determination** and is the square of the sample correlation between the variables for this simple regression model. Unlike correlations that are often quoted by meteorologists, variances have the advantage of being additive and so provide a clear budget of how much of the total response can be explained. Note also that even quite high correlations (e.g. 0.5) mean that only a small fraction of the total variance can be explained (e.g. $(0.5)^2 = 0.25$).

The MINITAB output contains an ANalysis Of VAriance (ANOVA) table in which the sums of squares `SS` equal to n times the variance are presented for the regression fit $\hat{y}$, the residuals ϵ, and the total response y. ANOVA can be used to test the significance of the fit by applying F-tests on the ratio of variances. The p-value in the ANOVA table gives the statistical significance of the fit. When summarizing a linear regression, it is important to quote BOTH the coefficient of determination AND the p-value. With the small sample sizes often encountered in climate studies, fits can have substantial R^2 values yet can still fail to be significant (i.e. do not have a small p-value).

7.4 Model fit validation using residual diagnostics

In addition to the basic summary statistics above, much can be learned about the validity of the model fit by examining the left-over residuals. The linear model is based on certain assumptions about the noise term (i.e. independent and Gaussian) that should always be tested by examining the standardized

residuals. Resisduals should be tested for:

1. **Structure** The standardized residuals should be identically distributed with no obvious outliers. To check this, plot ϵ_i versus i and look for signs of structure. The residuals should appear to be randomly scattered (normally distributed) about zero.

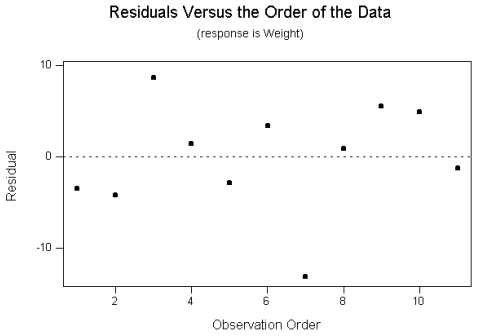


Figure 7.2: Residuals versus order of points for the regression of weight on height.

2. **Independence** The residuals should be independent of one another. For example, there should be no sign of runs of similar residuals in the plot of ϵ_i versus i. Autocorrelation functions should be calculated for regularly spaced residuals to test that the residuals are not serially correlated.

3. **Outliers** There should not be many standardised residuals with magnitudes greater than 3. Outlier points having large residuals should be examined in more detail to ascertain why the fit was so poor at such points.

4. **Normality** The residuals should be normally distributed. This can be examined by plotting a histogram of the residuals. It can be tested

by making a normal probability plot in which the normal scores of the residuals are plotted against the residual value. Straight line indicates normal distribution.

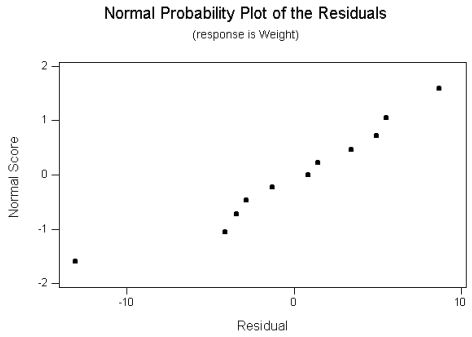


Figure 7.3: Normal probability plot of the residuals for the regression of weight on height.

5. **Linearity** The residuals should be independent of the fitted (predicted) values $\{\hat{y}_i\}$. This can be examined by making a scatter plot of ϵ_i versus $\{\hat{y}_i\}$. Lack of uniform scatter suggests that there may be a nonlinear dependence between y and x that could be better modelled by transforming the variables. For mutliple regression, with more than one explanatory variable, the residuals should be independent of ALL the explanatory variables.

In addition to these checks on residuals, it is also important to check whether the fit has been dominated by only a few **influential observations** far from the main crowd of points that can have **high leverage**. The leverage of a particular point can be assessed by testing the mean squared differences of all the predicted values to leaving out this point (known as **Cook's distances**).

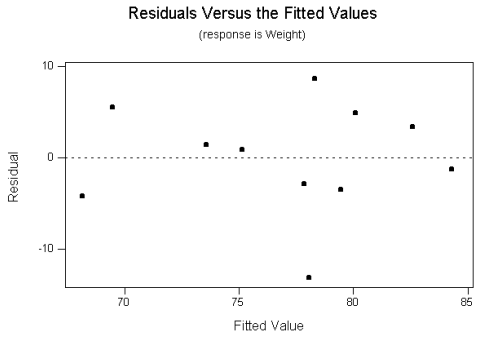


Figure 7.4: Residuals versus the fitted values for the regression of weight on height.

7.5 Weighted and robust regression

In the above **Ordinary Least Squares (OLS)** regression, it was assumed that the noise term was i.i.d. normal (identically and independently distributed normally). However, this assumption about the noise term is not always the most appropriate as can sometimes be noted in the residual diagnostics.

In cases where the variance of the noise is not identical at all points, it is better to perform a **General Least Squares** regression that gives less weight to y_i values that are more uncertain.

In cases where the noise is more extreme than that expected from a normal distribution (i.e. fatter tails), it is better to perform **robust regression**. This is appropriate if it is found that the standardized residuals have large magnitudes. Robust regression is also advisable when dealing with small samples as often occurs in climate studies. There are many different ways to do robust regression including Least Absolute Deviations (L_1 norm), M-Estimators, Least Median Squares, and Ranked Residuals. More details can be found in standard texts such as Draper and Smith (1998).

7.6 Further sources of information

The comprehensive book by Draper and Smith (1998) on applied regression analysis covers all these areas in much more depth and is well worth reading if you are involved with regression. StatSoft Inc.'s Electronic Statistics Textbook (www.statsoft.com/textbook/stathome.html) has a nice animated section on basic linear regression showing the danger of influential observations.

Chapter 8

Multiple and nonlinear regression

Aim: Linear regression can be extended to deal with cases having more than one explanatory variable (multiple regression), more than one response variable (multivariate regression), or non-linear dependencies. This chapter will briefly present all these possible extensions.

8.1 Multiple regression

It is often the case that a response variable may depend on more than one explanatory variable. For example, human weight could reasonably be expected to depend on both the height and the age of the person. Furthermore, possible explanatory variables often co-vary with one another (e.g. sea surface temperatures and sea-level pressures). Rather than subtract out the effects of the factors separately by performing successive iterative linear regressions for each individual factor, it is better in such cases to perform a single **multiple regression** defined by an extended linear model. For example, a mutliple regression model having two explanatory factors is given by

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i \tag{8.1}$$

This model can be fit to the data using least squares in order to estimate the three β parameters. It can be viewed geometrically as fitting a $q = 2$ dimensional hyperplane to a cloud of points in (x_1, x_2, y) space.

The multiple regression equation can be rewritten more concisely in matrix notation as

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{E} \tag{8.2}$$

where $\mathbf{Y}$ is a $(n \times 1)$ data matrix (vector) containing the response variable, $\mathbf{X}$ is a $(n \times q)$ data matrix containing the q factors, β is a $(q \times 1)$ data matrix containing the factor coefficients (model parameters), and $\mathbf{E}$ is a $(n \times 1)$ data matrix (vector) containing the noise terms.

The least squares solution is then given by the set of **normal equations**

$$(\mathbf{X}'\mathbf{X})\beta = \mathbf{X}'\mathbf{Y} \tag{8.3}$$

where $'$ denotes the transpose of the matrix. When $\mathbf{X}'\mathbf{X}$ is non-singular, these linear equations can easily be solved to find the β parameters. The β parameters can be used to determine unambiguously which variables are significant in determining the response.

As with many multivariate methods, a good understanding can be obtained by considering the **bivariate** case with two factors ($q = 2$). To make matters even simpler, consider the unit scaled case in which x_1 and x_2 have been standardized (mean removed and divided by standard deviation) before performing the regression. By solving the two normal equations, the best estimates for the beta parameters can easily be shown to be given by

$$\beta_1 = \frac{r_{1y} - r_{12}r_{2y}}{1 - r_{12}^2} \tag{8.4}$$

$$\beta_2 = \frac{r_{2y} - r_{12}r_{1y}}{1 - r_{12}^2} \tag{8.5}$$

where $r_{12} = cor(x_1, x_2)$ is the mutual correlation between the two x variables, $r_{1y} = cor(x_1, y)$ is the correlation between x_1 and y, and $r_{2y} = cor(x_2, y)$ is the correlation between x_2 and y. By rewriting the correlations in terms of the beta parameters

$$r_{1y} = \beta_1 + \beta_2 r_{12} \quad (8.6)$$

$$r_{2y} = \beta_2 + \beta_1 r_{12} \quad (8.7)$$

it can be seen that the correlations with the response consist of the sum of two parts: a direct effect (e.g. β_1), and an indirect effect (e.g. $\beta_2 r_{12}$) mediated by mutual correlation between the explanatory variables. Unlike descriptive correlation analysis, multiple regression is model-based and so allows one to determine the relative contribution from these two parts. Progress can then be made in discriminating important direct factors from factors that are only indirectly correlated with the response.

The MINITAB output below shows the results of multiple regression of height on weight and age for the sample of meteorologists at Reading:

```
The regression equation is
Weight = - 40.4 + 0.517 Age + 0.577 Height

Predictor       Coef    StDev       T       P
Constant      -40.36    49.20   -0.82   0.436
Age           0.5167   0.5552    0.93   0.379
Height        0.5769   0.2671    2.16   0.063

S = 6.655       R-Sq = 41.7%     R-Sq(adj) = 27.1%

Analysis of Variance

Source          DF      SS      MS      F       P
Regression       2      253.66  126.83  2.86    0.115
```



```
Residual Error   8      354.34   44.29
Total           10      608.00
```

It can be seen from the p-values and coefficient of determination that the inclusion of age does not improve the fit compared to the previous regression that used only height to explain weight. Based on this small sample, it appears that at the 10% level age is not a significant factor in determining body weight (p-value $0.379 > 0.10$), whereas height is a significant factor (p-value $0.063 < 0.10$).

8.2 Multivariate regression

Multiple regression can easily be extended to deal with situations where the response consists of $p > 1$ different variables. **Multivariate regression** is defined by the **General Linear Model**

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{E} \quad (8.8)$$

where $\mathbf{Y}$ is a $(n \times p)$ data matrix containing the response variables, $\mathbf{X}$ is a $(n \times q)$ data matrix containing the factors, β is a $(q \times p)$ data matrix containing the factor coefficients (model parameters), and $\mathbf{E}$ is a $(n \times p)$ data matrix containing the noise terms.

The least squares estimates for the beta parameters are obtained by solving the normal equations as in multiple regression. To avoid having large uncertainities in the estimates of the beta parameters, it is important to ensure that the matrix $\mathbf{X}'\mathbf{X}$ is well-conditioned. Poor conditioning (determinant of $\mathbf{X}'\mathbf{X}$ is small) can occur due to **collinearity** in explanatory variables, and so it is important to select only response variables that are not strongly correlated with one another. To choose the best model, it is vitally important to make a careful **selection of variables** when choosing the explanatory variables. Semi-automatic methods such as forward, backward, and stepwise selection have been developed to help in this complex process of model identification.

8.3 Non-linear response

While a linear response is justifiable in many situations, there are also occasions when the response is not expected to be linear. For example, a least squares regression of probability incorrectly implies that predicted probabilities can lie outside the acceptable range of 0 to 1. To deal with such situations, there are two basic approaches. Either you nonlinearly transform the response variable (see normalising transformations, chapter 2) and then do a linear regression using the transformed response, or you non-linearly transform the fitted values, which are a linear combination of the explanatory variables. For example, the widely applied **logistic regression** uses the **logit** transformation $y' = \log(y/(1-y))$ ("log odds"). The logarithm transformation is often used when dealing with quantities that are strictly positive such as prices, while the square root transformation is useful for transforming positive and zero count data (e.g. number of storms) prior to linear regression. In a ground-breaking paper, Nelder and Wedderburn (1972) introduced a formal and now widely used procedure for choosing the link function $g(y)$ known as **Generalized LInear Modelling GLIM** (note "Generalized" not "General" !).

8.4 Parametric and non-parametric regression

The response can also sometimes depend on a nonlinear function of the explanatory variables e.g. $y = f(x) + \epsilon$. For example, under realistic carbon emission scenarios, predicted future global warming is not expected to be a simple linear function of time and so a linear fit would be inappropriate.

In some cases, the expected form of the non-linear function is known and can be parameterised in terms of basis functions. For example, polynomial regression consists of performing multiple regression with variables $\{x, x^2, x^3, \ldots\}$ in order to find the polynomial coefficients (parameters). Note, however, that strong correlations between $\{x, x^2, x^3, \ldots\}$ can lead to collinearity and poor fits. A better approach is to use a basis set of orthogonal uncorrelated predictor functions such as Fourier modes. These types of regression are known

as **parametric regression** since they are based on models that require the estimation of a finite number of parameters.

In other cases, the functional form is not known and so can not be parameterised in terms of any basis functions. The smooth function can be estimate in such cases using what is known as **non-parametric regression**. Two of the most commonly used approaches to non-parametric regression are **smoothing splines** [1] and **kernel regression**. Smoothing splines minimise the sum of squared residuals plus a term which penalizes the roughness of the fit, whereas kernel regression involves making smooth composites by applying a weighted filter to the data. Both methods are useful for estimating the slow trends in climatic time series and avoid spurious features often obtained in simpler smoothing approaches.

8.5 Further sources of information

The comprehensive book by Draper and Smith (1998) on applied regression analysis covers linear and parametric regression in detail and provides many other useful references. Non-parametric regression and Generalized Linear Models are well-described in the book by Green and Silverman (1986). The pioneering article by Nelder and Wedderburn (1972) presents the original motivation for this important development in modern statistics.

[1] not to be confused with the common interpolating cubic splines as described in Numerical Recipes and elsewhere !

Chapter 9

Introduction to time series

Aim: The aim of this chapter is to provide a brief introduction to the analysis of time series with emphasis on the time domain approach.

9.1 Introduction

The variation in time of environmental quantities can be studied using the rich branch of statistics known as **time series analysis**. A **discrete** (as opposed to **continuous**) **time series**[1] is a sequence of observed values $\{x_1, x_2, \ldots, x_n\}$ measured at discrete times $\{t_1, t_2, \ldots, t_n\}$. Climatological time series are most often sampled at **regular** intervals $t_k = k\tau$ where τ is the **sampling period**.

The main aims of time series analysis are to **explore** and extract **signals** (patterns) contained in time series, to make **forecasts** (i.e. future predictions in time), and to use this knowledge to optimally **control** processes.

The two main approaches used in time series analysis are **time domain** and **spectral (frequency) domain**. The time domain approach represents time series directly as functions of time, whereas the spectral domain approach represents time series as spectral expansions of either Fourier modes or wavelets.

[1] NOTE: time series NOT timeseries !

9.2 Time series components

A lot can be learnt about a time series by plotting x_k versus t_k in a **time series plot**. For example, the time series plot in Figure 9.1 shows the evolution of monthly mean sea-level pressures measured at Darwin in northern Australia.

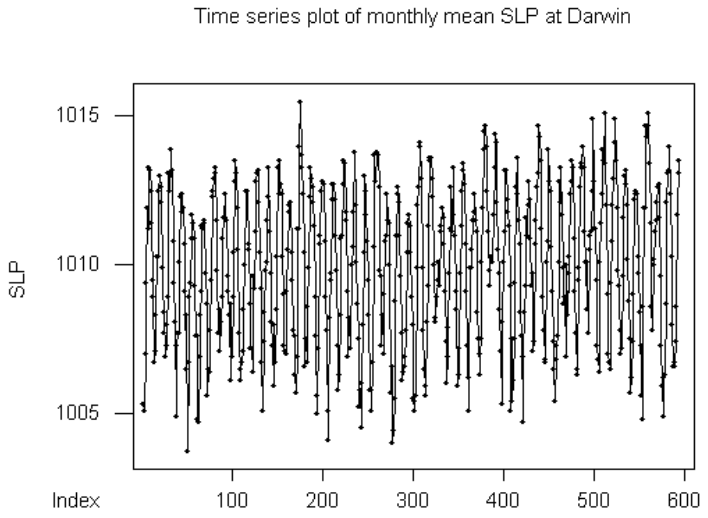


Figure 9.1: Time series of the montly mean sea-level pressure observed at Darwin in northern Australia over the period January 1950 to July 2000.

A rich variety of structures can be seen in the series that include:

- **Trends** - long-term changes in the mean level. In other words, a smooth regular component consisting primarily of Fourier modes having periods

longer than the length of the time series. Trends can be either **deterministic** (e.g. world population) or **stochastic**. Stochastic trends are not necessarily monotonic and can go up and down (e.g. North Atlantic Oscillation). Extreme care should be exercised in extrapolating trends and it is wise to always refer to them in the past tense.

- **(Quasi-)periodic signals** - having clearly marked cycles such as the **seasonal component** (annual cycle) and interannual phenomena such as El Niño and business cycles. For periodicities approaching the length of the time series, it becomes extremely difficult to discriminate these from stochastic trends.

- **Irregular component** - random or chaotic noisy residuals left over after removing all trends and (quasi-)periodic components. They are (second-order) **stationary** if they have mean level and variance that remain constant in time and can often be modelled as filtered noise using time series models such as ARIMA.

Some time series are best represented as sums of these components (**additive**) while others are best represented as products of these components (**multiplicative**). Multiplicative series can quite often be made additive by normalizing using the logarithm transformation (e.g. commodity prices).

9.3 Filtering and smoothing

It is often useful to either **low-pass filter** (**smooth**) time series in order to reveal low-frequency features and trends, or to **high-pass filter** (**detrend**) time series in order to isolate high frequency transients (e.g. storms).

Some of the most commonly used filters are:

- **Moving average MA(q)**
  This simple class of low-pass filters is obtained by applying a running mean of length q to the original series

$$y_t = \frac{1}{q}\sum_{k=-q/2}^{q/2} x_{t+k} \tag{9.1}$$

For example, the three month running mean filter MA(3) is useful for crudely filtering out intraseasonal oscillations. Note, however, that the sharp edges in the weights of this filter can causing spurious ringing (oscillation) and leakage into the smoothed output.

- **Binomial filters** $(1/2, 1/2)^m$

These smoother low-pass filters are obtained by repeatedly applying the MA(2) filter that has weights $(1/2, 1/2)$. For example, with $m = 4$ applications the binomial filter weights are given by $(1/2, 1/2)^4 = (1, 4, 6, 4, 1)/16$ which tail off smoothly towards zero near the edges. After many applications, the weights become Gaussian and the filtering approximates Gaussian kernel smoothing.

- **Holt exponential smoother**

This simple and widely used recursive filter is obtained by iterating

$$y_t = \alpha x_t + (1-\alpha) y_{t-1} \tag{9.2}$$

where $\alpha < 1$ is a tunable smoothing parameter. This low-pass filter gives most weight to most recent historical values and so provides the basis for a sensible forecasting procedure when applied to trend, seasonal, and irregular components (Holt-Winters forecasting).

- **Detrending (high-pass) filters**

High-pass filtering can most easily be performed by subtracting a suitably low-pass filtered series from the original series. The detrended residuals

$x_t - y_t$ contain the high-pass component of x. For example, the backward difference detrending filter $\Delta x = x_t - x_{t-1}$ is simply twice the residual obtained by removing a MA(2) low-pass filtered trend from a time series. It is very efficient at removing stochastic trends and is often used to detrend non-stationary time series (e.g. random walks in commodity prices).

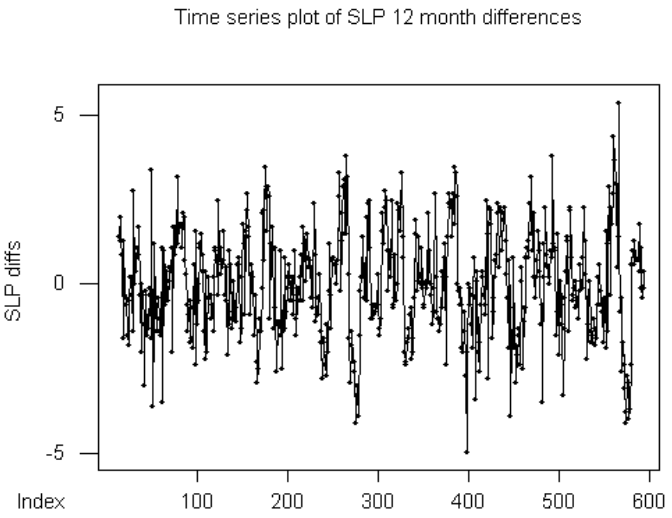


Figure 9.2: Time series plot of one-year backward differences in monthly mean sea-level pressure at Darwin from the period January 1951 to July 2000. The differencing has efficiently removed both the seasonal component and the long-term trend thereby revealing short-term interannual variations.

9.4 Serial correlation

Successive values in time series are often correlated with one another. This **persistence** is known as **serial correlation** and leads to increased spectral power at lower frequencies (**redness**). It needs to be taken into account when testing significance, for example, of the correlation between two time series. Among other things, serial correlation (and trends) can severely reduce the effective number of degrees of freedom in a time series. Serial correlation can be explored by estimating the sample **autocorrelation coefficients**

$$r_k = \frac{\frac{1}{n}\sum_{i=k+1}^{n}(x_i - \overline{x})(x_{i-k} - \overline{x})}{\frac{1}{n}\sum_{i=k+1}^{n}(x_i - \overline{x})^2} \tag{9.3}$$

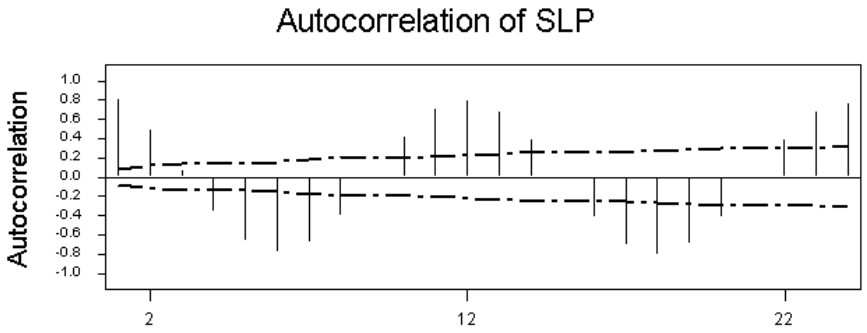
where $k = 0, 1, 2, \ldots$ is the **time lag**. The zero lag coefficient r_0 is always equal to one by definition, and higher lag coefficients generally damp towards small values with increasing lag. Only autocorrelation coefficients with lags less than $n/4$ are sufficiently well-sampled to be worth investigation.

The autocorrelation coefficients can be plotted versus lag in a plot known as a **correlogram**. The correlogram for the Darwin series is shown in Fig. 9.3. Note the fast drop off in the **autocorrelation function (a.c.f.)** for time lags greater than 12 months. The lag-1 coefficient is often (but not always) adequate for giving a rough indication of the amount of serial correlation in a series. A rough estimate of the **decorrelation time** is given by $\tau_0 = -\tau/\log(r_1)$ and the effective number of degrees of freedom is given by $n\tau/\tau_0 = -n\log(r_1)$. See von Storch and Zwiers (1999) for more details.

9.5 ARIMA(p,d,q) time series models

Auto-Regressive Integrated Moving Average (ARIMA) time series models form a general class of linear models that are widely used in modelling and forecasting time series (Box and Jenkins, 1976). The ARIMA(p,d,q) model of the time series $\{x_1, x_2, \ldots\}$ is defined as

$$\Phi_p(B)\Delta^d x_t = \Theta_q(B)\epsilon_t \tag{9.4}$$



Lag	Corr	T	LBQ	Lag	Corr	T	LBQ	Lag	Corr	T	LBQ	Lag	Corr	T	LBQ
1	0.80	19.61	386.62	8	-0.38	-3.83	1555.20	15	-0.01	-0.11	2719.90	22	0.39	2.58	3971.56
2	0.49	7.86	529.32	9	0.01	0.13	1555.31	16	-0.40	-3.11	2819.61	23	0.67	4.38	4253.18
3	0.07	0.98	532.01	10	0.41	4.06	1658.96	17	-0.69	-5.22	3110.16	24	0.77	4.83	4618.84
4	-0.34	-5.04	603.38	11	0.70	6.69	1956.57	18	-0.79	-5.72	3491.75				
5	-0.64	-9.04	852.56	12	0.80	7.11	2343.91	19	-0.68	-4.72	3780.65				
6	-0.76	-9.45	1200.63	13	0.68	5.64	2629.47	20	-0.39	-2.62	3876.55				
7	-0.66	-7.25	1467.48	14	0.38	3.01	2719.78	21	-0.00	-0.00	3876.55				

Figure 9.3: Correlogram showing the autocorrelations as a function of lag for the Darwin series

where B is the backward shift operator, $Bx_y = x_{y-1}$, $\Delta = 1 - B$ is the backward difference, and Φ_p and Θ_q are polynomials of order p and q, respectively. ARIMA(p,d,q) models are the product of an autoregressive part AR(p) $\Phi_p = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p$, an integrating part $I(d) = \Delta^{-d}$, and a moving average MA(q) part $\Theta_q = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q$. The parameters in Φ and Θ are chosen so that the zeros of both polynomials lie outside the unit circle in order to avoid generating unbounded processes. The difference operator takes care of "unit root" $(1 - B)$ behaviour in the time series and for $d > 0.5$ produces non-stationary behaviour (e.g. increasing variance for longer

time series).

An example of an ARIMA model is provided by the ARIMA(1,0,0) first order autoregressive model $x_y = \phi_1 x_{y-1} + a_y$. This simple AR(1) model has often been used as a simple "red noise" model for natural climate variability.

9.6 Further sources of information

A vast number of books and articles have been written on time series analysis. One of the clearest introductory guides is a little red book by Chatfield (1984) which is well worth reading. A much larger book by Brockwell and Davis (1991) goes into more details and is clearly illustrated. It also covers methods for multivariate time series analysis and forecasting. Bob Nau's "Statistical Forecasting" course at Duke University is a good online guide to how to go about forecasting time series (www.duke.edu/ rnau/411home.htm). Some humourous quotations about forecasting time series can be found at www.met.rdg.ac.uk/cag/STATS/quotes.html.