University of Oxford

Introduction to Statistics for Medical Students

Kanishka Bhattacharya

Wellcome Trust Centre for Human Genetics,
Roosevelt Drive,
Oxford OX3 7BN.

These notes are written by Dr YY Teo. I am using these solely for the purpose of teaching and with his due permission.

Foreword

In compiling this set of notes, I have drawn heavily from materials and resources by Brian Ripley, Jonathan Marchini, James McBrown and many other references which I am unable to attribute in full. I am however exceptionally grateful to Stephen Goss for his suggestions and patient editing.

It should be highlighted that this set of notes serves to provide an easy source of quick reference, and is by no means comprehensive for all the topics discussed within. Students are reminded that there is a constant need to refer to specific references for detailed discussions and expositions of the statistical and mathematical information.

The advancement of statistics has necessarily resulted in some of the methods discussed here to be outdated, and more efficient methods is and/or will be available. However, I have chosen to focus on the fundamentals for each of the areas discussed, forming the foundations necessarily for further reading and understanding of any novel or more complex methods. It should however be reminded also that often, the simplest form of statistics will suffice in most well-designed studies, and a revision of the study design and data collected should first occur before the use of more apparently sophisticated statistical analysis.

Hopefully this set of notes will be sufficiently clear and concise for an understanding of basic statistics.

YY Teo
First Edition *December 2004* Second Edition *November 2005*

Contents

Chapter 1

Overview of Statistics

In this chapter, we provide a brief overview of statistics in terms of the phases of statistical analysis and discuss briefly the role of computers in modern statistical analysis. We shall discuss sample selections and the different types of data and distributions which physiologists and biologists often encounter.

1.1 A Brief Introduction to Statistics

Statistics plays a very important role in all areas of science. Statistical models and methods allow us to take samples of data from our area of focus and answers questions about the population of objects or events that we are interested in studying.

Before turning to a discussion of the reasons for collecting and analysing data for a representative sample, we should reiterate the important division between a population and a sample. The term *population* means the entire collection of units or measurements of those units about whom information is available, or more formally, the set of values for one or more variables taken by all units. The term *sample* denotes the subset of the population selected for study or the values of the variable(s) recorded for those selected units. An example may be that we can consider the whole faculty of physiology students to be the population, while a randomly selected group of 10 physi-

ology students will represent a sample of the population.

Usually we are interested in learning about certain attributes or properties of a population, such as its parameters, structure or distribution. However, in most cases, we cannot observe a population's attributes because doing so would require analysing the whole population, which is generally not possible. Instead, a sample from the population is selected, and information are obtained for the samples which may or may not be generalised to the entire population. For example, to know the mean height of the population of physiology students, we can either collect the height of all the students, or we could select a sample and calculate the mean height of the sample. We could then calculate certain statistics of the sample to provide some information about how accurate or precise the mean height of the sample is for estimating the mean height of the population.

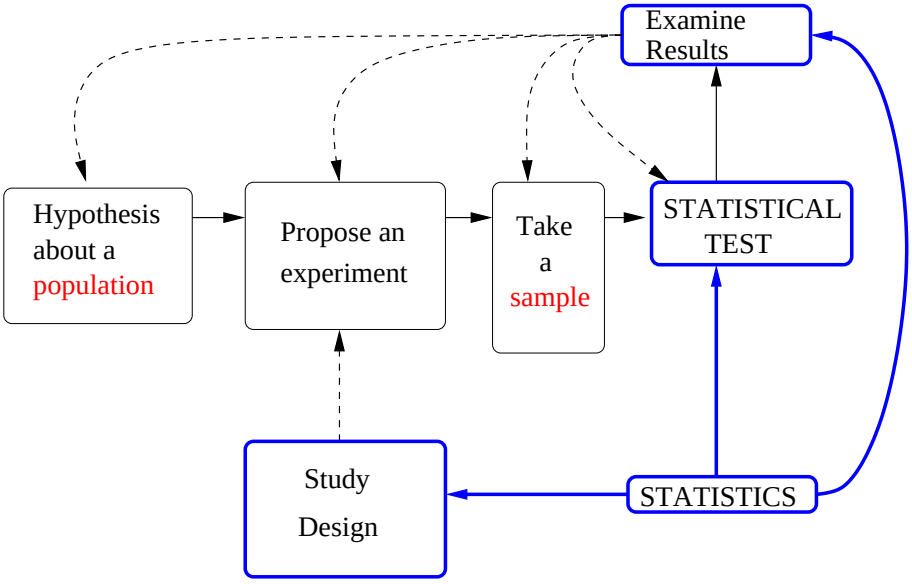


Figure 1.1: The scientific process and role of statistics.

1.2 Phases of Statistical Analysis

Initial Data Manipulation
This involves putting the data in the right format to carry out checks of data quality before commencing any initial analysis of the data. This step is an important one and should include reviewing the method(s) of data collection in order to look for possible sources of bias that might invalidate conclusions drawn from the data, checking for discrepant observations (usually due to measurement errors or errors in entering the data), and searching for missing observations.

Exploratory Data Analysis (EDA)
Simple analysis of the data should always be done in order to clarify their general form, to check for discrepant or extreme observations, to suggest possible directions for more complicated analysis, and to investigate assumptions required by the subsequent definitive analysis. Usually, this step involves producing tabular, graphical and numerical summaries of the data.

Definitive Analysis
This phase entails using formal and sometimes informal techniques of statistical inference in order to draw conclusions about certain attributes of interest for the underlying population.

Presentation of Conclusions
This step involves presenting the graphical and numerical results from above in an accurate and concise form.

1.3 Role of Computes in Statistical Analysis

Computers can be used in each of the four phases of statistical analysis, as well as for data collection and entry. More specifically, computers can be used for *data collection*, *data entry*, *data checking*, *data screening*, *definitive analysis* and *presentation of results*.

Using a computer to perform statistical analysis results in numerous advantages, from increased accuracy and speed, to the versatility and ability in handling large amount of data. Informative graphics can also be produced easily, and data can be manipulated easily in terms of mathematical operations and transformations.

There are, unfortunately, disadvantages associated with the use of computers in statistical analysis. One of the most common problem is that the versatility offered by statistical softwares makes it easy to use an inappropriate statistical procedure, often when the researcher performing the analyses do not understand the statistics and rationale behind the analysis. It often leads to *data dredging* as well, which refers to the search for significant relationships by performing a large number of analyses, often without a properly formulated hypothesis.

As a result, caution must be exercised when using computers to perform statistical analysis. More importantly, before using a statistical software, one should first understand what needs to be done and whether the software is performing the relevant analysis. In addition, researchers should guard against the temptation to produce large amounts of computer output not required by the planned analysis.

1.4 Sample Selection

In this course, we shall be focusing on data which consists of the values of $p(\geq 1)$ variables of interest for each of $n(\geq 1)$ units in a representative sample of the population of interest. This type of data is especially common in applications such as demographics, medicine, psychology, sociology and zoology, and it is relatively straightforward to draw inferences about its underlying population. For probabilistic reasons, we will assume throughout that the underlying population of interest is infinite, unless stated otherwise.

Data can be collected with either of two types of goals in mind:

Descriptive Inference: when the main objective is to describe a large group, using information from a sample from that group.

Analytical Inference: when the main objective is to study the properties of and relationships between variables using a small sample, assuming that the results from that sample can be generalised to a larger population.

Simple Random Sample

A simple random sample of size n is a subset containing n units from the population of interest. These units are chosen in such a way that every possible subset of size n has the same probability of being selected as any other; equivalently, every unit in the relevant population has the same probability of being selected for the sample.

Stratified Sampling

Sometimes the population to be sampled is divided into non-overlapping subpopulations called strata (eg. genotypic data, genetic structures, gender), across which it is suspected that the answers to the questions of interest may differ. Sampling can be done by collecting data across the strata in the population as this improves the precision of our estimates of the parameters of interest. In this stratified sampling method, the total desired sample size is divided between the strata in a manner that reflects the properties of the variables of interest within each stratum. Once the number to be sampled from each stratum has been determined, the appropriate number of units is selected randomly from each stratum.

Cluster Sampling

In cluster sampling, the population is again divided into non-overlapping groups or clusters. However in this case, the groups are not assumed to differ systematically, but rather each is assumed to be representative of the entire population. In such situations, only several of these clusters are first selected and then units are randomly sampled within each cluster to reduce the amount of data collected. An example of cluster sampling will be if a researcher hoped to estimate the mean body mass index (BMI) of Oxford

students, the researcher may assume that all colleges were more or less the same in this regard, he could select several colleges and then measure the body mass index for randomly selected students within these colleges only.

To contrast between stratified sampling and cluster sampling, we expect the parameters of interest to vary maximally *across* groups and minimally *within* groups in stratified sampling while we expect the parameters to vary minimally *across* groups and maximally *within* groups in cluster sampling.

Multi-stage Sampling
In some instances there may be several layers of clusters or strata. For example, the population of interest could be divided into health districts, each of which could then be subdivided into its component hospitals. If this was the case, then multi-stage sampling, in which groups, then subgroups, and finally units within subgroups are selected, could then be employed. An example may be to assume that hospitals in Oxfordshire admit patients with medical conditions representative of the entire UK population (cluster), and the focus is on the difference in incidence of diabetes for male and female patients (strata).

Multi-phase Sampling
Sampling seldom occurs in only one phase, an initial random sample from the overall population might be used to estimate certain properties of the variables of interest in each of the strata in that population. The resulting estimates could be employed in order to determine how the total number of units that will be sampled in the second phase should be divided amongst the strata for more in-depth sampling in that phase.

We should note that all of these sampling methods employ random sampling at some stages, usually at the final stages of sampling. The importance of choosing units randomly is twofold: it helps to avoid biases, and it is an explicit assumption of many statistical methods.

1.5 Types of Variables

There are different types of variables from measurements of variables of interest, which are determined by the set of values that they can take, and whether they are categorical in nature or numerical. When data for only one variable is analysed, the analysis is termed *univariate*, whereas an analysis that investigates the relationship or association between two or more variables is termed *multivariate*.

Nominal - Categorical
Nominal categories refer to data which can take on only a finite set of values, where these categories or levels have no intrinsic ordering. These are also commonly referred to as *factors*. Examples would include gender (male, female), dichotomous disease status (disease, unaffected), etc.

Ordinal - Categorical
Ordinal categories refer to data that can take on only a finite set of values, where the categories in this set do have an intrinsic ordering, but often not on a well-defined scale. An example would be the assessment of the quality of life: poor, decent, good, excellent.

Interval
This refers to a variable that can take on only a finite set of values, where the categories in this set have an intrinsic ordering, but also have numerical scores or labels (eg. quality of life as ranked in increasing scale of 1 - 5). These labels are often treated as category averages, means, or medians, and the differences between them can be used as a measure of the separation between two categories. This type of variable can also result from coarsely observing a numerical variable, for example when the possible range of values for a numerical variable is divided into a number of bins and only the bin location is observed for each unit.

Discrete
Discrete variables take on integer or counting number values which may be

the number of occurrences of some phenomenon. For example, the number of children, or the number of lectures attended in one term.

Non-ratio Continuous - Numerical
This refers to numerical variables that take values along a continuous scale. Variables of this type are fairly common, especially in physiological measures. However, do note that variables of this type may take values from 0, and can be allowed to be non-linear. Non-linearity means that the difference between 5 and 10 may mean differently from the difference between 80 and 85.

Ratio
This refers to a variable that takes on values along a continuous and well-defined scale. For variables of this type, a difference of one unit has the same interpretation at any part of the scale, and a value of 0 truly denotes the absence of the characteristic.

It should be noted that numerical variables on a continuous scale can always be reduced to interval categorical variables by grouping their values into bins, and this is often seen in medical research where numeric data is divided into tertiles (3 groups of approximately equal sizes), quartiles (4 groups) or quintiles (5 groups) for comparisons, often between the top group with the bottom group. For example, in genetics research, a possible way to evaluate the effects of genotypes on quantitative trait loci (QTL) may be to divide the data into quintiles, where comparison is made between the 1st and 5th quintile, signifying the subjects with the worst and best responses respectively.

1.6 Distributions

The measured values of variables will always vary across members of a population, or across members of a sample from that population. The pattern of occurrence of the various values of a variable is termed its *distribution*. Primarily, a distribution describes the possible values that a variable can take, and the relative frequency with which each of these different values occur.

The distribution of values for all units in the population is termed the *population distribution*, whereas the distribution of values for the units in a selected sample from the population is termed the *empirical distribution*. The population distribution is usually not known or not observed unlike the empirical distribution. In most situations, we assume that the empirical distribution is a good representation of the underlying population distribution.

Chapter 2

Exploratory Data Analysis

Exploratory data analysis (EDA) can help to reduce the information contained in a data set to a few key indicators that describe or summarise its main characteristics and therefore provide a better overall picture of the data. Although some particular features of the data may be lost by summarising the data, trends or patterns in the data may be revealed which may be relevant to the questions of interest. Certain EDA techniques will also highlight departures from these trends/patterns in the data set, providing an efficient graphical method to identify outliers. Outliers can be defined as data points that deviate remarkably from the majority of the sample. Although outliers may result from measurement or recording errors, they can also correspond to anomalous units. Finally EDA provides a graphical method to investigate the assumptions that are required for statistical inference.

For example, some methods of statistical inference require the assumption that the population underlying the data for one variable has a normal distribution. A histogram of the observations for the variable can easily provide graphical evidence of whether this assumption is reasonable.

EDA methods can be tabular, numerical (descriptive statistics) or graphical. The specific method of appropriate EDA depends on the type of data that are being investigated (univariate or multivariate, categorical or numeric).

2.1 Tabular EDA Methods

For a nominal or ordinal categorical variable, a frequency table or *one-way table*, is a possible way of describing the data. For each category, the table can show either its *absolute frequency* (the number of occurrences of the category), or its *relative frequency* (the number of occurrences of the category divided by the total number of occurrences of all categories). Contingency tables which are two-way or multi-way frequency tables are useful in describing the relationship or association between two or more ordinal/nominal variables.

	Marital Status		
Education	Married Once	Married more than once	**Total**
College	550	61	611
No College	681	144	825
Total	1,231	205	1,436

Consider the above example where an absolute frequency table relating number of marriages and education for a sample of 1,436 married women listed in Who's Who in 1949. Note that of the women who went to college, 10% had been married more than once as compared to 17% for those without college education. For women married more than once, 30% had a college education as compared to 45% for women who married only once. The table shows an association between having a college education and the increase in chances of being married only once.

2.2 Numerical EDA Methods

Sample Quantiles

Sample quantiles can be used to describe either categorical or continuous variables. The α-th sample quantile, denoted $\eta(\alpha)$, is the smallest value such that $(100 \times \alpha)\%$ of the observations for the variable take values which are less than or equal to $\eta(\alpha)$. For example, 5% of the observed values for a given

variable are smaller than its 5^{th} sample quantile.

An important set of quantiles is the *sample quartiles*, a set of values that divides the observed range of a variable into four intervals, each containing 25% of the observations. The sample quartiles are denoted by Q_1, Q_2 and Q_3, and are referred to as the lower quartile, median, and the upper quartile respectively. The quartiles are often used to calculate the *sample inter-quartile range* (IQR), which is defined as the difference between the upper and the lower quartile $(Q_3 - Q_1)$.

The sample quartiles are usually combined with the minimum and maximum value observed in the data, to produce a five-number summary of the dataset. For example, consider a dataset for the height of 120 randomly chosen male students from Balliol College, a five number summary for this dataset will be

Min	$\boldsymbol{Q_1}$	**Median**	$\boldsymbol{Q_3}$	**Max**
0	172	181	188	201

From the five number summary, it seems that there is a problem with the dataset, as the minimum height registered is zero. This was actually a respondent who failed to provide his height and the data was mistakenly entered as zero. Without performing this simple check, analysis of the height using the raw dataset will produce erroneous interpretations. Assuming that the rest of the four numbers in the table are correct, the IQR for this set of data is thus $16 = 188 - 172$.

Location

This property is concerned with finding the position of the value in the dataset that best characterises it. The sample *median*, *mean*, and *mode* can all be used. It should however be emphasized that although the mode has meaning for all the six types of variables, calculation of the median is only sensible for variables other than nominal, and calculation of the mean is only possible for variable types other than nominal and ordinal. In particular, these measures of location are not particular informative when the

empirical distribution for a given variable is not unimodal (i.e. the frequency distribution has more than one peak).

The sample median can be calculated by ranking the n observed values of the variable from smallest to largest where the middle value in this ordered list will be the median. If n is even, the median will be the average of the two middle values. The sample median indicates the centre of the empirical distribution of a given variable in the sense that half of the values are smaller than or equal to the sample median (and half of them are larger than or equal to it).

Another measure of location is the sample mode, which is simply the value of the variable that appears with the highest frequency in the dataset. The sample mode is however not necessarily unique because two different values may occur with the same highest frequency. For this reason, the sample mode is not always a good indicator of location.

The sample mean is the most widely used location measure, and is commonly denoted as $\bar{X}$ where

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Like the sample median, the sample mean also indicates the centre of the distribution, but in the sense of a centre of gravity.

Comparing between the sample median and the sample mean, if the empirical distribution of the variable is symmetric with respect to the mean, then the median and the mean have the same value. For instance, let us consider a dataset containing the measurements of length of the forearm (in inches) for 140 adult males. For this data, the median and the mean are 18.8 and 18.802, reflecting the symmetric nature of the distribution. However, the sample mean and median are not coincident when the empirical distribution of the variable is asymmetrical. This is especially true when outliers are present for a variable, since the sample mean is greatly affected by outliers whereas the sample median is not. Consider the height example of 120 stu-

dents, the median is 181 while the mean is 178. This discrepancy occurs due to the erroneous entry of zero for the student whose height was not reported. However this error does not affect the sample median because the median selects only the middle observed value and is thus not affected by outlying values at either extremes. For this reason, we say that the median exhibits *robustness against outliers*, which is often a desirable statistical property, especially in the case where outliers represent measurement or recording errors.

Spread or Dispersion

The spread or dispersion of a variable measures the degree to which the observed values for that variable are concentrated around a location measure. The smaller spread indicates that the observed values are more tightly clustered around the centre of the empirical distribution. Measures of spread include the *sample range*, the IQR, the *sample variance*, the *sample standard deviation*, and the *sample coefficient of variation*. As before for sample mean, it is not possible to calculate these quantities for nominal or ordinal variables.

The simplest measure of spread is the sample range, which is defined as the difference between the sample maximum and the minimum for a variable. However, since the sample range depends on only two observations at both extremes, it is highly sensitive to outliers and may not be a reliable indicator of spread. A similar measure which is less sensitive to outliers is the IQR, although if the dataset contains a large proportion of outliers, the IQR may be similarly unreliable.

Another measure of spread is the sample variance, often denoted by s^2 or σ^2. The sample variance is defined mathematically as

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Note that the variance is always non-negative, and is zero only if all of the observed values for a variable are identical (i.e. there is no variation). The sample variance is expressed in squared units, which can make it a difficult quantity to interpret intuitively. A common measure of spread is therefore

the sample standard deviation, often denoted as s or σ, which is defined as the (positive) square root of the sample variance.

It should be emphasized that the sample range, variance and standard deviation depend on the units in which a variable is measured. As a result, the empirical values of these measures can be misleading when comparisons are drawn across variables using different units of measurements. For example, consider the height example with 119 students (omitting the student with zero height), the data measures height in centimetres and the standard deviation is 9.94cm. If the data is instead entered in metres, the standard deviation will instead be 0.0994m. Comparisons of the two values will not be meaningful without careful observation of the units of measurements.

The sample coefficient of variation (CV) is another measure of spread which overcomes the above problem, as it uses the absolute size of the mean to adjust for the standard deviation. The sample coefficient of variation for a sample is defined with respect to the standard deviation and the mean as

$$CV = \frac{s}{\bar{X}}$$

This coefficient has no units, which allows us to use it for comparing dispersions of variables measured in different units. Note however if the sample mean for a variable happens to be very near 0, the value of the CV may be artificially inflated and therefore suggest a greater degree of dispersion than is actually present.

Consider the example below of a comparison between the sample standard deviation and coefficient of variation of blood pressure (systolic and diastolic) for a sample of subjects with hypertension, and suppose we observe the following values:

	Systolic BP	**Diastolic BP**
Mean	140	90
Standard Deviation	10	10

Although the standard deviation is the same for both systolic and diastolic blood pressure, the dispersion of blood pressure around the mean is clearly

more spread out for systolic BP than for diastolic BP. This fact is revealed by the sample coefficients of variations, which are 1/14 and 1/9 for systolic and diastolic BP respectively, reflecting a greater variability for the latter. This would not be apparent if comparisons were made using only the standard deviations.

Skewness

Skewness refers to deviations from symmetry with respect to a location measure. The quantity, often referred to as b_1, is commonly used as a measure of asymmetry and is represented mathematically by

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^3}{s^3}.$$

The resulting quantity is unit-free. If the distribution of a variable is symmetric around its sample mean, then b_1 has a value of 0. Positive values of b_1 indicate that the variable is right-skewed (i.e. there is a longer or fatter tail for values larger than the mean), while a negative value of b_1 provides evidence of a longer or fatter tail for values smaller than the mean (i.e. the variable is left-skewed).

Kurtosis

Kurtosis denotes the degree of peakedness of the distribution, often as compared to a Normal (Gaussian) distribution. The coefficient of kurtosis, usually referred as b_2, is represented as

$$b_2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^4}{s^4},$$

and is always non-negative and unit-free. This coefficient takes the value of 3 for the normal distribution, which is described as *mesokurtic*. A value of this coefficient that is smaller than 3 indicates a distribution that is *platykurtic* (i.e. a distribution that is less peaked than the normal distribution). For the flattest (least-peaked) of all distributions, the kurtosis for a uniform distribution takes a value of 1.8. Conversely, if b_2 has a value larger than 3, then this indicates a distribution that is *leptokurtic* (i.e. a distribution that is more peaked).

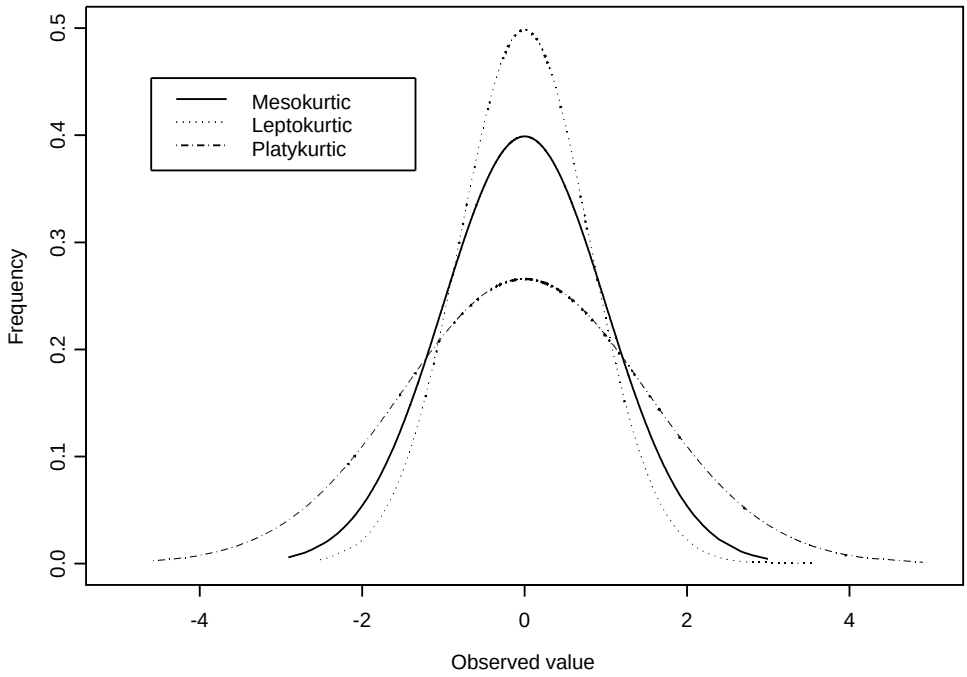


Figure 2.1: Graphical depiction of the different kinds of kurtosis. The solid line is generated using a Normal Distribution.

Covariance and Correlation

The degree of association between two numerical variables can be assessed using correlation coefficients, for which the Pearson's correlation coefficient is the most common.

Consider two numerical variables X and Y, the sample covariance for these two variables, which is defined as

$$\text{cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) \ ,$$

where X_i and Y_i are the observations of variables X and Y for unit i, and $\bar{X}$ and $\bar{Y}$ are the sample means of the variables. The covariance is zero for variables that are independent, and for dependent random variables, the co-

variance will be positive, if larger than average values of X occur typically when Y is above average. On the other hand, the covariance will be negative, if larger than average values of X occur typically when Y is below average. In the first case we say that X and Y are positively associated, in the second case we call them negatively associated.

Consider for instance the height and the weight of a randomly chosen person. Taller than average people will often also have higher than average weight. Therefore we can expect that the random variables "height" and "weight" are positively associated. On the other hand, the variables "age" and "running distance in 10 seconds" will usually be negatively associated for adults.

It would be also of interest to find out about the strength of association. Unfortunately, a large covariance cannot be interpreted as indicating a strong relationship. Just as the sample variance and standard deviation are affected by the units in which a variable is measured, the sample covariance will also reflect the absolute size of the units of measurements for the two variables.

Pearson's correlation coefficient, r, removes this dependence on the unit of measurement by scaling the sample covariance by the product of the sample standard deviations of X and Y.

$$r(X,Y) = \frac{\text{cov}(X,Y)}{s_x s_y}$$

The correlation coefficient takes on values between -1 and 1, where if r is approximately 0, then there is no evidence of linear correlation. On the other hand, a value of 1 indicates a perfect positive linear association while a value of -1 indicates a perfect negative linear association. It is essential to remember that Pearson's correlation coefficient assesses only the *linear* association of two variables, and is not a measure of non-linear relationships.

Pearson's correlation is not robust to outliers, given the dependence on standard deviations. A more robust measure of association will be to rank the values of each variable from smallest to largest (assigning scores from 1 to n in ascending ranks) for both variables, and calculate the Pearson's r using

the ranks from the two variables. If a variable has two identical values the usual procedure for assigning ranks is to assign the average of the two ranks to both values. The resulting correlation coefficient is known as Spearman's rank correlation coefficient and is robust to outliers.

2.3 Graphical EDA Methods

Graphical methods make it very easy to discover trends and patterns in a data set, and some of these methods in particular are extremely useful in identifying outliers, or departures from the trend.

Frequency Plots and True Histograms
It has been noted previously that it is possible to calculate absolute or relative category frequencies for a variable of the categorical variety, and the possible continuum of values for the variable when divided into intervals for variables of the numerical type. Instead of presenting these frequencies in tabular form, we could plot a frequency chart for the data. These absolute and relative frequency plots are often referred to as *histograms*.

Properly defined, a histogram is a bar graph in which each bar corresponds to a category created by grouping the values of the variable into intervals, classes, or bins (unless the variable is categorical to begin with), and where the height of each bar is proportional to the absolute (or relative) frequency of the corresponding class. It is important to distinguish between a histogram or frequency plot, and a true histogram. The latter is similar to a frequency plot, except that in a true histogram, the *area* of a bar, rather than the *height* of a bar, is proportional to the frequency of the interval class. More specifically, in a true histogram, the area of each bar is equal to the relative frequency of the interval class to which it corresponds. A frequency plot is identical to a true histogram only when relative frequencies and bins of equal length are used for the frequency plot.
For both frequency plots and true histograms, the number of bins used can greatly affect the appearance of both types of plots, and that the absolute

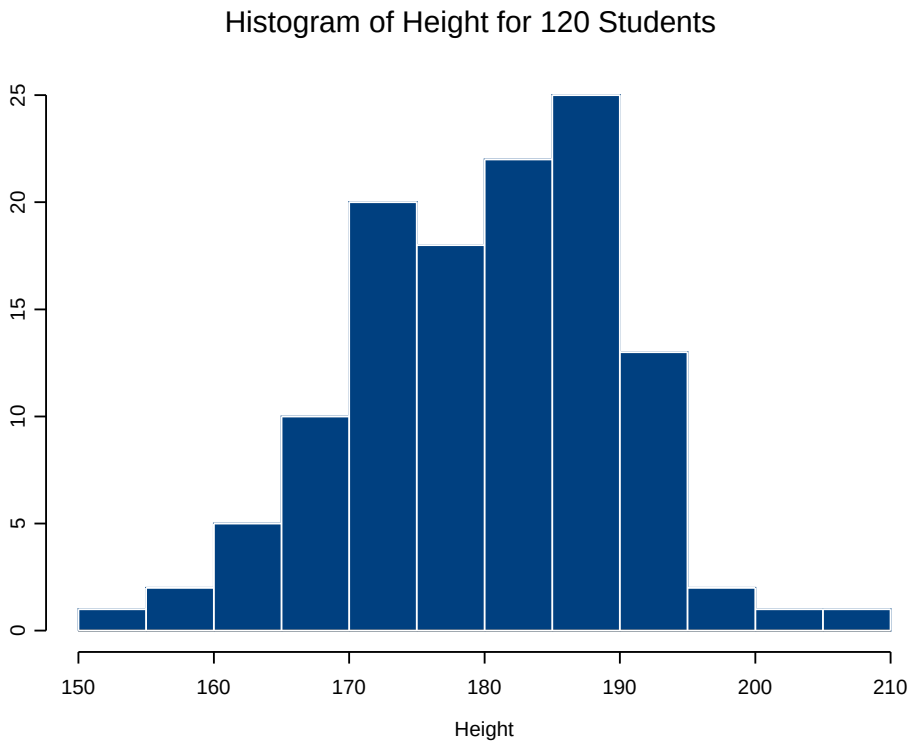


Figure 2.2: Histogram of the height distribution for 120 students. The histogram in this case is constructed such that the true histogram is identical to the frequency plot.

frequencies of the Y-axis changes for different number of bins. The more classes and bins there are, the more details of the data, but the sparser the counts. If the number of bins is too small the clumping effect will result in the loss of particular features of the data while using too many intervals or bins will result in many narrow and unnecessary details which may obscure the overall picture. Thus there is a trade-off with regards to the choice for the number of bins or identically, the width of the bins. Do note that different statistical packages have different rules in choosing the number of pins relative to the number of observations.

Histograms can also be influenced by where the breakpoints between interval classes are located, and by whether a value that occurs at a category break-

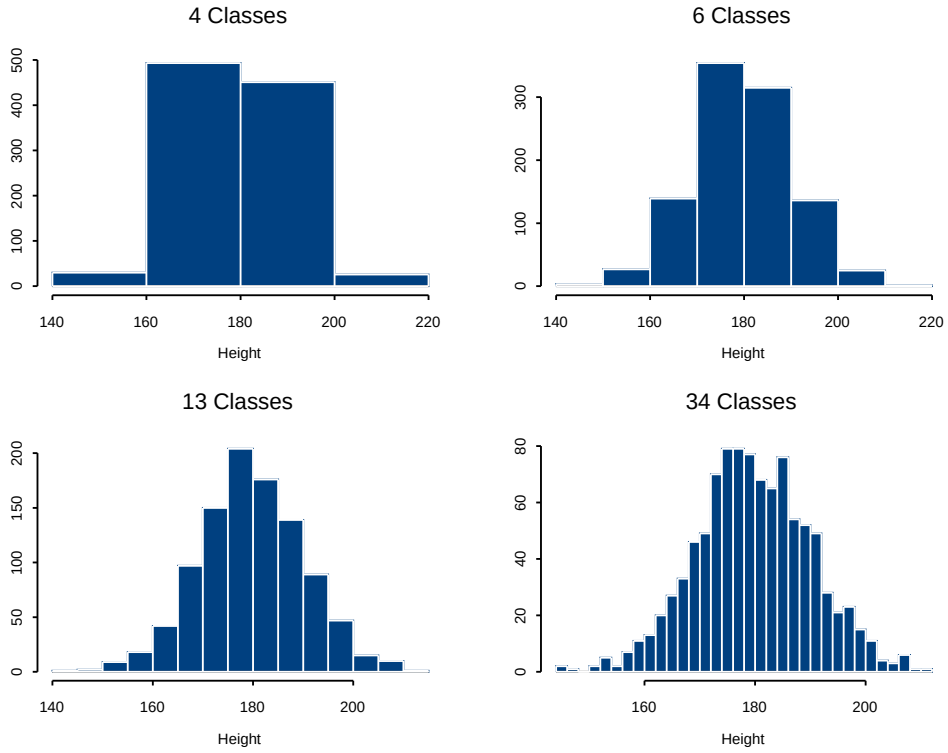


Figure 2.3: Histograms of the height distribution for 1000 students using 4 different number of bins.

point is considered to belong to the bin on the left or the right.

It is important to note that frequency plots and histograms are particularly useful for getting an idea of the distribution of a variable, and in particular, where the centre is located, the spread of the data, and whether it is symmetric, right or left skewed, and how fat and long the tails are. Histograms can also indicate whether the data is unimodal or multimodal.

Univariate Boxplots Boxplots are extremely useful graphical devices for describing interval and numerical variables, which is sometimes also known as a *box-and-whiskers* plot. This plot is based on the five number summary and is particularly useful for identifying outliers and extreme outliers, and for comparing the distributions of variables within two or more classes. The

ends of the box are the lower and upper sample quartiles, and thus the length of the box is the IQR for the variable. The sample median for the variable is marked by a line inside the box. The lines extending from the box (the 'whiskers') extend up to the smallest and largest observation within the interval $(Q_1 - 1.5\text{IQR}, Q_3 + 1.5\text{IQR})$. Points that fall within the interval $(Q_1 - 3\text{IQR}, Q_1 - 1.5\text{IQR})$ are designated as negative outliers, and points that fall in the interval $(Q_3 + 1.5\text{IQR}, Q_3 + 3\text{IQR})$ are designated as positive outliers. Those points located outside the interval $(Q_1 - 3\text{IQR}, Q_3 + 3\text{IQR})$ are considered to be extreme outliers.

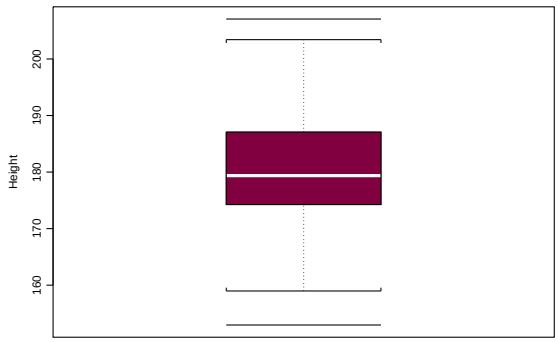


Figure 2.4: Boxplot for the height data of 120 students.

There are two outliers at either end for this data set. In addition, we can tell that the distribution is roughly symmetric since the sample median line is roughly in the middle of the box and also that the two whiskers are similar in length.

Multivariate Boxplots

The boxplot can also be used for bivariate analysis in the specific case where one desires to investigate the association between a categorical variable and a non-ordinal and non-nominal variable (i.e. an interval or numerical variable). Boxplots make it easy to compare the distributions of the variables within each of two or more classes (levels) of the first variable.

Consider the height example with additional height data for 120 female students from Balliol College.

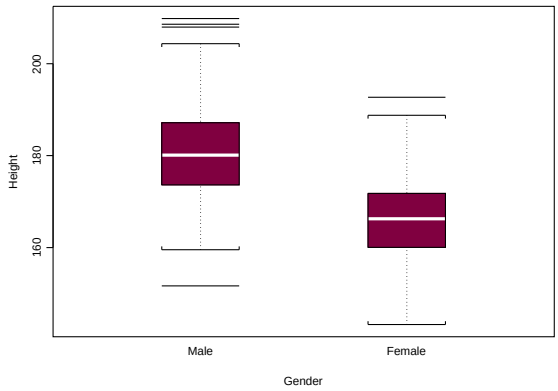


Figure 2.5: Boxplot for the height data of 240 students, categorised by their gender.

From the boxplots above, we see graphical evidence suggesting a difference in the distributions of height for male and female students. The spread seems similar across the two gender as suggested by the similar widths of the boxes on either side of the medians.

Scatterplots

The relationship between two numerical variables can be viewed graphically using a scatterplot. Furthermore, the relationships between two numerical variables and one categorical variable can be efficiently displayed through the use of a scatterplot in which different symbols indicate the various levels of the categorical variable. An important point to note is that the axes should always be clearly labeled, and if necessary, a legend provided to state the corresponding symbols for each categorical variable.

As an example, consider the height dataset where 120 male and 120 female students are sampled from Balliol College, where data on their height and weight are obtained.
From this figure, we see that in general, there exists linear relationships

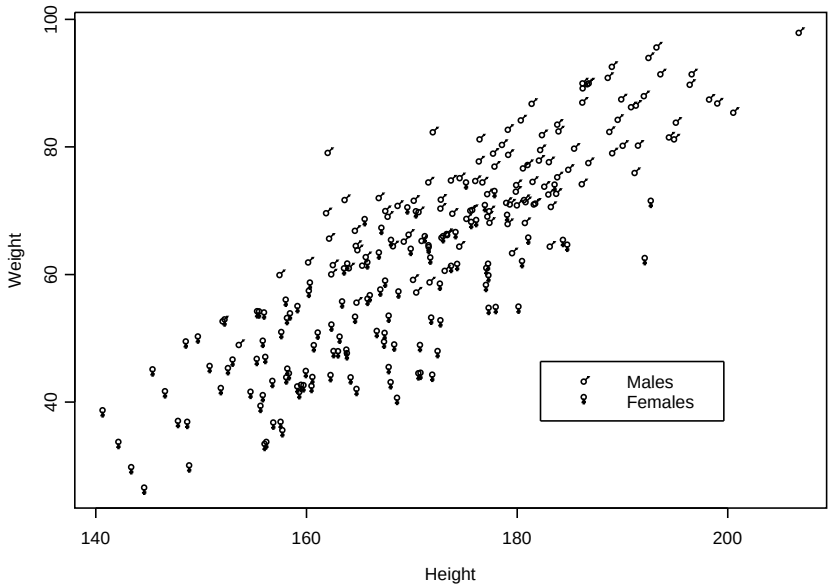


Figure 2.6: Scatterplot for the height and weight data for 120 male and 120 female students.

between height and weight for both males and females, although it seems graphically that the relationship seems stronger for males than females. It can be safely concluded that in general, increasing height corresponds to an increase in weight.

Transformations of Variables

Transforming a variable refers to applying the same mathematical function, such as $\ln(x)$, $\exp(x)$, or x^2, to all the observed values of a variable. Clearly, since the levels of categorical variables do not have a strict numerical interpretation, it is not possible or appropriate to apply mathematical transformations to these variables. For continuous variables, $\ln(x)$ is a particularly common transformation that is often applied to variables that only can or do take on positive values, such as height and weight.

There are a number of reasons to justify the need for transformations of

continuous variables. In some cases, there is theoretical motivation for such a transformation. For example, if we are examining population growth in a developing country over time, we might want to take the natural logarithm of the population size since we expect exponential population growth in an theoretical environment of no population control.

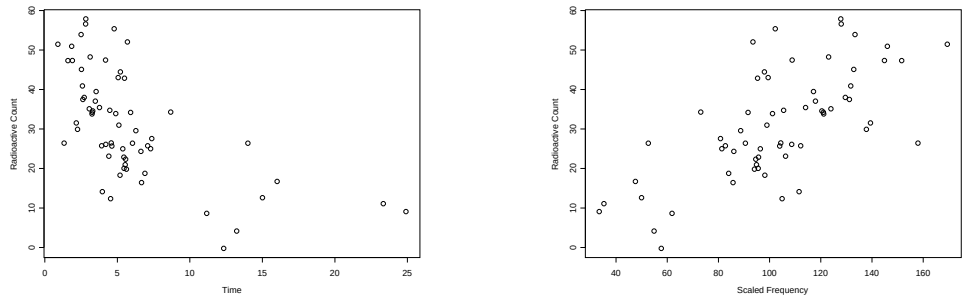


Figure 2.7: Identification of transformations from scatterplots.

Another possible reason for transforming a continuous variable may be that the transformations allow the data to satisfy the assumptions required by statistical inference methods. For instance, many methods of statistical inference assume that the variable of interest has an underlying distribution that is normal. A normal variable can theoretically take on any value in the interval $(-\infty, \infty)$, which is obviously not the case for some variables (such as population size), that can only take on positive values. Using a logarithmic transformation will take a number in the interval $(0, \infty)$ to another in the interval $(-\infty, \infty)$, which may result in a transformed variable for which the assumption of normality is more reasonable.

Chapter 3

Overview of Methods of Statistical Inference

Statistical inference can be divided into three areas: estimation, confidence intervals and hypothesis tests. These areas however should not be viewed as independent and isolated applications. In general, statistical inference takes the sequence of estimation of the attributes of interests to be used to form confidence intervals, and to perform hypothesis tests for these attributes. There is also a duality between confidence intervals and hypothesis tests that will be discussed.

Estimation techniques entail using the data to make a 'best guess' at the attribute(s) about which we are hoping to draw inference. This sample-based guess is selected because it is a good representative of the unknown population attribute(s) of interest. In general, 'guesses' or 'estimates' of underlying population attribute(s) are accompanied by "errors", which are used to give an indication of the precision and reliability of the estimates.

A confidence area for only one property of interest is referred to as a confidence interval, whereas a confidence area for two or more properties of interest is called a confidence region. The formation of a confidence interval (region) involves constructing a one-dimensional interval or a multi-dimensional region that, according to the data, is likely to contain the true unknown val-

ues of the attribute(s) of interest.

Performing a hypothesis test entails using the data to decide how likely it is that a certain hypothesis about the underlying property of interest (i.e. the null hypothesis) is true. More specifically, if the data that we observed in our sample would be extremely unlikely to occur if the null hypothesis were true, then we reject that hypothesis.

Generally, the statistical methods of inference introduced above can be divided into two classes: parametric and non-parametric. In parametric inference, a specific distributional family is assumed for the underlying variable(s) of interest, and the known statistical / probabilistic properties of that family are then used to design estimators, confidence intervals, and / or hypothesis tests. In non-parametric inference, these three entities are designed by employing various rules of probability, but no specific distributional family is assumed for the underlying population.

Methods of statistical inference rely on certain assumptions and if the assumptions required by a method are not valid for the population underlying the particular dataset, then it is possible that the conclusions reached may be invalid. It is important that the data analyst be aware of these assumptions, and to verify whether they are valid for the dataset. Assumptions can be investigated informally through graphical methods, or through formal statistical tests. For example, most standard statistical inference assumes that the dataset follows approximately a Normal distribution, and this can be assessed informally by plotting the histogram, or more formally through a goodness-of-fit test for a Normal distribution.

3.1 Estimation

Estimation techniques involve using the data to provide a suitable guess at the population attribute(s) we wish to deduce, and for the purpose of this course, we shall focus on deducing parameters (point estimation), rather than structural properties of the variables. Note that a parameter is a numerical

characteristic of the population of interest, or a numerical function of the random variable(s) of interest. Examples of parameters which we may be interested in are population median or mean. An estimator is therefore a representation of the parameter from the data.

As the estimators are constructed through the use of the data, each estimator will be associated with a degree of precision and reliability, that is often dependent on the size of the data sample. This degree of precision and reliability is often known as the *error* of the estimator, and in general we wish to minimise this error for greater precision in estimation.

We will assume henceforth that the population random variable has a distribution that is symmetric, where the location parameter of interest represents the 'centre' of the underlying distribution. Also, we assume that the underlying population is virtually infinite and our data sample is always randomly selected, and of size n.

Definition: A **statistic** is simply any mathematical function of the data in a sample, or any mathematical function of the realisations of the random variables in a sample. A statistic is a random variable as it is constructed from random variables. Examples of statistics include the sample mean, sample median, or sample skew.

Definition: An **estimator** is a statistic that is specifically designed to measure a particular population parameter. Since estimators are a special case of statistics, they are also random variables and have associated probability distributions.

Definition: An **estimate** is a realisation or value of an estimator that occurs once the sample data is evaluated in the estimator expression.

Note that $\bar{X}$ denotes the random variable for the sample mean, obtained from the random variables $\{X_1, X_2, \ldots, X_n\}$, while $\bar{x}$ denotes one realisation of the sample mean, obtained from the data sample $\{x_1, x_2, \ldots, x_n\}$. Alter-

natively, $\bar{X}$ can be seen to represent all possible values that the realisation $\bar{x}$ can take.

Common estimates in statistics are the mean and variance of the population, with the estimator for the population mean being represented by

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \ ,$$

and the estimator for the population variance being commonly represented as

$$S^2 = \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

3.1.1 Sample Distribution of an Estimator

The sampling distribution of a statistic is a probability distribution that describes the probabilities of the possible values for a specific statistic. The exact form of the sampling distribution for a given estimator will depend on the underlying population distribution from which the data were drawn. In general, knowing the sampling distribution for an estimator is useful and necessary for constructing confidence intervals and hypothesis testings.

Sampling Distribution of the Mean

To consider the sampling distribution of the sample mean for the variable X, we assume that X has a $N(\mu, \sigma^2)$ distribution [i.e. a Normal distribution with mean μ and variance σ^2].

For a sample of size n $(x_1, x_2, \ldots, x_n)$ from a population with mean μ and variance σ^2, the mean of this sample is

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Suppose a new sample of size n is obtained from the population and the sample mean calculated, and if this sampling process is repeated until all possible sample outcomes are accounted for, the sampling set of means for $\bar{x}$

yields a probability distribution and this distribution is termed the *sampling distribution of the sample mean*, commonly denoted as $\bar{X}$. It can be shown easily that $\bar{X}$ takes the form of $N(\mu, \sigma^2/n)$, and this means that the sampling distribution is distributed symmetrically around the true population mean, with a spread proportional to σ^2/n. As the spread is inversely dependent on the sample size, the precision of the estimation increases for larger sample sizes, resulting in the realisations to be more tightly clustered around the true mean value μ. The standard deviation of $\bar{X}$ is $\sigma/\sqrt{n}$, which is referred to as the standard error of $\bar{X}$. The variance of $\bar{X}$ is commonly defined as the error variance.
The above applies when X is known to have a $N(\mu, \sigma^2)$ distribution. When the underlying population distribution is unknown but symmetric, the *Central Limit Theorem* states that the sampling distribution of $\bar{X}$ is approximately the same as the above, provided that the sample size of X values is sufficiently large.

3.1.2 Central Limit Theorem

Let X be a random variable with mean μ and standard deviation σ (but not necessarily normally distributed). If $\bar{X}$ is the mean of a random sample of size n drawn from the distribution of X, then the distribution of $\bar{X}$ tends to a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$, provided n is sufficiently large.

In general, the sample size required depends on the degree of asymmetry of the original distribution of X, requiring larger n for greater degree of asymmetry.

3.1.3 Standard Error and Bias of Estimators

The bias of an estimator is defined as the difference between the expected value of an estimator and the corresponding population parameter it is designed to estimate. Further, an estimator is said to be unbiased for a param-

eter, θ, if its expected value is precisely θ.

Unbiasedness is generally a desirable property for an estimator, and often an estimator for a given population parameter is constructed such that it will be unbiased. (This explains why the sample variance has a denominator of $n-1$ rather than n).

Note that the bias of an estimator can be calculated analytically or through simulation techniques. For example, the bias of the sample mean can be calculated analytically since mathematical calculation and statistical theory can be used to derive an explicit expression for the expected value of the sample mean, although simulations may be required to obtain the bias of the sample median.

In general, the most common measure of the spread or dispersion of a distribution, relative to its mean, is the standard deviation. However for the sampling distribution of an estimator, the standard deviation is commonly referred to as the *standard error* of the estimator. The standard error is used as a measure of the precision of the estimator, and depends on the scale or units of measurement of the variable of interest. It is common to compare the value of an estimator with that of its standard error, using the ratio $s.e.(\hat{\theta})/\hat{\theta}$, which is unit free, and is termed the *coefficient of variation*. In general, the smaller the coefficient of variation, the more precise the estimate is.

Compromise Between Standard Error and Bias

Suppose there are two different estimators for the same population parameter. In general, one of the two possible estimators will have a smaller standard error than the other, and we define the estimator with the smaller standard error as being more *efficient*. In addition, we will define an estimator efficient if it achieves the smallest standard error possible for the estimation of a given parameter. The smallest possible standard error for an estimator of a certain parameter can be found using mathematical results such as the Cramer-Rao lower bound. Greater efficiency is a generally attractive property for an es-

timator to have since it means that the estimator has a small standard error and is a more precise estimator for the underlying parameter.

However, unbiasedness is also a desirable property for an estimator to possess. An unbiased estimator with the smallest standard error is the ideal estimator but quite frequently no such estimator exists and there may be trade offs between the bias and variance of the estimators. Often, the bias of an estimator for some population attribute of interest can only be decreased at the expense of increasing its standard error. There is no rule to allow us to

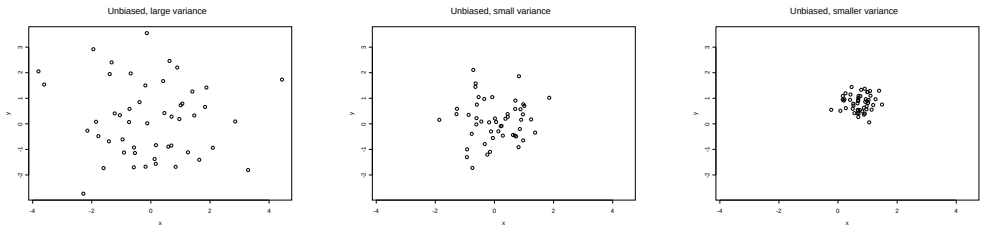


Figure 3.1: Three estimators with different properties.

choose between several estimators for the same population parameter since in different situations, a different choice of estimator might be preferable. A balance between bias and variance is the general rule for deciding between bias and variance in estimators.

As a side note, in addition to an estimator which is unbiased and efficient, we may also want our estimators to be robust to outliers in the data sample. For instance, the sample median may be preferable rather than the sample mean as an estimator for the population mean if the data contain extreme values, since the median is robust to outliers whereas the mean is not. Note however that robustness will often come at the expense of higher variances.

Also, it is true in general that the precision of an estimator increases (decreasing standard errors) as the sample size increases, or estimators being more precise when sample size is larger. In fact, having a larger sample is generally advantageous since large samples allow for detection of very small differences in means. However there is a need to balance the cost and time

required to collect a larger sample. There is usually a compromise between clinical efficiency and statistical efficiency.

3.2 Confidence Intervals

We can define a confidence interval (CI) as a region, constructed under the assumptions of a model, that contains the true value (the parameter of interest) with a specified probability. This region is constructed using particular properties of an estimator and is a statement regarding both the accuracy and precision of this estimate. There are two quantities associated with confidence intervals that we need to define:

Definition: The **coverage probability** refers to the probability that a procedure for constructing random regions will produce an interval containing, or covering, the true value. It is a property of the interval producing procedure, and is independent of the particular sample to which such a procedure is applied. We can think of this quantity as the probability that the interval constructed by such a procedure will contain the parameter of interest.

Definition: The interval produced for any particular sample, using a procedure with coverage probability p, is said to have a **confidence level** of p.

Note that the confidence level and coverage probability are equivalent *before* we have obtained our sample. After the sample has been obtained, the parameter is either in or not in the interval. Thus we would expect 95% of 95% CIs that are constructed to cover the true parameter under repeated sampling. It is a common (and natural) mistake, when given a 95% confidence interval, to interpret it as "the probability that the parameter lies between x and y is 0.95". The correct interpretation requires an inversion of the thinking process: instead of focusing on the probability of the parameter being in the interval, we need to focus on the probability of the interval containing the parameter. The difference is subtle but important, as parameters are regarded as fixed, unknown *constants*, and not random quantities.

Sample Size and Width of Confidence Intervals

For any given confidence level, an increase in sample size will yield narrower, or more precise, confidence intervals. One of the reasons is that the length of a CI depends on the standard error of an appropriate estimator, and the standard error has an inverse relationship with the sample size, decreasing as n gets larger. Intuitively, a larger sample will contain more information about the population parameter of interest, and therefore result in more precise estimations. Conversely, for any given sample size, an increase in confidence level will yield wider intervals. Intuitively this is so because a wider interval will result in greater confidence that the interval contains the true value. Importantly, there is a trade-off between precision (interval length) and accuracy (coverage).

3.2.1 One-Sided Confidence Intervals

Two-sided confidence intervals are used when we are interested in inferring two points between which the population quantity lies, and this is usually the form of CI constructed. If, however, there are very strong prior knowledge / beliefs regarding the process under investigation, we might consider constructing a one-sided confidence interval. These CIs are appropriate when we are interested in either an upper or lower bound for μ, but not both. Consider the following example:

(Pagano and Gauvreau, 1993) Consider the distribution of haemoglobin levels for the population of children under the age of 6 years who have been exposed to high levels of lead. Suppose this population is normally distributed with mean μ (and standard deviation σ), on which we wish to make inference. Under usual circumstances, we simply want to locate μ, with the focus of the inference being "between which two points does μ lie". In this circumstance, a two-sided CI is appropriate.

Suppose however that we have some extra information regarding the process under investigation. Specifically, suppose it is known that children with lead poisoning tend to have lower levels of haemoglobin than children who do not.

We might therefore be interested in an upper bound on μ. Hence we would construct a one-sided confidence interval for μ of the form $(-\infty, \mu_U)$, where μ_U denotes the upper bound. Note that we would truncate this interval below at 0, giving $(0, \mu_U)$, as haemoglobin levels cannot be negative.

3.2.2 Constructing Confidence Intervals

At the introductory level, we consider the confidence intervals for means of normally distributed populations which require the explicit assumptions that the data are independent and are approximately normally distributed. Note that if the assumptions are not satisfied, the procedure is not valid and should not be used. However, there may be alternative ways to satisfy such assumptions. For example, a transformation may induce normality in non-normal data, or taking differences of paired (dependent) data will yield independent observations.

For notational convenience, we introduce the quantity

$$\alpha = 1 - \text{confidence level}$$

which upon rearranging, we obtain the relationship in its more usual form

$$\text{confidence level} = 1 - \alpha.$$

We will introduce α in a more formal context subsequently when we discuss hypothesis tests, but in the current context, it defines the probability that the confidence interval does not contain the true parameter, and is thus a measure of inaccuracy, or error.

Recall that by the Central Limit Theorem, the sample mean $\bar{X}$ can be regarded as being normally distributed with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$ when n is sufficiently large. We can thus calculate the probability that $\bar{X}$ lies within a certain distance of μ (provided σ^2, the population variance is known) by using the standard normal, or Z transform. Although unlikely in most applications, we assume we know σ^2.

We can look up the value of z for a chosen confidence level of $1-\alpha$ (i.e. $z_{1-\alpha}$) and construct a confidence interval around the sample mean, by use of z and the standard error of the mean (s.e. $= \sigma/\sqrt{n}$), by considering

$$\bar{x} \pm z_{1-\alpha}\frac{\sigma}{\sqrt{n}}.$$

The confidence interval satisfies the expectation of identifying a range with a probability of $1-\alpha$ that it will include the true population mean. For example, at the 95% level of confidence with $z_{0.95} = 1.96$, we can calculate the 95% confidence interval as

$$\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}\ ,$$

where we would expect 95% of the cases where we shall have defined a range including the true population mean (See later chapter for more details).

Assumptions of Distributions and Variances

The above discussion assumes that we are able to calculate the standard error of the mean as the standard deviation of the population divided by the square-root of the sample size (i.e. $\sigma/\sqrt{n}$). However, most often the standard deviation of the population σ is unknown, and we must use a best estimate for it instead. The estimate used is typically S, the standard deviation actually observed in the sample. This introduces more error, and we must calculate wider ranges to achieve the same level of confidence. Instead of using the z-statistic, we use the Student's t statistic, and we calculate the range as

$$\bar{x} \pm t_{1-\alpha}\frac{S}{\sqrt{n}}.$$

Like z, the value of t can be looked up in a table, but the values of t depend on the sample size. For large sample sizes, t approximates to z, since we can expect that the standard deviation, S, of a large sample will be a good estimate of the true standard deviation, σ, of the population. For smaller samples, t is larger than z, and so we calculate a wider range for our confidence interval than if we had used z. Sample size is indicated in tables of

t as the *number of degrees of freedom* (or "*df*") where, for a simple set of observations,

$$df = n - 1.$$

Student's t distribution is symmetric with respect to 0 and has heavier tails than the standard normal probability density function (pdf). It approaches the standard normal distribution as the number of degrees of freedom (and hence sample size) increase.

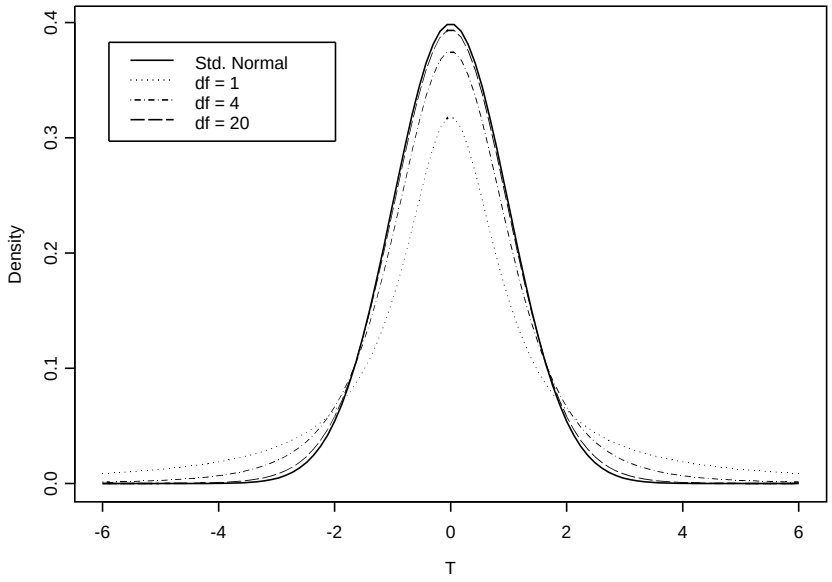


Figure 3.2: Student's t pdfs compared with a standard normal pdf.

3.2.3 Concluding Remarks on Confidence Intervals

Confidence intervals are highly underrated and underused in many areas of scientific research. However, interval estimation is an important area of statistical inference, because a random interval simultaneously provides information regarding both estimate accuracy and precision. It is unfortunate that, due to ignorance of this area, journals prefer the ubiquitous p-value. A few important notes:

1. The interval is random, not the parameter. Thus, we talk about the probability of the interval containing the parameter, and not the probability of the parameter lying in the interval.

2. The width of an interval is a measure of precision. The confidence level of an interval is a measure of accuracy.

3. The width of a CI depends on the size of the estimator's standard error (which depends on the sample size), and on the level of confidence we require (which depends on the sampling distribution of the particular statistic we use to construct the CI).

4. It is imperative that the required assumptions are satisfied before constructing confidence intervals using the formulae described here, as if the assumptions are not satisfied, different procedures must be used to construct the intervals.

3.3 Hypothesis Tests

An important area of statistical inference involves using observed data to decide between competing explanations, or hypotheses, about the process under investigation. Statisticians refer to this aspect of inference as hypothesis testing. Statistical hypothesis testing is a formal means distinguishing between probability distributions on the basis of random variables generated from one of the distributions. The general idea can be written as follows:

Prior to observing the data:

1. State a baseline hypothesis, H_0. This hypothesis is usually a statement of 'no change', or of maintaining the status quo. Hence this conjecture is often referred to as the **null hypothesis**.

2. State an **alternative hypothesis**, H_1. This is typically the hypothesis of interest to the researcher. That is, it is the hypothesis that the researcher wishes to demonstrate to be true.

After observing the data:

1. Decide how likely the observed data is, assuming the null hypothesis to be true.

2. Reject the null hypothesis in favour of the alternative if there is sufficient evidence to suggest doing so. Otherwise, do not reject the null hypothesis.

Example: Pregnancy Test Kit
Suppose a woman buys a pregnancy test kit off the counter from a pharmacy. She is interested to find out whether she is pregnant. The null hypothesis in this case, the status quo, is that she is not pregnant. The alternative hypothesis, the hypothesis of interest, is that she is pregnant. These hypotheses will be formulated prior to observing the data or the result.

Upon testing, the pregnancy test kit may show +ve, evidence that the woman may be pregnant; or -ve, evidence that the woman may not be pregnant. The degree of belief that the woman has in the pregnancy test kit is dependent on the sensitivity and specificity of the pregnancy test kit. This is because the results are seldom definitely correct, and there are associated errors which need to be understood.

3.3.1 Types of Error

For most simple hypothesis tests, there can be two factual scenarios. The first scenario is when the null hypothesis is truly correct, while the second scenario is when the alternative hypothesis is truly correct. As such, there are in general two types of error associated with simple hypothesis tests.

Type I Error
The null hypothesis may be rejected when it is true. We denote the probability of committing this type of error by α, and is commonly called the *significance level* of the test. In the pregnancy test kit example, a type I error can be thought of as the probability of obtaining a +ve result when the woman is in fact not pregnant.

Type II Error

The null hypothesis may be accepted when it is false. The probability of committing this error is denoted by β. In the pregnancy test kit example, a type II error can be thought of as the probability of obtaining a -ve result when the woman is in fact pregnant.

The probability that the null hypothesis is rejected when it is false is termed the *power* of the test, and is equal to $1 - \beta$. The power of a test measures how *sensitive* the test is in detecting deviations from the null hypothesis. Conversely, the *specificity* measures the rate of true negatives, or the probability that the null hypothesis is accepted when it is true, and is equal to $1 - \alpha$. The following table shows the relationship between the introduced terms:

	H_1 **actually true**	H_0 **actually true**
Data supports H_1	Sensitivity = 1 - β (Power)	Type I Error = α
Data supports H_0	Type II Error = β	Specificity = 1 - α

We would like to construct tests with α and β as small as possible. Indeed, since they are probabilities of error, we would like them to be equal to 0 ideally. However there exists a trade-off between these two quantities, where in order to decrease α, we must increase β and vice versa. For example, consider the following example:

A standard test for diabetes is based on glucose levels in the blood after fasting for a prescribed period. For healthy persons, the mean fasting glucose level is found to be 5.31 mmol/L with a standard deviation of 0.58 mmol/L. For untreated diabetes, the mean is 11.74 and the standard deviation is 3.50. In both groups, the levels appear to be approximately normally distributed. To operate a simple diagnostic test based on fasting glucose levels, we need to set a cutoff point C, so that if the patient's fasting glucose level is at

least C, we say they have diabetes. If it is lower, we say that they do not have diabetes. When C is set at 6.5 mmol/L, the sensitivity and specificity is 93.3% and 98.0% respectively, whereas when C is 5.7 mmol/L, the sensitivity and specificity is 95.8% and 74.9% respectively.

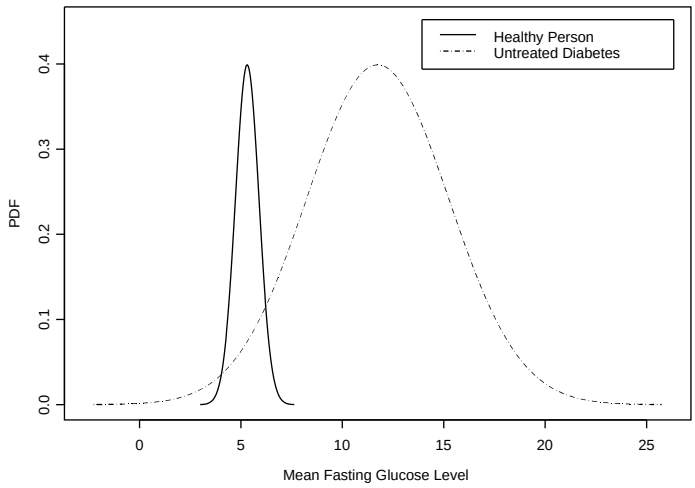


Figure 3.3: Plot of the distributions for mean fasting glucose.

3.3.2 P-values

The p-value is defined as the smallest value of α for which the null hypothesis would be rejected, given the data. That is, the p-value of a test is the probability of observing, by chance, a value of the test statistic as extreme as, or even more extreme than, the one we did observe, assuming that the null hypothesis was true.

If this probability is extremely small, then either H_0 holds and we have observed an extremely rare event; or H_0 is false (there is sufficient evidence to reject H_0 in favour of the alternative, H_1). Thus, the p-value can be seen as a measure of the 'risk' taken when, assuming H_0 is true, we decide to reject this hypothesis. If this 'risk' is sufficiently small, we can feel confidence that we are not observing a freak random event; rather, we are observing strong

evidence against the null hypothesis. We define sufficiently small to be values less than the level of significance of the test, α (typically set at 0.05 or 5%).

So the p-value is the probability of obtaining a false positive (i.e. data support alternative hypothesis when the null hypothesis is actually true). Refer to table on sensitivity-specificity.

P-values and Confidence Intervals
There is a direct correspondence between hypothesis tests and confidence intervals. Simply put, a 100(1-α)% confidence interval for a parameter contains all the values of that parameter for which the null hypothesis of a test would not be rejected at the α level of significance. Therefore, a hypothesis test can be performed through construction of a confidence interval for the parameter of interest. If the hypothesized value (under the null hypothesis) falls in this interval, there is insufficient evidence to reject the null. For example, suppose we are interested in testing the null hypothesis that the mean weight of females in this course is 55kg (i.e. $H_0 : \mu = 55$), and the obtained 95% confidence interval from the sample for the mean is (53.4, 58.6), as the CI contains the hypothesized value, there is insufficient evidence to reject the null hypothesis. If however, we are testing a null hypothesis that the mean weight of females is 50kg, then as this value falls outside the CI, we would reject the null hypothesis in favour of the alternative at 5% level of significance.

3.3.3 General Approach to Hypothesis Testing

Based on the ideas discussed above, we now describe an approach for constructing statistical hypothesis tests. In general, we could consider hypothesis tests which are either one sided or two sided. For instance, if we are interested in finding the mean weight of female students in Oxford, we may be interested in a two sided hypothesis that $\mu \neq 55$. However if we are interested in finding the glucose level in diabetic subjects, we might be interested in a one sided hypothesis that $\mu \geq 6.5$. The usual procedure for the construction of hypothesis tests can be written as follows:

Before data collection/observation:

1. State the hypotheses H_0 and H_1.
2. Choose and fix the significance level of the test, α.
3. Establish the critical region of the test corresponding to α. This region depends on the distribution of the test statistic T under the null hypothesis, and whether the alternative hypothesis is one or two sided.

After data collection/observation:

1. Calculate the value of T from the sample realisation, usually termed t.
2. Compare t with the null distribution of T in order to see whether or not it falls in the critical (rejection) region, or calculate the p-value associated with the observed t.
3. Make a decision about the hypotheses.

For example, to test the null hypothesis that a population mean has some specific value, μ_0, we need to see whether μ_0 lies within the interval calculated for the appropriate level of confidence about our sample mean. The direct way to do this is to calculate how many standard errors of the mean separate μ_0 and the sample mean. We calculate the test statistic T as

$$T = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

where T is distributed as Z. If $T \geq z_{0.95}$, then μ_0 lies outside the range in which 95% of cases should contain the population mean, and we can reject the null hypothesis, that the population mean is μ_0 at a 5% level of significance. In particular, if $T = z_{0.98}$, we can reject the null hypothesis at a 2% level of significance. (For a worked example, see the following chapter).

However if we do not know σ, we have to estimate the standard error of our mean as $S/\sqrt{n}$ instead. The value T calculated in this case is distributed as Student's t, and we must judge our null hypothesis by comparing T with t

for the appropriate sample size (i.e. for the appropriate number of df).

Note also that the test statistic typically comes in two parts:

1. a numerator which measures the separation between an observed mean and the value specified by the hypothesis
2. the divisor, which is the standard error of the mean (or our best estimate of the standard error).

3.3.4 Issues on Hypothesis Testing

Multiple Testing

Suppose we are testing a hypothesis regarding many parameters. For example, suppose we have data on the effects of four different types of drug treatments. We might want to test the hypothesis that the effects from all four treatments is equal, versus the alternative that at least one of the treatment differs from the others.

Suppose we carried out a test at the 0.05 level of significance, and rejected the null in favour of the alternative. The next step will naturally be to find out which treatment gives the best effect. In order to do this, we could test every pair of drug treatment to find which one gives significantly better results, resulting in 6 pairwise comparisons, and thus 6 tests to perform.

Recall that α is the probability of rejecting the null when it is in fact true. When we set this probability to, for example, 0.05, we are effectively making a claim that if the test were to be repeated 20 times, we would expect to make a type I error on one of the 20 tests. Thus when we use the same data to perform multiple comparisons (tests), we need to be aware that we are increasing the chance of drawing spurious conclusions, or effectively increasing the chance of making a type I error.

This problem can be avoided by making each individual comparison more conservative, i.e. by making the significance level (α_{ind}) smaller in order to

maintain an overall significance level of α. Note that α, the overall significance level, refers to the probability that *at least one* of the multiple null hypotheses tested will be rejected when they are all actually true, whereas α_{ind} refers to the probability that any individual null hypothesis will be rejected when it is true. There are many different methods that can be used to decide on a value of α_{ind}, but perhaps the most straightforward and commonly used technique is the *Bonferroni correction*. Simply, if we wish to make m tests on the same data at an overall significance level of α, we should set the significance level of each test at

$$\alpha_{ind} = \frac{\alpha}{m}.$$

Data Driven Hypotheses

Another commonly occurring mistake when undertaking hypothesis tests is to generate hypotheses based on the results of other hypothesis test on the same data. Such situations arise as follows. Suppose that before data collection, the alternative hypothesis was two-sided. After data collection and analysis, the p-value was found to be 0.08 (for example). This is larger than 0.05, and so we cannot reject the null hypothesis at the 0.05 level of significance. However, we know that if we had chosen a one-sided alternative hypothesis, the p-value would have been half that observed under the two-sided alternative, and thus giving a significant result. We therefore construct another (one-sided) test and obtain a significant result.

Unfortunately this should not be performed since it violates the principles of the hypothesis test, where the hypotheses were to be specified *before* observing the data. Strictly defined, this is considered a problem of multiple comparison as well.

It is therefore essential to note that hypotheses are always specified prior to observing the data. If a particular test reveals something interesting about the process under investigation, and hence generates another hypothesis, ideally another experiment should be conducted to test this subsequent hypothesis.

Parametric and Non-Parametric Tests

A parametric test requires that precise assumptions about the population distribution of the quantity of interest be satisfied in order to use it, whereas non-parametric tests do not require any assumptions. The nomenclature can be misleading, since both classes of tests refer to parameters. Usually non-parametric tests are applied to parameters such as the median, which although they are parameters in the broad sense of the term, they do not in general define a distribution, as compared to parameters such as the mean or the variance which often characterise a distribution.

Whenever possible, parametric tests are preferred because, as long as the assumptions required by the parametric tests are satisfied, parametric tests have a larger power than non-parametric analogues.

Chapter 4

Revision on Z-tests and t-Tests

Commonly, when we are interested in investigating the properties of continuous variables, we will usually be interested in making inferences about the population mean. The assumption of known variance is usually not fulfilled in practical applications, although in most situations, the test will assume that the underlying distribution of the variable is normal, or that the sample size is large for either a Student's t-distribution or Normal approximation to be appropriate.

4.1 Single Sample Test for Population Mean

Normality assumed with known variance
Consider the following hypothetical situation: From previous experience we know that the birth weights of babies in England are normally distributed with a mean of 3000g and a standard deviation of 500g. We think that the babies in Australia have a mean birth weight greater than 3000g and we would like to test this hypothesis.

Intuitively we know how to go about testing our hypothesis. We need to take a sample of babies from Australia, measure their birth weights and see if the sample mean is *significantly larger* than 3000g. More formally, we start by writing down our two competing hypotheses.

The main hypothesis that we are most interested in is the **research hypothesis**, denoted H_1, that the mean birth weight of Australian babies is greater than 3000g.

The other hypothesis is the **null hypothesis**, denoted H_0, that the mean birth weight is equal to 3000g.

We can write this compactly as

$$H_0 : \mu = 3000$$
$$H_1 : \mu > 3000$$

The null hypothesis is written first followed by the research hypothesis. The research hypothesis is often called the **alternative hypothesis** even though it is often the first hypothesis we think of.

Normally, we start with the research hypothesis and set up the null hypothesis to be directly counter to what we hope to show. We then try to show that, in light of our collected data, that the null hypothesis is false. The idea behind this approach is that we can never use a sample of data to prove a hypothesis is true but we can use a sample of data to prove a hypothesis is false.

Once we have set up our null and alternative hypothesis we can collect a sample of data. For example, we can imagine we collected the birth weights of 44 Australian babies, with a sample mean birth weight of $\bar{x} = 3275.955$. We now want to calculate the probability of obtaining a sample with mean as large as 3275.955 under the assumption of the null hypothesis H_0. To this we need to calculate the distribution of the mean of 44 values from a $N(3000, 500^2)$ distribution.

We know that if $X_1, X_2, \ldots, X_n$ are n independent and identically distributed random variables from a $N(\mu, \sigma^2)$ distribution, then

$$\bar{X} \sim N\big(\mu, \frac{\sigma^2}{n}\big).$$

Now we can calculate a test statistic T for the null hypothesis that the mean weight of Aussie babies is 3000g using the formula on page 46,

$$\begin{aligned} T &= \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \\ &= \frac{3275.995 - 3000}{500/\sqrt{44}} \\ &= 3.66 \end{aligned}$$

which corresponds to the z-value for 99.985%. The probability that we could observe a mean of 3275.995g in a sample of 44 babies if the true Australian population mean was only 3000g is only 1 - 0.99985 = 0.00015.

The obtained p-value of 0.00015 is very low, implying a very low probability of the data if we assume the null hypothesis to be true.

The convention within statistics is to choose a level of significance before the experiment that dictates how low the p-value should be before we reject the null hypothesis. In practice, many people use a significance level of 5% and conclude that there is significant evidence against the null hypothesis if the p-value is less than or equal to 0.05. A more conservative approach uses a 1% significance level and conclude that there is significant evidence against the null hypothesis if the p-value is less than 0.01.

In our current example, the p-value is 0.00015 which is much lower than 0.05. In this case we would conclude that

"there is significant evidence against the null hypothesis at the 5% level"

Another way of saying this is that

"we reject the null hypothesis at the 5% level"

If the p-value for the test is much larger, say 0.32, then we would conclude that

"the evidence against the null hypothesis is not significant at the 5% level"

or

"we cannot reject the null hypothesis at the 5% level".

Normality assumed with unknown variance

There are certain assumptions which must be satisfied for the use of the Z-test in hypothesis testing, essentially that of either known population variance and large sample sizes. If however the population variance is unknown, and the sample size is insufficiently large, we can use the Student's t-distribution to obtain the critical regions or p-values instead of the normal distribution, provided the distribution of the sample data is sufficiently normal.

Suppose that $X_1, X_2, \ldots, X_n$ is a sample of n observations of a random variable that is distributed as $N(\mu, \sigma^2)$, with σ^2 unknown. To test the hypotheses

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

where H_1 may be a one- or two-sided alternative, the optimum test statistic is

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

where S is the square root of the sample variance. Under H_0, the statistic T has a Student's t-distribution with $n - 1$ degrees of freedom.

Example: Consider the measurements of heights (in mm) of the ramus bone for a sample of 20 boys aged 8.5 years. The data are:

45.3, 46.8, 47.0, 47.3, 47.5, 47.6, 47.7, 48.5, 48.8, 48.9, 49.2, 50.0, 50.4, 50.8, 51.4, 51.7, 52.8, 53.0, 53.2, 54.6

We assume that these heights come from a normally distributed population, as linear biological measures often do (we could investigate this assumption using histograms / qq plots, or a formal goodness of fit test). However, we do not know the population variance.

Suppose that we are interested in testing the null hypothesis that the population mean is equal to 50mm against the alternative that it is not. That is, we want to test

$$H_0 : \mu = 50$$
$$H_1 : \mu \neq 50$$

at the $\alpha = 0.05$ level of significance.

For the data, we have that $\bar{x} = 49.63, s = 2.54$ and $s.e.(\bar{X}) = 0.568$. The observed value of T is $t = -0.664$, and we can compare this value with a Student's t-distribution with $n - 1 = 19$ degrees of freedom.

Note that, since the alternative is two-sided, the critical region for this test will consist of two parts, each with $\alpha/2$ of probability. This region consists of values of T below -2.09 and above 2.09. These values were obtained using statistical tables for the Student's t-distribution, or could alternatively be obtained using SPSS.

Since t does not fall into the critical region, we conclude that there is insufficient evidence to reject H_0 at a significance level of 0.05. The p-value for this test is 0.516. Note that, if our alternative had been one-sided (i.e. $\mu < 50$), then our p-value would have been half that (i.e. 0.258), and the critical region would have included only points below -1.73 as this is the value which accumulate 0.05 of probability under a t_{19} distribution respectively.

4.2 Independent Two Sample Tests for Means

4.2.1 Variances known

Suppose our research hypothesis is that the mean birth weight of boys is greater than the mean birth weight of girls. Suppose we know that the standard deviation of boys weights is 500g and the standard deviation of girls weights is 400g. We want to test our research hypothesis using a significance

level of 5%.

Consider the following steps:

Step 1 Our research / alternative hypothesis can be written as

$$H_1 : \mu_{\text{boys}} > \mu_{\text{girls}}$$

and we set our level of significance to be 5%. This dictates that we will carry out a one-tailed test.

Step 2 We set up our null hypothesis to be directly counter to our research hypothesis

$$H_0 : \mu_{\text{boys}} = \mu_{\text{girls}}$$

Step 3 In this example we will assume that we collected data from Australian babies, and we have $n_{\text{boys}} = 26$ boys and $n_{\text{girls}} = 18$ girls.

Step 4 We base our test statistic on the difference between the sample means of the boys and girls. Under the null hypothesis, we know that

$$\bar{X}_{\text{boys}} \sim N\Big(\mu, \frac{500^2}{26}\Big)$$
$$\bar{X}_{\text{girls}} \sim N\Big(\mu, \frac{400^2}{18}\Big)$$

We need to test H_0, which is $\bar{X}_{\text{boys}} - \bar{X}_{\text{girls}} = 0$. We need to calculate a test statistic to compare the difference of the two means with 0. It can be shown that the standard error for the difference of the two means is

$$\sqrt{s.e._{\text{boys}}^2 + s.e._{\text{girls}}^2}.$$

Therefore under the null hypothesis, we know that

$$\bar{X}_{\text{boys}} - \bar{X}_{\text{girls}} \sim N\Big(0, \frac{500^2}{26} + \frac{400^2}{18}\Big)$$

Thus we can construct a test statistic as

$$Z = \frac{\bar{X}_{\text{boys}} - \bar{X}_{\text{girls}}}{\sqrt{\frac{500^2}{26} + \frac{400^2}{18}}} \sim N(0, 1)$$

Suppose we have $\bar{x}_{\text{boys}} = 3375.308$ and $\bar{x}_{\text{girls}} = 3132.444$, we obtain $Z = 1.785$.

In general, to test for a difference between two means (with σ_1 and σ_2 known) from n_1 and n_2 observations from two groups, under a null that the difference is μ_0, we use the test statistic

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - \mu_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1).$$

4.2.2 Variances unknown

Suppose we have two independent samples of sizes n_1 and n_2. Further, suppose each sample is from normally distributed random variables X_1 and X_2. Consider a null hypothesis that the difference in the variables means is μ_0. Often μ_0 is often 0 in practice (test of no difference between the two population means). For this test, we use the statistic

$$T = \frac{\bar{X}_1 - \bar{X}_2 - \mu_0}{s.e.(\bar{X}_1 - \bar{X}_2)}$$

There is a need to distinguish between situations where the samples share a common variance and situations where they do not. The formulae for the standard error of the difference of the means, as well as the degrees of freedom for the null distribution of T are different depending which of these situations holds.

Equal Variances

Assume that both samples have equal (unknown) variances. In this case, the estimated value of the quantity in the denominator of T is

$$s.e.(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{(n_1 + n_2)[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]}{n_1 n_2 (n_1 + n_2 - 2)}}$$

where σ_1^2 and σ_2^2 are the unbiased estimates for the variances in each sample. The null distribution of T then follows a Student's t-distribution with $(n_1 + n_2 - 2)$ degrees of freedom.

Unequal Variances

If the variances of the two samples are different, we need to use a different estimated standard error of the mean and to calculate the degrees of freedom using an approximation. In this situation, the estimated standard error of the mean is

$$s.e.(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

and the degrees of freedom may be approximated by

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

4.3 Paired Two-Sample Tests

In certain situations we might be interested in comparing the effect of a particular treatment on pairs of observations. These pairs can either come from the same individual measured before and after the treatment (self-pairing) or from pairs of similar individuals (e.g. pairs of patients of the same sex, age, etc.) given different treatments. Pairing is used in an attempt to make comparisons between treatments more accurate. It does this by making members of any pair as similar as possible in all areas except treatment category. Thus, any difference we do see can be attributed (in theory) to treatment effects.

Denote the paired sample as $(X_i, Y_i), i = 1, 2, \ldots, n$. We assume that each pair element is individually normally distributed with means μ_X and μ_Y respectively, and unknown variances. The null hypothesis is $\mu_X - \mu_Y = \mu_0$, against a one-sided or two-sided alternative. Note that the value μ_0 is the hypothesized mean (population) difference between treatments. Often, the researcher will be interested in testing the hypothesis that $\mu_0 = 0$ (the hypothesis that there is no difference between treatments).

Note that paired data are statistically dependent, and thus violate the assumption of independence. In order to remove this dependency, we take pairwise differences, and use these differences as the data sample. Thus, let

$D_i = X_i - Y_i$ (i.e. the difference for the ith pair). The statistic for this test is then

$$T = \frac{\bar{D} - \mu_0}{s.e.(\bar{D})}$$

where $\bar{D}$ is the mean of the differences and its estimated standard error is given by

$$s.e.(\bar{D}) = \frac{s.d.(D)}{\sqrt{n}} = \frac{1}{\sqrt{n}}\sqrt{\frac{\sum_{i=1}^{n}(D_i - \bar{D})^2}{n-1}}.$$

The null distribution of T is a Student's t-distribution with $n-1$ degrees of freedom. If the sample size is large, the approximate validity of the test follows from the Central Limit Theorem (even if the underlying distribution is not normal) and the normal distribution may be used as an approximation.

4.4 Tests for the Population Proportion

4.4.1 One Sample Test

Often, hypotheses may be formulated to test the population proportion of 'success' events. Such formulations are usually warranted in situations where dichotomous outcomes ('success' and 'failure') are possible, and the researcher is interested to know the proportion of 'success' events. Occasionally, the researcher will consider one of the possible values of p to be of special interest. In such cases, the researcher can perform a hypothesis test with the null hypothesis stating that p equals the special value of interest, p_0.

First, we should note that, in this special case, when there are only two possible outcomes, e.g. 'success' or 'failure', then the mean proportion of occurrence of success, p, has a standard error of

$$\sqrt{\frac{p(1-p)}{n}}.$$

The proportion of successes in the sample of n trials, $\hat{p} = \frac{x}{n}$, is the point estimate of the proportion of successes in the population. Formulate the hypothesis as

$$H_0 : p = p_0 \; vs. \; H_1$$

where H_1 is a one- or two-sided alternative. The optimum test statistic is

$$T = \frac{\hat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

For large values of n and under H_0, T follows an approximate normal distribution.

Note that the variance for hypothesis testing of population proportion, $\frac{p_0(1-p_0)}{n}$ is different from that used for constructing confidence intervals of the population proportion, $\frac{\hat{p}(1-\hat{p})}{n}$. In hypothesis testing, we have the hypothesized population proportion p_0, and hence the variance must be constructed from p_0. In constructing confidence intervals for p, we do not know the population proportion and hence p is estimated by the unbiased estimate $\hat{p}$.

4.4.2 Two Sample Tests of Proportions

In certain situations, instead of comparing population means of two samples, we are interested in comparing the proportions of specific events within each sample or strata. This is equivalent to finding the magnitude of $p_1 - p_2$, where p_i indicates the proportion for population i. However, in comparisons of proportions, there are a few possible scenarios in which we have to differentiate between:

1. Difference in proportion from independent samples.
2. Difference in proportion for several response categories from a single sample.
3. Difference in proportion for different dichotomous responses from a single sample.

The general test statistic in testing for differences in population proportions is given by

$$T = \frac{(\hat{p_1} - \hat{p_2}) - (p_1 - p_2)}{s.e.(p_1 - p_2)}$$

where in testing for no difference in population proportion, the second term in the numerator equals 0. Under the null hypothesis, T follows an approximate normal distribution. However it is important to note the following different standard errors for different scenarios. We will briefly mention each scenario in turn.

Proportions from Independent Samples
This scenario is most commonly seen in estimating the response rates for the same categorical variable/question across different strata/population.

For example, comparing the difference in proportions of patients recovered after receiving a homogenous drug treatment in Oxford and Abingdon. So we assume that the proportion of recovered patients in Oxford p_1 is independent from the proportion of recovered patients in Abingdon p_2.

Suppose sampling is performed with n_1 and n_2 patients sampled from Oxford and Abingdon respectively, and with corresponding sample proportions of $\hat{p}_1$ and $\hat{p}_2$ of patients recovering correspondingly.

The standard error of $p_1 - p_2$ is thus

$$s.e.(p_1 - p_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}$$

Proportions from Single Sample with Multiple Responses
This is most commonly seen in situations where a single categorical variable can choose from more than two responses, and estimation of proportions are within the same sample. An easy way to recognize this scenario is that the summation of all the response proportions must be equal to unity.

For example, 200 patients from a hospital in Oxford is randomly sampled and asked to rate the service standards of the hospital staff, with 4 possible ratings, "Very poor", "Moderately poor", "Moderately good", "Very good".

Suppose the collated results are correspondingly 12, 44, 68 and 76, hence yielding the proportion of 0.06, 0.22, 0.34 and 0.38. Testing for difference in the response proportions for category i and j, $p_i - p_j$, will not be a test between two independent samples, since there is only one sample in consideration. Hence the previous formulation will not be applicable. The standard error has to capture the additional information that the two categories are obtained from the same sample and that there is a certain level of dependence between the two proportions p_i and p_j (since when one changes, there is a likelihood that the other will change correspondingly since the summation of the proportions equal unity).

The standard error of $p_1 - p_2$ is thus

$$s.e.(p_1 - p_2) = \sqrt{\frac{\hat{p}_1 + \hat{p}_2 - (\hat{p}_1 - \hat{p}_2)^2}{n}}$$

Proportions from Single Sample with Dichotomous Responses to Multiple Factors

This is most commonly observed in situations where the same sample is required to assess the outcome of many dichotomous responses. A feature of this form of testing is that there are multiple questions for a single sample, and that each question has only two possible outcomes.

For example, 200 students from the cohort of students taking a course are randomly sampled to answer a few questions, with only 2 possible responses of either "yes" or "no" to choose from. The questions include

1. Do you enjoy the lectures?
2. Do you find the content easy?
3. Do you find the lecturer is going too slowly?

Out of the 200 responses, 180 answered yes to the first question, 170 answered yes to the second question and 120 answered yes to the third question. So

the sample proportions of the students who answered "yes" for questions 1, 2 and 3 are thus 0.90, 0.85 and 0.60. Thus we can establish unbiased estimates of $p_i, i = 1, 2, 3$ as $\hat{p}_1 = 0.90$ and hence $\hat{q}_1 = 0.10$, $\hat{p}_2 = 0.85$ and $\hat{q}_2 = 0.15$, and finally $\hat{p}_3 = 0.60$ leading to $\hat{q}_3 = 0.40$. In testing for the difference in proportions of students who answered "yes" for question 1 and 2, we need to capture the interdependence between the questions in that we are testing within the same sample.

The standard error of $p_1 - p_2$ is thus

$$s.e.(p_1 - p_2) = \sqrt{\frac{\min(\hat{p}_1 + \hat{p}_2, \hat{q}_1 + \hat{q}_2) - (\hat{p}_1 - \hat{p}_2)^2}{n}}$$

Chapter 5

ANOVA and Chi-Square Tests

In the previous chapter we have discussed parametric tests for testing the differences of means between two populations. It is possible to generalise the analysis for comparing the means from two independent, normally distributed populations, to comparing the means from more than two independent populations. Also we extend our repertoire of statistical tests by studying chi-square tests, which allow us to test whether a sample of data is consistent with a specific theoretical distribution, and also allow us to test for an association between two categorical variables. Such tests of associations are extremely useful and very common in both Physiology, Psychology and Human Sciences.

5.1 Analysis of Variance

Analysis of variance, or commonly known as ANOVA, at its simplest level allow for a comparisons of the means of two or more independent populations. The technique compares these means by examining certain components of variance, and hence the name.

Consider the following situation: Suppose we have random samples from each of k normally distributed populations. We write the null hypothesis that all the means are equal (of homogeneity of means) as

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

This hypothesis is tested against the *general* alternative that at least one of the population means is different from the others. Note that the preceding statement defines several alternative hypotheses. For instance, suppose that we are analysing 3 samples. It may be that only one of the means is different while the other two means might be equal (and therefore form a homogeneous group). Additionally, we would also want to reject H_0 if the three means were all different from each other. If we do not reject the null hypothesis, we simply conclude that there are no differences between the means. If we do reject the null, we must still explore the possible patterns that led to this rejection. This exploration is known as *post-hoc analysis* and will be discussed later.

The basic idea underlying ANOVA relies on an important theorem in mathematical statistics which states that the total variance of the pooled sample can be divided into two components: the *within groups variance* and the *between groups variance*.

The within groups variance is simply the sum of the variances calculated for each individual group. The between groups variance is the variance obtained using the means of each group as data. The within groups variance represents the internal variability of the groups; the between groups variance measures how well separated these groups are. If the groups are well separated from each other, the ratio of between group variance to within group variance should be large. In order to decide how large the value of this ratio should be in order to be considered significant, we use the F distribution. The degrees of freedom are $(k-1)$ for between groups variance, and $(n-k)$ for within groups variance, where n is the total sample size and k the number of groups.

Note that the two-sample t-test (with equal variance) is a particular case of ANOVA (with $k = 2$). Thus, we would expect that the assumptions of normality and homogeneity of variance (required by the t-test) are also required for comparing k independent samples, and we should test that these assumptions are satisfied before commencing on any analysis.

Example: Consider the following example, where we are interested to compare the means of the daily consumption of the anti-psychotic drug clozapine (CPZ) for subjects of the 3 different genotypes for the dopa-responsive gene (DRD). This is a typical example where we are comparing the means of three groups for differences, with the null hypothesis of no differences between all three groups. SPSS produces the following output (Fig. 5.1), with a p-value of 0.420, indicating no evidence to reject the null hypothesis. Therefore we conclude that there is no difference in the daily intake of CPZ for the three genotype groups.

ANOVA

Daily CPZ equivalent(Mg)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	590721.4	2	295360.724	.869	.420
Within Groups	1.17E+08	343	340020.766		
Total	1.17E+08	345			

Figure 5.1: One-way ANOVA of the daily consumption of CPZ for the 3 different genotype groups.

Non-parametric Approach

When using the ANOVA procedure, we assume that the data from each group follow a normal distribution and that the groups we are comparing have homogeneous variances. If the variances are not all equal, then the conclusions about the group means drawn from ANOVA analysis may not be valid since the observed ANOVA p-value will be smaller than the one we would have obtained if the assumption of equal variance was satisfied. This means that ANOVA will yield an anti-conservative p-value (that is, an increase in the probability of Type I error) if the homogeneity of variances is not satisfied. Therefore, it is important to test, either formally or informally, that the homogeneity of variance assumption is satisfied, as has been stated before.

If this assumption does not appear to be satisfied, transforming the data

(perhaps using a logarithmic transformation) can sometimes help, as we will see in the Post-hoc analysis section below. In addition to homogenising the variances, using a transformation may sometimes induce approximate normality in the data.

However, if we cannot find a transformation that appears to homogenise the variance or normalise the variables, then we should consider using a non-parametric test. Non-parametric tests do not require specific assumptions regarding the distribution of the underlying population, and the non-parametric analogue of the independent two-sample t-test is the **Mann-Whitney test**. The non-parametric equivalent of ANOVA is the **Kruskal-Wallis test**, which is a generalisation of the Mann-Whitney test for more than two groups. For both the general test and its two-sample version, the null hypothesis is that the **medians** are equal, against the general alternative that at least one differs from the others. Note that it makes sense to compare the medians, rather than the means, because if the data are skewed, as they probably would be if we are using a non-parametric test, then the value of the mean will be either artificially inflated or deflated. The Mann-Whitney and the Kruskal-Wallis tests test this null hypothesis by transforming the data into pooled ranks (i.e. they start by assigning rank 1 to the smallest observation in the pooled sample, and so on) and then calculating a test statistic from these ranks. Both tests appear in the non-parametric submenu of SPSS.

5.1.1 Post-Hoc Analysis

In the use of ANOVA or the Kruskal-Wallis test, when we investigate how the centre of the continuous variable changes across the k groups, our initial null hypothesis is that the population means (or medians) are all homogenous. If we fail to reject this hypothesis, the analysis ends there. If we do reject the initial null hypothesis, then we will have to establish the reason(s) for doing so. For example, it may be that only one group mean differs from the rest, or that there is a particular pattern in which the groups appear to be separated.

One way of finding significant differences between the means is to make all possible pairwise comparisons (i.e. test if each pair of means is equal). Note that we can use either the two-sample t-tests to make these pairwise comparisons if we assume that the populations are normal, or the Mann-Whitney test to make these pairwise comparisons. In either case, making these pairwise comparisons leads to the problem of *multiple comparisons*, as we saw in Chapter 3. As before, one way to ensure an overall significance level of α is to use the Bonferroni correction, whereby each individual test is conducted at the

$$\alpha* = \frac{\alpha}{\frac{k(k-1)}{2}}$$

level of significance. Note that $\frac{k(k-1)}{2}$ is the number of possible pairwise comparisons between k groups. There are other techniques for adjusting the significance level used for multiple comparisons, and statistical software such as SPSS gives several possibilities, including some which correct for unequal population variances. We will only concentrate on Bonferroni correction in these notes.

Example: Cesare et al (1990) conducted a role-play experiment of the effect of physical disabilities on interviewer ratings. There were 14 assessments by different 'interviewers' for each of the 5 groups, control and four types of disability (Amputee-, Crutches-, Hearing-, Wheelchair-disability). This is an example of a comparison of the means scores with several groups ($k = 5$ here). The p-value obtained after performing an ANOVA is 0.027, which implies that there is evidence to suggest that the means are not all equal.

In order to investigate the patterns in the means, we perform a post-hoc analysis using the Tukey correction (another method for correction) to adjust for multiple comparisons. Figure 5.2 shows the confidence intervals after performing all possible pairwise comparisons, and it was found that there exists statistical evidence that the means for the groups Crutches and Hearing are significantly different, while there is no evidence to suggest that the other pairwise comparisons of means are significantly different. Therefore the p-significant value (of 0.027) for the ANOVA can be attributed mainly to the

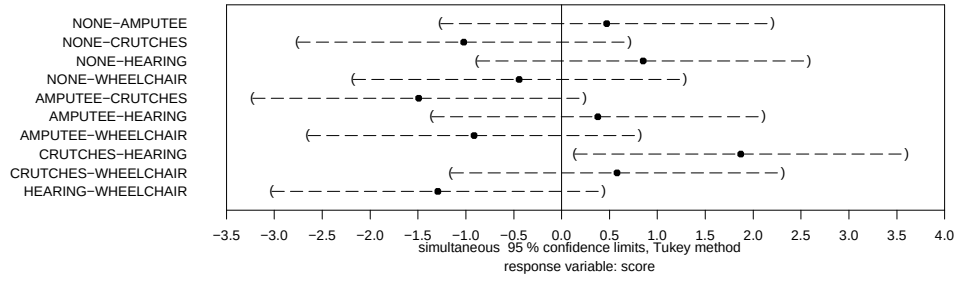


Figure 5.2: Pairwise comparisons of means using Tukey correction for multiple comparisons.

difference in means between the groups Crutches and Hearing.

5.2 Categorical Variable

Suppose now that the variable whose properties we hope to investigate is categorical rather than continuous. Further, suppose this variable has K (mutually exclusive) categories (i.e. each unit in the underlying population falls into exactly one of the K categories) and that we have a random sample of n units. In this case, we are not interested in the variable's mean, since the mean is not defined for ordinal and nominal variables. Instead we are interested in testing the null hypothesis that the proportions of the population in the K categories are $\pi_1, \ldots, \pi_K$ respectively, where the π_i sums to one. Often we will be interested in the particular hypothesis that the π_i's are equal (i.e. $\pi_i = 1/K$). However, it should be noted that the test presented below is quite general and does not simply apply to testing the equality of category proportions for a single categorical variable.

If we let π_i represent the hypothesized proportion of units falling into the i^{th} category, we would expect $E_i = n\pi_i$ occurrences of the objects in the i^{th} category if the null hypothesis were true. Letting O_i denote the observed number of units (out of n) in the i^{th} category, the following test statistic measures how far the hypothesized (expected) data is from the observed data. Intuitively, if this value is large, there is evidence against the null

hypothesis whereas if this value is small, the information contained in the sample provides little evidence against the null. The test statistic is

$$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} \ ,$$

which, under the null hypothesis, has a Chi-squared distribution with $K-1$ degrees of freedom. Therefore, the rejection region for this test consists of all values that are greater than the $(1-\alpha)$ quantile of the distribution, for an appropriate choice of α. The quantiles of the χ^2 distribution (with the appropriate number of degrees of freedom) can be easily calculated in the majority of statistical software packages or found in χ^2 tables in most statistics texts.

Example: Suppose we wish to test whether there is a seasonal effect of the number of births in Oxford, and we observe 1361 births in total, 334 of which were in spring, 372 in summer, 327 in autumn, and 328 in winter. If we let π_i be the proportions of births in each of the four seasons, then our hypothesis is

$$H_0 : \pi_1 = \pi_2 = \pi_3 = \pi_4 = 0.25$$
$$H_1 : \text{at least one proportion is different}$$

Consider the following table to see how the test works:

	Spring	**Summer**	**Autumn**	**Winter**
Observed	334	372	327	328
Expected	340.25	340.25	340.25	340.25
Obs-Exp^2/Exp	0.11	2.96	0.52	0.44

The observed value of the test statistic is 4.03, with 3 degrees of freedom. The rejection region, at the 0.05 level of significance, are values of the χ^2 statistic that are greater than $\chi^2_{0.05,3} = 7.82$. Hence, we would not reject the null hypothesis, or we conclude that there is insufficient evidence to suggest a seasonal effect on the birth rate.

5.2.1 Binary Response with Categorical Predictor

Suppose we have K populations from which we have taken samples of size $n_1, n_2, \ldots, n_K$ respectively. In addition, suppose that each of the K populations can be further divided into a 'success' category and a 'failure' category. Alternatively, we could describe this situation by saying that we have a categorical predictor variable with K levels and a binary (1 = success, 0 = failure) response variable. We are normally interested in testing whether the true population proportions of success in the K groups are equal to some hypothesized values $p_1, \ldots, p_K$, where each p_i is between 0 and 1. Frequently, the researcher may be interested in testing whether the probability of success is the same in each of the populations, in which case the K hypothesized proportions would all be equal. (i.e. $p_1 = p_2 = \cdots = p_K = p$, for some p between 0 and 1). However, this is not the only null hypothesis that can be investigated using the following test.

If the null hypothesis were true, then we would expect $E_i = n_i p_i$ successes in the i^{th} sample for $i = 1, \ldots, K$. Let O_i denote the observed number of successes in the i^{th} sample. To use these observed and expected frequencies to test the general null hypothesis specified above against an appropriate alternative, we employ the χ^2 test statistic, since it measures in some sense, the distance between the observed and expected category frequencies. However, in this situation, the test statistic has a χ^2 distribution with K, rather than $K - 1$ degrees of freedom under the null hypothesis.

5.2.2 Goodness-of-Fit Tests

Instead of testing hypotheses regarding a population parameter, we may be interested in hypotheses about the structure of the underlying population. For instance, in the heights example we saw previously in Chapter 3, we assumed the data were sampled from a normally distributed population. However, we might be interested in formally testing this assumption using the data in our sample. This is a non-parametric test since there are no specific assumptions on the distributional form of the data.

The χ^2 goodness-of-fit test is appropriate only for testing hypothesis regarding the distribution of discrete random variables (there is an analogous version for continuous variable in the form of the Kolmogorov-Smirnov goodness-of-fit test). The test compares the observed frequencies to what we would expect to see, assuming that the probability model specified by the null distribution is true. The test statistic is a measure of the distance between the observed and expected frequencies. The value of this statistic is then compared with the critical region for the test (which is constructed from the null distribution).

Suppose we observe a sample of size n and we obtain the frequencies for m different values or classes (that is, we count how many of the n observations fall into each class). The test statistic is

$$\chi^2 = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i}$$

where O_i and E_i are the observed and expected frequencies for the i^{th} class respectively. The closer the observed frequencies are to the expected frequencies for each class, the less likely we are to reject H_0. If there are discrepancies between the data and the expected frequencies under H_0, then χ^2 will be large and we are more likely to reject the null.

The degrees of freedom for this test is

$$df = m - 1 - p.e.$$

where $p.e.$ is the number of parameters estimated in order to calculate the expected frequencies E_i. The critical region of the test consists of large values of χ^2, as this indicates discrepancies between the data and the hypothesized model. For this test, the p-value is the probability that a chi-square distribution with df degrees of freedom will take a value larger than or equal to the calculated value of χ^2.

Example: A retrospective research studying the temporal distribution of the outbreaks of major flu epidemics over the period 1500 to 1931 is performed, counting the number of outbreaks each year. The random variable of interest counts the number of outbreaks occurring in each year of that 432 year period. The observed frequencies indicate, for instance, that there were 223 years with 0 outbreaks of major flu epidemics.

The null hypothesis is that this data is a realisation of a Poisson distributed random variable, X (i.e. the random variable has a Poisson distribution). We decided on this distribution for the null hypothesis since it is a good model for random events occurring at a constant rate. In this distribution

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x} \ ,$$

where x denotes the number of outbreaks in any one year, and λ the mean number of outbreaks per year (i.e. total outbreaks / total number of years).

In order to calculate the expected frequencies under the null hypothesis, we need to estimate the parameter λ of the Poisson distribution for this data. Using the sample mean, we estimate its value to be 0.6921. The expected values are found by plugging this value for λ into the probability mass function for the Poisson distribution for each category, and multiplying these probabilities by the total number of events (432). The observed and expected frequencies are

X	0	1	2	3	4	≥ 5
observed	223	142	48	15	4	0
expected	216.23	149.65	51.79	11.95	2.07	0.28

The agreement between the data and the values produced by the model is quite good, and the value of χ^2 is 0.1047, on 6 - 1 - 1 = 4 degrees of freedom (since we had to estimate one parameter - the mean of the Poisson distribution), and the p-value is 0.99. That is, we would be almost certainly wrong to reject the null hypothesis of a Poisson model for these data.

5.2.3 Testing for Independence

We now consider how to analyse statistically whether two or more categorical variables are independent of each other.

One way of representing the relationship between two categorical variables is via a two-way table. Let one of the variables have r levels and the other have c levels. Then, a two-way table is a table with r rows and c columns, with each cell containing the observed number of objects falling into that crossed category. Let N_{ij} denote the number of observations in the i^{th} row and j^{th} column; let $i = 1, \ldots, r; j = 1, \ldots, c$, C_j denote the column totals; R_i denote the row totals; and n the total sample size. Then, the general form of a two-way table is

				Column			
Row	**1**	**2**	...	**j**	...	**c**	**Total**
1	N_{11}	N_{12}	...	N_{1j}	...	N_{1c}	R_1
2	N_{21}	N_{22}	...	N_{2j}	...	N_{2c}	R_2
⋮	⋮	⋮		⋮		⋮	⋮
i	N_{i1}	N_{i2}	...	N_{ij}	...	N_{ic}	R_i
⋮	⋮	⋮		⋮		⋮	⋮
r	N_{r1}	N_{r2}	...	N_{rj}	...	N_{rc}	R_r
Total	C_1	C_2	...	C_j	...	C_c	n

We are interested in testing whether the row and column variables are independent: the null hypothesis is that there is no relationship between the row and column classifications. There are two possible scenarios that might lead us to test this hypothesis. In the first, we have a sample of n units drawn from a population. We believe that each unit in the sample can be classified according to two categorical variables. In this case, the variables have a symmetric relationship; this is not the case in the second scenario, in which there are c populations, and sample of sizes $C_1, \ldots, C_c$ are drawn from each. Each unit is then classified according to a categorical variable with r possible levels.

The difference between these two situations lies in how the data is collected. In the first case, the researcher sets the sample size, n, and then classifies

each unit into one of the rc cells. In the second case, the column totals are the sample sizes selected at the design stage. The first situation is known as *multinomial sampling*, and the second as *product multinomial sampling*. Although these are two completely separate scenarios, the χ^2 testing procedure described below is the same for both.

The statistic that tests the null hypothesis in an $r \times c$ table compares the observed counts with the expected counts, the latter being calculated under the assumption that the null hypothesis is true. The expected count in the ij^{th} cell of the table is given by

$$E_{ij} = \frac{R_i C_j}{n} \ ,$$

and the χ^2 test statistic is

$$\chi^2 = \sum_{ij} \frac{(N_{ij} - E_{ij})^2}{E_{ij}}.$$

This statistic has a χ^2 distribution on $(r-1)(c-1)$ degrees of freedom. Therefore, we reject the null hypothesis that the rows and columns are independent if the observed value of the χ^2 statistic is greater than $\chi^2_{\alpha,(r-1)(c-1)}$, the $(1-\alpha)$ quantile of the chi-squared distribution with $(r-1)(c-1)$ degrees of freedom.

Note that these results are based on approximations, and thus there are certain important assumptions that need to be satisfied in order to use this test. To apply this theory, we must assume that we have a sufficiently large sample such that

1. The smallest expected count is 1 or more.

2. At least 80% of the cells have an expected count of 5 or more.

Example: It was thought that there might be a genetic predisposition to myopia and a case-control study is carried out to find out whether there is any evidence to suggest genetic association with the incidence of severe myopia. The following table shows the genotypic distribution among the 2 groups.

Meir Genotype	Myopes	Normal	Total
AA	97	62	159
AG	91	56	147
GG	20	13	33
Total	208	131	339

We are interested in testing whether the genotypes of the Meir gene is associated with the onset of severe myopia. In doing so, we find the χ^2 test statistic to be 0.0347, which we compare to a χ^2 distribution with $(3-1)\times(2-1) = 2$ degrees of freedom. This produces a p-value of 0.983, presenting insufficient evidence to suggest any association between the Meir gene and the onset of severe myopia.

5.3 Additional Aspects of Categorical Analysis and χ^2

5.3.1 Equality of Variance

Previously, we have seen that particular forms of parametric tests for more than one group assume homogeneity of variances between the groups. We are able to test the assumption of variance equality by the use of the **F-test**. If two independent random variables, X_1 and X_2, each having χ^2 distributions on ν_1 and ν_2 degrees of freedom respectively, then the random variable

$$F = \frac{X_1/\nu_1}{X_2/\nu_2}$$

has an **F-distribution** with parameters ν_1 and ν_2, denoted $F(\nu_1, \nu_2)$.

It can be shown that if $X_1, \ldots, X_n$ is a random sample from a $N(\mu, \sigma^2)$ population.

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \ ,$$

where S^2 denotes the unbiased sample variance. Therefore, to test the hy-

pothesis

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$$

against a one- or two-sided alternative, we use the statistic

$$F = \frac{S_1^2}{S_2^2}$$

where S_i^2 is the unbiased sample variance estimate for sample $i, i = 1, 2$. From the result stated above, the null distribution of the test if $F(n_1 - 1, n_2 - 1)$, and this can be used to construct a rejection region appropriate to the form of the alternative hypothesis. Most statistical packages will perform this test although it is important to note that this result depends on the assumption that each sample is normally distributed and independent.

5.3.2 Odds Ratio

Consider a case-control study, classified by the presence or absence of a qualitative factor (i.e. smoker or non-smoker) and a dichotomous disease status (i.e. presence or absence of lung cancer), and consider the data in the following tabular form:

	Lung Cancer (+)	**Lung Cancer (-)**	
Smoking (+)	n_1	n_2	$n_1 + n_2$
Smoking (-)	n_3	n_4	$n_3 + n_4$
	$n_1 + n_3$	$n_2 + n_4$	n

Association between the disease status and the qualitative factor can be determined through the use of a χ^2 test of independence, provided the assumptions are satisfied. We are also interested in the effect of the factor, and the ratio of odds, defined as

$$OR = \frac{n_1/n_3}{n_2/n_4} = \frac{n_1/n_2}{n_3/n_4} = \frac{n_1 n_4}{n_2 n_3}$$

is a common measure used.

The odds ratio is usually reported together with the confidence interval, as a representation of the statistical significance of the odds ratio. Obtaining

the confidence interval involves calculating the variance of the odds ratio and obtaining the variance of the odds ratio directly is non-trivial. However, the variance may be easily obtained through a logarithmic transformation of the odds ratio, as

$$\text{Var}(\log OR) = \frac{1}{n_1} + \frac{1}{n_2} + \frac{1}{n_3} + \frac{1}{n_4}.$$

It is therefore easier to obtain the confidence interval for the log odds ratio, and by taking exponentials, obtain the confidence intervals of the odds ratio.

Example: Numbers of children with malignant disease and their controls whose mothers (retrospectively) reported influenza during the relevant pregnancy (Oxford survey of Childhood Cancers)

	Influenza	No Influenza	
Cancer	96	8766	8862
No Cancer	64	8798	8862

By performing a χ^2 test of independence, we obtain $\chi^2_1 = 6.061$, yielding a p-value of 0.0138, suggesting that association exists between the onset of cancer and the presence of influenza. The odds ratio is $\frac{96/8766}{64/8798} = 1.51$, suggesting that children whose mothers have influenza during pregnancy is 1.5 times more likely than children whose mothers do not have influenza during pregnancy to suffer from malignant cancer. The variance for the log odds ratio is

$$\text{Var}(\log OR) = \frac{1}{96} + \frac{1}{8766} + \frac{1}{64} + \frac{1}{8798} = (0.16208)^2$$

and the 95% confidence interval for the log odds ratio is

$$\Big(\log(1.51) - 1.96 \times 0.162,\ \log(1.51) + 1.96 \times 0.162\Big) = (0.09, 0.73)$$

yielding the 95% confidence interval for the odds ratio as (1.10, 2.07).

Chapter 6

Linear Regression

Regression is a tool for exploring relationships between variables and the most common form of statistical analysis for the case in which both the response and the predictors are continuous is known as **linear regression**. Linear regression explores relationships that are readily described by straight lines, or their generalisations to many dimensions. A surprisingly large number of problems can be analysed using techniques of linear regression, and even more can be analysed by means of transformations of original variables that result in linear relationships among the transformed variables. Linear regression is a parametric procedure, as the response variable is assumed to follow a normal distribution. A complete description of linear regression is beyond the scope of this course, however sufficient details are given to have a basic understanding of the concepts behind linear regression, and the statistical extensions that are possible.

6.1 Linear Model

6.1.1 Simple Linear Regression

In simple linear regression, we test for a linear relationship between two variables, one of which is defined as the *explanatory variable*, predictor or the regressor (x), and the other defined as the *dependent variable*, response or the regressand (y). Linear regression models the linear relationship between

the response (y) and the predictor (x) as

$$E(y|x) = \alpha + \beta x,$$

where α denotes the *intercept* parameter and β the *slope* parameter. Statistical techniques such as maximum likelihood, or numerical techniques such as least squares, are used to estimate these parameters and provide their standard errors. By testing the hypothesis that $\beta = 0$, we can test whether y is linearly related to x.

Another manner of looking at the above relationship is

$$y = \alpha + \beta x + \epsilon$$

for an error ϵ which is normally distributed with mean 0 and variance σ^2. One of the crucial assumption of linear regression is *homoscedasticity*, or constant variance for ϵ for all cases.

Linear regression is performed by all statistical software packages, with most giving the estimate of the regressions parameters and their standard errors, the test statistic (t) for testing the hypothesis of no relationship between the response and the predictors, and the p-value of the test.

6.1.2 Multiple Regression

Suppose there are p numeric explanatory variables $x_1, \ldots, x_p$, a linear model postulates

$$Ey = \alpha + \beta_1 x_1 + \cdots + \beta_p x_p$$

or

$$y = \alpha + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

for $\epsilon \sim N(0, \sigma^2)$.

In general, each predictor contributes a single *term* in the model formula, and a single term may contribute more than one coefficient to the fit.

Example: Consider the environmental study of the concentration of Ozone in New York that measured four variables ozone, solar radiation, temperature and wind speed for 111 consecutive days from Chambers and Hastie, 1992. The diagram shows a scatter plot of ozone against temperature. From the

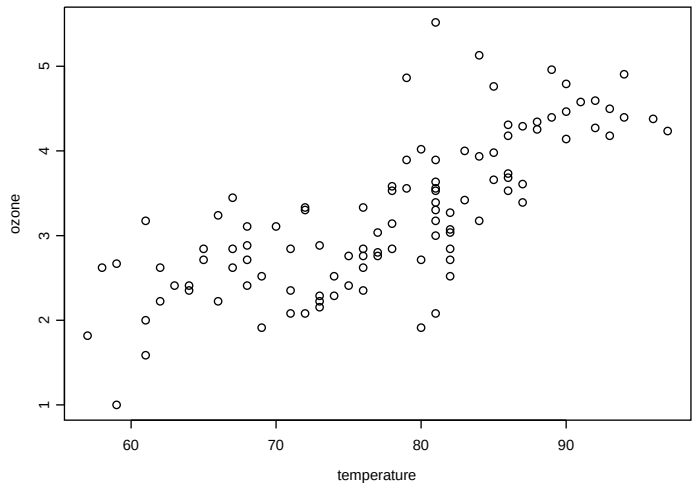


Figure 6.1: Scatterplot of ozone against temperature.

scatterplot, we hypothesise a linear relationship between temperature and ozone concentration. We choose ozone as the response and temperature as the single predictor. The choice of response and predictor variables is driven by the subject matter in which the data arises, rather than statistical considerations.

After fitting the regression model, we obtain the value of the intercept as -2.226 and the coefficient for temperature to be 0.070, thus yielding the model

$$\text{Ozone} = -2.226 + 0.070 \times \text{temperature} + \epsilon$$

An interpretation of the model will be that a one degree increase in temperature will increase the ozone concentration by 0.07. Note however that a common mistake in interpreting the result often arises when the model is extrapolated to values not within the range of the regressors. In this example,

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-2.226	.461		-4.824	.000
	TEMP	7.036E-02	.006	.753	11.951	.000

a. Dependent Variable: OZONE

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	49.462	1	49.462	142.828	.000[a]
	Residual	37.747	109	.346		
	Total	87.209	110			

a. Predictors: (Constant), TEMP

b. Dependent Variable: OZONE

Figure 6.2: Linear regression of ozone concentration against temperature, output from SPSS.

it will be erroneous to state that at zero temperature (temperature = 0), the ozone concentration is -2.226! Care must be taken that linear regressional models are not used for extrapolation beyond the range of the regressors, in this case the range of the temperature is between 57 degrees and 97 degrees Fahrenheit.

6.1.3 Prediction vs. Explanation

There are two main reasons for wishing to construct a linear relationship between the predictors and the response:

1. To explain our data.

2. To predict the values of y at a new value of the x_i's.

Note that these are considerably different objectives. An explanation needs to be intelligible and should generally be as simple as possible (while capturing the essence of the data), whereas a prediction can be complicated as accuracy is the only criterion.

Therefore, the objective should always be clear before a model is constructed, and the choice and number of predictors included in the model should reflect the required objective.

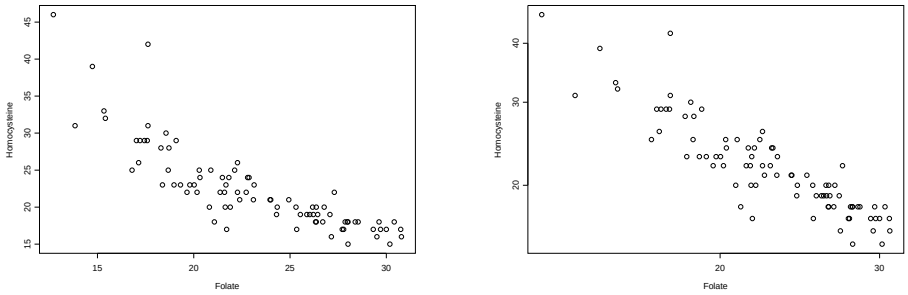
6.2 Transformations

There are many reasons to transform data as part of a regression analysis:

- to achieve linearity.
- to achieve homogeneity of variance (constant variance) about the regression equation.
- to achieve normality or symmetry about the regression equation.

A transformation that achieves one of these goals often ends up achieving all three. This sometimes happens because when a data have a multivariate normal distribution, the linearity of the regression and homogeneity follow automatically. So anything that makes a set of data look multivariate normal in one respect often makes it look multivariate normal in other respects. However, it is not necessary that data follow a multivariate normal distribution for multiple regression to be valid. For standard tests and confidence intervals to be reliable, the response should be close to normally distributed with constant variance about their predicted values. The values of the predictors need not be a random sample from any distribution. They may have any arbitrary joint distribution without affecting the validity of fitting regression models.

Here are some data where the values of both variables were obtained by sampling. They are the homocysteine (HCY) and folate (as measured by CLC) levels for a sample of individuals. Both variables are skewed to the right



and the joint distribution does not have an elliptical shape. If a straight line was fitted to the data with HCY as a response, the variability about the line would be much greater for smaller values of folate and there is a suggestion that the drop in HCY with increasing folate is steeper at lower folate levels.

When logarithmic transformations are applied to both variables, the distributions of the individual variables are less skewed. A straight line seems like a reasonable candidate for describing the association between the variables and the variances appear to be roughly constant about the line.

Often both variables will not need to be transformed and, even when two transformations are necessary, they may not be the same. When only one variable needs to be transformed in a simple linear regression, there is an issue of whether the transformation should be applied to the response or the predictor. Consider a data set showing a quadratic effect between y and x. There are two ways of removing the non-linearity by transforming the data. One is to square the predictor while the other is to take the square root of the response. The general rule is to transform the response variable (y) to achieve homoscedasticity (constant variance), before transforming the predictors to achieve linearity.

6.3 Simple Regression Diagnostics

There are various plots which are obtained whenever a linear model is fitted, and it is essential that these plots are analysed as they reveal the adequacy and the appropriateness of the fitted model. We first define the **residuals** as the difference between what is fitted and what is observed, defined as

$$r_i = y_i - \hat{y}_i$$

where y_i and $\hat{y}_i$ denote the observed and the fitted response for the i^{th} observation respectively.

Different data points contribute differently to the fitted model, with data points at either end of the predictors having a greater influence to the fitted model. Every point influences the fitted line by pulling the fitted line toward itself, and points at either ends have greater tendencies to 'slant' the fitted line. This measure of influence is known statistically as **leverage**. As the calculation of leverage is mathematically tedious, the precise mathematical representation will not be included in this course, however interested readers are pointed to pg. 774 of *Classical Inference and the Linear Model* by Stuart, Ord and Arnold (1999).

The residuals tell us whether a point has been explained well by the model, but if it has not, they do not tell us what the size of the effect on the fitted coefficients of omitting the point might be. A badly fitted point in the middle of the design space will have much less effect on the predictions than one at the edge of the design space (c.f. leverage).

Cook (1977) proposed a measure that combines both the effect of leverage and that of being badly fitted, and the Cook's statistic is proportional to the product of the leverage and the squared residuals. Although several small modifications have been proposed, Cook's statistic is still one of the key measure of the degree of influence of each point. In general, we will take note of leverages with two or three times the value of p/n, where p is the number of fitted parameters and n is the number of data points.

The following four plots of different types of residuals, leverage and Cook's distance are often produced when considering regression diagnostics:

1. Plot residuals against the index of the dataset.
   This will show up observations with large residuals, implying possible outliers. It can also show effects from the time ordering of the measurements (Fig. 6.3).

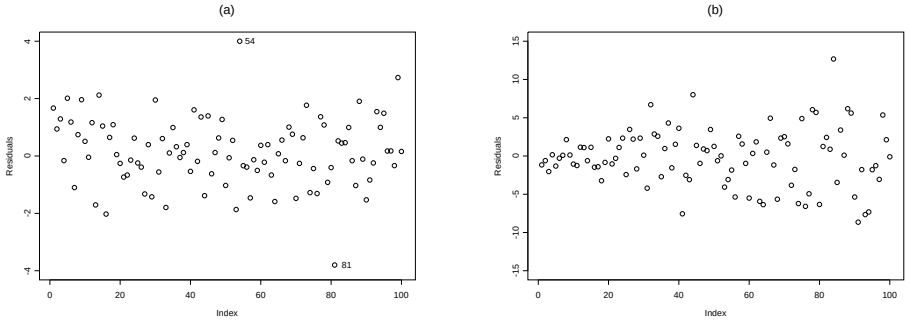


Figure 6.3: (a) Plot of residuals against the index, identifying possible outliers at points 54 and 81. (b) Plot of residuals against the index, identifying possible effects from the ordering of the measurements as the residuals increase as the indices increase.

2. Plot residuals against x (if one-dimensional), or any regressor.
   This can show up patterns in the residuals which indicate non-linearity: for example, that the relationship is with x^2 rather than with x. It can also demonstrate that a potential extra regressor will be useful (Fig. 6.4).

3. Plot residuals against the fitted values of the y.
   This can show up heteroscedasticity, where the variance is not constant over the whole range. This plot is done against $\hat{y}$ rather than y as the residuals are correlated with y but not with $\hat{y}$ (Fig. 6.5).

4. Leverage plots against index will show which points *may* have large influence.
   Such points may or not be outliers and plotting Cook's statistic will draw attention to points which seem to be influential (Fig. 6.6).

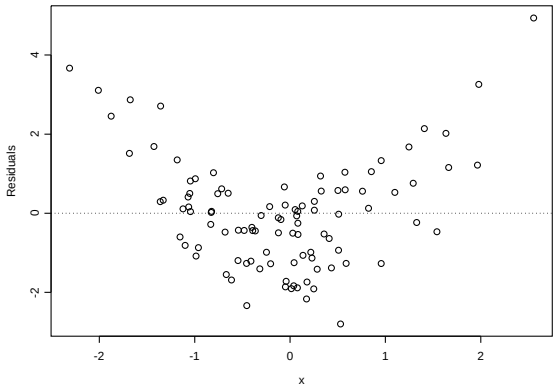


Figure 6.4: Plot of residuals against the regressor x, identifying possible quadratic structure in the residuals, suggesting that the relationship is quadratic rather than linear.

Caution: If there is more than one outlier, these methods may fail to show any of them, as we only consider the effect of omitting one point at a time.

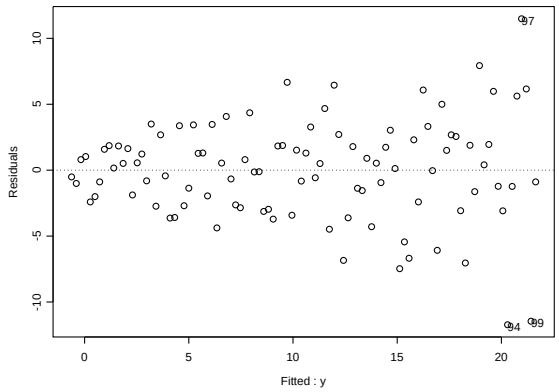


Figure 6.5: Plot of residuals against the fitted y, identifying a trend of increasing variance suggesting heteroscedasticity.

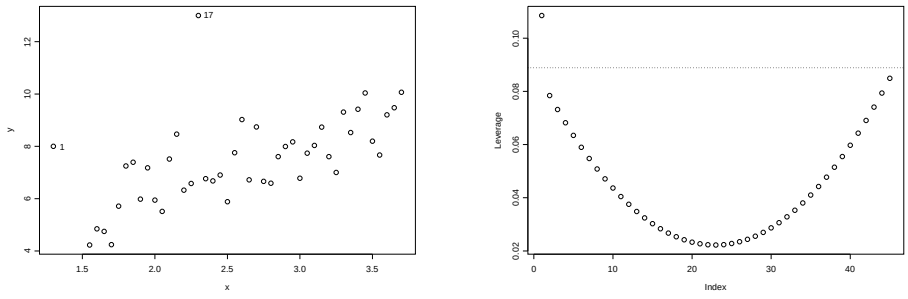


Figure 6.6: Plot of response y against regressor x, with corresponding plot of leverages against index.