LONDON
SCHOOL of
HYGIENE
&TROPICAL
MEDICINE

Foundations of Medical Statistics

2038

Analytical Techniques

Tim Collier, Medical Statistics Department, LSHTM

tim.collier@LSHTM.ac.uk

Table of Contents

Chapter 1: Exploratory and Descriptive Methods

Objectives

By the end of this session, you should:

- Understand the concept and importance of workflow in research
- Understand the importance of exploratory analysis and summary statistics in the workflow of data analysis
- Understand the distinction between continuous, discrete, and categorical data
- Know how to display such data in tables and figures
- Understand commonly used summary measures of location, dispersion, skewness, and kurtosis

1.1 The workflow of research

This session

Much of the focus of the medical statistics programme is on statistical inference and statistical methods. The goal of statistical inference is to learn about a population from a sample drawn appropriately from that population. Usually this is done through one or more summary statistic calculated from the observed sample. We are, therefore, usually interested in summarising and describing observed sample data only because of what we can infer from it about the unobserved population from which it is drawn.

However, to arrive at correct inferences, it is important to understand the workflow of research, and the important role that exploratory analysis and summary statistics play in that process. Here we briefly consider the workflow of research as a whole and the workflow of data analysis (see figure 1.1).

1.1.1 Study design and conduct

Correct inferences are built upon good study design and conduct. A poorly designed or executed study will not produce reliable results. It is important that statistical input is given early in the study design process, influencing formulation of the study hypotheses, helping determine the required sample size, shaping the definition of outcomes and decisions about what data should be collected to enable the study questions to be answered. Electronic case report forms and data dictionaries are useful tools to help the data collection process, as well as the later data preparation and statistical analysis.

Although a statistician will not usually be involved in the day-to-day conduct of the study, they do have a role to play, along with data managers, in ensuring the accuracy and completeness of the data being collected. Checks should be built into data collection systems e.g., plausible ranges, completeness of required fields, consistency checks, etc. Interim data summaries can be helpful in identifying and resolving problems early. It may be too late to rectify problems with data if they are only identified at the end of the study.

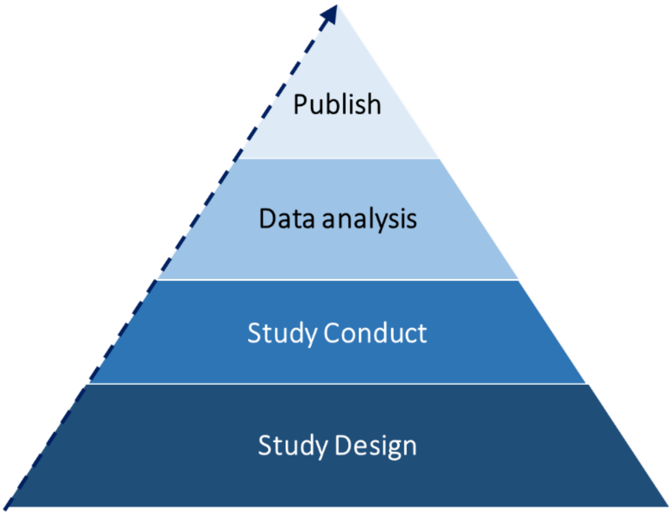


Figure 1.1: The workflow of research

1.1.2 Workflow of data analysis

From a research perspective the most important output from the study design and conduct is the study data. Generally, it is from the point at which the study database is complete that the statistician will be most heavily involved. Within this data analysis stage of the research process there is an important workflow that should be understood and followed. On receiving data, it is tempting to jump straight into statistical modelling and hypothesis testing. However, there are essential steps that should be followed to ensure the accuracy and reproducibility of any results.

The first step is that of ***preparing*** the data for analysis. This involves data cleaning and data processing. The goal of this step is to produce the final analysis dataset i.e., a clean dataset containing all the observations and variables required for the statistical analysis.

Data cleaning is the critically important process of detecting and rectifying errors in a dataset. Data quality should not be taken for granted. In fact, I recommend that you should always worry about data quality. No matter how sophisticated the statistical analysis might be, if the data are flawed the results will be wrong – garbage in, garbage out. It is therefore critical that the process of data cleaning is taken seriously and that it should not be rushed. Summary statistics and data visualisation (tables and graphs) as well as simply looking at the raw data, are the key tools for detecting errors. My recommendation is to look at the data before doing anything else. There are data problems that can only be detected by actually looking at the data.

Data errors generally arise during the data entry process (both manual or machine). Numbers can be mistyped, decimal points missed, data entered in the wrong field, a number mistakenly read from a form, dates entered in wrong format, and many more. Such data entry errors will often (though not always) result in implausibly small or large values. These are generally fairly easy to detect using summary statistics e.g., minimum and maximum values or standard deviations, and data visualisation e.g., histograms. However, very small or very large values are not necessarily errors. It can be helpful to get advice from a subject knowledge expert on ranges of plausible values for a variable. Mistakes in data entry can result in plausible values. Such errors are more difficult to detect since they do not obviously stand out.

Errors such as duplicated values or inconsistent values require more careful inspection. Most statistical packages have tools that can check for duplicated values across one or more variable. Cross tabulations or scatter plots can help identify inconsistencies. Dates can easily be recorded in the wrong format e.g., 12-09-2022 could be 12th September or 9th December depending on the format. Checking that dates lie within the time-period of the study and that later visits follow earlier visits can help identify issues like this.

Laboratory data are often collected and recorded in different units. This isn’t erroneous data, but such instances must be identified, and the results translated into a consistent unit.

Having identified implausible values it is important that such values are corrected if possible, or removed i.e., changed to missing values. Observations that are outliers but not implausible may be left in the dataset but need to be identified and dealt with carefully e.g., the sensitivity of results to the inclusion of such observations might be assessed.

It is rare that a study or clinical trial will collect data in a single data file. Often data collected at different times (e.g., baseline, follow-up visits, end of study) or from different centres, or related to different aspects of the study (medical history, demographics, vital signs, laboratory data, adverse events, study outcomes) will be collected in separate datasets. Data processing is the process of importing and combining these “raw” datasets to create a dataset containing all the variables required. This will also involve the creation of derived variables e.g., body mass index may be derived from weight and height, age at entry may be derived from dates of birth and entry to the study. An important point to make here is that a copy of the original datasets, should always be kept unchanged.

When working on a randomised controlled trial all the data cleaning and preparation should take place blind to treatment allocation. The treatment codes which reveal the treatment allocation for each participant should be kept separate from the database until the data preparation is complete – this is referred to as database lock.

It is important that data cleaning and processing is well documented, recording what errors were found, how they were resolved, how derived variables were created etc. This is important in enabling datasets and results to be replicated.

The second key step in data analysis is that of ***statistical analysis and reporting*** – which could be considered as two separate steps. We won’t spend much time on this here since this is the covered in detail throughout the masters programme. We just emphasise that the statistical analysis step comes after careful data preparation. Additional data errors may be discovered during the analysis, which may require returning to the data preparation step. It is also worth emphasising that statistical analysis should always begin with simple descriptive and summary statistics before moving to statistical modelling and inference.

Effective reporting of results in tables, figures, text and in verbal presentations are key skills for a statistician to develop. We recommend looking at how results are presented in tables, figures and described in text in good medical journals e.g., NEJM, the Lancet. There is a session devoted to presenting results in term 1 and you will develop these skills in module assessments and particularly in the research project.

1.1.3 Documentation

Careful and comprehensive documentation of the data preparation and statistical analysis steps is critical to enabling results to be reproduced. This will include documenting data sources (e.g., which version of the database was used), careful definition of derived variables (e.g., there may be several different definitions of a variable in the literature – which was used?), were any errors detected and if so, how were they dealt with, and any selection of the analysis population (e.g., were any patients excluded and if yes why?). The documentation should be sufficient to enable someone else to work their way from the raw study data to the final analysis dataset. Methods of statistical analysis and statistical software versions used should also be documented. Some of the documentation may be in the form of annotations and comments added to program files e.g., a comment may be recorded in a data processing program describing how a derived variable was defined, or how cut-points were decided, or why a correction was made.

The final (and often overlooked) step in the workflow of data analysis is that of ***archiving***. This will include archiving of data (the raw data and derived datasets), statistical programs (data processing and statistical analysis) and other important study documents that would enable the statistical analysis to be replicated.

1.1.4 Developing a workflow system

A good workflow system does not just happen. It needs to be carefully planned and executed. There isn’t a single perfect workflow system – the important thing is to appreciate the importance of workflow and to develop a good system that works for you.

A good workflow system should enable you to produce accurate and reproducible result. It should also enable you to work more efficiently (e.g., avoid wasting time trying to find files, data, variables, results), and make it easier to return to a project after a few months away or to pass on your work to another researcher. It should also make your work more portable i.e., enable you to work from different locations without difficulty. In short it should make your research life easier. Here are a few ideas to help you think about developing your own system.

Develop a standard system for structure of electronic folders. Figure 1.2 shows a schematic of a structure that I use for almost all my projects. This can of course be varied as required but having a consistent and logical structure is very helpful – I know where a particular document or dataset will be stored, which is not true for many people who have come to me for help with statistical analysis. I have an empty template that I just copy-and-paste when I start a new project – it's generally just the main project folder that needs a new name.

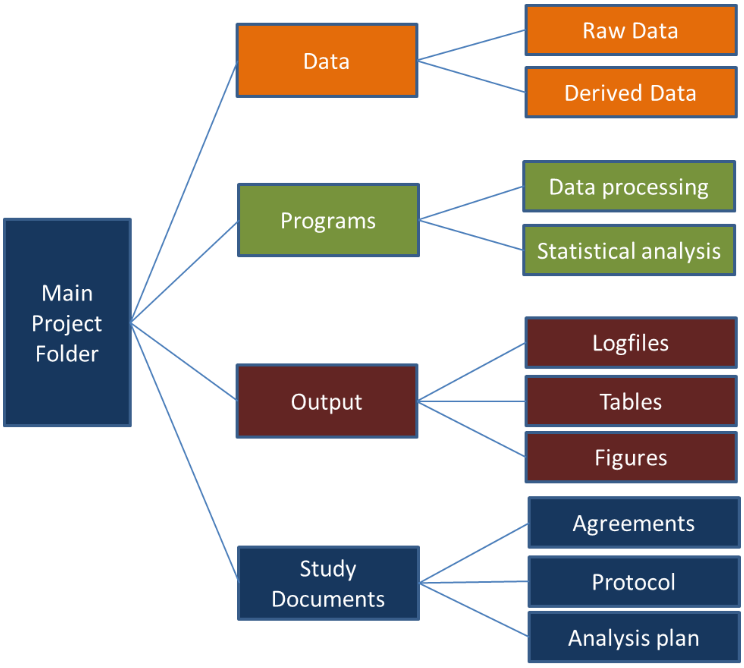


Figure 1.2 Schematic of a possible folder structure for a project

Develop a consistent naming convention for data files and program files, which will enable you to identify a file without having to open it, e.g., bl_demog would be a good name for a dataset containing baseline demographic data. I always keep data management and statistical analysis programming in separate files. All my data management files begin dm#_ where # will be a number indicating the order in which the program files should be executed. The underscore will be followed by an abbreviated term to give information on what that program does e.g., dm0_import.

Use a sensible and consistent naming convention for variables. I aim to keep variable names as short as possible (e.g., sbp rather than systolic_blood_pressure) since the variable names will need to be typed in later programs. This can cause a problem with clarity, e.g., suppose I have a variable named wc; that's nice and short, but what does wc stand for? This problem can be overcome by attaching variable labels which provide clarification on what the variable is and give details of units e.g., "Waist circumference-cm". I always use lower case characters for variable names – that means I never have to think about case (so saves time) and, since many statistical packages are case sensitive, it also avoids any confusion from having two variables with same name but different case e.g., age and Age. Gains can be made from being consistent with naming of similar variables. For example, using _dt as a suffix for any date variables or using the same prefix for the same variable measured at different time e.g., sbp1, sbp3, sbp6, sbp12 for systolic blood pressure measured at 1, 3, 6 and 12 months respectively.

I recommend you avoid building very long (i.e., many lines) program files for either data processing or statistical analysis – long program files become very difficult to navigate and edit. My preference is to split data processing tasks into a number of program files containing related tasks. These can be called one after the other as needed. For example, the first data processing program for most of my projects will be one that imports the raw data and saves it in Stata format (the software which I invariably use). Then I may have separate program files for processing baseline data, laboratory data, etc. The same applies to statistical analysis programs. As part of the documentation process it is helpful to have a master file detailing what each file does, along with what are the inputs and outputs. This does take time to produce but good documentation can save time in the long run.

Another very important feature of a good workflow system is having your files saved in a secure location that is backed-up regularly. Losing files that have involved many hours of work is one of the most painful experiences in research.

There are many other things to consider, and as mentioned above, there is not a single right or perfect system. But I think even a mediocre system is better than working ad hoc.

1.2 Types of data

The guiding principles of statistics are mostly the same whatever the nature of the data we are analysing. However, their precise implementation i.e., the statistical methods we use, depends crucially on the form which the observations take.

As we will see later in this chapter, different graphical and numerical summaries will be appropriate for different types of data. We will also see in later sessions and throughout the MSc program that the statistical method of analysis often crucially depends on the type of *outcome* variable with which we are dealing.

There are various ways of classifying data, e.g., quantitative versus qualitative, numeric versus categorical (see the recommended textbooks). Here we distinguish the following three main types of observation: *continuous*, *discrete* and *categorical*.

1.2.1 Continuous data

Continuous data consist of observations which usually arise from a measurement carried out with an instrument (hence sometimes referred to as metric data). Such observations are intrinsically numeric.

Examples of continuous data include height, weight, distance walked in a six-minute walk test, haemoglobin concentration, blood pressure. In theory the observations can (at least approximately) take any value in a range, though in reality all such measurements are limited to the precision of the instrument being used e.g., a tape measure, weighing scales. However, acting as though these are measured with infinite accuracy usually provides a very good (and convenient) approximation in practice.

Example 1: Body mass index (kg/m^2) in a sample of patients undergoing major abdominal surgery (with and without diabetes).

Without diabetes					**With diabetes**				
23.51	35.00	20.80	35.00	20.96	22.59	57.92	37.38	26.57	28.23
39.14	26.78	38.48	24.11	25.53	33.65	27.77	47.75	31.22	28.65
27.24	25.73	26.62	25.21	26.56	28.73	32.46	28.73	20.82	23.07
34.58	30.69	22.48	24.75	21.78	39.71	22.31	33.46	23.03	32.47
24.01	29.37	27.04	22.31	24.21	30.51	37.72	22.86	41.50	
34.78	24.30	27.27	45.54		32.41	50.42	22.75	27.44	
28.72	26.35	31.98	21.22		29.38	36.81	36.68	25.07	
21.31	29.69	27.64	42.91		40.63	48.11	30.96	22.15	
24.69	29.52	22.77	22.92		31.63	30.85	25.24	32.88	
29.21	25.25	34.38	22.10		23.33	36.68	23.24	18.99	

1.2.2 Discrete data

These are observations that usually arise from a counting process and, as with continuous data, are intrinsically numeric. Each observation therefore belongs to one of a set of positive integer values. Examples include number of infections in a hospital ward in a year, number of siblings in a family, days from surgery to discharge, number of blood transfusions.

Example 2: Number of siblings in 25 families.

2, 3, 3, 2, 0, 3, 2, 3, 2, 1, 2, 3, 0, 5, 1, 4, 2, 4, 1, 3, 2, 2, 3, 4, 4.

Although discrete data can only take certain numeric values i.e., positive integer values, in practice it may often be appropriate to treat such observations as though they were continuous.

1.2.3 Categorical data

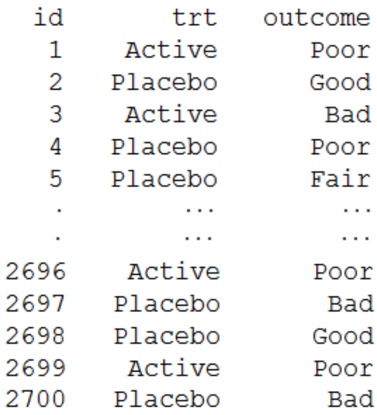
In such data each observation belongs to one of a specified set of classes. They differ from discrete data in that the observations are not intrinsically numeric. Examples include questionnaire answers (agree, disagree), crude colour assessment (red, blue, green); occupation (professional, manual worker *etc.*).

Although such observations are not intrinsically numeric, numeric values may be assigned to each class for the purpose of data collection, storage and analysis.

There are several subtypes of categorical data:

- **binary**: an important and commonly encountered type of categorical variable which has only two classes, e.g., dead/alive, pregnant/not pregnant. Often indicates presence or absence of some condition.
- **nominal**: the categories have no sense of ordering, e.g., ethnicity, blood type.
- **ordinal**: the categories have a real order, e.g., some questionnaire answers (agree/indifferent/disagree) or assessments of outcome (good/mild/moderate/poor).

Example 3: Outcome at 1-year in 2,700 patients following surgery in a randomised trial



```
  id       trt   outcome
   1    Active      Poor
   2   Placebo      Good
   3    Active       Bad
   4   Placebo      Poor
   5   Placebo      Fair
   .       ...       ...
   .       ...       ...
2696    Active      Poor
2697   Placebo       Bad
2698   Placebo      Good
2699    Active      Poor
2700   Placebo       Bad
```

It is quite common practice to “categorize” continuous data i.e., to group each observation into one of a specific set of classes, where the classes are defined using some cut-off values. For example, body mass index (kg/m^2) which is a continuous variable, is often grouped into underweight (<18.5), normal weight (18.5-24.9), overweight (25-29.9) and obese (30+). Haemoglobin (g/dl), another continuous variable, is often used to diagnose anaemia (yes/no) with recognised cut-off points <120 g/l in females and <130 g/l in males.

There are pros and cons for using this approach. The process always results in loss of information e.g., in the BMI example above a person with a BMI of 25.0 kg/m^2 is treated the same as a person with a BMI of 29.9 kg/m^2. There is also a danger of this process being abused with cut-off values being specially selected to accentuate or attenuate some association. This is a particular danger when there are no recognised or established cut-off values. It is best practice to pre-specify the cut-off values in a statistical analysis plan i.e., before seeing the data, or to use cut-offs such as quartiles.

1.3 Displaying data

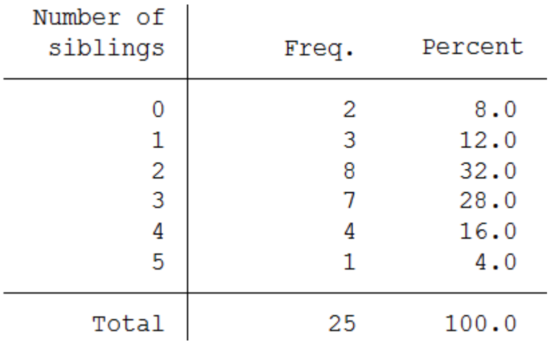
It is generally difficult to grasp the patterns in, or distribution of, a set of observations simply by looking at the raw numbers. Tabular and graphical displays provide one of the most important ways of exploring and communicating key features of data.

As explained above, the type of the variable we are dealing with, generally determines the appropriate graphical or tabular display. Similar methods can be used for displaying discrete and categorical data, although if a discrete variable takes many values, it may appropriately be treated as a continuous variable.

1.3.1 Displaying discrete or categorical data

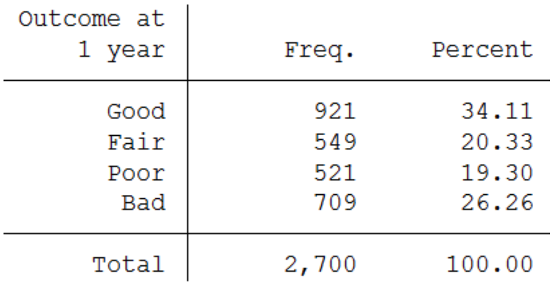
A simple summary of the number and percentage of observations in each category can be displayed in a frequency table. For example, the data from examples 2 and 3 above, could be displayed in frequency tables as follows:

Table 1.1: Number of siblings in 25 families



```
  Number of |
   siblings |      Freq.     Percent
------------+-----------------------
          0 |          2        8.0
          1 |          3       12.0
          2 |          8       32.0
          3 |          7       28.0
          4 |          4       16.0
          5 |          1        4.0
------------+-----------------------
      Total |         25      100.0
```

Table 1.2: Outcome at 1 year following surgery



```
 Outcome at |
     1 year |      Freq.     Percent
------------+-----------------------
       Good |        921       34.11
       Fair |        549       20.33
       Poor |        521       19.30
        Bad |        709       26.26
------------+-----------------------
      Total |      2,700      100.00
```

Table 1.1 is an example of a discrete variable i.e., the observations are intrinsically numeric but can only take integer values. The table is ordered according to these numeric values. If the variable takes many values of then it may be necessary to combine some values, or to have a greater than top category e.g., 4+.

Table 1.2 is an example of an ordered categorical variable. This table is ordered according to underlying numeric values that have been assigned to the outcomes “Good”, “Fair”, “Poor” and “Bad”. It would not be appropriate in this instance to order the table according to alphabetical order. The underlying values are in some respects arbitrary but need to reflect the ordering of the categories i.e., they could be 0 “Good”, 1 “Fair”, 2 “Poor” and 3 “Bad” or 1 “Good”, 2 “Fair”, 3 “Poor” and 4 “Bad”.

The distribution of discrete or categorical variables can also be helpfully displayed graphically in a bar chart. The bar chart can be used to show the frequency or percentage. Again, using the data from examples 2 and 3 above:

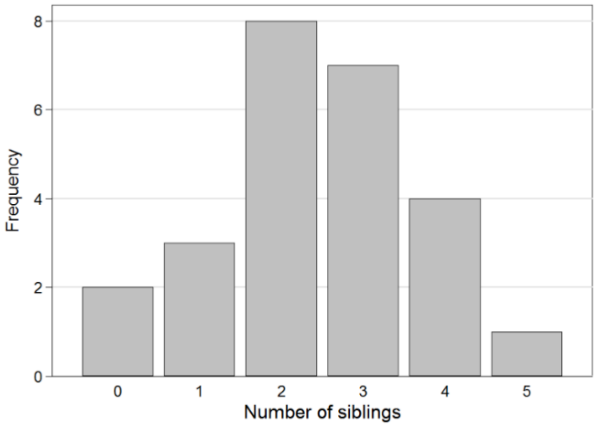


Figure 1.3: Bar chart displaying the number of siblings in 25 families.

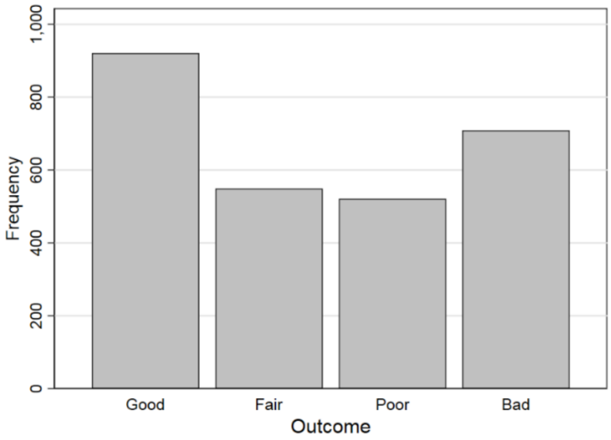
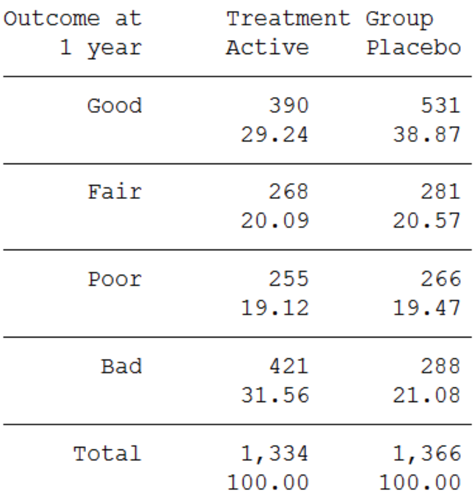


Figure 1.4: Bar chart displaying distribution of outcome at 1-year following surgery

As with the frequency table 2 above the ordering of the outcome at 1 year is determined by underlying assigned numeric values. In both examples the y-axis displays the frequency – it is also possible to display percentages. When comparing distributions across groups e.g., by treatment group or by some outcome, percentages may be preferred particularly if the groups are of very different sizes.

Frequency tables and bar charts are often used to compare the distribution of a categorical or discrete variable across levels of another variable.

Table 1.3: Outcome at 1-year by treatment group



```
Outcome at      Treatment Group
    1 year      Active    Placebo
---------------------------------
      Good         390        531
                 29.24      38.87
---------------------------------
      Fair         268        281
                 20.09      20.57
---------------------------------
      Poor         255        266
                 19.12      19.47
---------------------------------
       Bad         421        288
                 31.56      21.08
---------------------------------
     Total       1,334      1,366
                100.00     100.00
```

The table above is copied straight from the output of a statistical package. Generally, some formatting of the output is required for tables for reports and publications.

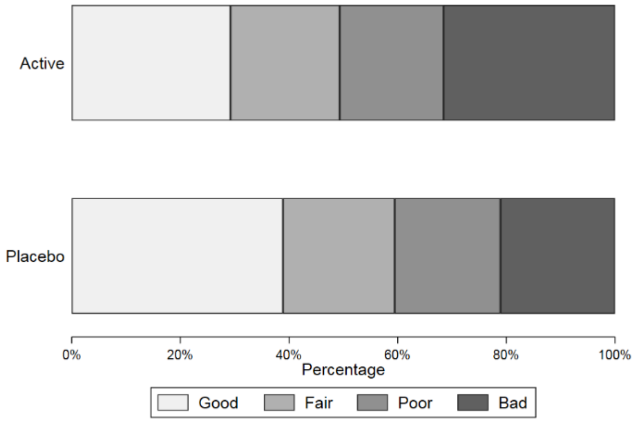


Figure 1.5: Stacked bar chart showing distribution of outcome at 1 year by treatment group

A stacked bar chart is one good way of graphically comparing the distribution of a categorical outcome between 2 or more groups. In figure 1.5 (which shows the same data as presented in table 1.3 above) we can see clearly that there is a shift towards better outcomes in the placebo group. The percentage of patients with fair and poor outcomes is similar in the two groups, but the percentage with good and bad outcomes is lower and higher respectively in the active group.

1.3.2 Continuous Data

Generally, it does not make sense to produce a frequency tabulation of the raw values of a continuous variable since the observations can take many values. Indeed, for a variable such as body mass index the observed values might be unique for all individuals in a study. It is quite common however to present frequency tabulations of categorized values of a continuous variable. For example, using the body mass index data presented above:

Table 1.4: Categorised BMI in a sample of patients undergoing major abdominal surgery by presence or absence of diabetes

BMI	Diabetes		
(kg/m^2)	No	Yes	Total
<25	17	11	28
25-29.9	17	10	27
30-39.9	9	17	26
40+	2	6	8
Total	45	44	89

The most commonly used plots for displaying continuous variables are the histogram and box plot. The defining feature of the histogram is area. The area (not the height) represents the count in the interval defined by the class boundaries. Intervals are usually, but not necessarily, equal in width.

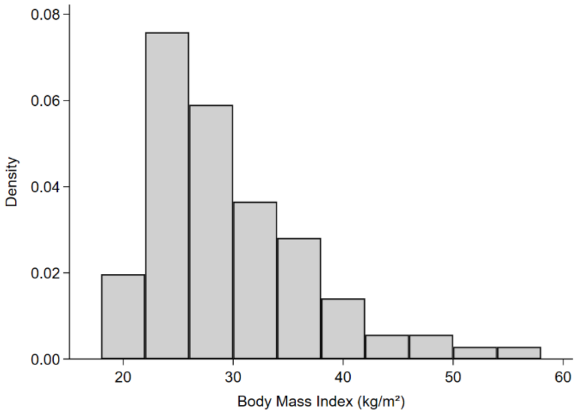


Figure 1.6: Histogram of the body mass index data

Another view of the shape of a distribution of numbers is provided by the box-and-whisker plot, usually shortened to box plot. The aim of the box plot is to display the median, quartiles, range and any outliers in the data. A rectangular box, on a suitable scale that covers the whole of the dataset, marks the interquartile range (*i.e.,* containing the middle 50% of the data), with the median marked appropriately in the box; the 25th and 75th percentiles indicated by the lower and upper sides of the (vertical) box, as shown below, are sometimes called lower and upper *hinges*.

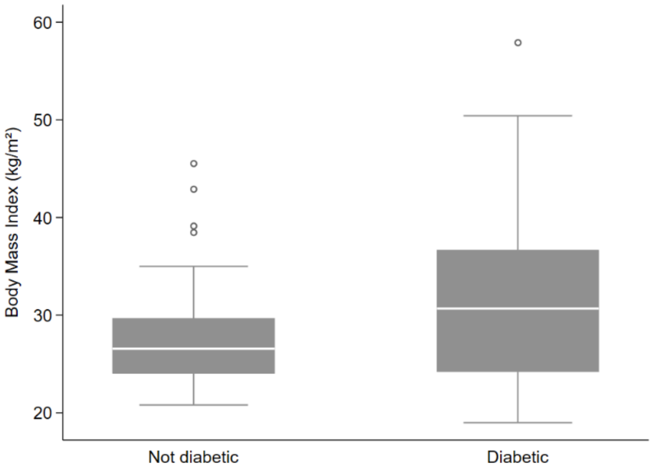


Figure 1.7: Box plot of the body mass index data

In boxplots created using Stata, the whiskers reach out above and below the box hinges to: i) the highest value not greater than the 75th percentile plus 1.5 times the interquartile range and ii) the lowest value not less than the 25th percentile minus 1.5 times the interquartile range.

Any observations not covered by the box and whiskers are then marked as separate points. Such observations might (depending on the circumstances) be regarded as atypical observations (outliers).

1.4 Summary statistics

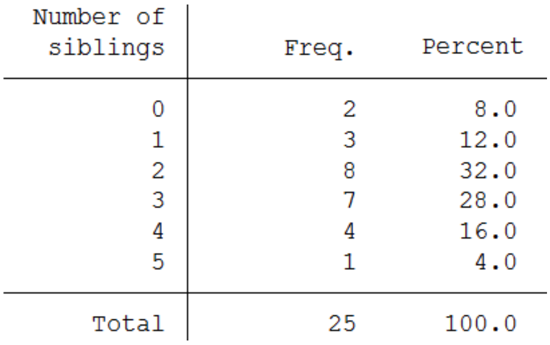
For much the same reason that we need plots to help picture what is contained in a collection of numbers we need quantities that summarise the main features of such collections: these can measure key aspects of the form of a distribution of data. Such quantities play a central role in all aspects of statistical work. Here we consider summary measures of location, dispersion, skewness, and kurtosis.

1.4.1 Measures of location

The location of a distribution refers in some way to its centre. Centre can mean many things; hence we have several different measures of location. Here we briefly consider the mode, median and mean.

1.4.1.1 The mode

The mode is the most commonly occurring value. Usually this is only relevant for categorical data. In the example above of the number of siblings in a sample of 25 families the mode is 2.



```
  Number of |
   siblings |      Freq.     Percent
------------+-----------------------
          0 |          2        8.0
          1 |          3       12.0
          2 |          8       32.0
          3 |          7       28.0
          4 |          4       16.0
          5 |          1        4.0
------------+-----------------------
      Total |         25      100.0
```

1.4.1.2 The median

The median is the halfway point of the distribution. 50% lies above and 50% lies below this point. Consider the following simple sample:

10, 16, 12, 5, 22, 14, 19

The median is easy to find if the sample has an odd number of members. It is the middle member of the *ranked* sample:

5, 10, 12, 14, 16, 19, 22

Here the median = 14.

With an even number of members, the mean of the middle two observations is taken. Suppose an additional observation of 38 is added to the previous example:

5, 10, 12, 14, 16, 19, 22, 38

In this example the median = (14 + 16)/2 = 15.

1.4.1.3 The arithmetic mean

The arithmetic mean is, in colloquial terms, what is usually meant by the average. Formally, for any set of *n* numbers $X_1, \ldots, X_n$ the arithmetic mean is expressed as:

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

For the example with 7 data points seen above the arithmetic mean is:

$$\frac{10 + 16 + 12 + 5 + 22 + 14 + 19}{7} = \frac{98}{7} = 14.0$$

The arithmetic mean is much more *sensitive* to “outlying” values than the median. For example, adding an observation of 38 to the data produces a new mean of:

$$\frac{10 + 16 + 12 + 5 + 22 + 14 + 19 + 38}{8} = 17.0$$

a change of 3. The corresponding medians are 14 and 15 respectively.

The mean has convenient mathematical properties that give it a central role in much statistical work. This does not imply however that it is always the appropriate measure of location to use.

1.4.1.4 Other means

There are other means, which can be interpreted as arithmetic means on different scales.

The *geometric* mean is the exponential (e^z) of the arithmetic mean of the log transformed values:

$$\sqrt[n]{\prod_{i=1}^{n} X_i} = exp\left[\frac{1}{n}\sum_{i=1}^{n} log_e(X_i)\right]$$

This is often used with positively skewed biological data.

The *harmonic* mean is the inverse (or reciprocal) of the arithmetic mean of the inverses:

$$\frac{1}{\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{X_i}\right)}$$

The harmonic mean is most often used to calculate the average of rates or ratios.

1.4.2 Measures of dispersion

Dispersion refers to the "spread" or variability of a set of numbers. Variables with the same central location may differ markedly in terms of dispersion. For example, the figure below shows two variables both of which have a mean of 50, but which are very different in terms of variability about the centre.

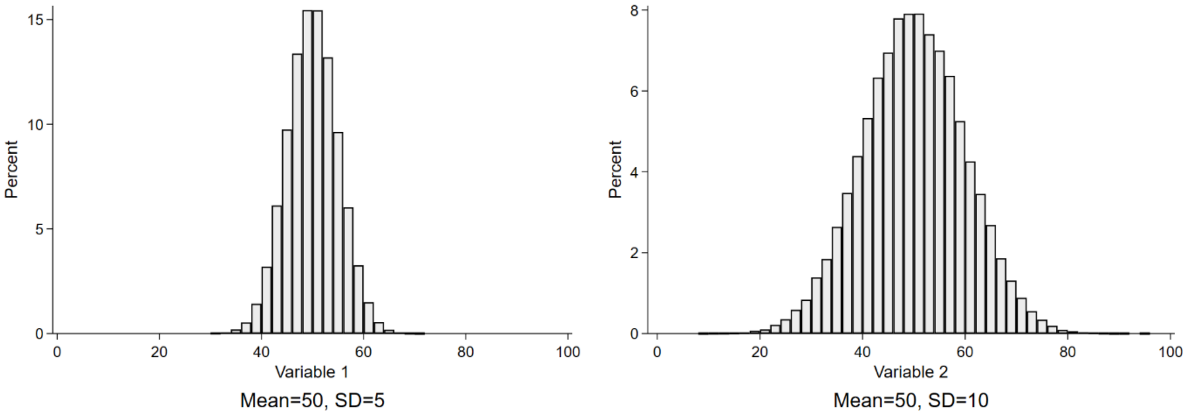


Figure 1.8: Distributions with similar central location but different dispersion

Here we briefly consider the range, interquartile range, variance, and standard deviation.

1.4.2.1 The range

The range is the difference between the largest and the smallest observations. For the original set of seven data points shown above this would be:

$$22 - 5 = 17$$

And for the extended set of eight observations:

$$38 - 5 = 33$$

The range has the disadvantage that it is highly dependent on sample size and outliers. In tables and reports the range is often reported as two values i.e., the minimum and maximum value [5, 22], rather than the difference between the two values.

1.4.2.2 The interquartile range

The inter-quartile range (IQR) is the range of values that contains the middle 50% of the sample. It can be expressed as:

Upper Quartile − Lower Quartile.

The quartiles are calculated as follows. For an **odd sized sample**, remove the median, and then calculate the "medians" of the remaining upper and lower halves. For example:

$$5, \underline{10}, 12, 14, 16, \underline{19}, 22.$$

The lower quartile (Q_L) is 10 and the upper quartile (Q_U) is 19 and the IQR is 9. For an **even sized sample** calculate the "medians" of each half:

$$5,\ \underline{10,\ 12},\ 14,\ 16,\ \underline{19,\ 22,}\ 38.$$

$$Q_L = (10+12)/2 = 11 \quad Q_U = (19+22)/2 = 20.5$$

Here the IQR is 9.5.

In fact, these are slightly simplified versions of the exact calculations. In practice differences between the methods are very small but may be seen when comparing hand calculations with those provided by a computer package.

As with the range, the IQR is often reported in descriptive tables as two numbers e.g., [11, 20.5] rather than the difference between them.

1.4.2.3 The variance and standard deviation

Another way of viewing spread is to consider the distance (or deviation) between each value and the arithmetic mean:

$$D_i = X_i - \bar{X}$$

A large spread implies that some of these (at least) will be large. A summary measure of these distances is needed. They cannot be averaged because by definition:

$$\frac{1}{n}\sum_{i=1}^{n} D_i = 0$$

The average absolute value ($|D_i|$) is sometimes used but is difficult to work with mathematically. Instead, the average squared deviation is conventionally used:

$$\frac{1}{n}\sum_{i=1}^{n} D_i^2$$

This is called the **variance** and plays a central role in many statistical procedures. It is not on the same scale as the original sample data however, so we also often use its square root, the **standard deviation (SD)** when presenting it in combination with results such as the mean:

$$\mathrm{SD} = \sqrt{\tfrac{1}{n}\textstyle\sum_{i=1}^{n} D_i^2}$$

Note, in passing, that these are the appropriate formula for the mean and standard deviation of an observed set of data. As will be seen in Inference a slightly modified formula (replacing *n* by *n*-1) is used when we wish to **estimate** a population variance from a sample.

There are several ways of writing the variance formula. Replacing D_i by $X_i - \bar{X}$gives the well-known formula:

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

which can also be written as:

$$\frac{1}{n}\left[\left(\sum_{i=1}^{n} X_i^2\right) - n\bar{X}^2\right]$$

Like the mean, the variance (and SD) is sensitive to the effects of outliers. For the sample of 7 observations the variance is equal to:

$$\frac{10^2 + \ \ldots \ + 19^2 - (7 \times 14^2)}{7} = \frac{194}{7} = 27.71$$

Hence the SD is $\sqrt{27.71}$ = 5.26. With the extra 38, the variance becomes 87.25 with corresponding SD 9.34: a large difference. Compare this with the comparative stability of the inter-quartile range.

1.4.3 Skewness

A distribution is symmetric if the left and right halves (with respect to the median or mean) have exactly the same shape. The normal distribution is symmetric. An asymmetric distribution is said to be skewed to the right (or positively skewed) if the right tail is the more "elongated", and skewed to the left (or negatively skewed) if the opposite is true.

Skewness is based on the ratio of the deviations about the mean cubed, to the variance, scaled to be unit free:

$$\frac{\frac{1}{n}\sum_{i=1}^{n} D_i^3}{\left(\frac{1}{n}\sum_{i=1}^{n} D_i^2\right)^{3/2}} .$$

Note that D_i^3 has the same sign as D_i so negative values reduce the skewness and vice versa. In a symmetric distribution this measure takes the value zero, it is positive/negative if the distribution is positively/negatively skewed.

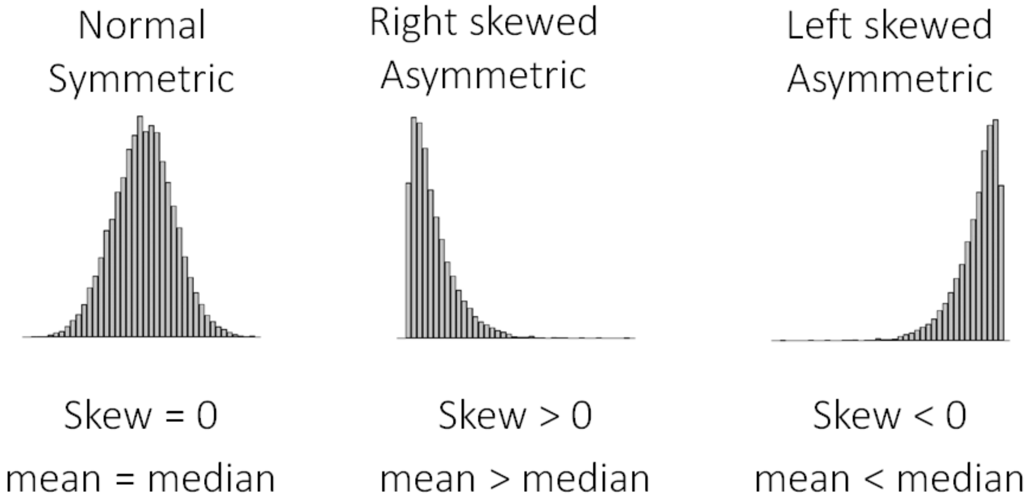


Figure 1.9: Relationship between skew and measures of location

For a symmetric distribution the mean and median coincide. In general, the choice of which of these to use as a measure of location will depend on the aims of the analyst and other features of the data. The mean has a central role in much statistical work because of its convenient mathematical properties; the median is by comparison rather awkward to work with theoretically. For this reason, it is probably true that the mean is used rather more often than it should be in practice.

In a skewed distribution the mean is pulled away from the median in the direction of the skewness. By definition the mean keeps its property of being the arithmetic average, but if a measure of location is wanted that lies in what we might call intuitively the “centre” of the distribution then the median is more appropriate. If (positive) data are skewed to the right then the geometric mean may well lie closer to the median than the mean, indicating that the log transformed values probably have a more symmetric distribution. Such transformations of data have many uses; we return to this in Analytical Techniques 6.

The skewness of the distribution also affects how we should approach the dispersion. The standard deviation treats (squared) deviations in both directions the same way. But in a skewed distribution, deviations to the right and left have different patterns. So, the standard deviation should be used with caution when data are very skewed, and the type of information provided by quartiles may be more valuable in this situation.

1.4.4 Kurtosis

The last feature of a distribution that we will consider is kurtosis. Kurtosis measures the “heaviness” or “lightness” of the tails of a distribution. In a normal distribution we expect 5% of observations to lie in the tails – approximately 2.5% of observations will lie below $\mu - 2\sigma$ and approximately 2.5% of observations to lie above $\mu + 2\sigma$. The kurtosis measures departures from this i.e., either fewer or more observations than expected.

Kurtosis is based on the fourth moment of the mean and is defined as:

$$\frac{\frac{1}{n}\sum_{i=1}^{n} D_i^4}{\left(\frac{1}{n}\sum_{i=1}^{n} D_i^2\right)^2}$$

For a normally distributed variable the kurtosis takes the value three. Where there are fewer observations than expected in the tails of a distribution (i.e., “light tailed”) the kurtosis will be less than 3. Where there are more observations than expected in the tails of a distribution (i.e., “heavy tailed”) the kurtosis will be greater than 3 (see figure 1.10).

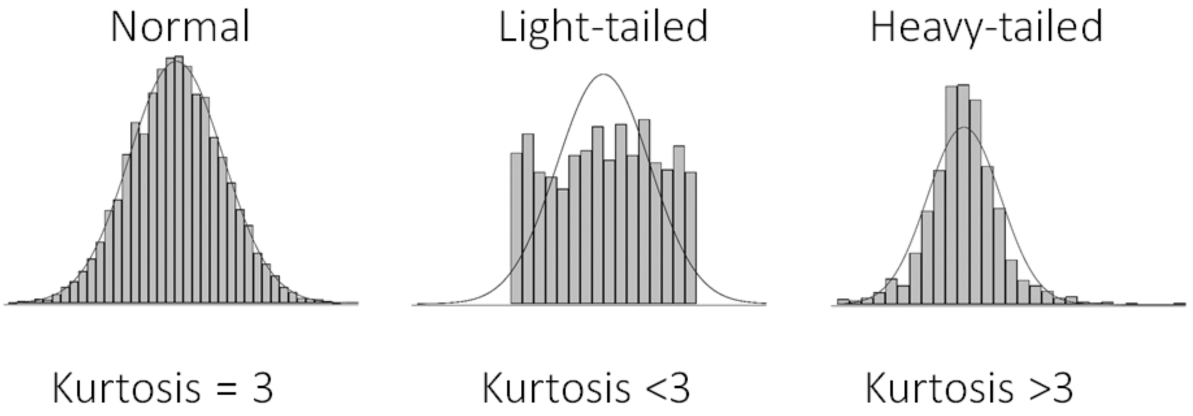


Figure 1.10: Examples of normal, light-tailed, and heavy-tailed distributions

A good example of a heavy tailed distribution is the *t* distribution, which will make several appearances throughout the course. The *t* is a family of distributions, each member indexed by an integer (the so-called degrees of freedom), from 1 upwards. As the degrees of freedom tend to infinity the distribution tends to the standard normal. The *t* distribution has heavier tails than the normal. This is illustrated in figure 1.11, in which the *t* distributions with 5 and 10 degrees of freedom are compared with a normal distribution with mean 40 and SD 10.

It is worth noting here that a very high kurtosis can be indicative of outlying values.

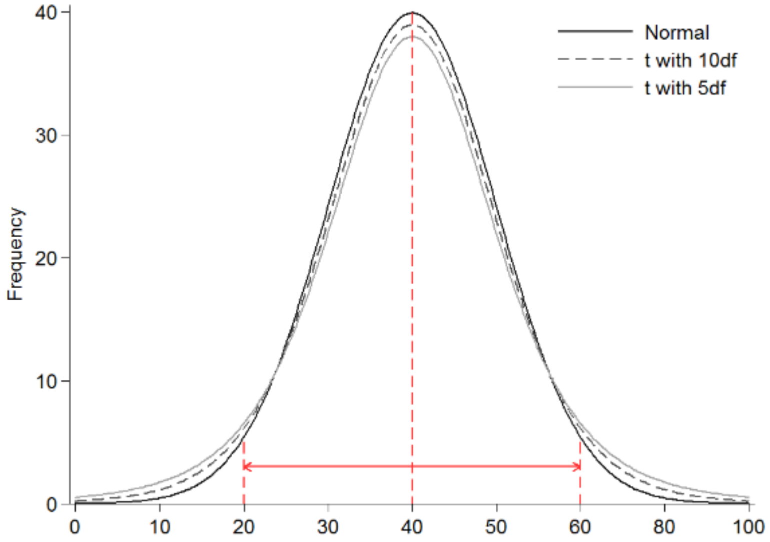


Figure 1.11: *t* distributions with 5 and 10 degrees of freedom compared with a standard normal distribution

Chapter 2: Confidence Intervals

Objectives

By the end of this session students will be able to:

- Understand the general definition of a confidence interval
- Understand the motivation for confidence intervals
- Understand the principles underlying the construction of confidence intervals
- Know how to interpret a confidence interval
- Construct confidence intervals for parameters in a number of simple settings

2.1 Introduction and general definition

Confidence Intervals (CIs) were introduced in Inference sessions 1 and 2. They provide a measure of how close a parameter estimate is likely to be to the population value. A CI is itself a statistic, consisting of a **pair** of values (*L, U*), say, where *L* < *U*. Just like other statistics the CI has a sampling distribution and the confidence level associated with an interval is based on this sampling distribution.

In general the statistic (*L, U*) is a 100(1-α)% CI for a population parameter μ if:

$$\text{Prob}\{\mu \in (L, U)|\mu\} = (1\text{-}\alpha) \qquad (1)$$

For example (*L, U*) is a 95% CI for μ if:

$$\text{Prob}\{\mu \in (L, U)|\mu\} = 0.95 \qquad (2)$$

These equations are probabilistic statements about the interval (*L, U*) *conditional* on the population parameterμ (not probabilistic statements about μ conditional on (*L, U*)). Equation (2) states that, for any value of μ, the probability that the population parameter is contained within the CI is 95%. Formally it can be stated that the **coverage** of the CI is 95%.

In the rest of the notes for this chapter we will simplify the notation by omitting the $|\mu$ from equations analogous to (1) and (2). However, it is important to understand that frequentist probability statements are always conditional on population parameters.

CIs are usually symmetric in probability terms. So for a 100(1-α)% CI we define (*L, U*) such that:

$$\text{Prob}\{\mu \le L\} = \text{Prob}\{\mu \ge U\} = \alpha/2 \qquad (3)$$

L is often termed the lower confidence limit and *U* the upper confidence limit.

2.2 Construction of CIs using the sampling distribution of an estimator

The key to the construction of CIs is the sampling distribution of the estimator. Suppose$\hat{\mu}$ is an estimator for μ whose sampling distribution is known and that there exist monotonically strictly increasing functions of μ, $A(\mu)$ and $B(\mu)$ such that:

$$\text{Prob}\{\hat{\mu} \leq A(\mu)\} = \text{Prob}\{\hat{\mu} \geq B(\mu)\} = \alpha/2 \qquad (4)$$

It therefore follows that

$$\text{Prob}\{A^{-1}(\hat{\mu}) \leq \mu\} = \text{Prob}\{B^{-1}(\hat{\mu}) \geq \mu\} = \alpha/2 \qquad (5)$$

Hence,$A^{-1}(\hat{\mu})$ and $B^{-1}(\hat{\mu})$ are respectively upper and lower 100(1-α)% confidence limits for μ, *i.e.,* $A^{-1}(\hat{\mu}) = U$ and $B^{-1}(\hat{\mu}) = L$ in equation (3).

Provided that appropriate functions *A* and *B* can be identified, construction of CIs follows this principle. In the following sections this is illustrated in some simple settings. It is also shown how the procedure is modified in some more complex settings.

2.3 Case Study 1: CI for the population mean of a normally distributed random variable with known variance

Let Y_i (i=1,2,..,n) be a sample of n independent observations from a population having a normal distribution with mean μ and ***known*** variance σ^2; *i.e.,* E(Y_i) = μ and Var(Y_i) = σ^2 for all *i*'s. As is justified in Inference 3 a natural estimator for μ is the sample mean $\bar{Y}$, *i.e.,* $\hat{\mu} = \bar{Y}$. With each Y_i being normally distributed with known variance, the distribution of $\bar{Y}$is known to also be normal:

$$\bar{Y} \sim N(\mu, \sigma^2/n) \Leftrightarrow Z = \frac{\bar{Y}-\mu}{\sqrt{\sigma^2/n}} \sim N(0,1) \qquad (6)$$

It follows that since $\text{Prob}\{Z \leq z_{\alpha/2}\} = \text{Prob}\{Z \geq z_{1-\alpha/2}\} = \alpha/2$

$$\text{Prob}\left\{\bar{Y} \leq \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\} = \text{Prob}\left\{\bar{Y} \geq \mu + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right\} = \alpha/2 \qquad (7)$$

Hence we have identified monotonically increasing functions of μ that satisfy equation (4). These functions are

$$A(\mu) = \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \text{and} \quad B(\mu) = \mu + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Note that due to the symmetry of the standard normal distribution $z_{\alpha/2} = -z_{1-\alpha/2}$ and so usually we write $A(\mu) = \mu - z_{1-\alpha/2}\,\sigma/\sqrt{n}$.

In this simple situation the functions *A* and *B* respectively involve the subtraction and addition of a known constant, $z_{1-\alpha/2}\, \sigma/\sqrt{n}$. As explained above it is necessary to invert these functions (by addition and subtraction of the constant respectively) to construct the CI, *i.e.*

$$U = A^{-1}(\bar{Y}) = \bar{Y} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \text{ and } L = B^{-1}(\bar{Y}) = \bar{Y} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad (8)$$

This gives values for *U* and *L* that satisfy equation (3). For succinctness the property of the CI $\bar{Y} \pm z_{1-\alpha/2}\, \sigma/\sqrt{n}$ can be expressed as

$$\text{Prob}\left\{\bar{Y} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{Y} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right\} = (1-\alpha) \qquad (9)$$

This asserts that the probability that the interval $\left(\bar{Y} - z_{1-\alpha/2}\, \sigma/\sqrt{n},\ \bar{Y} + z_{1-\alpha/2}\, \sigma/\sqrt{n}\right)$ will contain the population mean (μ) is (1-α). This interval is called the 100(1-α)% CI for the population mean. The ends of the interval are called the 100(1-α)% confidence limits. For 95% confidence limits, the most conventional value, we use $z_{0.975}$= 1.96. For other confidence levels, use the following values:

$Z_{0.90}$	=	1.28	80% level
$Z_{0.95}$	=	1.645	90% level
$Z_{0.995}$	=	2.58	99% level
$Z_{0.9995}$	=	3.29	99.9% level.

Notice that the width of the CI $\bar{Y} \pm z_{1-\alpha/2}\, \sigma/\sqrt{n}$ increases with increasing standard deviation and decreases with increasing sample size. Both of these results are intuitive: the larger the variability in *Y*, and the smaller the sample size, the less precise is the estimate of the mean.

It is important to appreciate that the formula for the 95% CI ($\bar{Y} \pm z_{1-\alpha/2}\, \sigma/\sqrt{n}$) differs from that for an estimated 95% reference range ($\bar{Y} \pm z_{1-\alpha/2}\sigma$). The reference range is the interval within which 95% of observations are expected to lie: accordingly, its width does not change with increasing sample size. The CI indicates the precision of the estimated mean. As the sample size increases the precision increases, and hence the width of the CI decreases.

2.4 Interpretation of CIs

As explained above and emphasised below it is important to understand that the probability statement being made in equation (9) and analogous expressions is about the *interval* conditional on the population parameter. For α = 0.05 there is a 95% probability that the interval$\left(\bar{Y} - 1.96\, \sigma/\sqrt{n},\ \bar{Y} + 1.96\, \sigma/\sqrt{n}\right)$ will contain μ.

Having calculated a particular 95% CI it is very tempting to simply state that μ has a probability of 0.95 of being inside these limits. This, however, would be incorrect because then we would implicitly be making a probability statement about μ conditional on the CI.

We have already emphasised that (in the frequentist paradigm) μ is not a random variable. In any particular case μ either is or is not in the interval. What we are doing when we construct a CI is to imagine that a series of repeated samples were taken from the same population and for each sample, we calculate a CI; in the long run 19 out of 20 of these confidence limits will include μ. If, in a particular problem we calculate a CI, we may happen to be unlucky in that this may be one of the 5% of intervals that does not include μ.

In order to make the direction of the conditionality apparent use wording along the lines of "a CI constructed in this way has a 95% probability of containing the population mean (or, in common parlance, "true mean")" when interpreting a particular 95% CI.

As mentioned in Inference 1 CIs can also be interpreted in terms of hypothesis tests. We will return to this in Analytical Techniques 3.

2.5 Case Study 2: CI for the population mean of a normally distributed random variable with unknown variance

As in 2.4 above, let Y_i (i=1,2,..,n) be a sample of n independent observations from a population having a normal distribution with mean μ but with an unknown variance σ^2; *i.e.,* E(Y_i) = μ and Var(Y_i) = σ^2 for all i's. Again a natural estimate for μ is the sample mean $\bar{Y}$. However the distribution of $\bar{Y}$ now depends on the unknown parameter σ^2. If we use the sample estimator$\hat{\sigma}^2$ to estimate σ^2 then the distribution of the following random variable T follows a t -distribution with n-1 degrees of freedom.

$$T = \frac{\bar{Y}-\mu}{\sqrt{\hat{\sigma}^2/n}} \sim t_{n-1}$$ (this result is formally justified in Inference 8).

CIs can be constructed using a process exactly analogous to that in section 2.4, with $\hat{\sigma}^2$ replacingσ^2, T replacing Z and t_{n-1} replacing z.

It follows that (analogously to the result in equation (8)) the lower and upper confidence limits are given by:

$$U = \bar{Y} + t_{n-1,1-\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}} \text{ and } L = \bar{Y} - t_{n-1,1-\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}.$$

In particular the 95% CI for μ can be written as:

$$\bar{Y} \pm t_{n-1,0.975}\frac{\hat{\sigma}}{\sqrt{n}} \qquad (10)$$

Example 2.1: 15 cases of Sudden Infant Death (SIDS) occurred in London during the first three months of 2017. Birth certificates were obtained for these cases and birthweights (g) were as follows:

2013	3827	3090	3260	4309	3374	3544	2835
3487	3289	3714	2240	2041	3629	3345	

We wish to calculate a 95% CI for the mean SIDS birthweight. The sample mean and standard deviation are 3200 and 663g respectively. Since $t_{14,\,0.975}$ = 2.1448, the 95% CI for the mean is:

$$\left(3200 - 2.1448\frac{663}{\sqrt{15}}, 3200 + 2.1448\frac{663}{\sqrt{15}}\right) = (2833, 3567)$$

The safest way to describe this result would be merely to say that the CI for the population mean extends from 2833 to 3567g, perhaps also stating that in general (under repeated sampling) the probability that a 95% CI constructed in this way includes the population mean is 95%. However, it would be **wrong** to state that 'there is a 95% probability that the true mean lies between 2833 and 3567g' since this statement implies that the population mean follows a probability distribution.

2.6 Case Study 3: CI for the population variance of a normally distributed random variable

Let Y_i (i=1,2,..,n) be a sample of n independent observations from a population having a normal distribution with mean μ and unknown varianceσ^2; *i.e.,* E(Y_i) = μ and Var(Y_i) = σ^2 for all i's. An unbiased estimator of σ^2 is the sample variance defined as follows:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}$$

As in previous examples, in order to calculate a CI for σ^2 it is necessary to know the sampling distribution of this estimator. It has been shown in Inference 2, that:

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

It follows that:

$$\text{Prob}\left\{\hat{\sigma}^2 \leq \frac{\sigma^2}{n-1}\chi^2_{n-1,\alpha/2}\right\} = \text{Prob}\left\{\hat{\sigma}^2 \geq \frac{\sigma^2}{n-1}\chi^2_{n-1,1-\alpha/2}\right\} = \alpha/2$$

Hence we have identified monotonically increasing functions of σ^2 that satisfy equations of the type given (for μ) in equation (4). These (multiplicative) functions are

$$A(\sigma^2) = \frac{\sigma^2}{n-1}\chi^2_{n-1,\alpha/2} \text{ and } B(\sigma^2) = \frac{\sigma^2}{n-1}\chi^2_{n-1,1-\alpha/2}.$$

As in previous examples it is necessary to apply the inverse of these functions to the estimated variance in order to construct the CI, *i.e.*

$$U = A^{-1}(\hat{\sigma}^2) = \frac{(n-1)\hat{\sigma}^2}{\chi^2_{n-1,\alpha/2}} \text{ and } L = B^{-1}(\hat{\sigma}^2) = \frac{(n-1)\hat{\sigma}^2}{\chi^2_{n-1,1-\alpha/2}} \qquad (11)$$

Example 1 (continued): We wish to calculate a 95% CI for the variance of the SIDS birthweight. Since$\chi^2_{14,0.025} = 5.629$and $\chi^2_{14,0.975} = 26.119$ (which can be obtained using Stata or from statistical tables) a 95% CI for σ^2 is:

$(14 \times 663^2/26.119, 14 \times 663^2/5.629)$ = (235613g^2 , 1093261g^2) and thus, taking square-roots, a 95% CI for $\sigma$is (485, 1046)g.

2.7 Approximate CIs for parameters estimated using large samples

As explained in Inference 2 the arithmetic mean of a random sample of size *n* follows an approximate normal distribution even if the random variable is not itself normally distributed, provided *n* is large. Indeed, by the Central Limit Theorem, as sample sizes tend to infinity the sampling distribution of most typically encountered parameter estimators tends to normal.

This means that the formulae used in section 2.4 can be used to construct approximate 95% CIs for a population mean even if the variable is not normally distributed, provided the sample size is large. Further the approach adopted in that section requires only minor modification in order to construct CIs for any parameter estimators that follow an approximate normal distribution when sample sizes are large. In the following sections these approximate CIs, together with alternatives that are more appropriate when sample sizes are small, are introduced in two further common settings.

2.8 Case Study 4: CI for the true population proportion

2.8.1 General principles

For a sample of size *n* let *R* denote the number of ‘successes’. Provided that the *n* observations are independent then *R* ~ Binomial(*n*,*π*). The sample proportion *P*, where$P = R/n$ is an unbiased estimate of the population parameter, the probability of success π. Suppose we wish to construct 95% confidence limits for π defined here as (π_L, π_U). It is natural to use *R* to construct the CI. However because *R* is discrete (taking only *n* + 1 possible values) it is not possible to construct a CI that has 95% coverage for all of the (infinite) possible values of π. The most usual approach (Clopper and Pearson, *Biometrika* 26:404-413, 1934) calculates confidence limits that have the property:

$\boldsymbol{Prob(\pi_L > \pi) \leq 0.025}$ for all $\boldsymbol{\pi}$ with exact equality for some values of $\boldsymbol{\pi}$ (12)

$\boldsymbol{Prob(\pi_U < \pi) \leq 0.025}$ for all $\boldsymbol{\pi}$ with exact equality for some values of $\boldsymbol{\pi}$ (13)

To compute the CI two different approaches can be used. The first uses the fact the sampling distribution of *R* is Binomial(n,π) to calculate the CI, the second uses a normal approximation to this binomial distribution. Somewhat confusingly the CIs from the first approach are often termed 'exact' CIs. They are exact in the sense that they use the 'exact' distribution of *R*, but not that they have coverage probabilities exactly equal to 0.95 for all values of π. However as the sample size increases the coverage properties of both methods improve, approaching 95% for all values of π.

2.8.2 Using exact methods based on the binomial distribution

Suppose that *R* takes the value *r* in the sample*.* We use the sampling distribution of *R* (Binomial(n,π)) to find the value of $\pi_L$which satisfies the following probability statement:

Prob($R \geq r$ | π=π_L) = 0.025 (14)

Similarly we find the upper limit π_U such that

Prob($R \leq r$ | π=π_U) = 0.025 (15)

Such confidence limits satisfy equations (12) and (13). See Agresti (*An introduction to categorical data analysis* (1996)), p18-20 for further discussion of the coverage properties of these CIs.

Example 2.2: Consider an example where *n* = 20, *r* = 5 and therefore *P* = 0.25. It is necessary to find $\pi_L$and π_U such that:

Prob($R \geq 5$ | π=π_L) = 0.025

Prob($R \leq 5$ | π=π_U) = 0.025

To solve these requires obtaining the solution to a complex polynomial equation, which is not straightforward by hand. However from Neave's table of cumulative probabilities of the binomial distribution for *n* = 20 (page 11), we can see that:

$$Prob(R \geq 5|\,\pi = 0.08) = 0.0183$$

$$Prob(R \geq 5|\,\pi = 0.09) = 0.0290$$

Therefore, π_L must be a value between 0.08 and 0.09.

Similarly from Neave's tables we can also see that:

$$Prob(R \leq 5|\,\pi = 0.45) = 1 - 0.9447 = 0.0553$$

$$Prob(R \leq 5|\,\pi = 0.50) = 1 - 0.9793 = 0.0207$$

Therefore π_U must be a value between 0.45 and 0.50. We cannot go further than this using Neave's binomial tables.

Stata can be used to calculate the exact 95% CI as follows:

```
. cii proportions 20 5
                                                         -- Binomial Exact --
    Variable |        Obs   Proportion     Std. Err.      [95% Conf. Interval]
-------------+---------------------------------------------------------------
             |         20          .25      .0968246      .0865715     .4910459
```

The figures below show the sampling distributions of *R* given $\pi =$ 0.0866 and given $\pi =$ 0.4910.

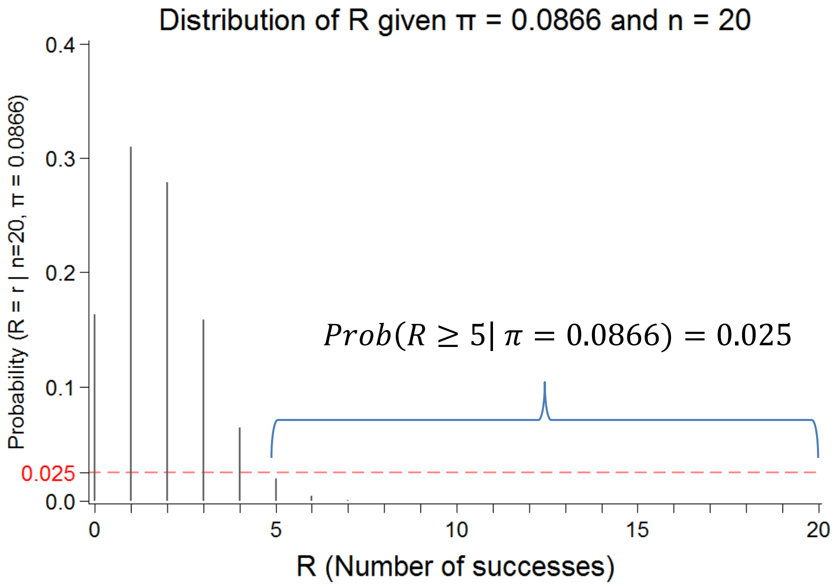


Figure 2.1: Sampling distribution of number of successes out of 20 (*R*) conditional on the probability of success being 0.0866.

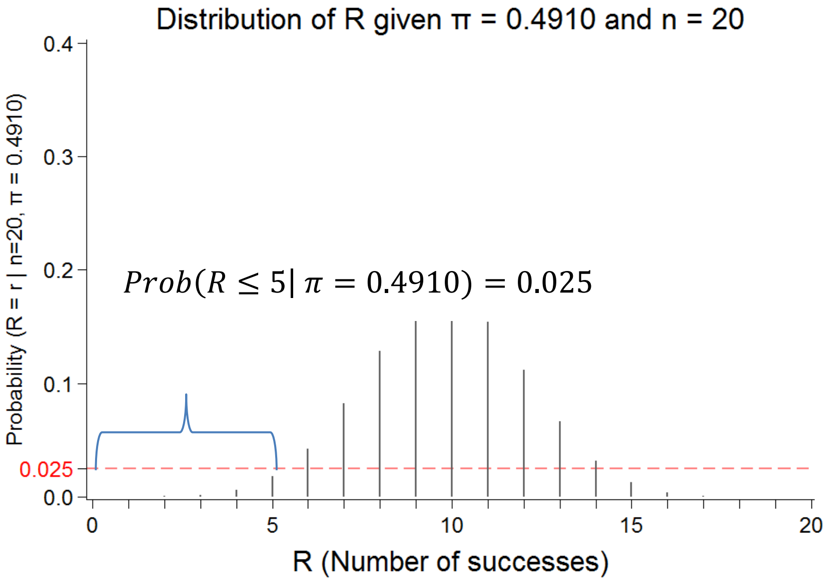


Figure 2.2: Sampling distribution of number of successes out of 20 (*R*) conditional on the probability of success being 0.4910.

2.8.3 Using a normal approximation to the binomial distribution (when sample size is large)

When *n* is large, *P* follows an approximate standard normal distribution, whose expectation and variance are both functions of π.

$$P \sim N(\pi, \sigma^2) \text{ where } \sigma^2 = \pi(1-\pi)/n$$

The variance of *P* depends on π and hence is unknown. However, the variance can be approximated by replacing π by its estimated value *p* (the observed value of *P*). Hence, approximately, for large samples

$$P \sim N(\pi, \hat{\sigma}^2) \text{ where } \hat{\sigma}^2 = p(1-p)/n$$

CIs can be constructed using a process exactly analogous to that in section 2.4, with *P* replacing $\bar{Y}$, π replacing μ and$\hat{\sigma}^2$replacingσ^2. It follows that a 100(1-α)% CI for π is given by

$$P \pm z_{1-\alpha/2}\sqrt{\frac{P(1-P)}{n}} \qquad (16)$$

Hence approximate 95% confidence limits for π are given by:

$$P \pm 1.96\sqrt{\frac{P(1-P)}{n}} \qquad (17)$$

This method of constructing approximate 95% CIs has the advantage of simplicity.

However, it should be noted that modifications of this formula (such as those due to Wilson and to Jeffreys) have been shown to have better coverage properties.

See Brown LD, Cai TT and DasGupta A, *Statistical Science*, 16:101-133, 2001 for a review.

2.8.4 Examples of confidence limits for binomial proportions

- *n*=10, *r*=4, *p*=0.4

Exact 95% CI for π:	(0.122, 0.738)
Normal approximation:	$0.4 \pm 1.96\sqrt{\frac{0.4\times0.6}{10}}$ = (0.096, 0.704)

- *n*=50, *r*=20, *p*=0.4

Exact 95% CI for π:	(0.264, 0.548)
Normal approximation:	$0.4 \pm 1.96\sqrt{\frac{0.4\times0.6}{50}}$ = (0.264, 0.536)

- *n*=1000, *r*=400, *p*=0.4

Exact 95% CI for π:	(0.369, 0.431)

Normal approximation: $0.4 \pm 1.96\sqrt{\frac{0.4\times0.6}{1000}}$ = (0.370, 0.430)

Note: As *n* increases, the CI shrinks (*i.e.,* the estimate becomes more precise), and the normal approximation is more accurate. In the first example the normal approximation does not work well.

2.9 Case Study 5: CI for a rate

2.9.1 Using exact methods based on the Poisson distribution

Suppose that *Y* is a count, of the number of events occurring during a certain period of time (*t*). Provided that events occur independently then $Y \sim \text{Poisson}(\mu t)$. The sample rate *R*, where$R = Y/t$ is an unbiased estimate of the population parameter, the event rateμ.

The probability that *Y*= *y* is given by $\frac{(\mu t)^y e^{-\mu t}}{y!}$ for y = 0, 1, 2, ..., ∞ .

Suppose that *Y* takes the value *y* in the sample and that we wish to construct 95% confidence limits for μ defined here as $(\boldsymbol{\mu_L}, \boldsymbol{\mu_U})$. In an analogous way to the approach in 2.9.2 we can use the sampling distribution of *Y* to find the value of$\mu_L$which satisfies

Prob($Y \geq y$ | μ =μ_L) = 0.025

Similarly we can find the upper limit $\mu_U$such that:

Prob($Y \leq y$ | μ =μ_U) = 0.025

These equations can be solved by hand using an iterative procedure (*i.e.,* trying several values for μ_L and μ_U until the above equalities are satisfied) or using a computer with appropriate functions (*e.g., cii* in Stata).

Example 2.3: In Millom rural district, next to the Sellafield nuclear reprocessing plant, 6 people aged below 24 died from leukaemia in a ten-year period starting from 1968, a rate of 0.6 deaths per year. To construct a 95% CI for the rate of deaths for this age group we need to find μ_L such that: Prob(6 events or more | μ =μ_L) = 0.025 and find μ_U such that: Prob(6 events or fewer | μ =μ_U) = 0.025.

Using Stata demonstrates that for *y* = 6, the exact 95% CI for μ is (0.2202, 1.3059 deaths per year). The syntax and output is as follows:

```
cii means 10 6, poisson
                                                 -- Poisson  Exact --
    Variable | Exposure  Mean    Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------------
             |       10    6      .244949      .2201894    1.305947
```

Note that it is more usual in epidemiology to express rates as the number of events per person-year of follow-up, rather than per year of follow-up as in this simple example.

2.9.2 Using a normal approximation to the Poisson distribution (when sample size is large)

When *n* is large, *Y* follows an approximate standard normal distribution, whose expectation and variance are both equal to μt.

$$Y \sim N(\mu t, \sigma^2) \text{ where } \sigma^2 = \mu t \, .$$

As in section 2.8.3, the variance of the estimator (*Y*) is a function of the unknown parameter. However approximating the variance of *Y* by *Y* itself and following the approach in section 2.8.3 it follows that a 95% CI for μ is given by :

$$\left(Y \pm 1.96\sqrt{Y}\right)/t$$

Chapter 3: Hypothesis tests

Objectives

By the end of this session students will be able to:

- Understand the motivation for hypothesis tests (revision).
- Perform hypothesis tests in a number of simple settings.
- Correctly interpret the result of a hypothesis test.

3.1 General principles and motivating example

Hypothesis testing has been introduced in the Inference course. This is one of the most important and commonly used techniques of statistical inference. Suppose a series of observations are selected at random from a population. We might be interested in a certain hypothesis, usually called the null hypothesis, which specifies a value (or values) for some parameter(s) of the population. The question then arises: do the observations in the sample throw any light on the plausibility of the hypothesis?

Some samples will be reasonably typical of those which might be expected by sampling theory if the null hypothesis was true. Other samples will have certain features which would be unlikely to arise if the null hypothesis is true; if such a sample is observed it would give reason for suspecting that the null hypothesis is untrue.

For example, let us assume that we toss a coin 10 times and observe a tail only in one tossing; how compatible is this result with the null hypothesis that the coin in question is unbiased? In this simple example, the population from which we have sampled the 10 observations is a notional one, nominally a population involving an infinite number of times in which this coin is tossed. The population parameter of interest is the probability that this coin, when tossed, would show a tail. The null hypothesis in question states that this probability is equal to 0.5. Intuitively the sample drawn seems unlikely to arise if the null hypothesis is true and may give us reason to suspect its truthfulness. However, we need to formalise this in a more appropriate way and define a rule as to what should be considered a “likely” or an “unlikely” sample given that the null hypothesis is true. This is the role of what is known as *hypothesis testing*.

After defining the null and alternative hypotheses and drawing our sample, we need therefore to calculate what is known as the *p*-value. This is defined as ***the probability of observing the sample under consideration or more extreme ones given that the null hypothesis is true***.

What is done is to divide the sample distribution into two parts.

One part involves all samples that are as extreme, or more extreme, than the observed sample given that the null hypothesis is true. For the above example, this would involve the 4 possible samples of size 10 that show 1 tail, 9 tails, 0 tails or 10 tails.

The second part involves all other samples that are more likely to occur than the observed data given that the null hypothesis is true. For the above example this means the 7 remaining possible samples of size 10 that show 2, 3, 4, 5, 6, 7, or 8 tails. Since the cumulative probability of the sample distribution is always equal to 1, then the *p*-value as defined above will always be between 0 and 1. The smaller the *p*-value the more untenable is the null hypothesis.

Often the test is constructed through consideration of a statistic, usually an estimator of a parameter of interest (or a function of such a parameter). In the above example such a statistic is the number of tails. When divided by 10 this random variable is an estimator of the probability of obtaining a tail (denoted by π).

Let *R* denote the number of times out of 10 that a tail is observed. *R* is then a random variable which follows the binomial distribution with probability of success equal to π. Suppose that *R* takes the value *r* in our sample then we can define the *p*-value as follows:

$$\boldsymbol{Prob\{R\ as\ or\ more\ extreme\ than\ r \mid \pi = 0.5\}}$$

$$\text{i.e., } \boldsymbol{Prob\{|R-5| \geq |r-5| \mid \pi = 0.5\}} \qquad (1).$$

where 5 is the expected value of R under the null hypothesis.

If the null hypothesis is true and $\pi = 0.5$, then $R \sim \text{Bin}(10, 0.5)$. The distribution of R under the null hypothesis is shown below.

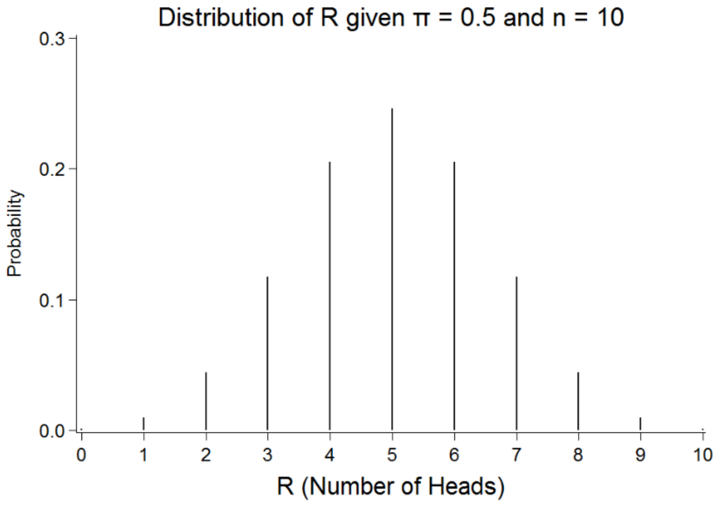


Figure 3.1: Binomial distribution *n*=10, π=0.5.

In the example above we observed 1 tail in 10 tosses. The probability of observing exactly 1 tail in 10 tosses under the null hypothesis is:

$$\binom{10}{1} \times 0.5^{10} = 0.009766$$

Recall that the *p*-value is the probability of observing our sample (or a more extreme sample that favours the alternative hypothesis) when the null hypothesis is true. In this example, values of R that are as or more extreme that that observed given that π = 0.5, are 0 , 9 , or 10 tails. Therefore, the p-value is:

$$= \left[\binom{10}{0} + \binom{10}{1} + \binom{10}{9} + \binom{10}{10}\right] \times 0.5^{10} = 0.021$$

3.1.1 One and two-sided tests

In the coin tossing example considered above obtaining one head from ten tosses is considered to be as 'extreme' as obtaining nine heads from ten tosses. This is an example of a 'two-sided' hypothesis test because results as far from the null hypothesis in the opposite direction from the observed result are considered to be as extreme as the observed result. A 'one-sided' hypothesis test only considers the probability of results that are as extreme, or more extreme, in a particular direction.

In the above example if one wishes to test the null hypothesis that the coin is fair (H_0: π=0.5) against the alternative hypothesis that heads are favoured (H_1: π<0.5) the one-sided *p*-value is defined as:

$$\mathbf{Prob}\{\boldsymbol{R} \ \leq \boldsymbol{r}|\boldsymbol{\pi} = \mathbf{0.5}\}.$$

Since, for our sample *r* = 1 this is given by

$$\text{Prob}\{\text{R=0}|\ \pi = 0.5\} + \text{Prob}\{\text{R=1}\ |\ \pi = 0.5\}\text{=0.011 .}$$

If the alternative hypothesis is that tails are favoured (H_1: π>0.5) the formula becomes:

$$\mathbf{Prob}\{\boldsymbol{R} \ \geq \boldsymbol{r}|\boldsymbol{\pi} = \mathbf{0.5}\}.$$

Since, for our sample *r* = 1 this is given by

$$\mathbf{Prob}\{\boldsymbol{R} \ \geq \mathbf{1}|\boldsymbol{\pi} = \mathbf{0.5}\} = \mathbf{1} - \mathbf{Prob}\{\boldsymbol{R} \ = \mathbf{0}|\boldsymbol{\pi} = \mathbf{0.5}\} = \mathbf{0.999}.$$

Commonly the one-sided *p*-value is defined to be the probability of results that are as extreme, or more extreme, ***in the same direction as the observed result***.

For our sample this probability is 0.011. As in many, but not all, situations, the *p*-value calculated from this one-sided test is exactly half that from a two-sided test. Clearly it is paramount to specify a priori (*i.e.,* before the data is being analysed) whether a one-sided or two-sided test is being used.

In the above example there is no logical justification for defining a biased coin as a coin with $\pi < ½$ and not considering $\pi > ½$ as bias. Similarly, in most clinical situations there is rarely any clinical justification for a one-sided test. Even when we have strong prior belief that a new treatment, for example, cannot be worse that an old one, we cannot be sure that we are right and thus a two-sided test is usually the appropriate one. In rare cases, for example when we can be certain that a real difference can only occur in one direction, and therefore that an observed difference in the opposite direction can only be due to chance, a one-sided test can be appropriate, but should be specified a priori.

3.1.2 Interpretation of *p*-values

Hypothesis tests are often thought of as a means to make decisions with an acceptance or rejection of the null hypothesis on the basis of the *p*-value. A cut-off is chosen and if the *p*-value is smaller than it, the null hypothesis is rejected. If the *p*-value is above the cut-off then the null hypothesis is accepted. Usually the cut-off is taken to be 0.05 but sometimes other values, such as 0.01, are used.

Much medical research is indeed concerned with decision making. For example, we may wish to carry out a randomised controlled trial of a new medication in order to assess whether or not this medication should be made routinely available to patients. However much medical research is not, at least directly, concerned with decision making. For example, if we carry out an epidemiological study in which we relate risk of a particular disease to gender, we do this because we are interested in understanding the aetiology of the disease, not because we want to assess whether to modify gender. For this reason, many researchers regard *p*-values as a measure of strength of **evidence** against the null hypothesis, rather than as an aid to decision making. It should, however, be pointed out that this is an area of some philosophical controversy in statistics, with not all statisticians convinced that *p*-values do measure evidence.

There are a number of different ways of reporting the results of a hypothesis test. If a *p*-value is less than 0.05 it is formally correct to state that the result is 'statistically significant at the 5% level'. More usually in the medical literature authors might report that there was 'evidence at the 5% level' that the hypothesis being tested was incorrect (*e.g.,* 'evidence that the coin was biased (p<0.05)'). One term to avoid is 'there was evidence that the result was statistically significant'.

If a *p*-value is greater than 0.05 it is appropriate to say that there was 'no (or insufficient) evidence' against the null hypothesis. In this situation it is incorrect to say that there is evidence that the null hypothesis is true, because 'absence of evidence' is not the same thing as 'evidence of absence'. To understand this, suppose we toss a coin three times and get three heads. The probability of this happening by chance if the coin is fair is 0.125 which is bigger than 0.05. We cannot claim that we have statistically significant evidence of bias, but it would be absurd to claim that our result provides evidence that the coin is fair! The correct interpretation is (the much weaker statement) that we do not have evidence that the coin is biased.

In the opinion of many statisticians a slavish focus on whether or not a *p*-value is above or below 0.05 is a rather simplistic approach to the interpretation of the results of an analysis. Accordingly, many statisticians will use words that describe the magnitude of the *p*-value. For example, a test which gives *p*=0.0001 will be interpreted as providing 'strong evidence' whilst tests that give *p*=0.06 or *p*=0.04 will both be interpreted as being 'borderline statistically significant' and providing 'some evidence'.

Exercise 3.1: Write a sentence, suitable for publication in a scientific journal, interpreting the result of the hypothesis test for the data in the motivating example.

3.2.3 Link between hypothesis tests and CIs

In AT 2 we discussed CIs and their interpretation. In particular we defined a symmetric 100(1-α)% CI for a parameter μ as a statistic, consisting of a **pair** of values(*L, U*), where

$$= Prob\{L \geq \mu \mid \mu\} = \ Prob\{U \leq \mu \mid \mu\} = \alpha/2 \text{ for all values of } \mu \qquad (2).$$

There is an important link between hypothesis testing and CIs. This link is illustrated as follows. Suppose we wish to test the null hypothesis $H_0: \mu = \mu_0$ (a particular value of μ) against the one-sided alternative $H_1: \mu > \mu_0$. Suppose $\hat{\mu}$ is an estimator for μ whose sampling distribution is known. It was shown in 2.3 that the key to the construction of CIs is the identification of monotonically strictly increasing functions of μ, *A*(μ) and *B*(μ) such that

$$\text{Prob}\{\hat{\mu} \leq A(\mu)\} = \text{Prob}\{\hat{\mu} \geq B(\mu)\} = \alpha/2 \qquad (3).$$

Provided that such functions can be identified,$U = A^{-1}(\hat{\mu})$ and $L = B^{-1}(\hat{\mu})$ are respectively upper and lower 100(1-α)% confidence limits for μ since

$$\text{Prob}\{A^{-1}(\hat{\mu}) \leq \mu\} = \text{Prob}\{B^{-1}(\hat{\mu}) \geq \mu\} = \alpha/2 \qquad (4).$$

Now suppose that $\hat{\mu}$ takes the value *m* and that *L* and *U* take the values *l* and *u* in our sample. The one-sided *p*-value is given by:

$$p = Prob\{\hat{\mu} \geq m \mid \mu = \mu_0\}$$

$$= Prob\{B^{-1}(\hat{\mu}) \geq B^{-1}(m) \mid \mu = \mu_0\} \text{ since B is monotonic increasing}$$

$$= Prob\{L \geq l \mid \mu = \mu_0\} \qquad (5).$$

Since, from (2), $Prob(L \geq \mu_0 \mid \mu = \mu_0) = 0.025$ it follows from (5) that

$$p = 0.025 \ \ if \ l = \mu_0$$

$$p < 0.025 \ \ if \ l > \mu_0$$

and $p > 0.025 \;\; if \; l < \mu_0$

Analogously, to test the null hypothesis against the one-sided alternative $H_1: \mu < \mu_0$ the one-sided *p*-value satisfies:

$p = 0.025 \;\; if \; u = \mu_0$

$p < 0.025 \;\; if \; u < \mu_0$

and $p > 0.025 \;\; if \; u > \mu_0$

In other words, the values of μ inside the confidence limits are precisely those that would not be significantly contradicted by a two-sided test at the α% significance level. Values of μ outside this interval would all be rejected by a test at the α% significance level.

3.2 Case Study 1: Exact hypothesis tests for the binomial distribution

If for a sample of size *n*, *R* successes are observed, then the sample proportion *P*, where$P = R/n$ is an unbiased estimator of the population parameter (probability of success π). Suppose we wish to test the null hypothesis that $\pi = \pi_0$ against $\pi \neq \pi_0$and observe *R* = *r*. We need to find the probability of observing *r* (*i.e., r* successes), or a more extreme proportion of successes, under the null hypothesis that $\pi = \pi_0$.

If $r < n\pi_0$, the **one-sided** *p*-value is given by:

p = Prob{*r* or fewer successes out of *n*| $\pi = \pi_0$}

$= P_0 + P_1 + \cdots + P_r$

where $P_x = \binom{n}{k}{\pi_0}^{x}(1-\pi_0)^{n-x}$.

If $r > n\pi_0$, the **one-sided** *p*-value is given by:

p = Prob{At least *r* successes out of *n*| $\pi = \pi_0$}

$= P_r + P_{r+1} + \cdots + P_n$

Now, to calculate the two-sided *p*-value, we can either, for simplicity, double the one-sided *p*-value or add to it the probabilities that are at least as small as P_r in the opposite tail of the sampling distribution.

3.3 Large sample tests

When the sample size is large, we usually perform an approximate test instead of an exact one. The original distribution of the random variable under investigation is no longer important, but the knowledge of its expected value and variance is essential for defining the large sample test. Let us assume that the random variable used to calculate our *p*-value is denoted by *R* and that it has an expected value and variance denoted by E(*R*) and Var(*R*), then to calculate our large sample test we define a random variable *Z* as follows:

$$Z = \frac{R-E(R)}{\sqrt{\mathrm{Var}(R)}} = \frac{R-E(R)}{\mathrm{SE}(R)}.$$

In almost all situations, as *n* becomes large, the distribution of *Z* tends to N(0, 1). The standard normal distribution can then be used to calculate the associated two-sided *p*-value.

3.4 Case Study 2: Approximate hypothesis tests for the binomial distribution

If *n* is large, it is tedious to carry out exact hypothesis tests without the aid of a computer. Instead, we can use the normal approximation to the Binomial distribution. Given that the null hypothesis is true and $\pi = \pi_0$ then

E(*R*) = $n\pi_0$ and Var(*R*) = $n\pi_0(1-\pi_0)$.

Under the null hypothesis H_0 the following random variable *Z* approximately follows a standard normal distribution.

$$Z = \frac{R-E(R)}{\sqrt{\mathrm{Var}(R)}} = \frac{R-n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}} = \frac{P-\pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} \text{ where } P=R/n$$

If $r < n\pi_0$ (*i.e., p=r/n <* π_0) the **one-sided** *p*-value to test the null hypothesis that $\pi = \pi_0$ is

Prob(*r* or fewer successes out of *n*, given $\pi = \pi_0$) $\approx \mathrm{Prob}\left(Z < \frac{p-\pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}\right)$.

To find this *p*-value we need to calculate the observed *z* and refer to statistical tables for a standardised normal deviate to find the above probability. Due to the symmetry of the standard normal distribution the two-sided *p*-value equals twice the above probability.

Similarly if $r > n\pi_0$ (*i.e., p >* π_0) we can calculate the **one-sided** *p*-value (defined below) and again multiply by two to get the two-sided *p*-value:

p = Prob (*r* or more successes out of *n*, given $\pi = \pi_0$)

$$\approx \mathrm{Prob}\left(Z > \frac{p-\pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}\right)$$

3.4.1 Continuity Correction

In the above section we approximate the distribution of a discrete random variable by the continuous normal distribution. The approximation can be improved if we use what is known as the continuity correction which means that we need to redefine *Z* as follows:

$$Z = \frac{|R-n\pi_0|-\frac{1}{2}}{\sqrt{n\pi_0(1-\pi_0)}} \quad \text{or} \quad \frac{|P-\pi_0|-\frac{1}{2n}}{\sqrt{\pi_0(1-\pi_0)/n}}$$

Armitage, Berry and Matthews suggest that approximate hypothesis tests should be avoided if $n\pi_0 < 10$ or $n(1-\pi_0) < 10$ i.e., for very small or very large π_0. The null binomial distribution in such cases is quite skewed and using the normal approximation would not therefore be appropriate. The continuity correction could be dropped for large *n* (e.g., *n* >100).

3.5 Case Study 3: Hypothesis testing for the true population mean, when the variance is known

In this section we discuss hypothesis testing regarding the mean of a population when the **population variance is known**. This situation may seem artificial, but sometimes it can occur and it is useful for setting ideas. We assume that we are interested in the population mean μ of a random variable *Y* (say the blood pressure of a specific group of patients) that has a known variance σ^2. A random sample of size *n* ($Y_1,\ldots, Y_n$) is taken to test the null hypothesis H_0 that:

H_0: $\mu = \mu_0$ against the alternative hypothesis H_1: $\mu \neq \mu_0$.

The natural estimate of the population mean is the sample mean $\bar{Y}$. Provided that *Y* is normally distributed, then $\bar{Y}$ is also normally distributed with the same mean μ and variance σ^2/n. Even if *Y* is not normal but *n* is large then $\bar{Y}$ is approximately normal through the central limit theorem. Hence under H_0 the following random variable *Z* approximately follows a standard normal distribution:

$$Z = \frac{\bar{Y}-E(\bar{Y})}{\sqrt{\mathrm{Var}(\bar{Y})}} = \frac{\bar{Y}-\mu_0}{\sqrt{\sigma^2/n}}$$

For an observed value of $\bar{Y} = \bar{y}$ that is $< \mu_0$ the **one-sided** *p*-value that tests the null hypothesis that $\mu = \mu_0$ is given by:

$$p = \mathrm{Prob}(\bar{Y} \leq \bar{y} \mid \mu = \mu_0)$$

$$= \mathrm{Prob}\left(Z < \frac{\bar{y}-\mu_0}{\sqrt{\sigma^2/n}}\right)$$

$$=\Phi\left(\frac{\bar{y}-\mu_0}{\sqrt{\sigma^2/n}}\right)$$ where$\Phi$is the distribution function for a N(0,1) distribution.

To find this *p*-value we need to calculate the observed *z* and refer to statistical tables to find the above probability. Due to the symmetry of the standard normal distribution the two-sided *p*-value is equal to twice the above probability.

Similarly, if $\bar{y} > \mu_0$ we can calculate the **one-sided** *p*-value $\mathrm{Prob}\left(Z > \frac{\bar{y}-\mu_0}{\sqrt{\sigma^2/n}}\right)$ and again multiply by two to get the two-sided *p*-value.

This two-sided *p*-value can be written more succinctly as 2*Prob$\left\{Z > \frac{|\bar{y}-\mu_0|}{\sqrt{\sigma^2/n}}\right\}$ or equivalently as $2\left[1 - \Phi\left(\frac{|\,\bar{y}-\mu_0|}{\sqrt{\sigma^2/n}}\right)\right]$ or as $2\Phi\left(\frac{-|\,\bar{y}-\mu_0|}{\sqrt{\sigma^2/n}}\right)$.

Example 3.1: Over a 12-month period 78 cases of Sudden Infant Death (SIDS) occurred in London. Birth certificates were obtained for these cases and birthweights were recorded. Let *Y* be the birthweight in grams. For these 78 cases, the mean birthweight was equal to 2995.5g. From a listing of all the birthweights during this year, it is known that the standard deviation of birthweight during this year was 800g and that the mean birthweight was 3500g. The question of interest is whether or not these data support the hypothesis that the mean birthweight of SIDS cases was equal to the mean birthweight of the whole birth population.

Let μ be the mean birthweight of the SIDS population during 12-month period, we wish to test the null hypothesis H$_0$: μ = 3500g against the alternative H$_1$: $\mu \neq$ 3500g.

DATA: *n* = 78 $\bar{y} = 2995.5$g and σ = 800g

One-sided *p*-value = $\text{Prob}\left\{Z < \frac{2995.5-3500}{\sqrt{800^2/78}}\right\} = \text{Prob}\{Z < -5.581\}$ = 0.000000012

The two-sided *p*-value is < 0.0001. We may conclude therefore that we have strong evidence against the null hypothesis. The data is not consistent with the hypothesis that the mean birthweight of SIDS cases is equal to the mean birthweight of the whole birth population.

3.6 Case Study 4: Hypothesis testing for the true population mean, when the variance is unknown

In this section we discuss hypothesis testing regarding the mean of a population when the population variance is unknown. We are interested in the population mean μ of a random variable *Y* that is normally distributed and has an unknown variance σ^2. A random sample of size *n* is taken that will be used to test the null hypothesis that:

H$_0$: $\mu = \mu_0$ against the alternative hypothesis: H$_1$: $\mu \neq \mu_0$.

The natural estimate of the population mean is still the sample mean $\bar{Y}$. However, the distribution of $\bar{X}$ now depends on the unknown parameter σ^2. If we estimate this parameter by its sample estimate $\hat{\sigma}^2$, then the random variable *T* = $\frac{\bar{Y}-\mu}{\sqrt{\hat{\sigma}^2/n}}$) follows a *t*-distribution with *n*-1 degrees of freedom (see Inference 8) under the null hypothesis.

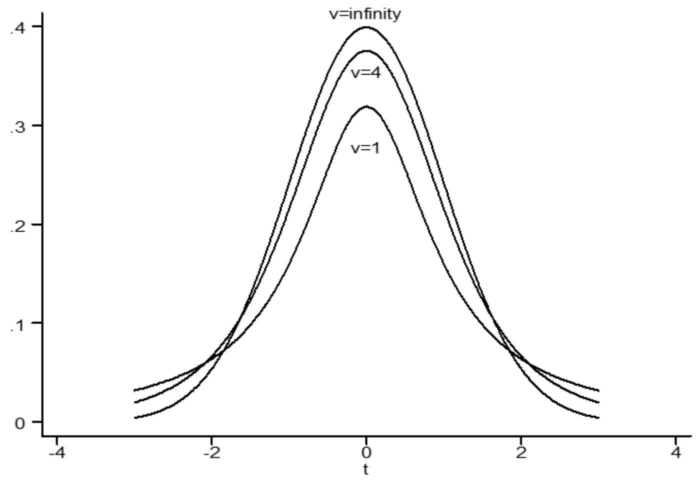


Figure 3.2: Student *t*-distributions with one, four and infinity degrees of freedom.

The t_n distribution is very similar to the normal distribution, but somewhat more heavy tailed (see figure 3.2) especially for smaller sample sizes. Like the standard normal distribution, it is bell shaped and symmetrical about zero. We denote the αth centile of the *t*-distribution by the symbol $t_{df,\,\alpha}$ (see Neave p20). We can use these centiles to calculate approximate *p*-values. In particular, if the observed value of *T* is greater than $t_{df,0.975}$, then the associated two-sided *p*-value is less than 0.05.

Example 3.2: Consider the SIDS data analysed in Analytical techniques 2. The birth weight in grammes of this group were as follows:

2013	3827	3090	3260	4309	3374	3544	2835
3487	3289	3714	2240	2041	3629	3345	

The mean and standard deviation of this sample are 3200g and 663g respectively. Without assuming that the population standard deviation is known, suppose we wish to test the null hypothesis that the mean birth weight in the SIDS cases is equal to that in the whole population (known to be 3500g).

The null hypothesis can be written algebraically as:

H_0: μ = 3500g with the alternative hypothesis H_1: $\mu \neq$ 3500g.

DATA: n = 15 $\bar{y} = 3200\text{g}$ and $\hat{\sigma}$= 663 g

$$T = \frac{3200-3500}{\sqrt{663^2/15}} = -1.752$$

From the *t*-distribution with 14 DF, we can see that $t_{14,\,0.95}$ = 1.7613 and since the observed $|T|$ is smaller than this value, then the one-sided *p*-value is > 0.05 and thus the two-sided *p*-value > 0.10.

We may conclude therefore that we do not have evidence against the null hypothesis. This data is consistent with the hypothesis that the mean birth weight of SIDS cases is equal to the mean birth weight in the whole birth population.

Note that this result is consistent with the CI for the mean (2833, 3567) grammes calculated in 2.6. 3500g lies within the 95% CI and so $p > 0.05$.

3.7 Case Study 5: Hypothesis testing for a difference between population means in a matched study

A matched comparison study is one in which the design of the study is such that observations made under one condition are individually linked to one or more observations made under one or more other conditions. In a matched comparison study the set of matched observations is sometimes referred to as a block.

For example, consider a study in which fifty subjects are given an anti-hypertensive drug designed to lower their diastolic blood pressure. Diastolic blood pressure is measured before (Y_{i1}) and after (Y_{i2}) administration of the drug. Here the blocks are the subjects. In a study in which twenty patients with Alzheimer's disease and their unaffected spouses each carry out a memory test the blocks are the husband-and-wife pairs.

Such studies can be analysed by first calculating the difference between observations made on the same block (*i.e.,* $D_i = Y_{i2} - Y_{i1}$) and then carrying out a one-sample t-test of the null hypothesis that the mean difference is zero. This procedure is referred to as the paired t-test.

Chapter 4: Association

Objectives

By the end of this session students will be able to:

- Understand the concept of an association between two random variables.
- Calculate and interpret Pearson's correlation coefficient.
- Test associations between two categorical variables using a chi-squared test.
- Measure associations between two binary variables using the odds ratio.

4.1 Introduction

Two variables are associated with one another if the distribution of one is dependent on the value taken by the other and vice-versa. The way in which such associations are quantified depends on the nature of the two variables, with important distinctions between categorical and continuous variables. In this session ways of quantifying and testing the statistical significance of such associations in a number of common settings are explored.

Relationships between pairs of variables can also be explored using regression as well as with techniques assessing association and there are a number of parallels between the two approaches which will be explored here and in the Regression Course. The distinction between association and regression is that association concerns the joint distribution of two variables whereas regression concerns the distribution of one variable conditional on the other. Regression is asymmetrical in the sense that the regression of *Y* on *X* is not the same as that of *X* on *Y.* Association is symmetric in the sense that the extent of association between *X* and *Y* is the same as that between *Y* and *X.*

Of course one must always be aware that association does not imply causation. See the Basic Epidemiology course for discussion of this point.

4.2 Association between pairs of Continuous Variables

4.2.1 Definition of the Pearson Correlation Coefficient

The Pearson correlation coefficient (ρ) is a measure of the linear association between two continuous variables. It is defined as follows:

$$\rho = \frac{E[(X - E(X))(Y - E(Y))]}{\sqrt{E(X - E(X))^2 E(Y - E(Y))^2}} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \qquad (1)$$

Suppose $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$ and $\boldsymbol{y} = \{y_1, y_2, \ldots, y_n\}$ are the values of the variables measured on a **random** sample of *n* subjects. Then ρ may be estimated by:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} = \frac{S_{xy}}{S_x S_y} \qquad (2)$$

Here s_{xy} is the sample covariance of *X* and *Y*, s_x is the sample SD of *X* (sometimes written as $\hat{\sigma}_x$) and s_y is the sample SD of *Y*. *r* is referred to as the Pearson or Product-Moment estimator of ρ.

4.2.2 Properties of the Pearson Correlation Coefficient

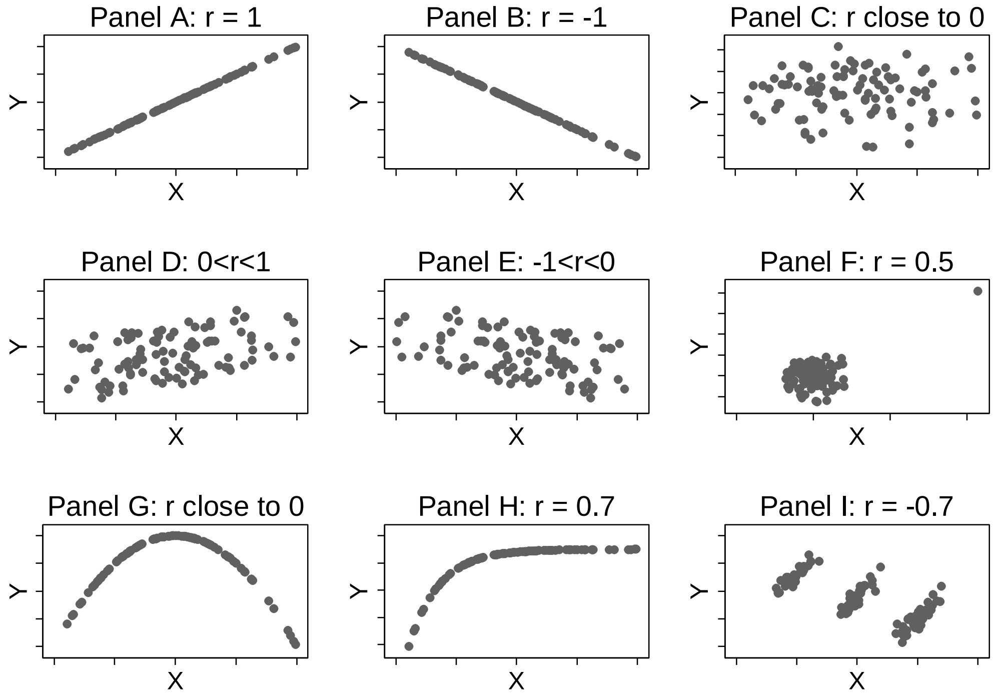


Figure 4.1: Examples of Pearson correlation coefficients

The figure above (after Fisher and van Belle (1993) p 373) illustrates the value of *r* in nine different datasets.

Key properties of *r* are as follows.

1) *r* takes values in the range -1 to 1 ($-1 \leq r \leq 1$).
2) In expectation, *r* takes positive values if increasing *X* is associated with increasing *Y* and negative values if increasing *X* is associated with decreasing *Y*.
3) $|r|$ = 1 if and only if there is an exact straight line relationship between the two variables.
4) *r* tends to be close to 0 if there is no linear relationship between the two variables.
5) The formula for *r* is symmetric in *X* and *Y*. This contrasts with regression formulae.

6) *r* is scale and location invariant. If the values of *X* and/or *Y* are multiplied by a constant factor, or have a constant added, *r* is unchanged. This contrasts with regression coefficients.

A number of the panels in figure 4.1 merit special consideration. Panel F demonstrates that a single outlying value can have a strong impact on *r.* If the point at the top right was omitted then the value of the correlation coefficient would be much lower. Panels G and H demonstrate that *r* measures linear relationships. If *Y* can be predicted exactly from *X*, but the relationship is non-linear then the magnitude of *r* will not be equal to 1. Indeed in Panel G the observed correlation coefficient is almost zero.

Panel I illustrates that attempting to describe a complex relationship with a single correlation coefficient can be misleading. Here there are evidently three clusters of points, within each of which there is a positive association between the two variables. However across clusters there is a negative association.

4.2.3 Testing the null hypothesis that $\rho = 0$

The relationship between the simple linear regression model and the Pearson correlation coefficient is discussed in the Regression course. It is explained there that the following test statistic can be used to test the null hypothesis H_0: $\rho = 0$.

$$T = r\sqrt{\frac{n-2}{1-r^2}} \sim t_{n-2} \qquad (3)$$

p-values are calculated using the fact that under H_0: $T \sim t_{n-2}$.

This test is valid provided that either the regression of *Y* on *X* satisfies the assumptions of the simple regression model (conditional on *X, Y* follows a normal distribution with constant variance and mean that is a linear function of *X*) and/or the regression of *X* on *Y* satisfies the assumptions of the simple regression model.

Exercise 4.1: Is there a statistically significant association between Erythrocyte adenosine triphosphate (ATP) levels measured in oldest and youngest sons in 17 families?

Data:	Oldest	Youngest
Mean	490 ng/ml	470 ng/ml
Variance	82000 ng/ml^2	42000 ng/ml^2
Covariance	35000 ng/ml^2	

4.2.4 Confidence Intervals for ρ

In order to calculate confidence intervals (CIs) for ρ it is necessary to know the form of the joint distribution of *X* and *Y.* Valid CIs can be calculated using Fisher's *Z*-transformation provided that *X* and *Y* follow a bivariate normal distribution.

Figure 4.2 illustrates data that follows a bivariate normal distribution. Contours linking points having equal probability density are elliptical. Formally, *X* and *Y* follow a bivariate normal distribution provided that ***both*** the regression of *Y* on *X* and that of *X* on *Y* satisfy the assumptions of the simple regression model (these are specified in 4.2.3 and in the Regression course).

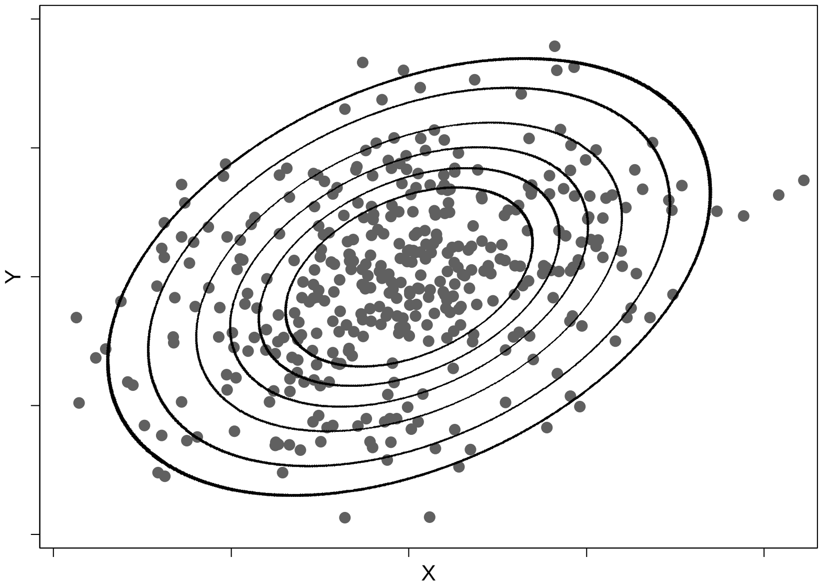


Figure 4.2: Data from a bivariate normal distribution. Contours linking points having equal probability density are elliptical.

If $\rho \neq 0$ the distribution of *r* is not symmetric and in many situations is far from approximately normal. However Fisher showed that, if *X* and *Y* are bivariate normal, then

$$Z_r = \frac{1}{2} log_e \left(\frac{1+r}{1-r}\right) = \tanh^{-1}(r) \qquad (4)$$

follows an approximate normal distribution as follows:

$$Z_r \sim N\left(\frac{1}{2} log_e \left(\frac{1+\rho}{1-\rho}\right), \frac{1}{n-3}\right) \qquad (5)$$

This formula can be used to calculate a CI for Z_ρ (= $\tanh^{-1}(\rho)$) which can then be back transformed to a CI for ρ using the inverse transformation

$$\rho = \frac{exp(2Z_\rho)-1}{exp(2Z_\rho)+1} = tanh(Z_\rho) \qquad (6)$$

Exercise 2: Calculate a 95% CI for the Pearson correlation coefficient estimated as part of exercise 1.

4.2.5 Testing for equality between two correlation coefficients

Often in statistics we wish to make comparison between two parameters. This subject is covered in detail in Analytical Techniques 5. Here we briefly consider the comparison of two correlation coefficients. For example, we might wish to test whether the magnitude of the association (as measured by a correlation coefficient) between salt intake and blood pressure is the same in males and females.

Fisher's *Z*-transformation can be used to formally compare two independent correlation coefficients (under the assumption of bivariate normality in both populations). Suppose r_1 is an estimate of ρ_1 in a sample of size n_1 and that r_2 is an estimate of ρ_2 in a sample of size n_2. To test H_0: $\rho_1 = \rho_2$ refer the following test statistic to the *N*(0,1) distribution.

$$\text{Test statistic} = \frac{Z_{r_2} - Z_{r_1}}{\sqrt{\frac{1}{n_2-3}+\frac{1}{n_1-3}}}$$

In practice this procedure is of limited value. It is much more useful to use an appropriate regression model to compare the slope of the relationship in the two populations. See material on interactions in the Regression course for details.

4.2.6 Use and misuse of correlation coefficients

Use of the Pearson correlation coefficient is extremely common in the medical literature. Unfortunately, much of this use is less than ideal because other procedures (such as regression) would be more informative, and some is incorrect. Some of the misuses are detailed below. See Altman (1991), chapter 11, sections 1-9 for more discussion of these and other limitations of *r.*

1) Valid inference about ρ is dependent upon having an unrestricted random sample of subjects. If subjects are selected on the basis of values taken by either of the variables of interest this will have an impact on the expected value of *r.* Figure 4.3 illustrates that reducing the range of *X* values tends to reduce *r*, whilst omitting points from the centre of the distribution tends to increase it. In contrast, the expectation of parameter estimates in a linear regression model relating *Y* to *X* are unchanged by such restrictions on the range of *X.*

2) A spurious correlation between an initial measurement (X_1) and a change in that measurement (X_2 - X_1) will tend to arise even when the true values of these variables are unrelated. This phenomenon, known as regression to the mean, arises because positive random error in the measurement of X_1 will be reflected as negative random error in X_2 - X_1 and vice-versa. This problem can be avoided by investigating the correlation between the difference in the measurements (X_2 - X_1) and their mean (X_2+X_1)/2.

3) The Pearson correlation coefficient is often used to assess agreement. This is inappropriate. A high correlation between two variables does not imply that they agree. For example, the correlation between *X* and 2**X* is 1, but they do not agree. See the course on generalised linear models for the correct approach.

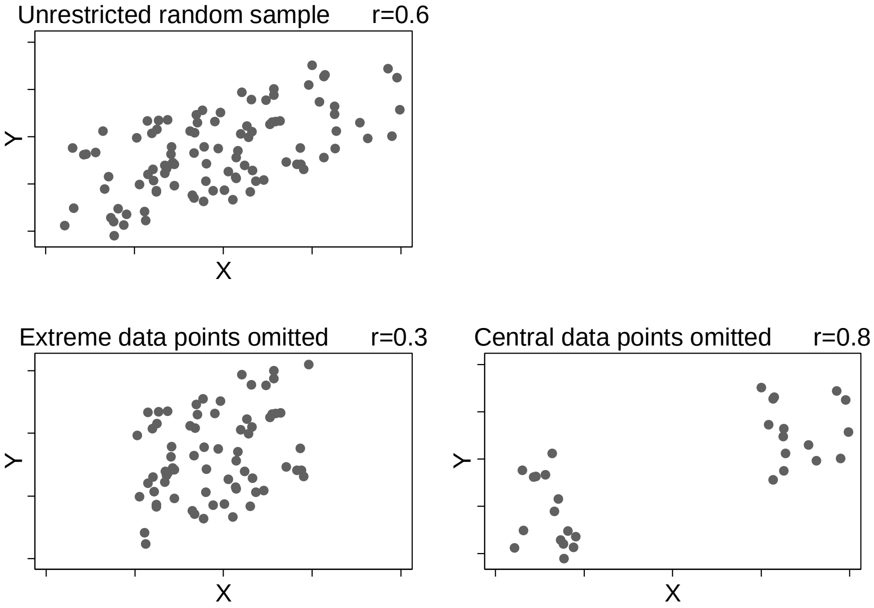


Figure 4.3: Effect of data restrictions on the Pearson correlation coefficient.

In general ***regression models*** are much more informative than Pearson correlation coefficients because they describe relationships on the scales the original measurements are made on and can be used to predict one variable from the other.

4.2.7 Obtaining Pearson correlation coefficients using Stata

Figure 4.4 illustrates the association between age and height in a sample of 190 children from the Gambia aged 6-36 months.

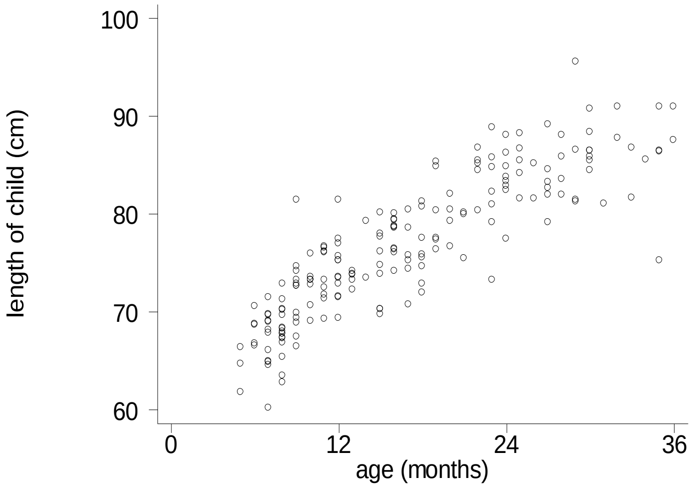


Figure 4.4: Association between age and height in children from the Gambia aged 6-36 months.

Stata can be used to calculate the Pearson correlation coefficient as follows.

```
. corr len age
(obs=190)

             |      len      age
-------------+------------------
         len |   1.0000
         age |   0.8676   1.0000
```

Exercise 4.3: Comment on the limitations of the correlation coefficient in this example.

4.3 Association between pairs of binary variables

The odds ratio is a measure of association between two binary variables (*X* and *Y*, each taking the values 0 and 1). Like Pearson's correlation coefficient the odds ratio is symmetric in that the odds ratio relating *X* to *Y* is the same as that relating *Y* to *X*. To illustrate this let π_{ij} be the probability that *X* = *i* and *Y* = *j.*

Table 4.1: Population parameters in a 2x2 contingency table

	Y = 0	*Y* = 1	Total
X = 0	π_{00}	π_{01}	$\pi_{0.}$
X = 1	π_{10}	π_{11}	$\pi_{1.}$
Total	$\pi_{.0}$	$\pi_{.1}$	1

The odds that Y_i = 1 given X_i = 0 is π_{01}/π_{00} and the odds that Y_i = 1 given X_i = 1 is π_{11}/π_{10}. Hence the odds ratio relating *Y* to *X* is:

$$\Psi = \frac{\pi_{11}/\pi_{10}}{\pi_{01}/\pi_{00}} = \frac{\pi_{11} \times \pi_{00}}{\pi_{10} \times \pi_{01}} \qquad (7)$$

A similar calculation gives the odds ratio relating *X* to *Y* as:

$$\Psi = \frac{\pi_{11}/\pi_{01}}{\pi_{10}/\pi_{00}} \qquad \text{which also} = \frac{\pi_{11} \times \pi_{00}}{\pi_{10} \times \pi_{01}}$$

For data in a 2x2 contingency table the estimated odds ratio is obtained by substituting the population proportions in the table above by the observed proportions. Denote the observed counts in the table using the following notation.

Table 4.2: Observed data in a 2x2 contingency table

	$Y = 0$	$Y = 1$	Total
$X = 0$	O_{00}	O_{01}	$O_{0.}$
$X = 1$	O_{10}	O_{11}	$O_{1.}$
Total	$O_{.0}$	$O_{.1}$	$O_{..}$

Substituting population proportions by their estimates in equation (7) gives the following formula for the estimated odds ratio.

$$\hat{\Psi} = \frac{\hat{\pi}_{11}\times\hat{\pi}_{00}}{\hat{\pi}_{10}\times\hat{\pi}_{01}} = \frac{(O_{11}/O_{..})\times(O_{00}/O_{..})}{(O_{10}/O_{..})\times(O_{01}/O_{..})} = \frac{O_{11}\times O_{00}}{O_{10}\times O_{01}} \qquad (8)$$

Example 4.1: The following 2×2 contingency table shows the prevalence of smoking (smoker v never smoker) in 2705 people with (case) and without (control) lung cancer:

Table 4.3: Observed lung cancer - smoking data (2x2 contingency table).

	Never Smoker	**Smoker**	**Total**
Control	467	1038	1505
Case	50	1150	1200
Total	517	2188	2705

from Remen, T. et al BMC Cancer 18, 1275 (2018)

Here the estimated odds ratio for is $\frac{1150 \times 467}{50 \times 1038} = 10.35$

4.3.1 CI for the odds ratio

Since the odds ratio is a multiplicative measure (with 2 being the inverse of 0.5) it is natural to construct symmetric CIs for the logarithmic transformation of the odds ratio and then back transform these.

To construct a 95% CI for the population odds ratio we therefore use the sampling distribution of log($\hat{\Psi}$), which for large sample sizes is normally distributed. It can be shown that the standard error of log($\hat{\Psi}$) is given in approximation by:

$$SE\left(\log\left(\widehat{\Psi}\right)\right) = \sqrt{\frac{1}{N\pi_{00}} + \frac{1}{N\pi_{01}} + \frac{1}{N\pi_{10}} + \frac{1}{N\pi_{11}}}$$

Each of the$N\pi_{ij}$terms in this expression is estimated by an observed count in the contingency table, facilitating construction of the CI.

Example (continued):

$$log\left(\widehat{\Psi}\right) = log\left(\frac{267 \times 541}{213 \times 281}\right) = 2.337$$

$$\mathrm{SE}\left\{log\left(\widehat{\Psi}\right)\right\} = \sqrt{\frac{1}{467} + \frac{1}{1038} + \frac{1}{50} + \frac{1}{1150}} = 0.1548$$

Therefore a 95% CI for log(Ψ) is 2.337 ± 1.96 x 0.1548 = (2.0333, 2.6403) and a 95% CI for Ψ is (exp(2.0333), exp(2.6403)) = (7.64, 14.02).

4.3.2 Hypothesis tests of H_0: Ψ = 1

If there is no association between two binary variables the odds ratio is 1. This null hypothesis can be tested through calculation of the logarithm of the odds ratio and its standard error as described above. However it is more usual to carry out either the χ^2 goodness of fit test or Fisher's exact test as described in the following sections.

The null hypothesis here can either be expressed in terms of the odds ratio being equal to 1, or, equivalently, as independence of the distributions of *X* and *Y.* This can be written as:

$$H_0: \pi_{ij} = \pi_{i.}\pi_{.j} \text{ for all [i, j]}$$

That is, $\frac{\pi_{i1}}{\pi_{.1}} = \frac{\pi_{i0}}{\pi_{.0}} = \pi_{i.}$ the probability that X = *i* does not depend on Y

Exercise 4: Prove that this implies that the odds ratio equals unity.

The alternative hypotheses here is:

$$H_1 : \pi_{ij} \neq \pi_{i.}\pi_{.j} \text{ for at least one } \{i, j\}.$$

4.3.3 The chi-squared test for comparing two proportions

Provided that the entries in the contingency table are reasonably large the null hypothesis of no association can be tested with a χ^2 test. The test statistic is as follows:

$$\chi^2 = \sum_i \sum_j \left(\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \right) \qquad \text{where } \mathrm{E_{ij}} = \mathrm{O_{i.}O_{.j}/O_{..}} \qquad (9)$$

Under H_0: $\chi^2 \sim \chi_1^2$. In order to calculate the *p*-value it is necessary to only consider the upper tail of the χ_1^2distribution. This is so because all departures from the null hypotheses lead to expected values of χ^2 that are greater than 1 (the expectation under the null hypothesis). Despite this, this and all other χ^2 tests for contingency tables are classified as two-sided hypothesis tests, because the alternative hypothesis is two-sided, not one-sided.

Example (continued): The table below shows the expected cell counts (*E_{ij}*) under the null hypothesis of no association between the probability of lung cancer and smoking. Note: 287.6 = 517 x 1505/2705 etc.

Table 4.4: Expected counts under the null hypothesis.

	Never Smoker	**Smoker**	**Total**
Control	287.6	1217.4	1505
Case	229.4	970.6	1200
Total	517	2188	2705

Equation (9) gives:

$$\chi^2 = \left[\frac{(467-287.6)^2}{287.6} + \frac{(1038-1217.4)^2}{1217.4} + \frac{(50-229.4)^2}{229.4} + \frac{(1150-970.6)^2}{970.6} \right] = 311.6$$

Comparison with tables of χ_1^2gives $p < 0.0001$. There is very strong evidence that the probability of lung cancer differs between smokers and non-smokers.

It is possible to use a continuity correction to allow for the fact that the observed values can only be integers, but that the chi-square distribution is continuous. Subtracting 0.5 from the absolute difference between observed and expected cell counts will give *p*-values that are typically closer to those obtained from Fisher's exact test (see below). The required modification to the formula is:

$$\chi^2 = \sum_i \sum_j \left(\left(\left| O_{ij} - E_{ij} \right| - 0.5 \right)^2 / E_{ij} \right) \qquad (10)$$

4.3.4 Fisher's "exact" test

The chi-square test relies on an approximation that is not valid in small samples. It is often suggested that *expected* cell counts should all be greater than 5 for the test to be reliable. For small sample sizes, we can use Fisher's exact test. It is a conditional test that assumes that the row and column totals are fixed. For fixed row and column totals all other counts in the table can be calculated provided that a single count is known. Tables can therefore be uniquely indexed by a single count. Most commonly the count in the top left cell (O_{00}) is used for this purpose.

To perform Fisher's exact test, we need to calculate the probability of the observed arrangement of the counts within the table (or more extreme ones) when the null hypothesis is true. Fisher showed that under the null hypothesis, the probability of the observed table (given the marginal totals are fixed) is:

$$P_{O_{00}} = \mathrm{Prob}(O_{00}, O_{01}, O_{10}, O_{11} | O_{0.}, O_{1.}, O_{.0}, O_{.1}) = \frac{O_{0.}!O_{1.}!O_{.0}!O_{.1}!}{O_{..}!O_{00}!O_{01}!O_{10}!O_{11}!}.$$

Therefore the one-sided p-value can be calculated as follows:

If $\widehat{\Psi} < 1$ then calculate $P_- = P_0 + P_1 + \cdots + P_{O_{00}}$.

If $\widehat{\Psi} > 1$ then calculate $P_+ = P_{O_{00}} + P_{O_{00}+1} + \ldots + \min(P_{O_{0.}}, P_{O_{.0}})$.

There are a number of different approaches to calculating the two-sided p-value. The simplest is to double the one-sided value. An alternative is to calculate the probability of all possible tables and sum all the probabilities which are less than or equal to$P_{O_{00}}$.

As explained above, Fisher's exact test calculates probabilities conditional on the row and column totals in the table. This is arguably not unreasonable since the row and column totals carry no information about the magnitude of the association between the two variables. However it has been shown (Lydersen S, Fagerland MW and Laake P, *Statistics in Medicine*: 28:1159-1175, 2009) that when row and/or column totals are not fixed by design then this test lacks statistical power compared with other, more computationally complex, procedures that take account of the sampling variability in these marginal totals.

Example 4.2: In a pilot study comparing a new treatment (NT) versus standard of care (SOC) in a potentially fatal condition the results after 30 days were are follows.

	Group		
Outcome	SOC	NT	**Total**
Survived	1	5	6
Died	8	2	10
Total	9	7	16

First calculate a 1-sided p-value for the above table. Since $(1 \times 2)/(5 \times 8) < 1$, calculate P_- = $P_0 + P_1$, where:

1	5
8	2

gives$P_1 = \frac{6!10!9!7!}{16!1!5!8!2!}$ =0.0236 and

0	6
9	1

gives P_0 =0.00087. (This is a "more extreme" arrangement with the same marginal totals)

Sum these two probabilities to get Fisher's one-tailed *p*-value of 0.02447. Simply doubling this gives a two-sided *p*-value of 2 × 0.02447 = 0.0489.

To calculate the exact two sided *p*-value, look at all possible tables and their corresponding probabilities.

O_{00}	O_{01}	O_{10}	O_{11}	Probability
0	6	9	1	0.00087
1	5	8	2	0.02360
2	4	7	3	0.15734
3	3	6	4	0.36713
4	2	5	5	0.33042
5	1	4	6	0.11014
6	0	3	7	0.01049

The exact *p*-value = 0.02360 + 0.00087 + 0.01049 = 0.03496. We conclude that there is evidence at the 5% significance level of an association between treatment and outcome, with those on the new treatment more likely to survive.

4.4 Association between pairs of (un-ordered) categorical variables

The chi-squared test introduced above can be extended to test for associations between categorical variables with more than two possible values.

Consider a contingency table in which two categorical variables (*X* and *Y)* are tabulated against each other. Denote by O_{ij} the number of occasions in which *X* takes the *i*th value (*i = 1,2,..,m*) and *Y* takes the *j*th value (*j = 1,2,..,n*)*.* The data can be displayed as follows:

Table 4.5: Observed data in a *M*x*N* contingency table

	Y = 1	*Y = 2*	.	*Y = n*	Total
X = 1	O_{11}	O_{12}	.	O_{1n}	$O_{1.}$
X = 2	O_{21}	O_{22}	.	O_{2n}	$O_{2.}$
.	.	.	.	.	.
$X_1 = m$	O_{m1}	O_{m2}	.	O_{mn}	$O_{m\cdot}$
Total	$O_{.1}$	$O_{.2}$	.	$O_{.n}$	$O_{..}$

The null hypothesis of no association between the two categorical variables is expressed in terms of the true classification probabilities - π_{ij} :

Table 4.6: Population parameters in an *M*x*N* contingency table

	Y = 1	*Y = 2*	.	*Y = n*	Total
X = 1	π_{11}	π_{12}	.	π_{1n}	$\pi_{1.}$
X = 2	π_{21}	π_{22}	.	π_{2n}	$\pi_{2.}$
.	.	.	.	.	.
$X_1 = m$	π_{m1}	π_{m2}	.	π_{mn}	$\pi_{m\cdot}$
Total	$\pi_{.1}$	$\pi_{.2}$	.	$\pi_{.n}$	1

The null and alternative hypotheses can be written as:

$$H_0\colon \pi_{ij} = \pi_{i.}\pi_{.j} \text{ for all } \{i,j\}.$$
$$H_1\colon \pi_{ij} \neq \pi_{i.}\pi_{.j} \text{ for at least one } \{i,j\}.$$

The null hypothesis of no association is tested with a χ^2 test.

$$\chi^2 = \sum_i \sum_j \left(\left(O_{ij} - E_{ij}\right)^2 / E_{ij} \right) \; where \; E_{ij} = O_{i.}O_{.j}/O_{..}$$

$$\text{Under } \mathrm{H}_0\colon \chi^2 \sim \chi^2_{(m-1)(n-1)}$$

Notice that the alternative hypothesis does not specify that $\pi_{ij} \neq \pi_{i.}\pi_{.j}$ for all categories of *X* and *Y.* This means that conclusions from a statistically significant test are non-specific about which categories of the two variables are non-independent.

Example 4.3: Frequency of ABO blood groups in three patient groups (Armitage and Berry, 2nd ed., p 376).

	Peptic Ulcer	Gastric Cancer	Controls	Total
A	983	383	2892	4258
B	679	416	2625	3720
O	134	84	570	788
Total	1796	883	6087	8766

The key question here is 'Does the *proportion* of individuals with each blood group differ between patient groups?'

$$\chi^2 = \frac{\left(983 - \frac{1796 \times 4258}{8766}\right)^2}{\left(\frac{1796 \times 4258}{8766}\right)} + 8\, more\ terms$$

$$= 40.54\ \ (p < 0.001\ from\ \chi^2_{4df}\ table$$

Exercise 4.5: What do you conclude from this test? What might you do next?

Note: The above general chi-squared test is not recommended for use when categories have an ordering (*e.g.,* 'mild', 'moderate', 'severe' disease). With this type of data the general chi-squared test lacks statistical power against plausible alternative hypotheses indicative of a trend in proportions. See the course on generalised linear models for further details.

Chapter 5: Comparisons

Objectives

By the end of this session students will be able to:

- Appreciate the fundamental role that comparison plays in medical statistics.
- Be able to use statistical techniques to compare two population means.
- Be able to use statistical techniques to compare two population variances
- Be able to use statistical techniques to compare two population proportions.

5.1 Introduction

A great deal of research activity involves the comparison of two or more groups. Although methods of analyses that involve comparisons of more than two groups can be applied to two groups, it is convenient to consider the case of two groups separately as the methods can be simplified and there are fewer problems of interpretation.

Consider a typical example of a comparative study involving two groups. Suppose that two treatments for high blood pressure are to be investigated: the investigator selects a group of hypertensive patients and randomly assigns them to receive one of the two therapies. Both groups are followed for a certain period of time and their blood pressures at the end of this period are recorded. The aim of the analysis is to compare the mean blood pressure after treatment in those given therapy 1 to that in those given therapy 2. Note that the random assignment of treatment ensures that any difference seen between the two groups can be attributed to the treatments under investigation, rather than to some other factor.

One key assumption for all the methods in this section is independence of results conditional on the parameters implied by the design (in the study above the two parameters are the group specific means). Consider, for comparison, a different experiment in which once a patient is chosen to receive treatment 1, his/her sex, age and blood pressure are recorded and only an individual with similar, sex, age and initial blood pressure is then assigned to treatment 2. Obviously with such matching, two matched subjects would tend to have more similar blood pressures than two patients not so matched, thus violating the assumption of independence.

We will discuss comparative parametric methods that deal with two independent samples coming from the normal and binomial distributions. Most parametric methods are essentially similar in the sense that we use the sampling distributions of relevant sample statistics to construct CI (CI) for differences in means, proportions or rates – or perform hypothesis tests concerning specific values of these parameters.

As in earlier sessions we start by considering the comparison of two means when the population variances are known. This situation is somewhat artificial, but is useful for setting ideas.

5.2 Case Study 1: Comparing two population means

5.2.1 Comparison of two population means, when the outcome variable is normally distributed and when the population variances are known

Let Y_{1i} (i=1,2,..,n_1) and Y_{2i} (i=1,2,...,n_2) be two independent random samples selected from populations 1 and 2 respectively where each variable is normally distributed with mean μ_k and known variance σ_k^2.

E(Y_{ki}) = μ_k and Var(Y_{ki}) = σ_k^2 for k = 1, 2 and i = 1, 2, ..., n_k.

A natural estimate for μ_k is the sample mean$\bar{Y}_k$, where:

$$\bar{Y}_k \sim \mathrm{N}\left(\mu_k, \frac{{\sigma_k}^2}{n_k}\right) \qquad \text{for } k = 1, 2.$$

Given that the two samples are independent, we know that the distribution of the difference between the two means is also normally distributed with the following mean and variance.

$$\bar{Y}_2 - \bar{Y}_1 \sim N\left(\mu_2 - \mu_1, \frac{\sigma_2^2}{n_2} + \frac{\sigma_1^2}{n_1}\right) \qquad (1)$$

$$\Leftrightarrow Z = \frac{\bar{Y}_2 - \bar{Y}_1}{\sqrt{(\sigma_2^2/n_2) + (\sigma_1^2/n_1)}} \sim N\left(\frac{\mu_2 - \mu_1}{\sqrt{(\sigma_2^2/n_2) + (\sigma_1^2/n_1)}}, 1\right) \qquad (2)$$

The sampling distribution for $\bar{Y}_2$ - $\bar{Y}_1$ can then be used to construct a 100(1-α)% CI for $(\mu_2 - \mu_1)$ or to perform hypothesis tests of specific values for $(\mu_2 - \mu_1)$.

Using methods analogous to those from the session on CIs (section 2.4), we can find L and U so that Prob{$(\mu_2 - \mu_1) \in (L, U)$}= 100(1-α)%. These are:

$$L = (\bar{Y}_2 - \bar{Y}_1) + z_{\alpha/2}\sqrt{\frac{\sigma_2^2}{n_2} + \frac{\sigma_1^2}{n_1}} \quad \text{and } U = (\bar{Y}_2 - \bar{Y}_1) + z_{1-\alpha/2}\sqrt{\frac{\sigma_2^2}{n_2} + \frac{\sigma_1^2}{n_1}}.$$

Due to the symmetry of the standard normal distribution$z_{\alpha/2} = -z_{1-\alpha/2}$. Hence, a 100(1-α)% CI for $(\mu_2 - \mu_1)$ is given by:

$$(\bar{Y}_2 - \bar{Y}_1) \pm z_{1-\alpha/2}\sqrt{\frac{\sigma_2^2}{n_2} + \frac{\sigma_1^2}{n_1}} \qquad (3)$$

To carry out hypothesis tests an analogous approach to that in the lecture on hypothesis tests (section 3.6) can be adopted. Usually we wish to test the null hypothesis that:

H$_0$: $(\mu_2 - \mu_1)$ = 0 against the alternative hypothesis H$_1$: $(\mu_2 - \mu_1) \neq 0$.

If H$_0$ is true, then Z in equation (2) follows a standard normal distribution (with zero mean). Therefore (using the same approach as in 3.6), given observed values of $\bar{Y}_1 = \bar{y}_1$and $\bar{Y}_2 = \bar{y}_2$ the two-sided *p*-value is:

$$2\left[1 - \Phi\left(\frac{|\bar{y}_2 - \bar{y}_1|}{\sqrt{(\sigma_2^2/n_2) + (\sigma_1^2/n_1)}}\right)\right] \qquad (4)$$

whereΦ is the distribution function for a standard normal distribution.

In constructing CIs and performing hypothesis tests the following assumptions are made.

(1) The observations Y_{ki} in each group follow a normal distribution.

(2) All observations are independent.

(3) The two population variances $\sigma_1{}^2$ and $\sigma_2{}^2$ are known.

The violation of the first assumption is not serious as long as the sample size is large, since then, like $\bar{Y}_1$ and $\bar{Y}_2$, $(\bar{Y}_2 - \bar{Y}_1)$ will be approximately normally distributed. Violation of assumption (2) is usually more serious. In reality assumption (3) rarely holds: so we now consider the situation where the variances are not known.

5.2.2 Comparison of two population means, when the outcome variable is normally distributed and when the population variances are unknown, but are assumed equal

In the previous section adding the assumption that $\sigma_1{}^2$= $\sigma_2{}^2$= σ^2, equation (1) becomes:

$$\bar{Y}_2 - \bar{Y}_1 \sim N\left(\mu_2 - \mu_1, \sigma^2\left(\frac{1}{n_2} + \frac{1}{n_1}\right)\right).$$

However, we can no longer use this distribution to make inferences about $(\mu_2 - \mu_1)$ since it depends on the unknown parameterσ^2. In order to proceed we first need to estimate the common variance σ^2. When two samples are involved the best unbiased estimate of σ^2 is a weighted average of the two sample estimates$\hat{\sigma}_1^2$and$\hat{\sigma}_2^2$ as given below:

$$\hat{\sigma}^2 = \frac{(n_1 - 1)\hat{\sigma}_1^2 + (n_2 - 1)\hat{\sigma}_2^2}{n_1 + n_2 - 2} \qquad (5)$$

Since $\frac{(n_1-1)\hat{\sigma}_1^2}{\sigma^2} \sim \chi^2_{n_1-1}$ and $\frac{(n_2-1)\hat{\sigma}_2^2}{\sigma^2} \sim \chi^2_{n_2-1}$ it follows that:

$$\frac{(n_1+n_2-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2}.$$

It follows that the random variable *T* (defined below) follows a *t*-distribution with n_1+ n_2 - 2 degrees of freedom,

$$T = \frac{(\bar{Y}_2 - \bar{Y}_1) - (\mu_2 - \mu_1)}{\hat{\sigma}\sqrt{(1/n_2) + (1/n_1)}}.$$

This random variable can be used to perform hypothesis tests and to construct CIs for$(\mu_2 - \mu_1)$. The test is known as the two-sample t-test. Using the same principle of construction as in 2.6 it can be shown that a 100(1-α)% CI for $(\mu_2 - \mu_1)$ is as follows:

$$(\bar{Y}_2 - \bar{Y}_1) \pm t_{n_1+n_2-2,1-\alpha/2}\hat{\sigma}\sqrt{(1/n_2) + (1/n_1)}.$$

Example 5.1: Table 5.1 below, shows the age (in months) of 11 infants at the time when they started to walk alone. These infants, as new-borns, were randomised to two groups, an eight week active exercise group and an eight week control group.

Test the null hypothesis that the group means are the same and construct a 95% CI for the difference in mean age at walking between the two groups.

Table 5.1: Children's ages at time of first walking alone by randomisation group

	Age in months for walking alone	
	Active exercise group (*i*=1)	Eight week control group (*i*=2)
	9.00, 9.50, 9.75, 10.00, 13.00, 9.50	13.25, 11.50, 12.00, 13.50, 11.50
n_i	6	5
$\bar{Y}_i$	10.125	12.350
$\hat{\sigma}_i$	1.447	0.962

Test H_0:$(\mu_2 - \mu_1)$ = 0 against the alternative hypothesis H_1:$(\mu_2 - \mu_1) \neq 0$.

Assuming that the two variances are equal, the common sample estimate of the variance at time of first walking is:

$$\hat{\sigma}^2 = \frac{(6-1)(1.447)^2 + (5-1)(0.962)^2 +}{6+5-2} = \frac{14.172}{9} = 1.575.$$

If H_0 is true then the *T*-statistic will follow a *t*-distribution with 9 degrees of freedom. Here the test statistic is given by:

$$T = \frac{(12.350 - 10.125)}{\sqrt{1.575((1/5)+(1/6))}} = \frac{2.225}{0.76} = 2.928.$$

From Neave's table, page 20, $t_{9,0.975}$ =2.2622 and $t_{9,0.99}$ =2.8214. Since the observed |*t*/>$t_{9,0.99}$, we conclude that we can reject the null hypothesis with *p*-value < 0.02.

A 95% CI for $(\mu_2 - \mu_1)$ is given by:

$$(\bar{Y}_2 - \bar{Y}_1) \pm t_{9,0.975}\hat{\sigma}\sqrt{(1/n_2) + (1/n_1)} = 2.225 \pm 2.2622 \times 0.76 = (0.51, 3.94).$$

The procedures described above can be carried out in a number of different ways in Stata. The most straightforward (illustrated using the data from exercise 1) is the following.

```
ttest age, by(group)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       1 |       6      10.125     .590727     1.44698    8.606488    11.64351
       2 |       5       12.35    .4301163    .9617692    11.15581    13.54419
---------+--------------------------------------------------------------------
combined |      11    11.13636    .5015472    1.663444    10.01885    12.25388
---------+--------------------------------------------------------------------
    diff |              -2.225    .7597667               -3.943712   -.5062884
------------------------------------------------------------------------------
    diff = mean(1) - mean(2)                                      t =  -2.9285
Ho: diff = 0                                     degrees of freedom =        9

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0084         Pr(|T| > |t|) = 0.0168          Pr(T > t) = 0.9916
```

Notice that, in contrast to the presentation of the theory above, Stata reports inferences concerning $(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})$ not$(\boldsymbol{\mu_2} - \boldsymbol{\mu_1})$, *i.e.,* the estimated difference in means is -2.225 months, not 2.225 months.

An equivalent analysis (but giving inferences relating to$(\boldsymbol{\mu_2} - \boldsymbol{\mu_1})$) can be carried out using a linear regression model as follows. The link between the two approaches is discussed in the regression course.

```
. regress age i.group

      Source |       SS       df       MS              Number of obs =      11
-------------+------------------------------           F(  1,     9) =    8.58
       Model |  13.5017045     1  13.5017045           Prob > F      =  0.0168
    Residual |    14.16875     9  1.57430556           R-squared     =  0.4879
-------------+------------------------------           Adj R-squared =  0.4311
       Total |  27.6704545    10  2.76704545           Root MSE      =  1.2547

------------------------------------------------------------------------------
         age |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     2.group |    2.225   .7597667     2.93   0.017      .5062884    3.943712
       _cons |   10.125   .5122346    19.77   0.000      8.966245    11.28376
------------------------------------------------------------------------------
```

5.2.3 Comparison of two population means, when the outcome variable is normally distributed and when the population variances are unknown and not assumed equal

Knowledge of the outcome variable and/or a test for equality of variances (see 5.3) can suggest that the assumption of equal variances in 5.2.2 is not justified. A number of different approaches can be adopted.

One approach is to transform the data (for example for each Y_i calculate $X_i = \log(Y_i)$) so that the chosen transformation stabilises the variance and then apply the method described above to the transformed data (for more on transformations, see Analytical Techniques 6).

Alternatively, both unknown variances can be separately estimated ($\hat{\sigma}_1^2$and$\hat{\sigma}_2^2$) from their respective samples. Provided that n_1 and n_2 are large the distribution of the estimated difference in means divided by its estimated standard error $\left(\sqrt{(\hat{\sigma}_1^2/n_1) + (\hat{\sigma}_2^2/n_2)}\right)$ approximates to a standard normal distribution and so equations (3) and (4) can be used to construct CIs and perform hypothesis tests respectively, merely replacing both population variances by their estimated values.

However, if n_1 and n_2 are not large (<30 for example) then a better approximation, due to Satterthwaite, is that the estimated difference in means divided by its estimated standard error follows a *t*-distribution with (non-integer) degrees of freedom *n** equal to:

$$n * = \frac{\left(\frac{\hat{\sigma}_1^2}{n_1}+\frac{\hat{\sigma}_2^2}{n_2}\right)^2}{\left[\left(\frac{\hat{\sigma}_1^2}{n_1}\right)^2 \Big/ (n_1-1)\right]+\left[\left(\frac{\hat{\sigma}_2^2}{n_2}\right)^2 \Big/ (n_2-1)\right]}.$$

A further alternative is to use one of the techniques taught in the robust methods module.

Stata allows the two-sample t-test to be carried out using Satterthwaites's method of calculating the degrees of freedom. As illustrated below, for the data in example 1 the Satterthwaite degrees of freedom are 8.66.

```
. ttest age, by(group) unequal

Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       1 |       6      10.125     .590727     1.44698    8.606488    11.64351
       2 |       5       12.35    .4301163    .9617692    11.15581    13.54419
---------+--------------------------------------------------------------------
combined |      11    11.13636    .5015472    1.663444    10.01885    12.25388
---------+--------------------------------------------------------------------
    diff |              -2.225    .7307245               -3.887862   -.5621381
------------------------------------------------------------------------------
    diff = mean(1) - mean(2)                                      t =  -3.0449
Ho: diff = 0                     Satterthwaite's degrees of freedom =   8.6632

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0073         Pr(|T| > |t|) = 0.0145          Pr(T > t) = 0.9927
```

5.3 Case Study 2: Comparing two population variances

5.3.1 Hypothesis test

The t-test described in section 5.2.2 above depends on the assumption (among others) that the two population variances are the same. It is therefore important to be able to carry out a test that investigates the appropriateness of this assumption. The simplest is the F-test which examines the ratio of the two variances.

It can be either a two-tailed or a one-tailed test. The two-tailed version is a test against the alternative that the variances are unequal. The one-tailed version only allows for a difference in one direction: that is the variance in the first population is either greater than or less than (but not both) the variance in the second population. In most instances, as with tests comparing means, we are interested in two-tailed tests.

As above let Y_{1i} (i=1,2,..,n_1) and Y_{2i} (i=1,2,...,n_2) be two independent random samples selected from populations 1 and 2 respectively where each variable is normally distributed with mean μ_k and known variance σ_k^2. The test statistic is the ratio of the estimates of the two variances$F = \hat{\sigma}_1^2/\hat{\sigma}_2^2$. The more this ratio deviates from 1, the stronger the evidence for unequal population variances.

If two random variables follow independent χ^2 distributions, then their ratio (scaled by their respective degrees of freedom) follows an F-distribution. Formally this can be written:

$$\text{If } A\sim\chi_a^2 \text{ and } B\sim\chi_b^2 \text{ independently then } F = \frac{A/a}{B/b}\sim F_{a,b} \ .$$

For the variance ratio test $F = \hat{\sigma}_1^2/\hat{\sigma}_2^2 \sim F_{n_1-1,n_2-1}$ (an F-distribution with n_1 -1 and n_2 -1 degrees of freedom) under the null hypothesis that the variances are equal. To test the null hypothesis, compare the value of this ratio with the appropriate percentiles of the F-distribution.

In order to carry out the test, without loss of generality make the first sample the one with the **larger** estimated variance, so that F takes a value f that is >1. The one-sided p-value is simply the probability that F is greater than or equal to f under the null hypothesis. There are two ways of calculating the two-sided p-value. The first is to add the one-sided p-value to the probability that F is less than 1/f under the null hypothesis (these two probabilities are equal if and only if $n_1 = n_2$). The second is simply to double the one-sided p-value. This approach is that used by Stata. If carrying out the test by hand with the aid of Neave's tables, then it is reasonable to simply refer the test statistic to the 97.5th percentile of the appropriate F-distribution in order to determine whether or not $p < 0.05$ using a two-sided test.

Example 5.2: From Example 5.1 we have data from two groups; the first has n_1 = 6, and$\hat{\sigma}_1^2$=2.094. The second has n_2 = 5 and $\hat{\sigma}_2^2$=0.925, (degrees of freedom are 5 and 4 respectively).

Under H$_0$: $F = \hat{\sigma}_1^2/\hat{\sigma}_2^2$~ $F_{5,4}$.

Taking the larger variance in the numerator the test statistic is 2.093/0.925 = 2.264. From the Neave tables, the 97.5th percentile of $F_{5,4}$ is 9.364. Therefore there is no evidence to reject the null hypothesis (that the variance in the two groups is equal) at the 5% level.

Stata has a command, **sdtest**, that is able to carry out the variance ratio test. Its use with the data from exercise 1 is shown below.

```
. sdtest age, by(group)
Variance ratio test
------------------------------------------------------------------------------
   Group |  Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       1 |    6      10.125     .590727     1.44698    8.606488    11.64351
       2 |    5       12.35    .4301163    .9617692    11.15581    13.54419
---------+--------------------------------------------------------------------
combined |   11    11.13636    .5015472    1.663444    10.01885    12.25388
------------------------------------------------------------------------------
    ratio = sd(1) / sd(2)                                         f =   2.2635
Ho: ratio = 1                                  degrees of freedom =     5, 4

    Ha: ratio < 1               Ha: ratio != 1                 Ha: ratio > 1
  Pr(F < f) = 0.7756         2*Pr(F > f) = 0.4488          Pr(F > f) = 0.2244
```

The variance ratio test depends heavily on the assumption that the data come from a normal distribution. Stata also has a test that is more robust to non-normality, called Levene's test.

Robust tests for equality of variance include Levene's, but also modifications suggested by Brown & Forsythe. See "sdtest" in the Stata manual for more details.

5.3.2 CIs

It is also possible to obtain CIs for the ratio of the variances. The approach has parallels with that introduced in session 2 (section 2.7) for construction of a CI for a (single) variance. The key result that is needed for their construction is that the *F*-statistic introduced above, when divided by the ratio of population variances, follows an F-distribution. Specifically

$$\frac{F}{\sigma_1^2/\sigma_2^2} = \frac{\hat{\sigma}_1^2/\hat{\sigma}_2^2}{\sigma_1^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}.$$

Using an approach analogous to that in 2.6 it follows that a 95% CI for the ratio of the population variances (σ_1^2/σ_2^2) is

$$\left(\frac{F}{F_{n_1-1,n_2-1,0.975}}, \frac{F}{F_{n_1-1,n_2-1,0.025}}\right) \tag{6}$$

Note that the values of $F_{m,n,p}$ are only tabulated for values of $F \geq 1$. So you can only directly read off *p*-values corresponding to the right tail of the *F* distribution. To get lower tail *p*-values, observe that if$F \sim F_{a,b}$ then $1/F \sim F_{b,a}$. It follows that the required 95% CI in (6) can be re-written as

$$\left(\frac{F}{F_{n_1-1,n_2-1,0.975}}, F \times F_{n_2-1,n_1-1,0.975}\right)$$

Example 5.3: To construct a 95% CI for the example above, we use the upper 97.5th percentile points of the $F_{5,4}$ and $F_{4,5}$ distributions, obtaining

$$\left(\frac{2.264}{F_{5,4,0.975}}, 2.264 \times F_{4,5,0.975}\right) = \left(\frac{2.264}{9.364}, 2.264 \times 7.388\right) = (0.24, 16.73).$$

5.4 Case Study 3: Comparing two population proportions

In this section we are concerned with comparing population proportions, *i.e.,* investigating the extent to which some probability of 'success' differs between two groups. One approach to this situation is to investigate the association between the binary grouping variable and the binary outcome variable using the techniques introduced in Analytical Techniques 4. The χ^2 test or Fisher's exact test can be used to test for differences in the probability of 'success' and the odds ratio used to quantify the magnitude of the association.

In this section we shift the focus of the analysis onto i) differences in the probabilities of success between the groups and ii) the ratio of the probabilities of success, drawing parallels between this approach and that using odds ratios where appropriate.

5.4.1 Inference for the difference between two population proportions

Let R_1 and R_2 be the number of successes out of two independent samples of size n_1 and n_2 respectively. Similarly let, P_1 and P_2 be the proportion of successes in these two samples. Assuming that $R_k \sim \text{Bin}(n_k, \pi_k)$ for k=1,2 then we know that:

$$E(P_k) = \pi_k \text{ and } V(P_k) = \frac{\pi_k(1-\pi_k)}{n_k} \qquad \text{for } k = 1, 2 \text{ and } P_1 \text{ and } P_2 \text{ independent.}$$

When n_k is large, the distribution of each proportion can be approximated by the following normal distribution:

$$P_k \sim \mathrm{N}\left(\pi_k, \frac{\pi_k(1-\pi_k)}{n_k}\right) \quad \text{for } k = 1, 2.$$

Given that the two samples are independent, it is easy to show that the distribution of the difference between the two proportions is also approximately normally distributed, if both sample sizes are large enough:

$$P_2 - P_1 \sim \mathrm{N}\left(\pi_2 - \pi_1, \frac{\pi_2(1-\pi_2)}{n_2} + \frac{\pi_1(1-\pi_1)}{n_1}\right) \qquad (7)$$

In order to use this approximate sampling distribution for (P_2 –P_1) to construct a 100(1-α)% CI for $\pi_2 - \pi_1$, we need to estimate the variance of the sampling distribution, which depends on the unknown parameters π_1 and π_2. For large *n, P_k* provides a good approximation toπ_k. Thus a 100(1-α)% CI for (π_2 - π_1) is given by:

$$(P_2 - P_1) \pm z_{1-\alpha/2}\sqrt{\frac{P_2(1-P_2)}{n_2} + \frac{P_1(1-P_1)}{n_1}}.$$

Analogously suppose that we wish to test the null hypothesis.

H_0: $(\pi_2 - \pi_1)$ = 0 (or equivalently that $\pi_2 = \pi_1 = \pi$) against the alternative H_1:$(\pi_2 - \pi_1) \neq$ 0.

We can use the sampling distribution of the following random variable *Z* that follows an approximate standard normal distribution:

$$Z = \frac{P_2 - P_1}{\sqrt{P(1-P)((1/n_2)+(1/n_1))}}\ .$$

We estimate *P* by$(R_1 + R_2)/(n_1 + n_2)$, the estimate of the probability of success under the null hypothesis.

Exercise 5.1(difficult!): Show that the square of this test statistic (Z^2) is numerically equal to the test statistic from the χ^2 test for association and hence that the two approaches yield identical results.

Example 5.4: In a clinical trial of D-Penicillamine (D-P) for treatment of primary biliary cirrhosis, the following results were obtained:

	Treatment		
	Placebo	D-P	
Died	16	18	34
Alive	21	43	64
	37	61	98

Test the null hypothesis that there is "no difference in the probability of death between the groups".

P_2=18/61 = 0.2951, and P_1 = 16/37 = 0.4324 .

$$Z = \frac{(0.2951 - 0.4324)}{\sqrt{\frac{34}{98}\left(1-\frac{34}{98}\right)\left(\frac{1}{61}+\frac{1}{37}\right)}}$$ = -1.385 and hence the *p*-value = 0.17 .

Conclusion: There is no evidence to reject the null hypothesis of no association between treatment and death.

A 95% CI for the difference in the probability of death given treatment is:

$$(P_2 - P_1) \pm 1.96\sqrt{\frac{P_2(1-P_2)}{n_2} + \frac{P_1(1-P_1)}{n_1}} = 0.2951\text{-}0.4324 \pm 1.96(0.1002) = (-0.33, 0.06).$$

As anticipated from the result of the hypothesis test, the 95% CI does include the null value of zero. It also includes quite notable differences between the treatments.

5.4.2 Inference for the relative risk comparing the two population proportions (π_2/π_1)

In epidemiology, there is a special interest in comparing two groups with respect to the risk of some event. In a prospective study, groups of subjects with different characteristics are followed up to see whether an outcome of interest occurs. Many clinical trials are like this (see example 1 above), but so too are observational studies where it is not possible to randomise the feature of interest (*e.g.,* blood group). We can calculate the proportions having the outcome in each group, and the ratio of these two proportions is a measure of the raised risk in one group compared to the other. We usually term this ratio the *relative risk* (or the *risk ratio*). It must be remembered that although the statistical arithmetic is the same in a randomised controlled trial as in an observational study, the strength of evidence is not the same.

From our sample the estimator of the relative risk is *RR* = P_2/P_1. We have seen that when *n* is large, P_k is approximately normally distributed. However the ratio of two normal distributions is not itself normal. To construct a 95% CI for the population risk ratio therefore, we use instead the sampling distribution of log(P_2/P_1) which for large sample sizes is approximately normally distributed.

For Y_k = log(P_k) it can be shown (using techniques to be introduced in AT 6) that V(Y_k) = $(1-\pi_k)/n_k\pi_k$ and thus that an estimator of this variance is $(1-P_k)/R_k$.

Hence an approximate 95% CI for log(π_2 / π_1) is given by:

$$log\left(\frac{P_2}{P_1}\right) \pm 1.96\sqrt{\frac{(1-P_2)}{R_2} + \frac{(1-P_1)}{R_1}}.$$

Denoting this interval by (*L*, *U*) a 95% CI for (π_2 / π_1) is given by (exp(*L*) , exp(*U*)).

Example 5.5: Using the data from Example 5.4, provide a 95% CI for the relative risk of death between the treatment and placebo groups.

P_2/P_1 = 0.2951/0.4324 = 0.6825

An approximate 95% CI for log(π_2 / π_1) is given by:

$$\log\left(\frac{P_2}{P_1}\right) \pm 1.96 \times \sqrt{\frac{(1-P_2)}{R_2} + \frac{(1-P_1)}{R_1}} = -0.3820 \pm 1.96 \times \sqrt{\frac{0.7049}{18} + \frac{0.5676}{16}}$$

$$= (\text{-}0.9175,\ 0.1535)\ .$$

So an approximate 95% CI for (π_2/π_1) = (exp(-0.9175) , exp(0.1535)) = (0.400, 1.166).

We conclude that the risk of death in the D-Penicillamine group is 68.3% that in the Placebo group with 95% CI (40.0% to 116.6%). Alternatively we might say that the risk of death is reduced by 31.7% with 95% CI (60.0% reduction to 16.6% increase). The size of the CI shows that the data are compatible with very different relative risks ranging from a large benefit to a small hazard.

Chapter 6: Assumptions and Transformations

Objectives

By the end of this session students will be able to:

- Appreciate the importance of assumptions in statistical procedures.
- Explore the validity of assumptions in simple settings.
- Be aware of alternative approaches to analysis that can be used when assumptions made by 'standard' techniques do not hold.
- Understand the rationale for transformation of variables.
- Calculate the mean and variance of transformed variables using variance transformation formulae.

6.1 Introduction

Almost all the statistical procedures met in the course so far and those to be introduced in the future make assumptions. This raises several important questions which are considered in this session. First, it is important to know the extent to which the conclusions we draw from statistical procedures are contingent on these assumptions. Second, how can these assumptions be examined and tested? Third, what can be done if these assumptions do not hold?

Some examples of assumptions made by simple procedures are:

- The one sample t-test (AT2) makes the assumptions that observations are a) independent and b) normally distributed.
- Constructing a CI (CI) for a rate (AT3) makes the assumption that the number of events follows a Poisson distribution.
- The χ^2 test for comparing two proportions (AT 4 and 5) makes the assumption that the observations in each group follow a binomial distribution.

One approach that can be adopted when assumptions do not hold is to transform a variable. Transformations (*e.g.,* taking a logarithm of cholesterol values before performing a t-test) will affect an analysis in two ways. These are:

- Alter (ideally improve) the extent to which the assumptions of a procedure are satisfied.
- Change (ideally not worsen) the ease of interpretation of the results.

Both of these must be considered before deciding to adopt a particular transformation. For example, suppose that we have systolic blood pressure measurements in two groups and it is decided to take a square-root of blood pressure in order to make the assumption of normality more reasonable. A disadvantage of this approach is that the parameter we estimate is no longer the difference between two mean blood pressures measured in mmHg, but the difference between two means of the square-root of blood pressure with units$\sqrt{\text{mmHg}}$. If the primary aim of the analysis is estimation of a meaningful parameter, then this is a serious disadvantage. However, if the primary aim of the analysis is hypothesis testing then the fact that we do not have an easily clinically interpretable parameter estimate is less important.

6.2 Robustness

In practice it is rare for all the assumptions of a procedure to hold exactly. The extent to which this matters relates to the robustness of the procedure. Robustness is loosely defined by van Belle et al. (p253) as follows:

> 'A statistical procedure is robust if it performs well when the needed assumptions are not violated "too badly", or if the procedure performs well for a large family of probability distributions'.

A particular test is 'performing well' if the *nominal* significance level (as used to define the cut-off point for the test statistic) is close to the *actual* probability that the test statistic exceeds the cut-off. A particular technique for constructing CIs for a parameter is 'performing well' if the *nominal* probability that the CI includes the true value of the parameter is close to the *actual* probability that the CI includes the true value of the parameter.

6.2.1 A simple example of a non-robust procedure

Suppose we have a sample of size 4 (X_1, X_2, X_3, X_4), from a normal distribution with unknown mean and (known) unit variance: $X_i \sim N(\mu, 1)$ and we wish to test the null hypothesis $H_0{:}\mu = 0$ against the alternative $H_1{:}\mu \neq 0$.

The appropriate test statistic is$\bar{X} = (X_1 + X_2 + X_3 + X_4)/4$. $\bar{X} \sim N(\mu, 0.25)$ under H_0 and hence, using a *nominal* significance level of 5% we reject H_0 in favour of H_1 if $|\bar{X}| > \sqrt{0.25} \times 1.96 = 0.98$.

Suppose now that (X_1, X_2, X_3, X_4) take the discrete values 1 with probability π and -1 with probability 1- π *and* we wish to test the null hypothesis H_0: $\pi = 0.5$ against the alternative hypothesis H_1: $\pi \neq 0.5$. The above test might be considered appropriate because under H_0 the X_i have zero mean and unit variance as above. However, in this situation the test statistic will exceed 0.98 if and only if all of the X_i equal 1 or all equal -1 and hence the *actual* probability that the test statistic exceeds 0.98 is 0.125 (2×0.5^4) under H_0. This difference between the *nominal* and *actual* significance levels shows a lack of robustness here.

6.2.2 Robustness of some common procedures

- CIs and hypothesis tests that assume normality in their construction are not reliable when sample sizes are small and distributions are skewed.

- The comparison of two variance estimates (Analytical Techniques 5) suffers from a lack of robustness.
- The central limit theorem shows that, for large samples, the one sample t-test for a mean is robust. In general, due to the central limit theorem, procedures based on means are robust when sample sizes are large.

6.3 Investigating normality

In most situations the first thing to do in any data analysis is to look at the data. Simple plots will often reveal when assumptions are unlikely to hold. The following plots are useful for examining the distribution of a single continuous variable.

- Box and whisker plots (see AT 1)
- Histograms (see AT 1)
- Normal plots

Departures from normality can often be classified in one of the following ways:

- Outliers
- Skewness (see AT 1)
- Kurtosis (see AT 1)

When assessing normality, it is important to ensure that the distribution you examine is the one about which the assumption of normality is made. For example, if we wish to use a t-test to investigate whether a mean ***change*** in cholesterol differs from zero then it is the distribution of the changes, not that of cholesterol itself, which we need to investigate. Another example concerns the two-sample t-test: here the relevant assumption is that the variable in question is normally distributed within each group, not in the sample as a whole.

6.3.1 Normal plots

It can be difficult to assess visually whether normality assumptions are violated using histograms and/or box and whisker plots, particularly when the departures from normality are subtle. Departures from normality are most easily assessed visually using a normal plot.

There are a number of variants, but most commonly a normal plot consists of data points plotted against their expected values, given their rank position, under the assumption of normality. If the data follow a normal distribution, then we expect that the observed 10th percentile will be close to the theoretical 10th percentile (under normality), that the observed 20th percentile will be close to the theoretical 20th percentile *etc.* Hence if the assumption of normality is correct, we expect to see an approximate straight line with unit slope. Departures from a straight-line indicate departures from normality.

To construct a normal plot a formula for a sample of size *n* of a random variable *X*, first order the observed values. Let *x(i)* denote the *i*th largest value and let$\hat{\mu}$ and$\hat{\sigma}$ denote the estimated mean and standard deviation respectively. An expression for the expected value of *x(i)* is required. Two approximate expressions are as follows:

$$\xi_i = \hat{\mu} + \hat{\sigma} \times \Phi^{-1}\left(\frac{i}{n+1}\right) \quad \text{or} \quad \xi_i^* = \hat{\mu} + \hat{\sigma} \times \Phi^{-1}\left(\frac{i-3/8}{n+1/4}\right).$$

Here Φ is the probability distribution function of the standard normal distribution (*i.e.,* $\Phi^{-1}(0.025) = -1.96, \Phi^{-1}(0.5) = 0, \Phi^{-1}(0.8) = 0.842, \Phi^{-1}(0.975) = 1.96 etc.$). Now plot *x(i)* against ξ_i (orξ_i^*) together with the line of equality.

Figure 6.1 shows the expected appearance of a normal plot under normality. The plot was produced after randomly generating 250 points from a *N*(120, 64) distribution in Stata. Notice that there is some departure from a straight line, particularly in the tail of the distribution, even though the data does follow a normal distribution.

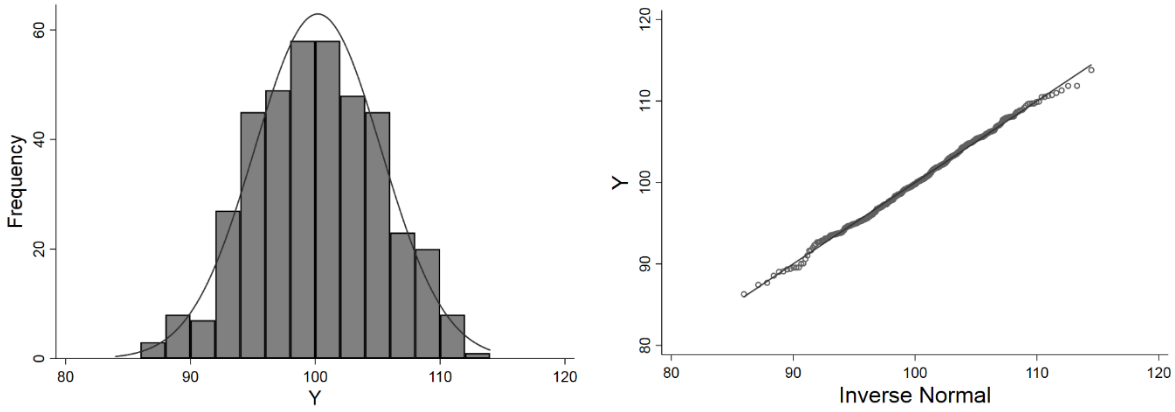


Figure 6.1: Appearance of histogram and normal plot for a normally distributed variable

Figures 6.2, 6.3, 6.4, 6.5 and 6.6 show histograms and normal plots for some common types of departure from normality.

6.3.1.1 Outliers

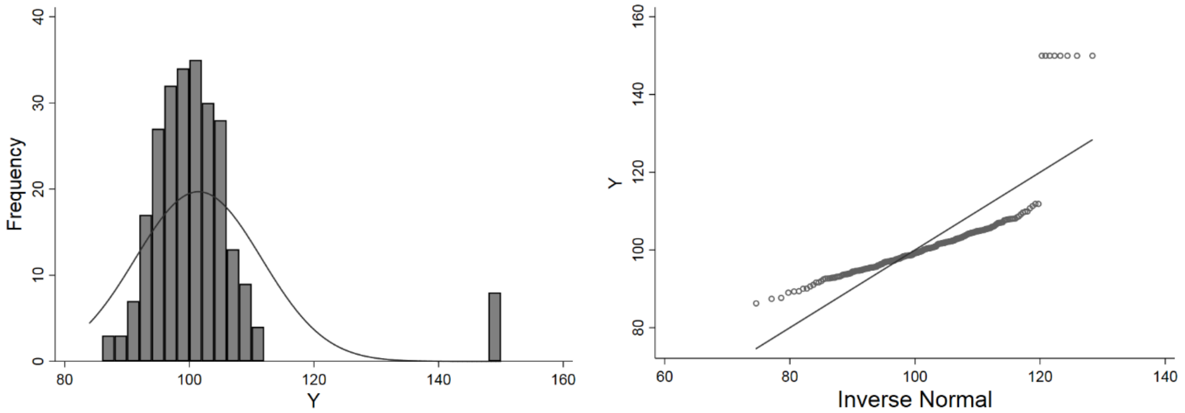


Figure 6.2: Appearance of histogram and normal plot for a variable with outlying values.

In figure 6.2 the data points in the normal plot do appear as an approximate straight line, with the exception of the 8 outlying points at the top right. The inflation of the standard deviation by the outlying points results in this observed straight line having a slope less than unity. With data like this it would be sensible to check whether the extreme points are genuine.

How would large outliers (as in figure 6.2) affect the skew and kurtosis?

6.3.1.2 Skewed distributions

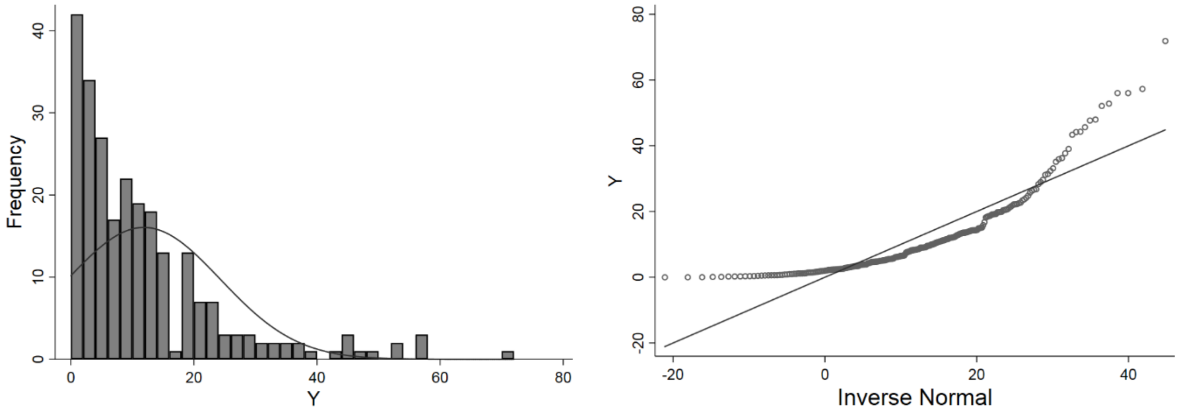


Figure 6.3: Appearance of histogram and normal plot for a variable exhibiting right-skewness.

In figure 6.3 there is clear departure from a straight-line relationship. The largest values of the variable are more extreme than would be expected under normality, whilst the smallest are not as extreme as would be expected. This is called right skewed or positively skewed.

Note that when producing histograms, it can be useful to superimpose a fitted normal distribution (with the same mean and standard deviation) on top of the observed distribution. However, as with superimposing “lines of best fit” (see Regression course) on scatter plots this can distract the reader from looking at the observed data, and so not all statisticians do this when displaying data in a histogram.

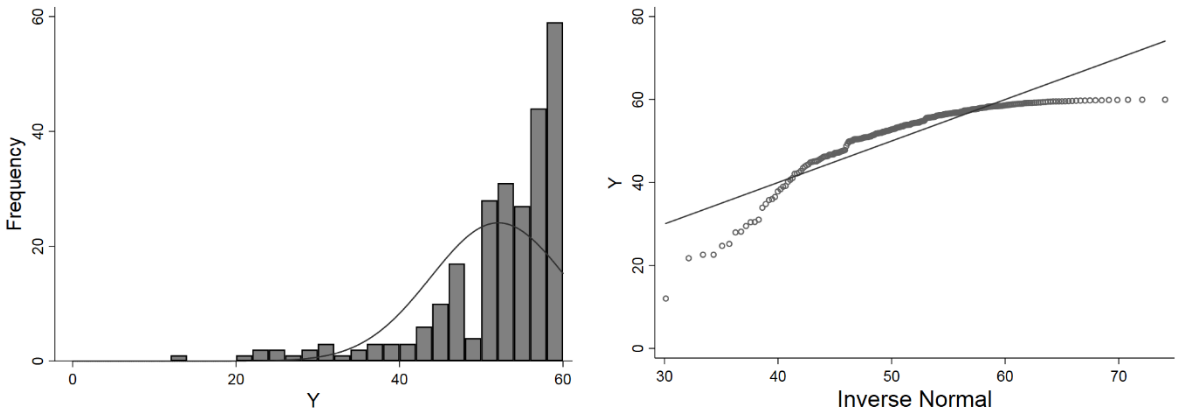


Figure 4: Appearance of histogram and normal plot for a variable exhibiting left-skewness.

Figure 6.4 also exhibits a clear departure from a straight-line relationship. The largest values of the variable are less extreme than would be expected under normality, whilst the smallest are more extreme. This is termed left-skewness or negative skewness.

6.3.1.3 Kurtosis

A variable exhibits skewness if the observations at one extreme of the distribution are more extreme, whilst those at the other end are less extreme than would be expected under normality. In contrast to this, kurtosis (see AT 1) describes symmetric departures from normality.

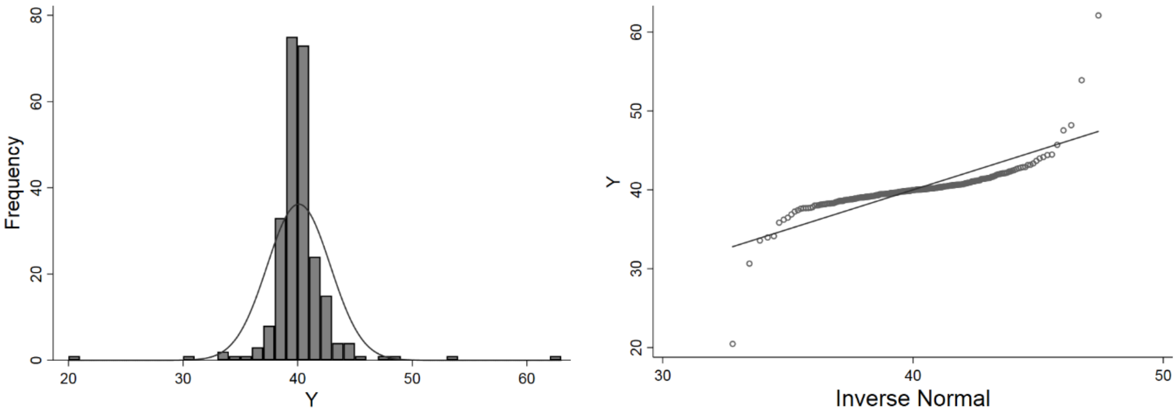


Figure 6.5: Appearance of histogram and normal plot for a heavy tailed variable.

In figure 6.5 there is an excess of points in the centre and extremes of the distribution as compared with a normal distribution having the same standard deviation. This 'peaked' distribution exhibits heavy-tailed kurtosis.

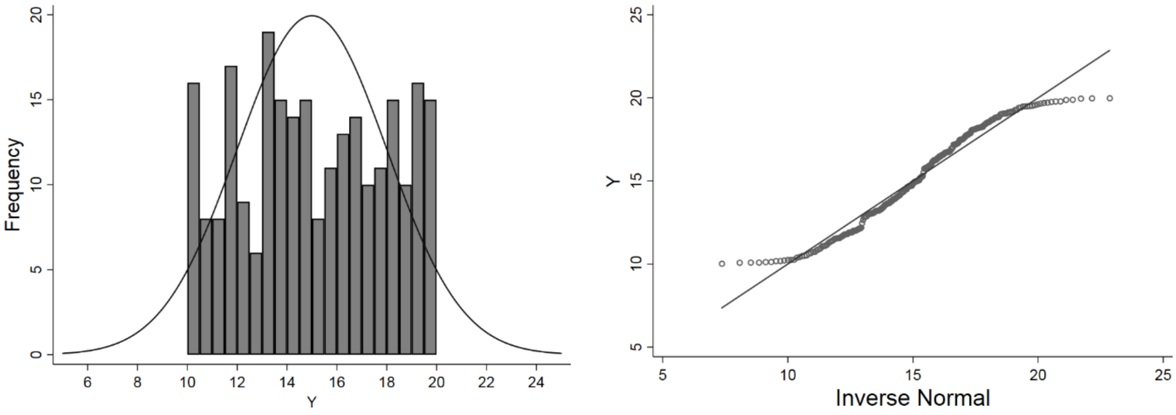


Figure 6.6: Appearance of histogram and normal plot for a variable exhibiting light-tailed kurtosis.

In figure 6.6 there are fewer points in the centre and extremes of the distribution as compared with a Normal distribution having the same standard deviation. This 'flat' distribution exhibits light-tailed kurtosis.

As mentioned above there are variations on the normal plot. One common variant is to plot *x(i)* against$\Phi^{-1}(i/n+1)$. In this case a straight line with intercept equal to the mean and slope equal to the standard deviation is expected under normality.

6.3.2 Statistical tests of normality

There are a number of statistical tests of normality. These include the Kolmogorov-Smirnov test and the Shapiro-Wilk test (see Armitage, Berry and Matthews for details). However, there are two main drawbacks to the use of these tests.

- With small datasets the tests lack statistical power and hence they may miss 'important' deviations from normality.
- With large datasets it is very common for 'small' deviations from normality to be statistically significant. For practical purposes these deviations are often unimportant, since with large datasets the central limit theorem means that statistical procedures are robust to departures from normality

If in doubt about the validity of a normality assumption many statisticians advise using a 'robust' technique in order to check the validity of an analysis, rather than relying on a statistical test of normality.

6.3.3 Analysis of continuous data that is not normally distributed

A number of different approaches can be used when data is not normally distributed. These include:

- Relying on the central limit theorem. Provided that sample sizes are large then hypothesis tests and CIs concerning parameters such as means can still be used because the central limit theorem ensures that the distribution of estimates of such parameters is approximately normal, even when variables are not normally distributed.
- Non-parametric approaches. These techniques are introduced in the Robust Methods course. However, there are a number of drawbacks to their use. First, they have limited utility in estimation and second they cannot be used to analyse more complex data (*e.g.,* there is no general non-parametric 'equivalent' for multiple regression).
- 'Robust' methods. There are a number of relatively computer intensive methods (*e.g.,* bootstrapping, 'sandwich' estimators of variance) of calculating standard errors that relax assumptions about normality and independence. See the Robust Methods course for details.
- Transformations. Endeavour to transform to a scale on which normality assumptions hold. There is, of course, no guarantee that such a transformation will exist

6.4 Transformations

6.4.1 The family of power transformations

The most commonly adopted transformations are those from the family of power transformations. These are:

$$\ldots, X^{-2}, X^{-1}, X^{-\frac{1}{2}}, \log(X), X^{\frac{1}{2}}, X^{1}, X^{2}, \ldots$$

Where $X^{1}$is the identity transformation i.e., no transformation.

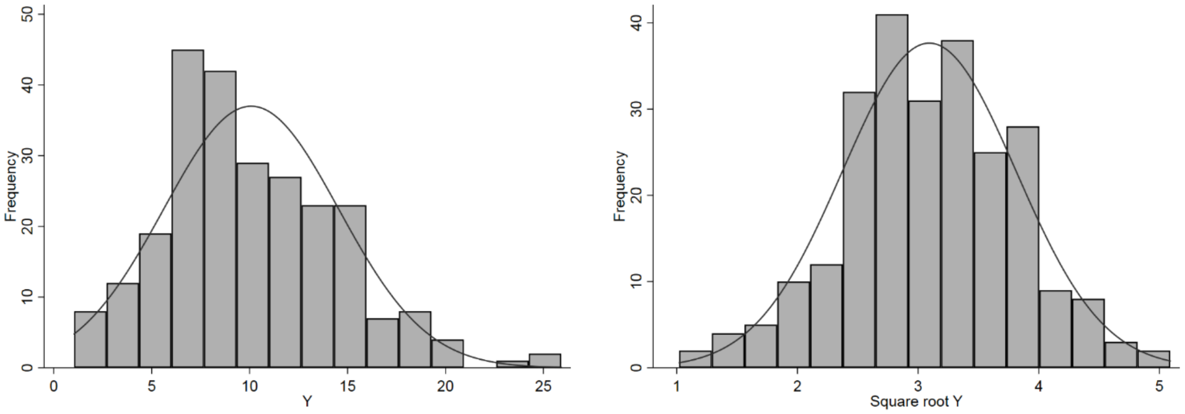


Figure 6.7: Using a square root transformation to reduce skewness.

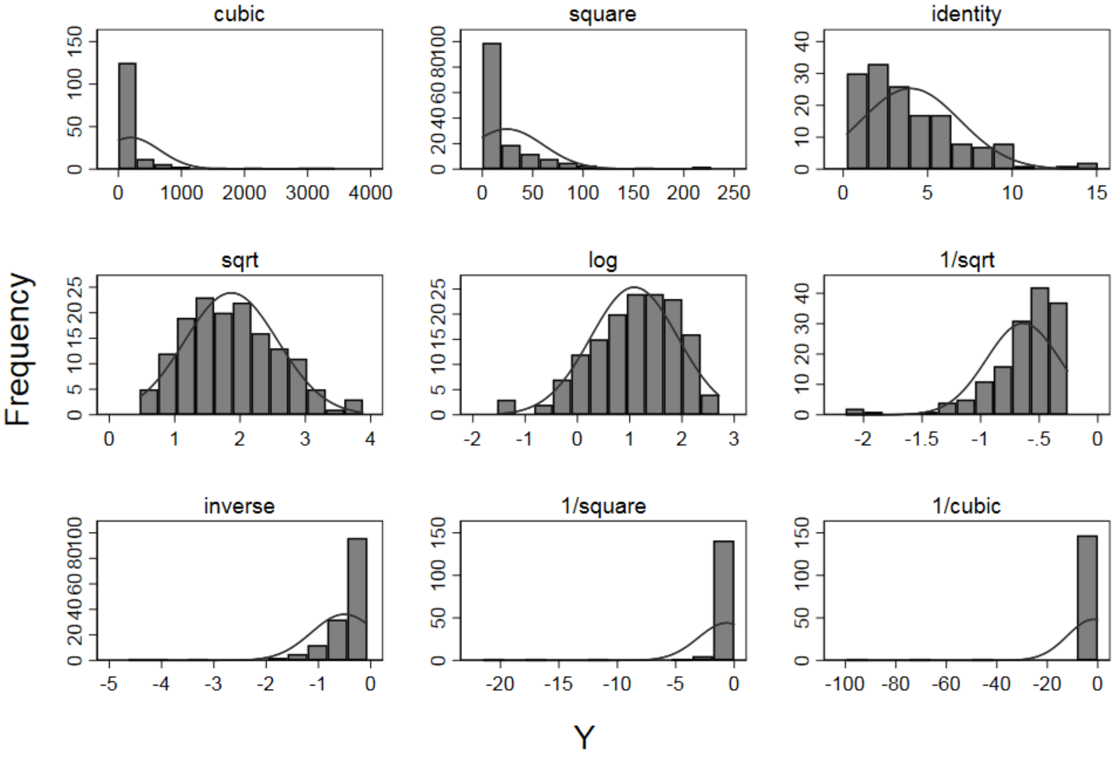
As illustrated in Figure 6.7 power transformations can often reduce skewness. The skew for Y is 1.43 compared to 0.43 for the square root of Y. However, this small gain in normality is probably outweighed by the complication of interpretation introduce by the square root transformation.

In contrast power transformations are not usually a useful way of reducing kurtosis. Usually the best way to deal with kurtosis is to use a non-parametric or robust procedure.

Routines such as Stata’s **ladder** or **gladder** commands can be used to inform the choice of power transformations. Output from both are shown below.

```
. ladder y

Transformation          formula              chi2(2)       P(chi2)
------------------------------------------------------------------
cubic                   y^3                        .         0.000
square                  y^2                        .         0.000
identity                y                      27.28         0.000
square root             sqrt(y)                 4.48         0.107
log                     log(y)                  7.95         0.019
1/(square root)         1/sqrt(y)              58.55         0.000
inverse                 1/y                        .         0.000
1/square                1/(y^2)                    .         0.000
1/cubic                 1/(y^3)                    .         0.000
```



Histograms by transformation

Figure 6.8: Output from Stata's ladder and gladder commands

In this example, although the square root transformation performs slightly better than the logarithmic transformation in terms of improving normality, we also need to consider ease of interpretation.

6.4.2 The logarithmic transformation

Of all these power transformations the logarithmic transform is the most widely used because data transformed in this way is relatively simple to interpret. When a result is expressed on a logarithmic scale it is almost always better expressed after back-transformation to the original scale. After back transformation the result is invariably expressed in a multiplicative manner.

Example: Suppose we wish to compare systolic blood pressures in men (Y_1) and women (Y_2) working for a particular company.

i) Analysis on an untransformed scale.

The difference between the two means is expressed as a difference in mmHg. For example, the difference ($\bar{Y}_2 - \bar{Y}_1$) might be -10mmHg.

ii) Analysis on a log transformed scale.

Here we first calculate $\overline{log(Y_2)} - \overline{log(Y_1)}$= *d* (say). Since exp $(log(a) - log(b)) = a/b$ the ratio of the geometric mean of Y_1 to that of Y_2 is exp(*d*). For example if *d* = -0.05 then the ratio of the geometric means is exp(-0.05) = 0.951. Loosely we could report the result by saying that blood pressure was lower ‘on average’ by 4.9% in women than in men.

6.4.3 Back-transformation of CIs

When a CI is calculated for a transformed variable it is usually most informative to back-transform the confidence limits as well as the point estimate to the original scale. After back transformation CIs will no longer be symmetric.

Example (continued): Suppose that analysis of log transformed blood pressure gives an estimated mean difference of -0.05 in log(blood pressure) with a 95% CI extending from -0.12 to 0.02.

A 95% CI for the ratio of the geometric means extends from exp(-0.12) to exp(0.02) = (0.887, 1.020) and hence we can say that, on average, blood pressure was lower by 4.9% with a 95% CI extending from a 11.3 % decrease to a 2.0% increase.

6.4.4 Base for logarithmic transformations

Usually logarithms are taken to base *e.* However sometimes it is desirable to work on a $\log_2$ scale rather than a $\log_e$ scale. On this scale a coefficient of 1 represents a doubling, a coefficient of 2 is two doublings (multiplying by four) *etc.*

6.4.5 The log-normal distribution

A random variable *X* follows a log-normal distribution if log(*X)* follows a normal distribution. As illustrated in Figure 6.9, log-normal distributions are skewed with the extent of the skewness determined by the variance of log*(X*).

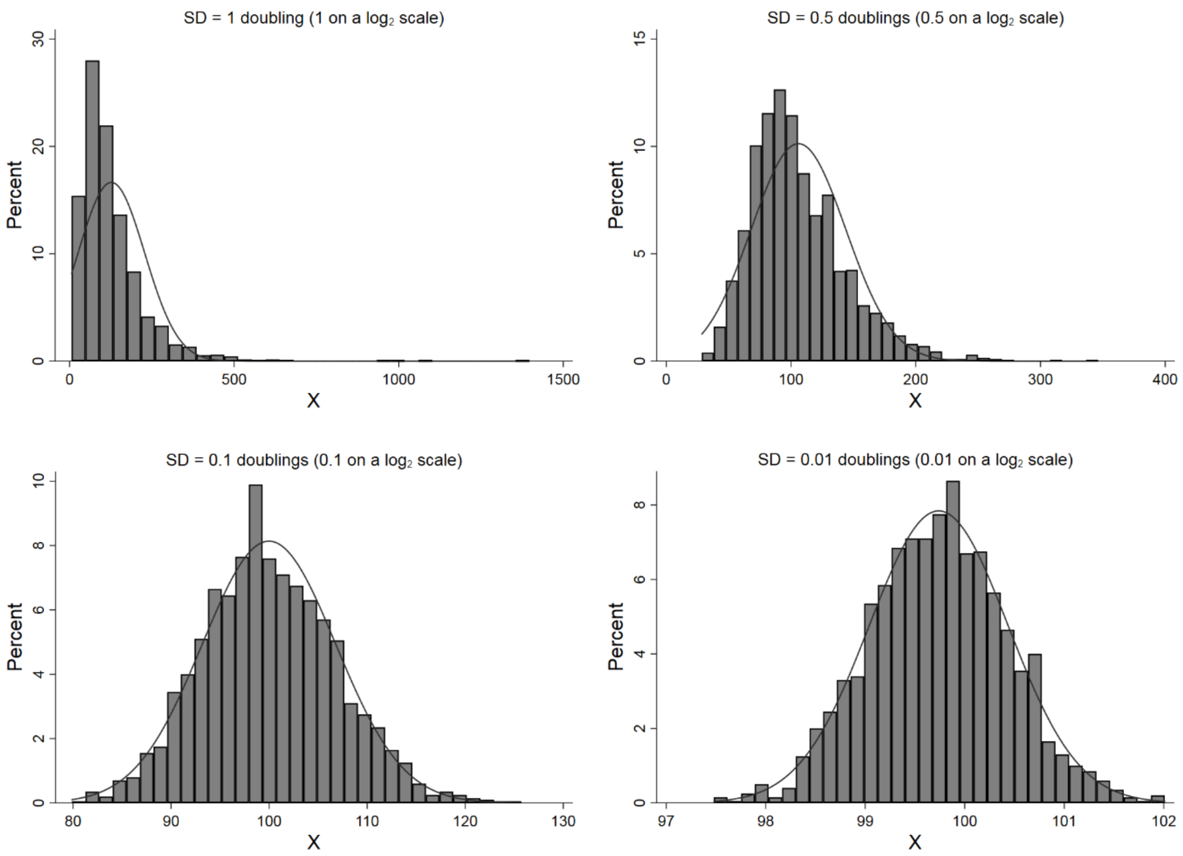


Figure 6.9: Examples of log-normal distributions with geometric mean of 100.

6.4.6 Transformations for proportions

Statistical procedures often utilise a transformation of one or more proportions, commonly to avoid difficulties arising from the fact that proportions are bounded in the range [0, 1]. Models for parameters transformed in this way are covered in detail in the Generalised Linear Models course. Common transformations are:

- A single proportion π is often expressed as an odds [$\pi/(1-\pi)$]. $\pi/(1-\pi)$ takes values in the range $[0,\infty)$.
- Odds are often log transformed to give a 'log odds' [$\log(\pi/(1-\pi))$]. The function $f(\pi) = \log(\pi/(1-\pi))$ is often termed the logit transformation.
- A comparison between two proportions π_1 and π_2 can be expressed as π_1/π_2 ('risk ratio') or $\log(\pi_1/\pi_2)$ ('log risk ratio').
- Alternatively proportions can be compared using $\pi_1(1-\pi_2)/(\pi_2(1-\pi_1))$ ('odds ratio') or the 'log odds ratio' = $\log[\pi_1(1-\pi_2)/(\pi_2(1-\pi_1))] = \log(\pi_1/(1-\pi_1))- \log(\pi_2/(1-\pi_2))$.

Exercise 1: What range of values can be taken by i) a 'log odds', ii) an 'odds ratio' and iii) a 'log odds ratio'.

Exercise 2: Suppose that an estimated log odds ratio (95% CI) for the association between smoking and lung cancer is 2.1 (1.5, 2.7). Calculate a 95% CI for the odds ratio.

6.5 Mean and variance of a transformed variable

Many estimators (*e.g.,* mean and variance from normally distributed samples) are exceptional in that they have relatively simple sampling distributions that can be derived analytically. However many situations are not so simple and approximations have to be used just to get the mean and variance.

Such problems are commonly addressed by using Taylor series expansions to obtain approximate formulae for the mean and variance of a transformation of an estimator whose properties are already known. The Taylor series expansion is as follows:

$$g(x) = g(\mu) + (x-\mu)g'(x)|_{x=\mu} + \tfrac{1}{2}(x-\mu)^2 g''(x)|_{x=\mu} + \dots \qquad (1)$$

Where $g'(x)|_{x=\mu}$ and $g''(x)|_{x=\mu}$ are the first and second derivatives of $g(x)$ evaluated at $x = \mu$.

Suppose *X* has mean μ and variance σ^2. Provided that $|\mu|$ is substantially larger than σ then the following formulae can be used to calculate the mean and variance of *g*(*X*).

$$E[g(x)] = g(\mu) + \tfrac{1}{2}(g''(x)|_{x=\mu})\sigma^2 \qquad (2)$$

$$V[g(x)] = \sigma^2 (g'(x)|_{x=\mu})^2 \qquad (3)$$

Exercise 3: Use the Taylor series formula (1) to derive formula (2).

6.5.1 Sketch Proof of Variance Formula

From (1) and (2)

$$g(X) - E\big(g(X)\big) \approx (X-\mu)\big(g'(x)|_{x=\mu}\big) + \tfrac{1}{2}((X-\mu)^2 - \sigma^2)g''(x)|_{x=\mu} + \dots \qquad (4)$$

Ignoring the quadratic and later terms in (4)

$$E\big[g(X) - E\big(g(X)\big)\big]^2 \approx E\big((X-\mu)\big)^2 \big(g'(x)|_{x=\mu}\big)^2$$

i.e. $\quad V[g(X)] \approx \sigma^2 \big(g'(x)|_{x=\mu}\big)^2$

6.5.2 Examples of variance transformation formulae

$$\text{a) If } g(X) = \log_e(X) \qquad g'(x) = \left(\frac{1}{x}\right)$$

$$\text{so} \qquad \mathrm{V}(\log_e(X)) \approx \sigma^2 \left(\frac{1}{\mu}\right)^2$$

$$= \frac{\sigma^2}{\mu^2}$$

$$\text{and} \qquad \mathrm{SD}(\log_e(X)) \approx \frac{\sigma}{\mu} = \mathrm{CV}(X)$$

The ratio of the standard deviation to the mean of a variable is known as the coefficient of variation (CV).

$$\text{b) If } g(X) = X^{1/2} \qquad g'(x) = \frac{1}{2}x^{-1/2}$$

$$\text{so} \qquad \mathrm{V}(X^{1/2}) \approx \sigma^2 \left(\frac{1}{4\mu}\right)$$

$$= \frac{\sigma^2}{4\mu}.$$

This demonstrates that the square-root transformation can be used to stabilise the variance of a variable whose variance is proportional to its mean. Before the widespread use of Generalised Linear Models this transformation was often used with variables that followed a Poisson distribution.