Statistics for Economists

Len Gill, Chris D. Orme & Denise Osborn

Contents

Preface

These notes are quite detailed **Course Notes** which, perhaps, look rather like a text book. They can not, however, be regarded as giving an exhaustive treatment of statistics (such as you might find in a proper text). In conjunction with any *recommended text*, you should use the material contained here to:

- supplement lecture notes (if applicable);
- provide a starting point for your own private study.

The material has been structured by Chapter, and section, rather than as a sequence of lectures, and aims to provide a coherent development of the subject matter. However, a programme of lectures could easily be constructed to follow the same development and include some, but not all, of the material contained in this document. Following an Introduction (which provides an overview) there are 15 substantive Chapters and at the end of each, and corresponding sequence of lectures (if appropriate), you should access any recommended additional learning resources in order to deepen your learning.

At the end of this document, in an Appendix, there are Standard Normal Tables and Student-t tables. Don't concern yourself with them at present, their use be described when appropriate.

These notes also contain *Exercises* consisting of a number of questions designed to promote your learning. These exercises also include work with EXCEL and are included to illustrate how many of the standard statistical techniques can be automated - thus removing much of the tedious calculation. Rudimentary knowledge of EXCEL is assumed.

Len Gill
Chris Orme
Denise Osborn
University of Manchester, 2003

Introduction

The subject of *statistics* is concerned with scientific methods for collecting, organizing, summarizing, presenting *data* (numerical information). The power and utility of statistics derives from being able to draw valid conclusions (inferences), and make reasonable decisions, on the basis the available data. (The term statistics is also used in a much narrower sense when referring to the data themselves or various other numbers derived from any given set of data. Thus we hear of employment statistics (% of people unemployed), accident statistics, (number of road accidents involving drunk drivers), etc.)

Data arise in many spheres of human activity and in all sorts of different contexts in the natural world about us. Such data may be obtained as a matter of course (e.g., meteorological records, daily closing prices of shares, monthly interest rates, *etc*), or they may be collected by survey or *experiment* for the purposes of a specific statistical investigation. An example of an investigation using statistics was the *Survey of British Births, 1970,* the aim of which was to improve the survival rate and care of British babies at or soon after birth. To this end, data on new born babies and their mothers were collected and analysed. The Family Expenditure Survey regularly collects information on household expenditure patterns - including amounts spent on lottery tickets.

In this course we shall say very little about how our data are obtained; to a large degree this shall be taken as given. *Rather, this course aims simply to describe, and direct you in the use of, a set of tools which can be used to analyse a given set of data.* The reason for the development of such techniques is so that evidence can be brought to bear on particular *questions* (for example),

- *Why do consumption patterns vary from individual to individual*?

- *Compared to those currently available, does a newly developed medical test offer a significantly higher chance of correctly diagnosing a particular disease*?

or *theories/hypotheses*, such as,

- "*The majority of the voting population in the UK is in favour of, a single European Currency*"
- "*Smoking during pregnancy adversely affects the birth weight of the unborn child*"
- *"Average real earnings of females, aged* $30-50$, *has risen over the past* 20 *years*"

which are of interest in the social/natural/medical sciences.

Statistics is all about using the available data to shed light on such questions and hypotheses. At this level, there are a number of similarities between a statistical investigation and a judicial investigation. For the statistician the evidence comes in the form of data and these need to be *interrogated* in some way so that plausible conclusions can be drawn. In this course we attempt to outline some of the fundamental methods of statistical interrogation which may be applied to data. The idea is to get as close to the *truth* as is possible; although, as in a court of law, the truth may never be revealed and we are therefore obliged to make reasonable judgements about what the *truth* might be based on the evidence (the data) and our investigations (analysis) of it.

For example, think about the following. Suppose we pick at random 100 male and 100 female, University of Manchester first year undergraduates (who entered with A-levels) and recover the A-level points score for each of the 200 students selected. How might we use these data to say something about whether or not, *in general*, (a) female students achieve higher A-level grades than males, or (b) female first year undergraduates are more intelligent than first year undergraduate males ? How convincing will any conclusions be?

We begin with some definitions and concepts which are commonly used in statistics:

Some Definitions and Concepts

- **DATA:** *body of numerical evidence* (i.e., numbers)
- **EXPERIMENT:** *any process which generates data*

  For example, the following are *experiments*:

  - *select a number of individuals from the UK voting population and how they will vote in the forthcoming General Election*
  - *flipping a coin twice and noting down whether, or not, you get a HEAD at each flip*
  - *interview a number of unemployed individuals and obtain information about their personal characteristics (age, educational background, family circumstances, previous employment history, etc)*

and the local and national economic environment. Interview them again at regular three-monthly intervals for 2 years in order to model (i.e., say something about possible causes of) unemployment patterns

- to each of 10 rats, differing dosage levels of a particular hormone are given and, then, the elapsed time to observing a particular (benign) reaction (to the hormone) in the each of the rats is recorded.

An experiment which generates data for use by the statistician is often referred to as *sampling,* with the data so generated being called a *sample* (of data). The reason for sampling is that it would be impossible to interview (at the very least too costly) all unemployed individuals in order to explain shed light on the cause of unemployment variations or, or all members of the voting population in order to say something about the outcome of a General Election. We therefore select a number of them in some way (not all of them), analyse the data on this sub-set, and then (hopefully) conclude something useful about the population of interest in general. The initial process of selection is called *sampling* and the conclusions drawn (about the general *population* from which the sample was drawn) constitutes *statistical inference*:

- **SAMPLING:** *the process of selecting individuals* (*single items*) *from a population.*
- **POPULATION:** *a description of the totality of items with which we are interested. It will be defined by the issue under investigation.*
- *Sampling/experimentation yields a* **SAMPLE** *(of items), and it is the sample which ultimately provides the data used in statistical analysis.*

It is tremendously important at this early stage to reflect on this and to convince yourself that *the results of sampling can not be known with certainty.* That is to say, although we can propose a strategy (or method) whereby a sample is to be obtained from a given population, we can not predict exactly what the sample will look like (i.e., what the outcome will be) once the selection process, and subsequent collection of data, has been completed. For example, just consider the outcome of the following sampling process: ask 10 people in this room whether or not they are vegetarian; record the data as 1 for *yes* and 0 for *no/unsure.* How many 1's will you get? The answer is uncertain, but will presumably be an integer in the range 0 to 10 (and nothing else).

Thus, the design of the sampling process (together with the sort of data that is to be collected) will rule out certain outcomes. Consequently, although not knowing exactly the sample data that will emerge we can list or provide some representation of what could possibly be obtained and such a listing is called a *sample space*:

- **SAMPLE SPACE:** *a listing, or representation, of all possible samples that could be obtained*

The following example brings all of these concepts together and we shall often use this simple scenario to illustrate various concepts:

- *Example:*
    - *Population*: a coin which, when flipped, lands either **H** (Head) or **T** (Tail)
    - *Experiment/Sampling*: flip the coin twice and note down **H** or **T**
    - *Sample*: consists of two items. The first item *indicates* **H** or **T** from the first flip; the second indicates **H** or **T** from the second flip
    - *Sample Space*: {(**H**,**H**),(**H**,**T**),(**T**,**H**),(**T**,**T**)}; list of 4 possible outcomes.

The above experiment yields a *sample of size* 2 (items or outcomes) which, when obtained, is usually given a numerical code and it is this coding that defines the data. If we also add that the population is defined by a *fair* coin, then we can say something about how *likely* it is that any one of the four possible outcomes will be *observed* (obtained) if the experiment were to be performed. In particular, elementary probability theory (see section 3), or indeed plain intuition, shows that each of the 4 possible outcomes are, in fact, equally likely to occur if the coin is fair.

Thus, in general, although the outcome of sampling can not be known with certainty we will be able to list (in some way) possible outcomes. Furthermore, if we are also willing to assume something about the population from which the sample is to be drawn then, although certainty is still not assured, we may be able to say how likely (or probable) a particular outcome is. This latter piece of analysis is an example of **DEDUCTIVE REASONING** and the first 8 Chapters in this module are devoted to helping you develop the techniques of statistical deductive reasoning. As suggested above, since it addresses the question of how likely/probable the occurrence of certain phenomena are, it will necessitate a discussion of *probability.*

Continuing the example above, suppose now that the experiment of flipping this coin twice is repeated 100 times. *If the coin is fair* then we have stated (and it can be shown) that, for each of the 100 experiments, the 4 possible outcomes are equally likely. Thus, it seems reasonable to predict that a (**H**,**H**) should arise about 25 times, as should (**H**,**T**), (**T**,**H**) and (**T**,**T**) - roughly speaking. This is an example of deductive reasoning. On the other hand, a question you might now consider is the following: *if when this experiment is carried out* 100 *times and a (***T,T***) outcome arises* 50 *times, what evidence does this provide on the assumption that the coin is fair?* The question asks you to make a judgement (an inference, we say) about the coin's fairness based on the observation that a (**T,T**) occurs 50 times. This introduces the more powerful notion of *statistical inference,* which is the subject matter of the sections $9 - 16$. The following brief discussion gives a flavour of what statistical inference can do.

Firstly, data as used in a statistical investigation are rarely presented for public consumption in *raw* form. Indeed, it would almost certainly be a meaningless exercise. Rather they are manipulated, summarised and, some would say, distorted! The result of such data manipulation is called a *statistic*:

- **STATISTIC:** *the result of data manipulation, or any method or procedure which involves data manipulation.*

Secondly, data manipulation (the production of *statistics*) is often performed in order to shed light on some *unknown* feature of the population, from which the sample was drawn. For example, consider the relationship between a *sample proportion* (p) and the *true* or actual population proportion (Π) , for some phenomenon of interest:

$$p = \Pi + error$$

where, say, p is the proportion of students in a collected sample of 100 who are vegetarian and Π is the proportion of all Manchester University students who are vegetarian. (Π is the upper case Greek letter *pi*; it is *not* used here to denote the number $Pi = 3.14159....$) The question is "what can p tell us about Π?", when p is observed but Π isn't.

For p to approximate Π it seems obvious that the *error* is required to be 'small', in some sense. However, the *error* is unknown; if it were known then an observed p would pinpoint Π exactly. In statistics we characterise situations in which we believe it is *highly likely* that the *error* is small (i.e., less than some specified amount). We then make statements which claim that it is highly likely that the observed p is close to the unknown Π. Here, again, we are drawing conclusions about the nature of the population from which the observed sample was taken; it is **statistical inference**. Suppose, for example, that based on 100 students the observed proportion of vegetarians is $p = 0.3$. The theory of statistics (as developed in this course) then permits us to *infer* that there is a 95% chance (it is 95% likely) that the interval $(0.21, 0.39)$ contains the unknown true proportion Π. Notice that this interval is symmetric about the value $p = 0.3$ and allows for margin of error of ± 0.09 about the observed sample proportion. The term margin of error is often quoted when newspapers report the results of political opinion polls; technically it is called a *sampling error* - the error which arises from just looking at a subset of the population in which we are interested and not the whole population.

For this sort of thing to work it is clearly important that the obtained sample is a fair (or typical) reflection of the population (not atypical): such samples are termed **(simple) random samples.** For example, we would not sample students as they left a vegetarian restaurant in order to say something about University student population as a whole! We shall signal that samples are random in this way by using expressions like: '*consider a random sample of individuals*'; '*observations were randomly sampled*'; '*a number of individuals were selected at random from an underlying population*', etc.

In order to understand the construction of statistics, and their use in inference, you need some basic *tools of the trade*, which are now described.

Notation and Tools of the Trade

Variables

- a variable is a label, with description, for an event or phenomenon of interest. To denote a variable, we use upper case letters. For example, X, Y, Z, etc are often used as labels.

- a lower case leter, x, is used to denote an observation obtained (actual number) on the variable X. (Lower case y denotes an observation on the variable Y, etc.)

- *Example*:

  Let X = *A-level points score. If we sample* 4 *individuals we obtain* 4 *observations and we label these 4 observations as* $x_1 =$; $x_2 =$; x_3= ; $x_4 =$. (You can fill in four numbers here.)
  x_1 denotes the first listed number (in this case, A-level points score), x_2 the second score, x_3 the third and x_4 the fourth.

In general, then, x_i is used to denote a *number* - the i^{th} observation (or value) for the variable X, which is read simply as "x" "i". The subscript i is usually a *positive integer* $(1, 2, 3, 4$, etc), although terms like x_0 and x_{-1} can occur in more sophisticated types of analyses. Thus, we also use y_i for values of the variable with label Y. Other possibilities are z_i, x_j, y_k etc. Thus, the label we may use for the variable is essentially arbitrary as is the subscript we use to denote a particular observation on that variable.

- the values $x_1, x_2, x_3, x_{4,} \ldots, x_n$ denote a sample of n observations (n numbers or values) for the variable X. The "dots" indicates that the sequence continues until the subscript n is reached (e.g., if $n = 10$, there are another 6 numbers in the sequence after x_4). For ease of notation we usually write this simply as $x_1, ..., x_n$.
  Similarly, $y_1, \ldots, y_m$ for a sample of m observations for the variable Y.

Summation notation

The summation notation is used to signify the addition of a set of numbers. Let an arbitrary set of numbers be doted x_1, ..., x_n.

- the symbol $\sum$ is used: the Greek letter capital *sigma.* English equivalent is S, for *sum.*

And we have the following definition of $\sum$:

- $x_1 + x_2 + \ldots + x_n \equiv \sum_{i=1}^{n} x_i$, or $\sum\limits_{i=1}^{n} x_i$

  which means add up the n numbers, x_1 to x_n. The expression $\sum_{i=1}^{n} x_i$ is read as "the sum from i equals 1 to n of x_i", or "the sum over i, x_i"

For example, the total A-level points score from the 4 individuals is:

$$x_1 + x_2 + x_3 + x_4 \equiv \sum_{i=1}^{4} x_i = \qquad .$$

Note that for any n numbers, $x_1, \ldots, x_n$, the procedure which is defined by $\sum_{i=1}^{n} x_i$ is *exactly* the same as that defined by $\sum_{j=1}^{n} x_j$, because

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n = \sum_{j=1}^{n} x_j .$$

You might also encounter $\sum_{i=1}^{n} x_i$ expressed in the following ways:

- $\sum\limits_{i} x_i$ *or* $\sum_i x_i$; $\qquad \sum x_i$; $\qquad$ *or even,* $\sum x$.

Moreover, since the labelling of variables, and corresponding observations, is arbitrary, the sum of n observations on a particular variable, can be denoted equivalently as $\sum_{k=1}^{n} y_k = y_1 + y_2 + \ldots + y_n$, if we were to label the variable as Y rather than X and use observational subscript of k rather than i.

Rules of Summation

- Let c be some fixed number (e.g., let $c = 2$) and let $x_1, \ldots, x_n$ denote a sample of n observations for the variable X:

$$\sum_{i=1}^{n}(cx_i) = cx_1 + cx_2 + \ldots + cx_n = c(x_1 + x_2 + \ldots + x_n) = c\left(\sum_{i=1}^{n} x_i\right)$$

In above sense c is called a constant (it does not have a subscript attached). It is constant in relation to the variable X, whose values (denoted x_i) are allowed to vary from across observations (over i).

Notice that when adding numbers together, the orderr in which we add them is irrelevant. With this in mind we have the following result:

- Let $y_1, \ldots, y_n$ be a sample of n observations on the variable labelled Y:

$$\begin{aligned}
\sum_{i=1}^{n}(x_i + y_i) &= (x_1 + y_1) + (x_2 + y_2) + \ldots + (x_n + y_n) \\
&= (x_1 + x_2 + \ldots + x_n) + (y_1 + y_2 + \ldots + y_n) \\
&= \left(\sum_{i=1}^{n} x_i\right) + \left(\sum_{i=1}^{n} y_i\right) \\
&= \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i.
\end{aligned}$$

Combining the above two results we obtain the following:

- If d is another constant then

$$\begin{aligned}
\sum_{i=1}^{n}(cx_i + dy_i) &= c\left(\sum_{i=1}^{n} x_i\right) + d\left(\sum_{i=1}^{n} y_i\right) \\
&= c\sum_{i=1}^{n} x_i + d\sum_{i=1}^{n} y_i.
\end{aligned}$$

$cX + dY$ is known as a linear combination (of the variable X and the variable Y) and is an extremely important concept in the study of statistics.

- And, finally

$$\begin{aligned}
\sum_{i=1}^{n}(x_i + c) &= (x_1 + c) + (x_2 + c) + ... + (x_n + c) \\
&= \left(\sum_{i=1}^{n} x_i\right) + (n \times c)
\end{aligned}$$

These sorts of results can be illustrated using the following simple example:

- *Example:*

i :	1	2	3	4
x_i :	3	3	4	1
y_i :	4	4	2	3
$c = 2$				
$d = 3$				

(i) $\sum_{i=1}^{4} cx_i = c\sum_{i=1}^{4} x_i$;

(ii) $\sum_{i=1}^{4}(x_i + y_i) = \sum_{i=1}^{4} x_i + \sum_{i=1}^{4} y_i$;

(iii) $\sum_{i=1}^{4}(cx_i + dy_i) = c\sum_{i=1}^{4} x_i + d\sum_{i=1}^{4} y_i$.

You should be able to verify these for yourselves as follows. Firstly, the left hand side of (i) is

$$\begin{aligned}\sum_{i=1}^{4} cx_i &= (2\times 3)+(2\times 3)+(2\times 4)+(2\times 1)\\ &= 6+6+8+2\\ &= 22\end{aligned}$$

and the right hand side of (i) is

$$\begin{aligned}c\sum_{i=1}^{4} x_i &= 2\times(3+3+4+1)\\ &= 2\times 11 = 22.\end{aligned}$$

Establishing (ii) and (iii) follow in a similar way (try it by working out separately the left hand side and right hand side of each of (ii) and (iii).

HEALTH WARNING! However, beware of "casual" application of the summation notation (think about what you're trying to achieve). In the above example

$$\begin{array}{ll}\text{(iv)} & \sum_{i=1}^{4} x_i y_i \neq \left(\sum_{i=1}^{4} x_i\right)\left(\sum_{i=1}^{4} y_i\right)\\ \text{(v)} & \sum_{i=1}^{4}\left(\dfrac{x_i}{y_i}\right) \neq \dfrac{\sum_{i=1}^{4} x_i}{\sum_{i=1}^{4} y_i}.\end{array}$$

In the case of the right hand side of (iv) is the "sum of the products"

$$\sum_{i=1}^{4} x_i y_i = 12+12+8+3 = 35,$$

whilst the right hand side is the "product of the sums"

$$\left(\sum_{i=1}^{4} X_i\right)\left(\sum_{i=1}^{4} Y_i\right) = (3+3+4+1)\,(4+4+2+3) = 11\times 13 = 143 \neq 35.$$

Now, show that the left hand side of (v) is not equal to the right hand side of (v).

Also, the square of a sum is not (in general) equal to the sum of the squares. By which we mean:

$$\left(\sum_{i=1}^{n} x_i\right)^2 \neq \sum_{i=1}^{n} y_i^2$$

where $x_i^2 = (x_i)^2$, the squared value of x_i. This is easily verified since, for example, $(-1+1)^2 \neq (-1)^2 + 1^2$. Or, using the preceeding example, $\sum_{i=1}^4 x_i^2 = 9+9+16+1 = 35$, whilst $\left(\sum_{i=1}^4 x_i\right)^2 = (3+3+4+1)^2 = 11^2 = 121$.

- *Example*: Consider a group of 10 students (i.e., $n = 10$) out celebrating on a Friday night, in a particular pub, and each by their own drinks. Let x_i denote the number of pints of beer consumed by individual i; y_i, the number of glasses of white wine; z_i, the number of bottles of lager. If only beer, wine and lager are consumed at prices (in pence) of a for a pint of beer, b for a glass of white wine, c for a bottle of lager, then the expenditure on drinks by individual i, denoted e_i, is: $e_i = ax_i + by_i + cz_i$. Whereas total expenditure on drinks is: $\sum_{i=1}^{10} e_i = a\sum_{i=1}^{10} x_i + b\sum_{i=1}^{10} y_i + c\sum_{i=1}^{10} z_i$.

Part I

Probability, Random Variables & Distributions

Chapter 1

BASIC DESCRIPTIVE STATISTICS

Raw data means collected (or sampled) data which have not been organised numerically. An example would be the recorded heights of 100 male undergraduates obtained from an anonymous listing of medical records. Data in this form are rarely (if ever) informative. In order to highlight patterns of interest, the data can be summarized in a number of ways. This is sometimes called *data reduction*, since the raw data is reduced into a more manageable form. The reduction process requires the construction of what are called *descriptive* or *summary statistics*:

- **Basic descriptive statistics** *provide an overview or* **summary** *of the numerical evidence* (**data**).

The construction of statistics involves manipulations of the raw data. The constructed statistic can be *pictorial* (a graph or diagram) or *numerical* (a table or number) and different statistics are designed to give different sorts of information. We shall consider some of the more commonly used (descriptive) statistics in this section. The calculations involved can be tedious, but are fairly mechanical and relatively straightforward to apply (especially if a suitable computer package is to hand, such as EXCEL). The lecture presentation shall discuss these in a somewhat brief manner, with a little more detail being contained in these notes which you can read at your leisure.

Although descriptive statistics (or summaries) can be both *pictorial* and *numerical*, their construction depends upon the *type* of data to hand

1.1 Types of data

Broadly speaking, by '**data**' we mean numerical values associated with some variable of interest. However, we must not be overly complacent about such a broad definition; we must be aware of different types of data that may need special treatment. Let us distinguish the following types of data, by means of simple examples:

- **NOMINAL/CATEGORICAL**
  - *Examples*? are given in the lecture
- **ORDINAL**
  - *Examples*? are given in the lecture
- **DATA WITH ACTUAL NUMERICAL MEANING**
  - *Examples*? are given in the lecture
  - **Interval scale data:** indicates rank and distance from an *arbitrary* zero measured in unit intervals. An example is temperature in Fahrenheit and Celsius scales.
  - **Ratio scale data:** indicates both rank and distance from a *natural* (or *common*) zero, with ratios of two measurements having meaning. Examples include weight, height and distance (0 is the lower limit and, for example, 10 miles (16km) is twice as far as 5 miles (8km)), total consumption, speed etc.

Note that the ratio of temperature (Fharenheit over Celsius) changes as the tempertaute changes; however, the ratio of distance travelled (miles over kilometres) is always the same whatever distance is trevelled (the constant ratio being about 5/8.)

Although one could provide examples of other sorts of data, the above illustrate some of the subtle differences that can occur. For the most part, however, we will be happy to distinguish between just two broad classes of data: *discrete* and *continuous.*

1.1.1 Discrete data

The variable, X, is said to be discrete if it can only ever yield isolated values some of which (if not all) are often repeated in the sample. The values taken by the variable change by discernible, pre-determined steps or jumps. A discrete variable often describes something which can be counted; for example, the number of children previously born to a pregnant mother. However, it can also be categorical; for example, whether or not the mother smoked during pregnancy.

1.1.2 Continuous data

The variable, Y, is said to be continuous if it can assume any value taken (more or less) from a continuum (a continuum is an interval, or range of numbers). A nice way to distinguish between a discrete and continuous variable is to consider the possibility of listing possible values. It is theoretically impossible even to *begin* listing all possible values that a continuous variable, Y, could assume. However, this is not so with a discrete variable; you may not always be able to finish the list, but at least you can make a start.

For example, the birth-weight of babies is an example of a continuous variable. There is no reason why a baby should not have a birth weight of 2500.0234 grams, even though it wouldn't be measured as such! Try to list all possible weights (in theory) bearing in mind that for any two weights that you write down, there will always be another possibility half way between. We see, then, that for a continuous variable an *observation* is recorded, as the result of applying some measurement, but that this inevitably gives rise to a rounding (up or down) of the *actual value.* (No such rounding occurs when recording observations on a discrete variable.)

Finally, note that for a continuous variable, it is unlikely that values will be repeated frequently in the sample, unless rounding occurs.

Other examples of continuous data include: heights of people; volume of water in a reservoir; and, to a workable approximation, Government Expenditure. One could argue that the last of these is discrete (due to the finite divisibility of monetary units). However, when the amounts involved are of the order of millions of pounds, changes at the level of individual pence are hardly discernible and so it is sensible to treat the variable as continuous.

Observations are also often classified as *cross-section* or *time-series:*

1.1.3 Cross-section data

Cross-section data comprises observations on a particular variable taken at a single point in time. For example: annual crime figures recorded by Police regions for the year 1999; the birth-weight of babies born, in a particular maternity unit, during the month of April 1998; initial salaries of graduates from the University of Manchester, 2000. Note, the defining feature is that there is no natural ordering in the data.

1.1.4 Time-series data

On the other hand, time-series data are observations on a particular variable recorded over a period of time, at regular intervals. For example; personal crime figures for Greater Manchester recorded annually over 1980-99; monthly household expenditure on food; the daily closing price of a certain stock. In this case, the data does have a natural ordering since they are measured from one time period to the next.

1.2 Some graphical displays

We shall now describe some simple graphical displays, which provide visual summaries of the raw data. We consider just 3 types: those which can be used with discrete data - the *relative frequency diagram*, or *bar chart*; those for use with continuous data - the *histogram*; and those which provide a summary of the possible relationship between two variables - the *scatter plot*. Each are introduced by means of a simple example.

1.2.1 Discrete data

Example: Consumption of beer, Mon-Fri (incl). A sample of $n = 100$ students is taken, and their individual consumption of pints of beer during a typical working week is recorded. By calculating the proportion, or percentage, of students who consume, respectively, 1, 2, 3, etc, pints of beer, the **relative frequency diagram** can be constructed, as depicted in Figure 2.1.

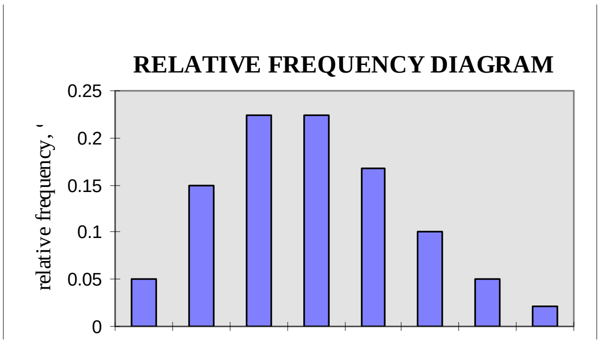


Figure 1.1: Beer Consumption (Source: fictitious)

This diagram, simply places a bar of height equal to the appropriate *proportion* (or percentage) for each pint. Notice that the bars are separated by spaces (i.e., they are isolated) which exemplifies the discrete nature of the data. The term 'relative frequency' simply means 'percentage, or proportion, in the sample'. Thus, we see from the diagram that the relative frequency of 1 pint is 0.15, or 15%; this means that 15 of the sample of $n = 100$ drank only 1 pint per week. Similarly, 5% did not drink any, whilst 42% of those students questioned drink at most two pints of beer during the week!

- **Relative frequency:** If a sample consists of n individuals (or items), and $m \leq n$ of these have a particular characteristic, denoted $\mathcal{A}$, then the relative frequency (or proportion) of characteristic $\mathcal{A}$ in the sample is calculated as $\frac{m}{n}$. The percentage of observations with characteristic $\mathcal{A}$ in the sample would be $\left(\frac{m}{n} \times 100\right)$ %. E.g. 0.65 is equivalent to 65%.

1.2.2 Continuous data

Example: we have a sample of $n = 87$ observations, which record the time taken (in completed seconds) for credit card customers to be served at Piccadilly station Booking Office. The 87 observations (the raw data) are listed as: $54, 63, 44, 60, 60, ...$etc. Due to rounding, some recorded times are repeated (but not very often) and some are never repeated. (These are continuous data in reality: it is possible to wait 55.78654 seconds, but it will not be recorded as such.) The data can be summarised, graphically, using a **histogram** which is constructed as follows:

- Group the raw data into intervals/classes, not necessarily of the same length:
  - the data are continuous, so there must (in general) be no spaces between intervals
  - take into account the fact that the data often rounded or, as in this case, measurements are recorded to the nearest second. That is, note that if $x = \textit{recorded time}$ and $t = \textit{actual time}$ then $x = 50$ implies that $50 \leqslant t < 51$ or $t \in [50, 51)$, meaning 't is greater than or equal to 50, but strictly less than 51'.
  - the number of intervals chosen is often a fine judgement: not too many, but not too small. Depending on the data 5 to 10 is often sufficient, but in the end you should choose the number of intervals to give you informative picture about the distribution of the data. Below I have chosen 6 intervals with the first being $[40, 50)$ where 40 is called the *lower class limit* and 50 is the *upper class limit.* The *class width* (of any interval) is the difference between the upper class limit and the lower class limit. For the first interval this is $50 - 40 = 10$.
  - record the *frequency* in each interval; i.e., record the number of observations which fall in each of the constructed intervals
- For each frequency, calculate the *relative frequency* (*rel. freq.*) which is "frequency divided by total number of observations".
- For each relative frequency construct a number called the density (of the interval) which is obtained as "relative frequency divided by class width".

Such manipulations give rise to the following *grouped frequency table*:

waiting time	*class width*	*mid-point*	*frequency*	*rel. freq.*	*density*
$[a, b)$	$(b - a)$	$(a + b)/2$			
$[40, 50)$	10	45	13	0.15	0.015
$[50, 55)$	5	52.5	12	0.14	0.028
$[55, 60)$	5	57.5	27	0.31	0.062
$[60, 65)$	5	62.5	22	0.25	0.050
$[65, 70)$	5	67.5	10	0.115	0.023
$[70, 75)$	5	72.5	3	0.035	0.007

Notice that the entries under *rel. freq.* sum to 1, as they should (why?). Using these ideas we construct a histogram, which conveys the impression of "*how thick on the ground*" observations are. Again, the graph is constructed from *bars,* but in such a way as to exemplify the underlying continuous nature of the data:

- construct bars over the intervals
- the bars must be connected - **no spaces -** this reflects the fact that the data are continuous (unless you think they are informative, avoid constructing intervals contain no observations)
- the **area** of each bar must be **equal to the relative frequency of that interval**

The **height** of each bar, so constructed, is called the **density** $= \frac{\text{relative frequency}}{\text{class width}}$, which gives rise to the last column in the grouped frequency table.

The resulting *histogram* looks like Figure 2.2 and you should be able to verify how it is constructed from the information given in the grouped frequency table.

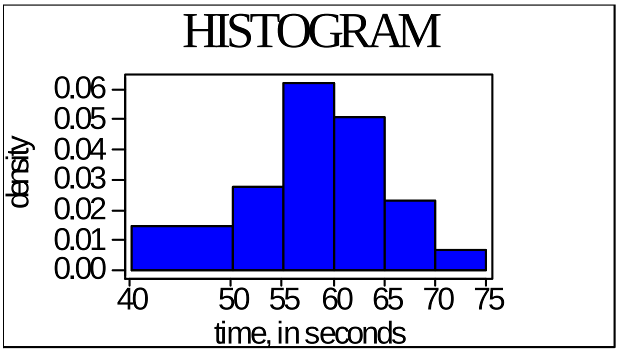


Figure 1.2: Waiting time of credit card customers

1.2.3 Two variables: Scatter diagrams/plots

The graphical summaries, introduced above, are for summarising just one variable. An interesting question is whether two (or more) characteristics of each member of a sample are inter-related. For example, one may be interested in whether an individual's weight (variable Y) is related to height (variable X).

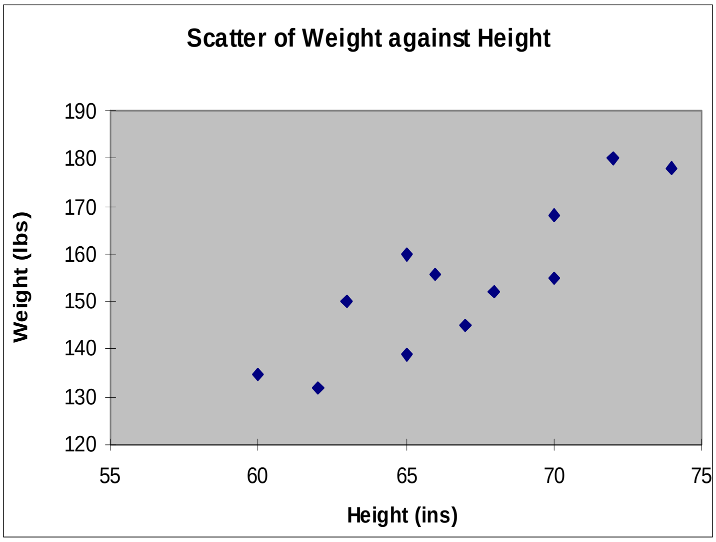


Figure 1.3: Weight (Y) against Height (X)

Consider a sample of $n = 12$, first year undergraduates, where for each individual (denoted i) we record the following information for individual i:

- y_i = observed weight measured in pounds (lbs); and x_i = observed height measured in inches.

These data naturally give a sequence of 12 co-ordinates, $\{x_i, y_i\}$, $i = 1, \ldots, 12$, which can be plotted to give the scatter diagram in Figure 2.3.

This sort of diagram should not be viewed as way of detecting the precise nature of the relationship between the two variables, X and Y, in general- a lot of common sense is required as well. Rather, it merely illuminates the simplest, most basic, relation of whether larger y values appear to be associated with larger (or smaller) values of the x variable; thereby signifying an underlying positive (respectively, inverse or negative) observed relationship between the two. This may be *suggestive* about the general relationship between height (X) and weight (Y), but is by no means conclusive. Nor does it inform us on general cause and effect; such as do changes in teh variable X cause changes in variable Y, or the other way around. Therefore care must be taken in interpreting such a diagram. However, in this case, if there is cause and effect present then it seems plausible that it will run from X to Y, rather than the other way around; e.g., we might care to use height as a predictor of weight.[1] In general, if cause and effect does run from the X variable to the Y variable, the scatter plot should be constructed with values of X, denoted x, on the horizontal axis and those for Y, denoted y, on the vertical axis.

1.3 Numerical summaries

In the previous sections we looked at graphical summaries of data. We now describe three numerical summaries. When these summaries are applied to a set of data, they return a number which we can interpret. Indeed, the numbers so computed provide summary information about the diagrams that could have be constructed.

Shall look at three categories of numerical summaries:

- location or *average*
- *dispersion*, spread or *variance*
- association, *correlation* or *regression*

1.3.1 Location

A measure of *location* tells us something about where the centre of a set of observations is. We sometimes use the expression *central location*, *central tendency* or, more commonly, *average*. We can imagine it as the value around which the observations in the sample are *distributed*. Thus, from Figure 2.1 we might say that the number of pints consumed are distributed around 3; or that the location of the distribution of waiting times, from Figure 2.2, appears to be between 55 and 60 seconds.

The simplest numerical summary (descriptive statistic) of location is the *sample (arithmetic) mean*:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{(x_1 + x_2 + \ldots + x_n)}{n}.$$

It is obtained by adding up all the values in the sample and dividing this total by the sample size. It uses all the observed values in the sample and is the most popular measure of location since it is particularly easy to deal

[1] When analysing the remains of human skeletons, the length of the femur bone is often used to predict height, and from this weight of the subject when alive.

with theoretically. Another measure, with which you may be familiar, is the *sample median.* This does not use all the values in the sample and is obtained by finding the middle value in the sample, once all the observations have been ordered from the smallest value to the largest. Thus, 50% of the observations are larger than the median and 50% are smaller. Since it does not use all the data it is less influenced by extreme values (or outliers), unlike the sample mean. For example, when investigating income distributions it is found that the mean income is higher than the median income; this is so because the highest 10% of the earners in the country will not affect the median, but since such individuals may earn extremely high amounts they will raise the overall mean income level.

In some situations, it makes more sense to use a *weighted sample mean*, rather than the arithmetic mean:

- weighted mean: $\sum_{i=1}^{n} w_i x_i = w_1 x_1 + w_2 x_2 + \ldots + w_n x_n$, where the weights $(w_1, \ldots, w_n)$ satisfy $\sum_{i=1}^{n} w_i = 1$.

Weighted means are often used in the construction of index numbers. (An example of where it might be a useful calculation is given in Exercise 1.) Note that equal weights of $w_i = n^{-1}$, for all i, gives the arithmetic mean.

All the above measures of location can be referred to as an *average.* One must, therefore, be clear about what is being calculated. Two politicians may quote two different values for the 'average income in the U.K.'; both are probably right, but are computing two different measures!

1.3.2 Dispersion

A measure of *dispersion* (or variability) tells us something about how much the values in a sample differ from one another and, more specifically, how closely these values are distributed around the central location.

We begin by defining a deviation from the arithmetic mean (**note the use of the lower case letter for a deviation**):

- deviation: $d_i = x_i - \bar{x}$

As an average measure of deviation from $\bar{x}$, we could consider the arithmetic mean of deviations, but this will always be zero (and is demonstrated in a worked exercise). A more informative alternative is the Mean Absolute Deviation (MAD):

- MAD: $\frac{1}{n}\sum_{i=1}^{n} |d_i| = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}| > 0$,

or the Mean Squared Deviation (MSD):

- MSD: $\frac{1}{n}\sum_{i=1}^{n} d_i^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 > 0$.

Like the arithmetic mean, the MSD is easier to work with and lends itself to theoretical treatment. The MSD is sometimes called the sample variance; *this will not be so in this course.* It is far more convenient to define the sample variance as follows (sometimes referred to as the "$n-1$" method)

- **Sample Variance** (*the* $n-1$ *method*): $s^2 = \dfrac{1}{n-1}\sum_{i=1}^{n} d_i^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 > 0$

The square root of this is:

- **Standard Deviation**: $s = +\sqrt{s^2} > 0.$

If we have a set of observations, $y_1, ..., y_n$, from which we calculate the variance, we might denote it as s_y^2 to distinguish it from a variance calculated from the values $x_1, ..., x_n$, which we might denotes as s_x^2.

Table 2.1, using the sample of data on heights and weights of sample of 12 first year students, illustrates the mechanical calculations. These can be automated in EXCEL.:

- *Example*

Let Y = Weight (lbs); X = Height (ins), with observations obtained as (y_i, x_i), $i = 1, ..., 12$.

i	y_i	x_i	$y_i - \bar{y}$	$x_i - \bar{x}$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})^2$	$(y_i - \bar{y}) \times (x_i - \bar{x})$
1	155	70	0.833	3.167	0.694	10.028	22.639
2	150	63	-4.167	-3.833	17.361	14.694	15.972
3	180	72	25.833	5.167	667.361	26.694	133.472
4	135	60	-19.167	-6.833	367.361	46.694	130.972
5	156	66	1.833	-0.833	3.361	0.694	-1.528
6	168	70	13.833	3.167	191.361	10.028	43.806
7	178	74	23.833	7.167	568.028	51.361	170.806
8	160	65	5.833	-1.833	34.028	3.361	-10.694
9	132	62	-22.167	-4.833	491.361	23.361	107.139
10	145	67	-9.167	0.167	84.028	0.028	-1.528
11	139	65	-15.167	-1.833	230.028	3.361	27.806
12	152	68	-2.167	1.167	4.694	1.361	-2.528
$\sum$	1850	802	0	0	2659.667	191.667	616.333

Table 2.1

Arithmetic means are: $\bar{y} = 1850/12 = 154.167$, i.e., just over 154 lbs; $\bar{x} = 802/12 = 66.833$, i.e., just under 67 inches.

Standard deviations are: for observations on the variable Y, $s_y = +\sqrt{2659.667/11} = 21.898$, i.e. just under 22 lbs; and for variable X, $s_x = +\sqrt{191.667/11} = 4.174$, i.e., just over 4 lbs.

1.3.3 Correlation and Regression

A commonly used measure of association is the *sample correlation coefficient,* which is designed to tell us something about the characteristics of a scatter plot of observations on the variable Y against observations on the variable X. In particularly, are higher than average values of y associated with higher than average values of x, and vice-versa? Consider again the scatter plot of weight against height. Now superimpose on this graph the horizontal line of $y = 154$ (i.e., $y \cong \bar{y}$) and also the vertical line of $x = 67$ (i.e., $x \cong \bar{x}$); see Figure 2.4.

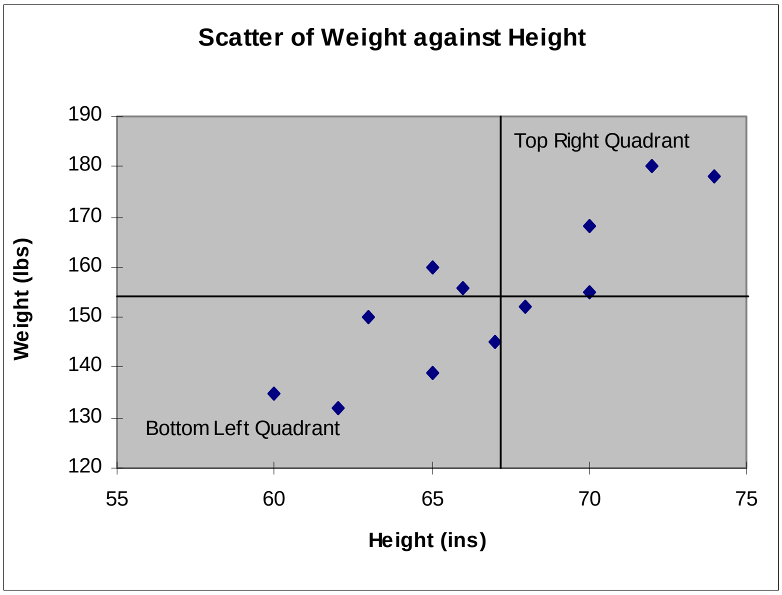


Figure 2.4: Scatter Diagram with Quadrants

Points in the obtained *upper right quadrant* are those for which weight is higher than average **and** height is higher than average; points in the *lower left quadrant* are those for which weight is lower than average **and** height is lower than average. Since most points lie in these two quadrants, this suggests that higher than average weight is associated with higher than average height; whilst lower than average weight is associated with lower than average height; i.e., as noted before, a positive relationship between the observed x and y. If there were no association, we would expect to a roughly equal distribution of points in all four quadrants. On the basis of this discussion, we seek a number which captures this sort of relationship.

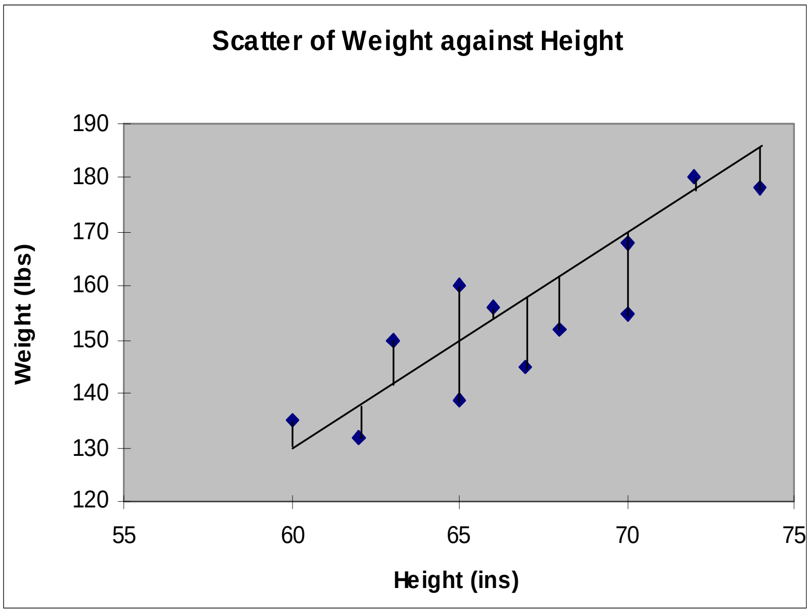


Figure 2.5: Scatter of mean deviations

Such a number is based, again, on the calculation and analysis of the respective deviations: $(x_i - \bar{x})$ and $(y_i - \bar{y})$, $i = 1, \ldots, 12$, and to help we plot these mean deviations in Figure 2.5. Consider the pairwise products of these deviations, $(x_i - \bar{x}) \times (y_i - \bar{y})$, $i = 1, \ldots, 12$. If observations all fell in either the top right or bottom left quadrants (a positive relationship), then $(x_i - \bar{x}) \times (y_i - \bar{y})$ would be positive for all i. If all the observations fell in either the top left or bottom right quadrants (a negative relationship), then $(x_i - \bar{x}) \times (y_i - \bar{y})$ would be negative for all i. Allowing for some discrepancies, a positive relationship should result in $(x_i - \bar{x}) \times (y_i - \bar{y})$ being positive on *average*; whilst a negative relationship would imply that $(x_i - \bar{x}) \times (y_i - \bar{y})$ should be negative on *average*. This suggests that a measure of association, or correlation, can be based upon $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$, the average product of deviations; note following standard mathematical notation the product $(x_i - \bar{x}) \times (y_i - \bar{y})$ is simply written as $(x_i - \bar{x})(y_i - \bar{y})$. The disadvantage of this is that it's size depends upon the units of measurement for y and x. For example, the value of $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ would change if height was measured in centimetres, but results only in a re-scaling of the horizontal axis on the scatter plot and should not, therefore, change the extent to which we think height and weight are correlated.

Fortunately, it is easy to construct an index which is independent of the scale of measurement for both y and x, and this is:

- the *sample correlation coefficient:*

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}; \qquad -1 \leq r \leq 1.$$

- the constraint that $-1 < r < 1$, can be shown to be true algebraically for any given set of data $\{x_i, y_i\}$, $i = 1, \ldots, n$.

For the above example, of weights and heights:

$$r = 616.333/\sqrt{(2659.667) \times (191.667)} = 0.863.$$

The following limitations of the correlation coefficient should be observed.

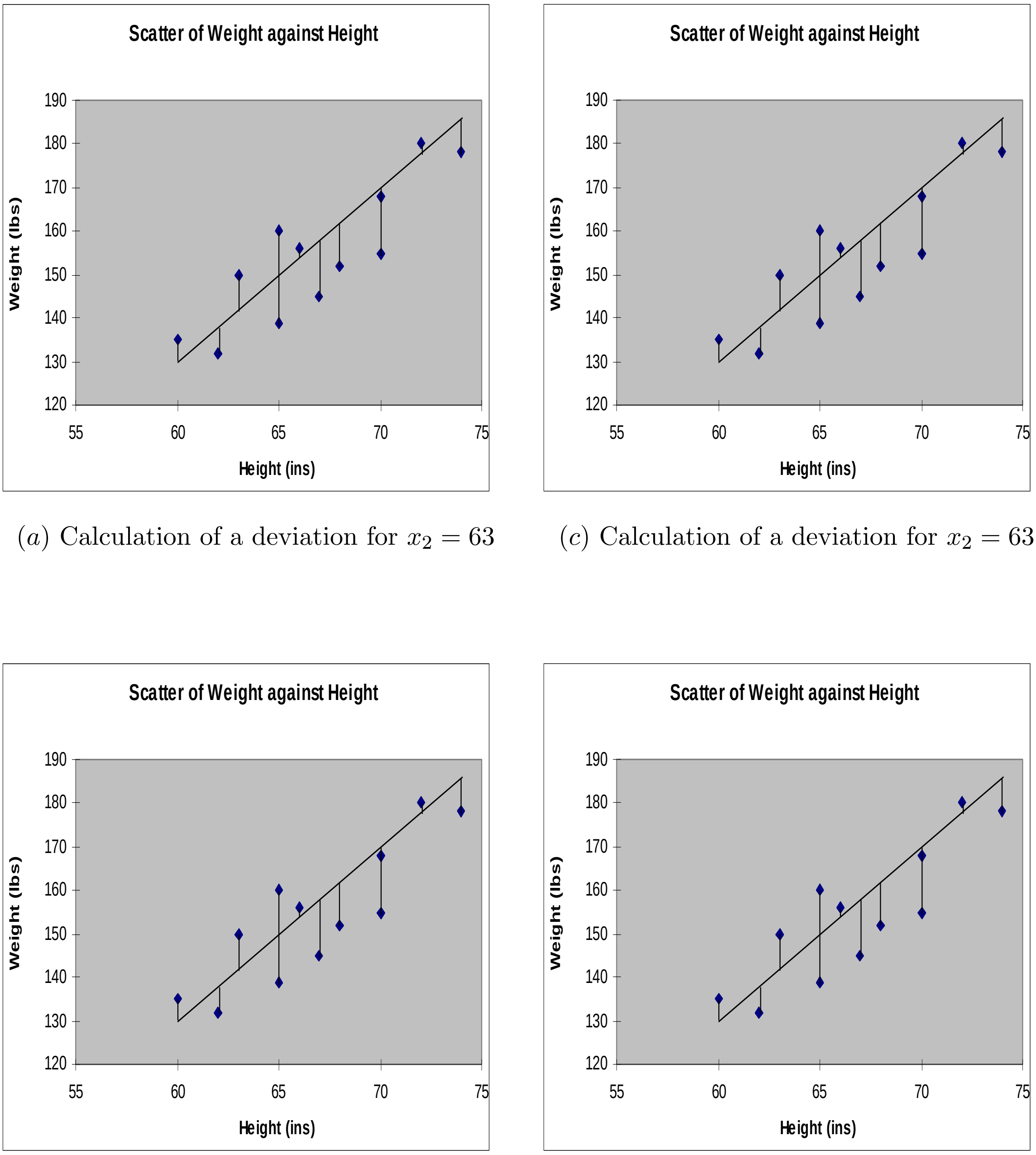
1. In general, this sort of analysis does not imply causation, in either direction. Variables may appear to move together for a number of reasons and not because one is causally linked to the other. For example, over the period 1945 – 64 the number of TV licences (x) taken out in the UK increased steadily, as did the number of convictions for juvenile delinquency (y). Thus a scatter of y against x, and the construction of the sample correlation coefficient reveals an apparent positive relationship. However, to therefore claim that increased exposure to TV causes juvenile delinquency would be extremely irresponsible.

2. The sample correlation coefficient gives an index of the apparent linear relationship only. It 'thinks' that the scatter of points must be distributed about some underlying straight line (with non-zero slope when $r \neq 0$). This is discussed further below, and in Exercise 1.

Let us now turn our attention again to Figures 2.3. and 2.4. Imagine drawing a straight line of *best fit* through the scatter of points in Figure 2.3, simply from 'visual' inspection. You would try and make it 'go through' the scatter, in some way, and it would probably have a positive slope. *For the present purposes, draw a straight line on Figure 2.3 which passes through the two co-ordinates of* $(60, 125)$ *and* $(75, 155)$. Numerically, one of the things that the correlation coefficient does is assess the slope of such a line: if $r > 0$, then the slope should be positive, and vice-versa. Moreover, if r is close to either 1 (or -1) then this implies that the scatter is quite closely distributed around the line of best fit. What the correlation coefficient doesn't do, however, is tell us the exact position of line of best fit. This is achieved using *regression.*

The line you have drawn on the scatter can be written as $z = 2x + 5$, where z, like y, is measured on the vertical axis; notice that I use z rather than y to distinguish it from the actual values of y. Now for each value x_i, we can construct a z_i, $i = 1, \ldots, 12$. Notice the difference between the z_i and the y_i, which is illustrated in Figure 2.6a for the case of $x_2 = 63$ (the second observation on height), where the actual weight is $y_2 = 150$. The difference, in this case, is $y_2 - z_2 = 150 - 131 = 19$. Correspondingly, differences also referred to as *deviations* or *residuals,* calculated for all values of x_i are depicted in Figure 2.6b. Note that the difference (deviation or residual) for $x_8 = 65$ is different from that for $x_{11} = 65$, because the corresponding values of y are different. The line $z = 2x + 5$ is just one of a number of possible lines that we could have drawn through the scatter. *But does it provide the "best fit"*? We would like the line of best fit to generate values of z which are close (in some sense) to the corresponding actual values of y, for all given values of x. To explore this further, consider another line: $z = 4x - 110$. This is illustrated in Figure 2.6c, where a deviation (or residual) is depicted for $x_2 = 63$; this deviation (or residual) differs from the corresponding one for the line $z = 2x + 5$. *But which of these two lines is better*? Intuitively, $z = 4x - 110$ looks better, but can we do better still? Formally, *yes*, and the idea behind "line of best fit" is that we would like the sum of squared deviations (or sum of squared residuals) between the z_i and the actual y_i to be as low as possible. For the line, $z = 2x + 5$, the deviation from the value y at any data value is $y_i - z_i$ (which is equal to $y_i - 2x_i - 5$) and the sum of the squared deviations (sum of squared residuals) is:

$$\begin{aligned}\sum_{i=1}^{12}(y_i - z_i)^2 &= \sum_{i=1}^{12}(y_i - 2x_i - 5)^2 = \\ &= (155 - 135)^2 + (150 - 121)^2 + \ldots + (152 - 131)^2 \\ &= 3844.\end{aligned}$$



(a) Calculation of a deviation for $x_2 = 63$

(c) Calculation of a deviation for $x_2 = 63$

(b) Deviations for each x_i

(d) Deviations for each x_i

Figure 2.6: Alternative lines and deviations

Now consider the other possibility: $z = 4x - 110$. Based on the same calculation, the sum of squared deviations (sum of squared residuals) between these $z_i$and the actual y_i is lower, being only 916;so, the second line is better. The question remains, though, as to what line would be "best"; i.e., is there yet another line ($a + bx$, for some numbers a and b) for which the implied sum of squared deviations is smallest? The answer is "yes there is" and the construction of such a line is now described.

The best line, $a + bx$,chooses the intercept a and slope $b$such that the implied average squared deviation between $\{a + bx_i\}$and $\{y_i\}$ is minimised. Properties of the line of best fit problem are summarised as:

- Line of best fit, or *regression equation*, has the form: $\hat{y} = a + bx$.
- The intercept, a, and slope, b, are obtained by *regressing* y on x.
- This minimises the *average squared deviation* between $\hat{y}_i$ and y_i, and constrains the line to pass through the point $(\bar{x}, \bar{y})$.

The technique of obtaining a and b in this way is also known as *ordinary least squares,* since it minimises the sum of squared deviations (sum of sqaured residuals) from the fitted line.

We shall not dwell on the algebra here, but the solutions to the problem are:

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad a = \bar{y} - b\bar{x};$$

see *Additional Worked Exercises* on the course website.

Applying the technique to the weight and height data yields: $b = 616.333/191.667 = 3.2157$, and $a = 154.167 - 3.2157 \times 66.833 = -60.746$, giving smallest sum of squares deviations as 677.753. This line is superimposed on the scatter in Figure 2.7.

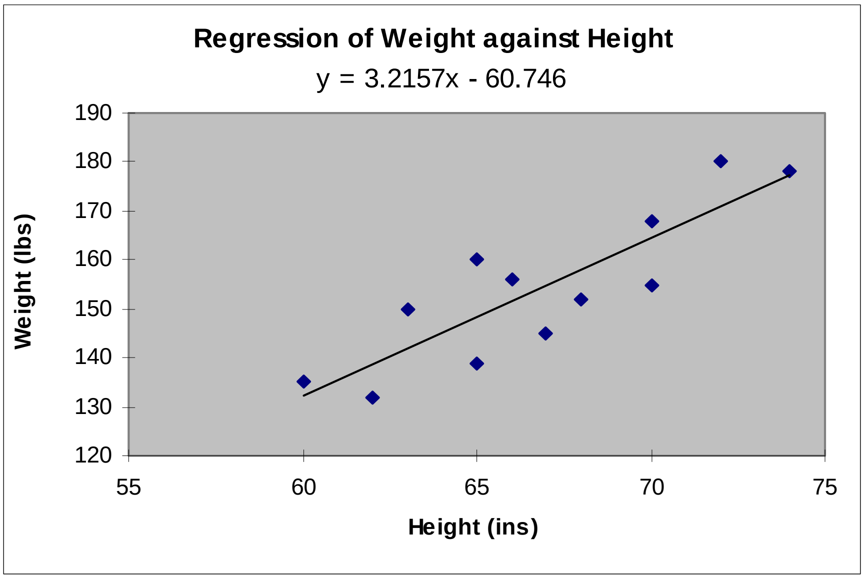


Figure 2.7: Scatter Plot with Regression Line

1.3.4 Interpretation of regression equation

Note that b is the slope of the fitted line, $\hat{y} = a + bx$; i.e., the derivative of $\hat{y}$ with respect to x :

- $b = d\hat{y}/dx = dy/dx + error$

  and measures the increase in y for a unit increase in x.

Alternatively, it can be used to impute an elasticity. Elementary economics tells us that if y is some function of x, $y = f(x)$, then the elasticity of y with respect to x is given by the *logarithmic derivative:*

- *elasticity*: $\dfrac{d\log(y)}{d\log(x)} = \dfrac{dy/y}{dx/x} \cong (x/y)b$

  where we have used the fact that the differential $d\log(y) = \dfrac{1}{y}dy$. Such an elasticity is often evaluated at the respective sample means; i.e., it is calculated as $(\bar{x}/\bar{y})b$.

- *Example:* In applied economics studies of demand, the log of demand (Q) is regressed on the log of price (P), in order to obtain the fitted equation (or relationship). For example, suppose an economic model for the quantity demanded of a good, Q, as a function of its price, P, is postulated as approximately being $Q = aP^b$ where a and b are unknown 'parameters', with $a > 0$, $b < 1$ to ensure a positive downward sloping demand curve. Taking logs on both sides we see that $\log(Q) = a^* + b\log(P)$, where $a^* = \log(a)$. Thus, if n observations are available, (q_i, p_i), $i = 1, ..., n$, a scatter plot of $\log(q_i)$ on $\log(p_i)$ should be approximately linear in nature. Thus suggests that a simple regression of $\log(q_i)$ on $\log(p_i)$ would provide a direct estimate of the elasticity of demand which is given by the value b.

1.3.5 Transformations of data

Numerically, transformations of data can affect the above summary measures. For example, in the weight-height scenario, consider for yourself what would happen to the values of a and b and the correlation if we were to use kilograms and centimetres rather than pounds and inches.

A more important matter arises if we find that a scatter of the some variable y against another, x, does not appear to reveal a linear relationship. In such cases, linearity may be retrieved if y is plotted against some function of x (e.g., $\log(x)$ or x^2, say). Indeed, there may be cases when Y also needs to be transformed in some way. That is to say, transformations of the data (via some mathematical function) may render a non-linear relationship “more” linear.

1.4 Exercise 1

1. Within an area of 4 square miles, in a city centre, there are 10 petrol stations. The following table gives the price charged at each petrol station (pre-Budget in pence per litre) for unleaded petrol, and the market share obtained:

Petrol Station	1	2	3	4	5	6	7	8	9	10
Price	96.0	96.7	97.5	98.0	97.5	98.5	95.5	96.0	97.0	97.3
Market Share	0.17	0.12	0.05	0.02	0.05	0.01	0.2	0.23	0.1	0.05

   (a) Calculate the sample (or arithmetic) mean and standard deviation (using the $n - 1$ method) of the price.

   (b) Calculate the weighted mean, using the market share values as the set of weights. Explain why the weighted and unweighted means differ.

   (c) Calculate the (sample) mean and standard deviation of the price per gallon.

2. Consider, again, the data on $X =$ *petrol prices* and $Y =$ *market share* given in the previous question. You must find the answers to the following questions “by hand” so that you understand the calculations involved; however, you should also check you answers using EXCEL and be prepared to hand in a copy of the EXCEL worksheet containing the appropriate calculations.

   (a) Show that the correlation coefficient between these two variables is -0.9552. Interpret this number and, in particular,is the sign as you would expect?

   (b) Use the data to fit the regression line, $\hat{y} = a + bx$; i.e., show that regressing y on x yields a value of $a = 6.09$ and $b = -0.0778$. Why would you expect the value of b to be negative given (a)?

   (c) Suppose, post-budget, the every price rises uniformly by 2 pence. Assuming that the market shares stay the same, write down what a regression of *market share* on *prices* would now yield for values of a and b.

3. You should use EXCEL to answer the following and be prepared to hand in a copy of the EXCEL worksheet containing the appropriate calculations.

Refer to the data given below in which you are given 11 observations on a variable labelled X and 11 observations on each of three variables Y, Z and W.

observation	x	y	z	w
1	10	8.04	9.14	7.46
2	8	6.95	8.14	6.77
3	13	7.58	8.74	12.74
4	9	8.81	8.77	7.11
5	11	8.33	9.26	7.81
6	14	9.96	8.10	8.84
7	6	7.24	6.13	6.08
8	4	4.16	3.10	5.39
9	12	10.84	9.13	8.15
10	7	4.82	7.26	6.42
11	5	5.68	4.74	5.73

(a) Use EXCEL to obtain three separate scatter diagrams of y against x, z against x and w against x.

(b) Show that the sample correlation coefficient between y and x is 0.82 and that this is the same as the corresponding correlation between z and x and also w and x.

(c) Using EXCEL, show that the three separate regressions of y on x, z on x and w on x **all** yield a "line of best fit" or regression equation of the form: $3 + 0.5x$; i.e., a line with intercept 3 and slope 0.5. Use EXCEL to superimpose this regression line on each of the three scatter diagrams obtained in part (a).

(d) To what extent do you feel that *correlation* and *regression* analysis is useful for the various pairs of variables?

4. Go to the module website (Exercises) and download the EXCEL spreadsheet, **tute1.xls**. This contains data on carbon monoxide emissions (CO) and gross domestic product (GDP) for 15 European Union countries for the year 1997.

(a) Using EXCEL, construct a scatter plot of carbon monoxide emissions against gross domestic product, construct the regression line (of CO on GDP) and calculate the correlation coefficient.

(b) Repeat the exercise, but this time using the natural logarithm of CO, ln(CO), and ln(GDP).

(c) What do you think this tells us about the relationship between the two variables?

1.4.1 Exercises in EXCEL

It is assumed that you know how to use EXCEL to perform the simple statistical calculations, as described in Sections 1 and 2 of these notes. Calculate sample means and standard deviations of the weight and height variables (Table 2.1). Also calculate the correlation between the two variables and obtain a scatter plot with regression line added (as in Figure 2.7).

Chapter 2

INTRODUCING PROBABILITY

So far we have been looking at ways of summarising samples of data drawn from an underlying population of interest. Although at times tedious, all such arithmetic calculations are fairly mechanical and straightforward to apply. To remind ourselves, one of the primary reasons for wishing to summarise data is so assist in the development of inferences about the population from which the data were taken. That is to say, we would like to elicit some information about the mechanism which generated the observed data.

We now start on the process of developing mathematical ways of formulating inferences and this requires the use of *probability*. This becomes clear if we think back to one of the early questions posed in this course: *prior to sampling is it possible to predict with absolute certainty what will be observed*? The answer to this question is *no*; although it would be of interest to know how *likely* it is that certain values would be observed. Or, what is the *probability* of observing certain values?

Before proceeding, we need some more tools:

2.1 Venn diagrams

Venn diagrams (and diagrams in general) are of enormous help in trying to understand, and manipulate probability. We begin with some basic definitions, some of which we have encountered before.

- **Experiment:** any process which, when applied, provides data or an outcome; e.g., rolling a die and observing the number of dots on the upturned face; recording the amount of rainfall in Manchester over a period of time.

- **Sample Space:** set of possible outcomes of an experiment; e.g., S (or Ω) $= \{1, 2, 3, 4, 5, 6\}$ or $S = \{x; x \geq 0\}$, which means '*the set of* *real non-negative real numbers*'.

- **Event:** a *subset* of S, denoted $E \subset S$; e.g., $E = \{2, 4, 6\}$ or $E = \{x; 4 < x \leq 10\}$, which means '*the set of real numbers which are strictly bigger than* 4 *but less than or equal to* 10'.

- **Simple Event**: just one of the possible outcomes on S

  – note that an event, E, is a collection of simple events.

Such concepts can be represented by means of a Venn Diagram, as in Figure 3.1.

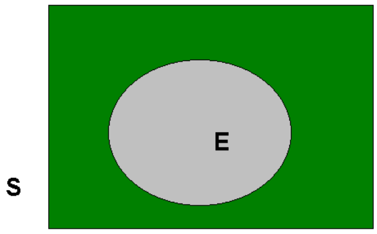


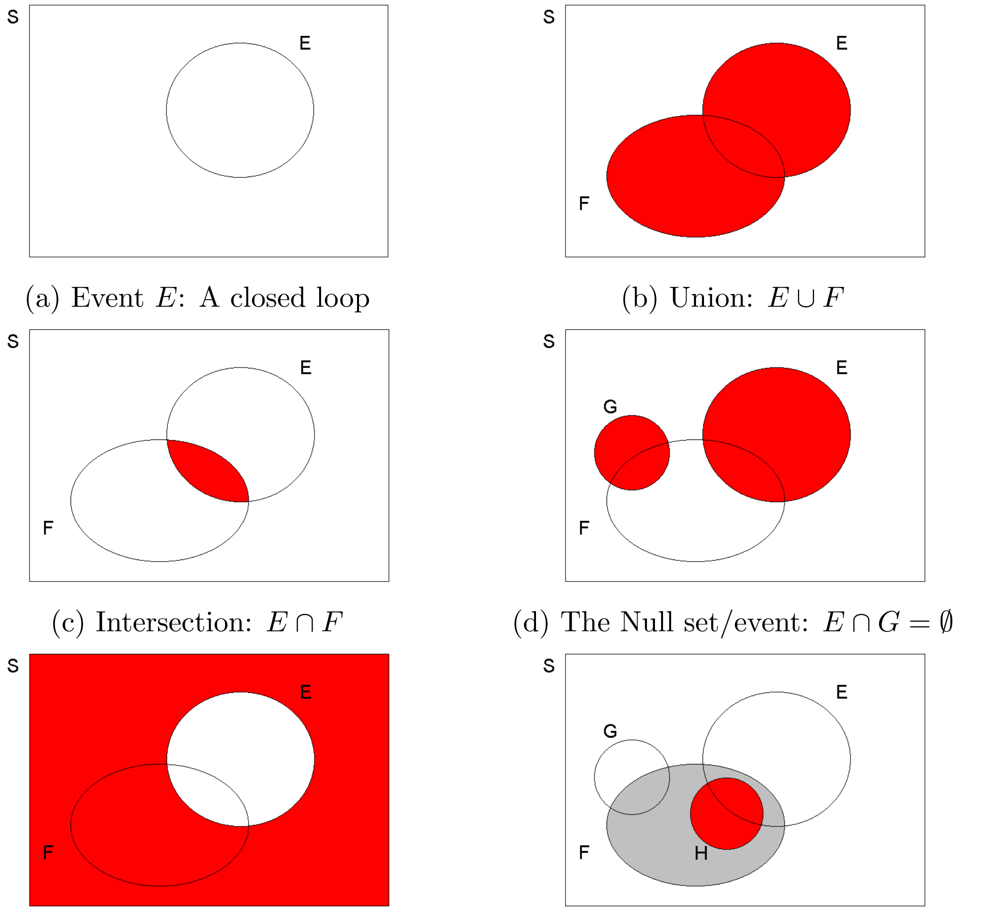
Figure 2.1: A Venn Diagram

The sample space, S, is depicted as a closed rectangle, and the event E is a closed loop wholly contained within S and we write (in set notation) $E \subset S$.

In dealing with probability, and in particular the probability of an event (or events) occurring, we shall need to be familiar with **UNIONS, INTERSECTIONS** and **COMPLEMENTS**.

To illustrate these concepts, consider the sample space $S = \{x; x \geq 0\}$, with the following events defined on S, as depicted in Figure 3.2:

$E = \{x; 4 < x \leq 10\}$, $F = \{x; 7 < x \leq 17\}$, $G = \{x; x > 15\}$, $H = \{x; 9 < x \leq 13\}$.



(a) Event E: A closed loop

(b) Union: $E \cup F$

(c) Intersection: $E \cap F$

(d) The Null set/event: $E \cap G = \emptyset$

(e) Complement of E: $\bar{E}$

(f) Subset of F: $H \subset F$ and $H \cap F = H$

Figure 3.2: Intersections, unions, complements and subsets

- The *union* of E and F is denoted $E \cup F$, with $E \cup F = \{x; 4 < x \leq 17\}$; i.e., it contains elements (simple events) which are either in E or in F or (perhaps) in both. This is illustrated on the Venn diagram by the dark shaded area in diagram (b).

- The *intersection* of E and F is denoted $E \cap F$, with $E \cap F = \{x; 7 \leq x \leq 10\}$; i.e., it contains elements (simple events) which are common to both E and F. Again this is depicted by the dark shaded area in (c). If events have no elements in common (as, for example, E and G) then they are said to be *mutually exclusive*, and we can write $E \cap G = \emptyset$, meaning the *null set* which contains no elements. Such a situation is illustrated on the Venn Diagram by events (the two shaded closed loops in (d)) which do not overlap. Notice however that $G \cap F \neq \emptyset$, since G and F

have elements in common.

- The *complement* of an event E, say, is everything defined on the sample space which is not in E. This event is denoted $\bar{E}$, the dark shaded area in (e); here $\bar{E} = \{x; x \leq 4\} \cup \{x; x > 10\}$.

- Finally note that H is a sub-set of F; see (f). It is depicted as the dark closed loop wholly contained within F, the lighter shaded area, so that $H \cap F = H$; if an element in the sample space is a member of H then it must also be member of F. (In mathematical logic, we employ this scenario to indicate that "H implies F", but not necessarily vice-versa.) Notice that $G \cap H = \emptyset$ but $H \cap E \neq \emptyset$.

2.2 Probability

The term *probability* (or some equivalent) is used in everyday conversation and so can not be unfamiliar to the reader. We talk of the probability, or chance, of rain; the likelihood of England winning the World Cup; or, perhaps more scientifically, the chance of getting a 6 when rolling a die. What we shall now do is develop a coherent theory of probability; a theory which allows us to combine and manipulate probabilities in a consistent and meaningful manner. We shall describe ways of dealing with, and describing, uncertainty. This will involve *rules* which govern our use of terms like probability.

There have been a number of different approaches (interpretations) of probability. Most depend, at least to some extent, on the notion of relative frequency as now described:

- Suppose an experiment has an outcome of interest E. The *relative frequency interpretation* of probability says that assuming the experiment can be repeated a large number of times then the relative frequency of observing the outcome E will settle down to a *number,* denoted $\Pr(E)$, $P(E)$ or $\mathrm{Prob}(E)$, called the **probability** of E.

This is illustrated in Figure 3.3, where the proportion of heads obtained after n flips of a fair coin is plotted against n, as n increases; e.g., of the first 100 flips, 55 were heads (55%). Notice that the plot becomes less 'wobbly' after about $n = 220$ and appears to be settling down to the value of $\frac{1}{2}$.

Due to this interpretation of probability, we often use observed sample proportions to approximate underlying probabilities of interest; see, for example, Question 4 of Exercise 2. There are, of course, other interpretations of probability; e.g., the subjective interpretation which simply expresses the strength of one's belief about an event of interest such as whether Manchester United will win the European Cup! Any one of these interpretations can

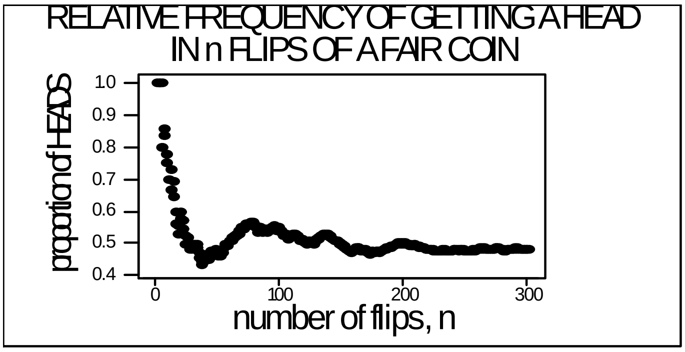


Figure 2.3: Relative frequency interpretation of probability

be used in practical situations provided the implied notion of probability follows a simple set of *axioms* or *rules.*

2.2.1 The axioms of probability

There are just *three* basic rules that must be obeyed when dealing with probabilities:

1. For any event E defined on S, i.e., $E \subset S$, $\Pr(E) \geq 0$; *probabilities are non-negative.*

2. $\Pr(S) = 1$; *having defined the sample space of outcomes, one of these outcomes must be observed.*

3. If events E and F are mutually exclusive defined on S, so that $E \cap F = \emptyset$, then $\Pr(E \cup F) = \Pr(E) + \Pr(F)$. In general, for any set of mutually exclusive events, $E_1, E_2, \ldots, E_k$, defined on S :

$$\Pr(E_1 \cup E_2 \cup \ldots \cup E_k) = \Pr(E_1) + \Pr(E_2) + \ldots \Pr(E_k)$$

   i.e., $\Pr\left(\bigcup_{j=1}^{k} E_j\right) = \sum_{j=1}^{k} \Pr(E_j)$.

In terms of the Venn Diagram, one can (and should) usefully think of the area of E, relative to that of S, as providing an indication of probability. (Note, from axiom 2, that the area of S is implicitly normalised to be unity).

Also observe that, contrary to what you may have believed, it is not one of the rules that $\Pr(E) \leq 1$ for any event E. Rather, this is an implication of the 3 rules given:

- **implications:** it must be that for any event E, defined on S, $E \cap \bar{E} = \emptyset$ and $E \cup \bar{E} = S$. By Axiom 1, $\Pr(E) \geq 0$ and $\Pr\left(\bar{E}\right) \geq 0$ and by Axiom 3 $\Pr(E) + \Pr(\bar{E}) = \Pr(S)$. So $\Pr(E) + \Pr\left(\bar{E}\right) = 1$, by Axiom 2. This implies that

1. (a) $0 \leq \Pr(E) \leq 1$

   (b) $\Pr(\bar{E}) = 1 - \Pr(E)$

The first of these is what we might have expected from probability (a number lying between 0 and 1). The second implication is also very important; it says that the probability of E not happening is '*one minus the probability of it happening*'. Thus when rolling a die, the probability of getting 6 is one minus the probability of getting either a 1, 2, 3, 4 or 5.

These axioms imply how to calculate probabilities on a sample space of equally likely outcomes. For example, and as we have already noted, the experiment of rolling a fair die defines a sample space of six, mutually exclusive and equally likely outcomes (1 to 6 dots on the up-turned face). The axioms then say that each of the six probabilities are positive, add to 1 and are all the same. Thus, the probability of any one of the outcomes must be simply $\frac{1}{6}$; which may accord with your intuition. A similar sort of analysis reveals that the probability of drawing a club from a deck of 52 cards is $\frac{13}{52}$, since any one of the 52 cards has an equal chance of being drawn and 13 of them are clubs; i.e., 13 of the 52 are clubs, so the probability of drawing a club is $\frac{13}{52}$. Notice the importance of the assumption of equally likely outcomes here.

In this, and the next section of notes, we shall see how these axioms can be used. Firstly, consider the construction of a probability for the *union* of two events; i.e., the probability that *either* E or F or (perhaps) *both* will occur. Such a probability is embodied in the *addition rule of probability*:

2.2.2 The addition rule of probability

When rolling a fair die, let E denote the event of an "odd number of dots" and F the event of the "number of dots being greater than, or equal, to 4". What is the probability of the event $E \cup F$? To calculate this we can collect together all the mutually exclusive (simple) events which comprise $E \cup F$, and then add up the probabilities (by axiom 3). These simple events are $1, 3, 4, 5$ or 6 dots. Each has a probability of $\frac{1}{6}$, so the required total probability is: $\Pr(E \cup F) = \frac{5}{6}$. Consider carefully how this probability is constructed and note, in particular, that $\Pr(E \cup F) \neq \Pr(E) + \Pr(F)$ since E and F have a simple event in common (namely 5 dots).

In general, we can calculate the probability of the union of events using the *addition rule of probability*, as follows.

- For any events, $E \subset S$ and $F \subset S$: $\Pr(E \cup F) = \Pr(E) + \Pr(F) - \Pr(E \cap F)$.

  So, in general, $\Pr(E \cup F) \leq \Pr(E) + \Pr(F)$.

This generalises to three events, E_1, E_2 and E_3 as

$$\begin{aligned}\Pr(E_1 \cup E_2 \cup E_3) &= \Pr(E_1) + \Pr(E_2) + \Pr(E_3) \\ &\quad - \Pr(E_1 \cap E_2) - \Pr(E_1 \cap E_3) - \Pr(E_2 \cap E_3) \\ &\quad + \Pr(E_1 \cap E_2 \cap E_3).\end{aligned}$$

We can demonstrate this as follows.
Note that

$$E \cup F = \left(E \cap \bar{F}\right) \cup (E \cap F) \cup \left(\bar{E} \cap F\right)$$

the union of 3 mutually exclusive events. These mutually exclusive events are depicted by the shaded areas **a**, **b** and **c**, respectively, in Figure 3.4.

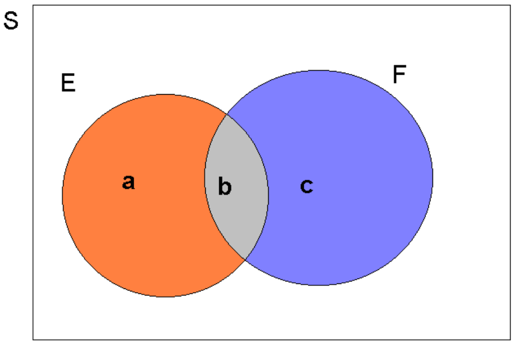


Figure 2.4: Decomposing $E \cup F$

Then by Axiom 3, and from the fact that the three events $\left(E \cap \bar{F}\right)$, $(E \cap F)$ and $\left(\bar{E} \cap F\right)$ are mutually exclusive so that the "area" occupied by $E \cup F$ is simply $\mathbf{a} + \mathbf{b} + \mathbf{c}$,

$$\Pr\left(E \cup F\right) = \Pr\left(E \cap \bar{F}\right) + \Pr\left(\bar{E} \cap F\right) + \Pr\left(E \cap F\right).$$

But also by Axiom 3, since $E = \left(E \cap \bar{F}\right) \cup (E \cap F)$, it must be that $\Pr(E) = \Pr\left(E \cap \bar{F}\right) + \Pr(E \cap F)$; similarly, $\Pr\left(\bar{E} \cap F\right) = \Pr\left(F\right) - \Pr\left(E \cap F\right)$. Putting all of this together gives

$$\Pr(E \cup F) = \Pr(E) + \Pr(F) - \Pr(E \cap F).$$

When E and F are mutually exclusive, so that $E \cap F = \emptyset$, this rule reveals Axiom 2: $\Pr(E \cup F) = \Pr(E) + \Pr(F)$.

- *Example:* What is the probability of drawing a Queen (Q) or a Club (C) in a single draw from a pack of cards? Now, 4 out of 52 cards are Queens, so $\Pr(Q) = \frac{4}{52}$, whilst $\Pr(C) = \frac{13}{52}$. The probability of drawing the Queen of Clubs is simply $\frac{1}{52}$; i.e., $\Pr(Q \cap C) = \frac{1}{52}$. What we require is a Club or a Queen, for which the probability is

$$\begin{aligned}\Pr(Q \cup C) &= \Pr(Q) + \Pr(C) - \Pr(Q \cap C) \\ &= \frac{4}{52} + \frac{13}{52} - \frac{1}{52} \\ &= \frac{16}{52} = \frac{4}{13}.\end{aligned}$$

- *Example:* Consider a car journey from Manchester to London via the M6 and M1. Let E = *heavy traffic somewhere on route* and F = *roadworks somewhere on route*. It is estimated that $\Pr(E) = 0.8$ and $\Pr(F) = 0.4$, whilst the probability of NOT encountering both is $\Pr(\overline{E \cap F}) = 0.6$. What is the probability of encountering heavy traffic or roadworks?

  We require $\Pr(E \cup F)$.

$$\begin{aligned}\Pr(E \cup F) &= \Pr(E) + \Pr(F) - \Pr(E \cap F) \\ &= \Pr(E) + \Pr(F) - (1 - \Pr(\overline{E \cap F})) \\ &= 0.8 + 0.4 - 1 + 0.6 \\ &= 0.8 = \Pr(E)\end{aligned}$$

  Notice that this implies, in this case, that $F \subset E$ (why?). This *model* then implies that when there are roadworks somewhere on route you are bound to encounter heavy traffic; on the other hand, you can encounter heavy traffic on route without ever passing through roadworks. (My own experience of this motorway inclines me towards this implication!)

Similar concepts apply when manipulating proportions as follows:

- *Example*: A sample of 1000 undergraduates were asked whether they took either Mathematics, Physics or Chemistry at A-level. The following responses were obtained: 100 just took Mathematics; 70 just took Physics; 100 just took Chemistry; 150 took Mathematics and Physics, but not Chemistry; 40 took Mathematics and Chemistry, but not Physics; and, 240 took Physics and Chemistry, but not Mathematics. What proportion took all three?

  This can be addressed with the following diagram:

  The shaded area contains the number who took all three, which can be deduced from the above information (since the total of the numbers assigned to each part of the Venn diagram must be 1000). The answer is therefore 30% (being 300 out of 1000).

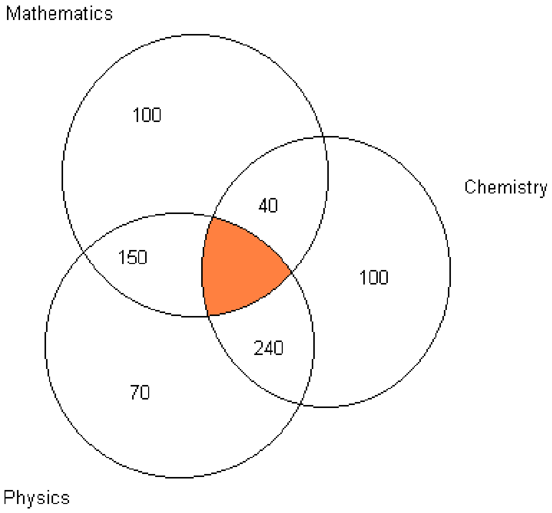


Figure 2.5: Venn Diagram for A-levels

- Two further results on unions, intersections and complements which are of use (and which are fairly easy to demonstrate using Venn diagrams) are **de Morgan Laws**:
  - $(\bar{A} \cap \bar{B}) = (\overline{A \cup B})$
  - $\bar{A} \cup \bar{B} = (\overline{A \cap B})$

Chapter 3

CONDITIONAL PROBABILITY

An important consideration in the development of probability is that of *conditional probability.* This refers to the calculation of updating probabilities in the light of revealed information. For example, insurance companies nearly always set their home contents insurance premiums on the basis of the postcode in which the home is located. That is to say, insurance companies believe the risk depends upon the location; i.e., the probability of property crime is assessed conditional upon the location of the property. (A similar calculation is made to set car insurance premiums.) As a result, the premiums for two identical households located in different parts of the country can differ substantially.

- In general, the probability of an event, E, occurring *given* that an event, F, has occurred is called the *conditional probability* of E given F and is denoted $\Pr(E|F)$.

As another example, it has been well documented that the ability of a new born baby to survive is closely associated with its birth-weight. A birth-weight of less than 1500g is regarded as dangerously low. Consider $E =$ *birth weight of a baby is less than* $1500g$, $F =$ *mother smoked during pregnancy*; then evidence as to whether $\Pr(E|F) > \Pr(E|\bar{F})$ is of considerable interest.

As a preliminary to the main development, consider the simple experiment of rolling a fair die and observing the number of dots on the upturned face. Then $S = \{1, 2, 3, 4, 5, 6\}$ and define events, $E = \{4\}$ and $F = \{4, 5, 6\}$; we are interested in $\Pr\left(E|F\right)$. To work this out we take F as known. Given this knowledge the sample space becomes restricted to simply $\{4, 5, 6\}$ and, given no other information, each of these 3 outcome remains equally likely. So the required event, 4, is just one of three equally likely outcomes. It therefore seems reasonable that $\Pr(E|F) = \frac{1}{3}$.

We shall now develop this idea more fully, using Venn Diagrams with the implied notion of area giving probability.

3.1 Conditional probability

Consider an abstract sample space, denoted by S, with events $E \subset S$, $F \subset S$. This is illustrated in Figure 4.1, where the important areas used in the construction of a conditional probability are highlighted as **a** and **b** :

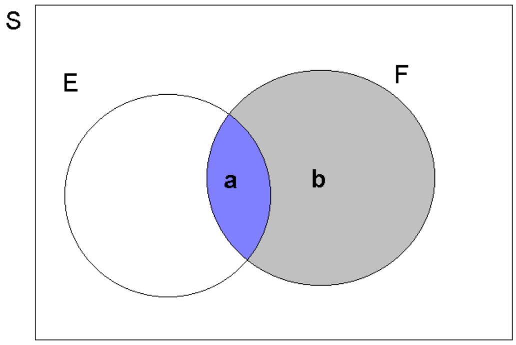


Figure 3.1: Areas used in constructing a conditional probability

In general, it is useful to think of $\Pr(E)$ as $\frac{area(E)}{area(S)}$; and similarly for $\Pr(F)$ and $\Pr(E \cap F)$ where $area(F) = a+b$ and $area(E \cap F) = a$. With this in mind, consider what happens if we are now told that F has occurred. Incorporating this information implies that the effective sample space becomes restricted to $S^* = F$, since F now defines what can happen and covers area $a+b$. On this new, restricted, sample space an outcome in E can only be observed if that outcome also belongs to F, and this only occurs in area a which corresponds to the event $E \cap F$. Thus the event of interest *now* is $E^* = E \cap F$, as defined on the *restricted* sample space of $S^* = F$.

In order to proceed with the construction of the conditional probability, $\Pr(E|F)$, let $area(S) = z$. Then, since the ratio of the area of the event of interest to that of the sample space gives probability, we have (on this restricted sample space)

$$\begin{aligned}\Pr(E|F) &= \frac{area\,(E \cap F)}{area\,(F)} \\ &= \frac{a}{a+b} \\ &= \frac{a/z}{(a+b)\,/z} \\ &= \frac{\Pr\,(E \cap F)}{\Pr\,(F)},\end{aligned}$$

which gives the required result as now formally defined:

- The probability that E occurs, given that F is known to have occurred, gives the **conditional probability** of E given F. This is denoted $Pr(E|F)$ and is calculated as

  $$\Pr(E|F) = \frac{\Pr(E \cap F)}{\Pr(F)}$$

  and from the axioms of probability will generate a number lying between 0 and 1, since $\Pr(F) \geq \Pr(E \cap F) \geq 0$.

- *Example:* A Manufacturer of electrical components knows that the probability is 0.8 that an order will be ready for shipment on time and it is 0.6 that it will also be delivered on time. What is the probability that such an order will be delivered on time given that it was ready for shipment on time?

  Let $R =$ READY, $D =$ DELIVERED ON TIME. $Pr(R) = 0.8, Pr(R \cap D) = 0.6$. From this we need to calculate $Pr(D|R)$, using the above formula. This gives, $Pr(D|R) = Pr(R \cap D)/Pr(R) = 6/8$, or 75%.

If we re-arrange the above formula for conditional probability, we obtain the so-called *multiplication rule of probability* for *intersections* of events:

3.1.1 Multiplication rule of probability

The multiplication rule of probability can be stated as follows:

- $\Pr(E \cap F) = \Pr(E|F) \times \Pr(F)$

Note that for any two events, E and F, $(E \cap F)$ and $(E \cap \bar{F})$ are mutually exclusive with $E = (E \cap F) \cup (E \cap \bar{F})$; this has been seen before. So the *addition rule* and *multiplication rule* of probability together give:

$$\begin{aligned}\Pr(E) &= \Pr(E \cap F) + \Pr(E \cap \bar{F}) \\ &= \Pr(E|F) \times \Pr(F) + \Pr(E|\bar{F}) \times \Pr(\bar{F}).\end{aligned}$$

This is an extremely important and useful result, in practice, as we shall see shortly.

3.1.2 Statistical Independence

If the knowledge that F has occurred does NOT alter our probability assessment of E, then E and F are said to be (statistically) *independent.* In this sense, F carries no information about E.

- Formally, E and F are **independent** events if and only if

$$Pr(E|F) = Pr(E)$$

which, in turn is true *if and only if*

$$Pr(E \cap F) = Pr(E) \times Pr(F).$$

3.1.3 Bayes' Theorem

One area where conditional probability is extremely important is that of clinical trials - testing the power of a diagnostic test to detect the presence of a particular disease. Suppose, then, that a new test is being developed and let $P =$ '*test positive*' and $D =$ '*presence of disease*', but where the results from applying the diagnostic test can never be wholly reliable. From the point of view of our previous discussion on conditional probability, we would of course require that $\Pr(P|D)$ to be large; i.e., the test should be effective at detecting the disease. However, if you think about, this is not necessarily the probability that we might be interested in from a diagnosis point of view. Rather, we should be more interested in $\Pr(D|P)$, the probability of correct diagnosis, and require this to be large (with, presumably, $\Pr(D|\bar{P})$ being small). Here, what we are trying to attach a probability to is a possible 'cause'. The observed outcome is a positive test result (P), but the presence or non-presence of the disease is what is of interest and this is uncertain. $\Pr(D|P)$ asks the question '*what is the probability that it is the presence of the disease which caused the positive test result*'? (Another recent newsworthy example would be the effect of exposure to depleted uranium on Gulf and Balkan war veterans. Given the presence of lymph, lung or brain cancer in such individuals (P), how likely is that the cause was exposure to depleted uranium weapons (D)? Firstly, is $\Pr(D|P)$ high or low? Secondly, might there being something else (F) which could offer a "better" explanation, such that $\Pr(F|P) > \Pr(D|F)$?)

The situation is depicted in Figure 4.2, in which there are two possible 'states' in the population: D (depicted by the lighter shaded area covering the left portion of the sample space) and $\bar{D}$. It must be that $D \cup \bar{D} = S$, since any individual in the population either has the disease or does not. The event of an observed positive test result is denoted by the closed loop, P. (Notice that the shading in the diagram is relatively darker where P intersects with D.)

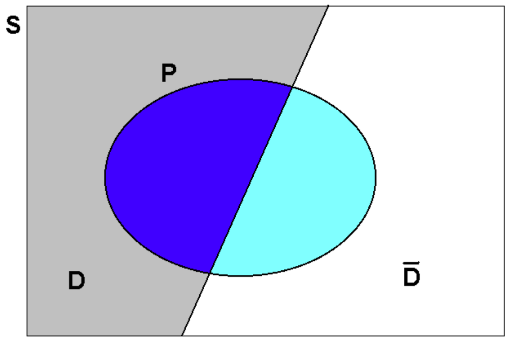


Figure 3.2: Diagram for explaining Bayes' Theorem

To investigate how we might construct the required probability, $\Pr(D|P)$, proceed as follows:

$$\begin{aligned}\Pr(D|P) &= \frac{\Pr(D\cap P)}{\Pr(P)} \\ &= \frac{\Pr(D\cap P)}{\Pr(P\cap D)+\Pr(P\cap\bar{D})},\end{aligned}$$

since $P=(P\cap D)\cup(P\cap\bar{D})$, and these are mutually exclusive. From the multiplication rule of probability, $\Pr(P\cap D)=\Pr(P|D)\times\Pr(D)$, and similarly for $\Pr\left(P\cap\bar{D}\right)$. Thus

$$\Pr(D|P)=\frac{\Pr(P|D)\times\Pr(D)}{\Pr(P|D)\times\Pr(D)+\Pr(P|\bar{D})\times\Pr\left(\bar{D}\right)},$$

which may be convenient to work with since $\Pr(P|D)$ and $\Pr\left(P|\bar{D}\right)$ can be estimated from clinical trials and $\Pr(D)$ estimated from recent historical survey data.

This sort of calculation (assigning probabilities to possible causes of observed events) is an example of *Bayes' Theorem.* Of course, we may have to consider more than two possible causes, and the construction of the appropriate probabilities is as follows.

1. Consider a sample space, S, where $E\subset S$ and A,B,C are three mutually exclusive events (possible causes), defined on S, such that $S=A\cup B\cup C$. In such a situation, A,B and C are said to form a **partition** of S. **Bayes' Theorem** states that:

$$\Pr(A|E)=\frac{\Pr(E|A)\times\Pr(A)}{\{\Pr(E|A)\times\Pr(A)\}+\{\Pr(E|B)\times\Pr(B)\}+\{\Pr(E|C)\times\Pr(C)\}}.$$

2. And, more generally, consider a sample space, S, where $E \subset S$ and $F_1, F_2, ..., F_k$ are k mutually exclusive events (possible causes), which form a partition of $S : S = \bigcup_{j=1}^{k} F_j$. **Bayes' Theorem** then states that:
$$\Pr(F_j|E) = \frac{\Pr(E|F_j) \times \Pr(F_j)}{\sum_{s=1}^{k} \{\Pr(E|F_s) \times \Pr(F_s)\}}.$$

From the above formula, you should be able to satisfy yourself that $\sum_{j=1}^{k} \Pr(F_j|E) = 1$. If this is not at first clear, consider case (1) and show that $\Pr(A|E) + \Pr(B|E) + \Pr(C|E) = 1$. The reason for this is that since A, B and C form a partition of S, they must also form a partition of any event $E \subset S$. In the above conditional probabilities, we are regarding E as the restricted sample space and therefore the probabilities assigned the mutually exclusive events (A, B, C) which *cover* this (restricted) sample space, E, must sum to 1.

- *Example*: Box 1 contains 2 red balls. Box 2 contains 1 red and 1 white ball. Box 1 and Box 2 are identical. If a box is selected at random and one ball is withdrawn from it, what is the probability that the selected box was number 1 if the ball withdrawn from it turns out to be red?

  Let A be the event of selecting Box 1 and B the event of drawing a red ball. Require $Pr(A|B)$.

$Pr(A|B) = Pr(A \cap B)/Pr(B)$;
$Pr(A \cap B) = Pr(A)Pr(B|A) = (1/2) \times 1 = 1/2$.
And,

$$\begin{aligned} Pr(B) &= Pr(A \cap B) + Pr(\bar{A} \cap B) \\ &= \Pr(A) \times \Pr(B|A) \quad + \quad \Pr(\bar{A}) \times \Pr(B|\bar{A}) \\ &= (1/2) \quad + \quad (1/2) \times (1/2) \\ &= 3/4. \end{aligned}$$

Therefore, $\Pr(A|B) = (1/2)/(3/4) = 2/3$.

3.2 Exercise 2

1. A and B are events such that $\Pr(A) = 0.4$ and $\Pr(A \cup B) = 0.75$.

   (a) Find $\Pr(B)$ if A and B are mutually exclusive.

   (b) Find $\Pr(B)$ if A and B are independent.

2. Events A, B and C are such that B and C are mutually exclusive and $\Pr(A) = 2/3$, $\Pr(A \cup B) = 5/6$ and $\Pr(B \cup C) = 4/5$. If $\Pr(B|A) = 1/2$ and $\Pr(C|A) = 3/10$, are A and C statistically independent?

3. Given the information in the example given in Section 3, about undergraduates taking Mathematics, Physics or Chemistry A-levels, calculate the following:

   (a) Of those who took Mathematics, what proportion also took Physics (but not Chemistry) and what proportion took both Physics and Chemistry?

   (b) Of those who took Physics and Chemistry, what proportion also took Mathematics?

4. The *Survey of British Births*, undertaken in the 1970s, aimed to improve the survival rate and care of British babies at, and soon after, birth by collecting and analysing data on new-born babies. A sample was taken designed to be representative of the whole population of British births and consisted of all babies born alive (or dead) after the 24th week of gestation, between 0001 hours on Sunday 5 April and 2400 hours on Saturday 11 April 1970. The total number in the sample so obtained was $n = 17,530$. A large amount of information was obtained, but one particular area of interest was the effect of the smoking habits of the mothers on newly born babies. In particular, the ability of a newly born baby to survive is closely associated with its birth-weight and a birth-weight of less than 1500g is considered dangerously low. Some of the relevant data are summarised as follows.

   For all new born babies in the sample, the proportion of mothers who:
   (i) *smoked before and during pregnancy was* 0.433
   (ii) *gave up smoking prior to pregnancy was* 0.170
   (iii) *who had never smoked was* 0.397.

   However, by breaking down the sample into mothers who smoked, had given up, or who had never smoked, the following statistics were obtained:
   (iv) 1.6% *of the mothers who smoked gave birth to babies whose weight was less than* 1500g,
   (v) 0.9% *of the mothers who had given up smoking prior to pregnancy gave birth to babies whose weight was less than* 1500g,
   (vi) 0.8% *of mothers who had never smoked gave birth to babies whose weight was less than* 1500g.

   (a) Given this information, how would you estimate the risk, for a smoking mother, of giving birth to a dangerously under-weight baby? What is the corresponding risk for a mother who has never smoked? What is the overall risk of giving birth to an under-weight baby?

(b) Of the babies born under 1500g, estimate the proportion of these (a) born to mothers who smoked before and during pregnancy; (b) born to mothers who had never smoked.

(c) On the basis of the above information, how would you assess the evidence on smoking during pregnancy as a factor which could result in babies being born under weight?

5. Metal fatigue in an aeroplane's wing can be caused by any one of three (relatively minor) defects, labelled A, B and C, occurring during the manufacturing process. The probabilities are estimated as: $\Pr(A) = 0.3$, $\Pr(B) = 0.1$, $\Pr(C) = 0.6$. At the quality control stage of production, a test has been developed which is used to detect the presence of a defect. Let D be the event that the test detects a manufacturing defect with the following probabilities: $\Pr(D|A) = 0.6$, $\Pr(D|B) = 0.2$, $\Pr(D|C) = 0.7$. If the test detects a defect, which of A, B or C is the most likely cause? (*Hint*: you need to find, and compare, $\Pr(A|D)$, $\Pr(B|D)$ and $\Pr(C|D)$ using Bayes Theorem.)

Chapter 4

RANDOM VARIABLES & PROBABILITY DISTRIBUTIONS I

The axioms of probability tell us how we should combine and use probabilities in order to make sensible statements concerning uncertain events. To a large extent this has assumed an initial allocation of probabilities to events of interest, from which probabilities concerning related events (unions, intersections and complements) can be computed. The question we shall begin to address in the next two sections is how we might construct *models* which assign probabilities in the first instance.

The ultimate goal is the development of tools which enable statistical analysis of *data*. Any data under consideration (after, perhaps, some coding) are simply a set of *numbers* which describe the appropriate members of a sample in meaningful way. Therefore, in wishing to assign to probabilities to outcomes generated by sampling, we can equivalently think of how to assign probabilities to the numbers that are explicitly generated by the same sampling process. When dealing with numbers, a natural line of enquiry would be to characterise possible *mathematical functions* which, when applied to appropriate numbers, yield probabilities satisfying the three basic axioms. A mathematical function which acts in this fashion is termed a mathematical or *statistical model*:

- *mathematical/statistical models*: mathematical functions which may be useful in assigning probabilities in gainful way.

If such models are to have wide applicability, we need a ‘general’ approach. As noted above, and previously, events (on a sample space of interest) are often described in terms of physical phenomena; see Sections 3 and 4. Mathematical functions require *numbers*. We therefore need some sort of *mapping* from the physical attributes of a sample space to real numbers, before we can begin developing such models. The situation is depicted in Figure 5.1, in which events of interest defined on a physical sample space, S, are mapped into numbers, x, on the real line. Note that there is only one number for each physical event, but that two different events could be assigned the same number. Thus, this mapping can be described by a *function*; it is this function, mapping from the sample space to the real line, which defines a *random variable.* A further function, $f(.)$, is then applied on the real line in order to generate probabilities.

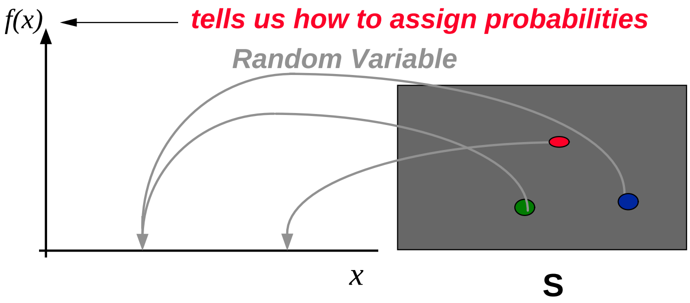


Figure 4.1: Mapping from S to the real line

The initial task, then, is the mapping from S to the real line and this is supplied by introducing the notion of a *random variable*:

4.1 Random Variable

For our purposes, we can think of a *random variable* as having **two** components:

- a label/description which defines the variable of interest
- the definition of a procedure which assigns numerical values to events on the appropriate sample space.

Note that:

- - often, but not always, how the numerical values are assigned will be implicitly defined by the chosen label
  - A random variable is **neither** RANDOM or VARIABLE! Rather, it is device which describes how to assign numbers to physical events of interest: "*a random variable is a real valued function defined on a sample space*".

  - A random variable is indicated by an *upper case* letter (X, Y, Z, T etc). The strict mathematical implication is that since X is a function, when it is applied on a sample space (of physical attributes) it yields a number

The above is somewhat abstract, so let us now consider some examples of *random variables*:

4.1.1 Examples of random variables

- Let X = '*the number of HEADs obtained when a fair coin is flipped* 3 times'. This definition of X implies a function on the physical sample space which generates particular numerical values. Thus X is a random variable and the values it can assume are:

$$X(\text{H,H,H}) = 3;\ X(\text{T,H,H}) = 2;\ X(\text{H,T,H}) = 2;\ X(\text{H,H,T}) = 2;$$
$$X(\text{H,T,T}) = 1;\ X(\text{T,H,T}) = 1;\ X(\text{T,T,H}) = 1;\ X(\text{T,T,T}) = 0.$$

- Let the random variable Y indicate whether or not a household has suffered some sort of property crime in the last 12 months, with $Y(\textit{yes}) = 1$ and $Y(\textit{no}) = 0$. Note that we could have chosen the numerical values of 1 or 2 for *yes* and *no* respectively. However, the mathematical treatment is simplified if we adopt the *binary* responses of 1 and 0.

- Let the random variable T describe the length of time, measured in weeks, that an unemployed job-seeker waits before securing permanent employment. So here, for example,

$$T(15\ \textit{weeks unemployed}) = 5,\quad T(31\ \textit{weeks unemployed}) = 31, \text{ etc.}$$

Once an experiment is carried out, and the random variable (X) is applied to the outcome, a number is *observed*, or *realised*; i.e., the value of the function at that point in the sample space. This is called a *realisation*, or possible outcome, of X and is denoted by a *lower case* letter, x.

In the above examples, the possible realisations of the random variable X (i.e., possible values of the function defined by X) are $x = 0, 1, 2$ or 3. For Y, the possible realisations are $y = 0, 1$; and for T they are $t = 1, 2, 3, ...$.

The examples of X, Y and T given here all applications of *discrete* random variables (the outcomes, or values of the function, are all integers). Technically speaking, the functions X, Y and T are not continuous.

4.2 Discrete random variables

In general, a *discrete random variable* can only assume discrete realisations which are easily listed prior to experimentation. Having defined a discrete random variable, probabilities are assigned by means of a *probability distribution.* A probability distribution is essentially a function which maps from x (the real line) to the interval $[0,1]$; thereby generating probabilities.

In the case of discrete random variable, we shall use what is called a *probability mass function*:

4.2.1 Probability mass function

The probability mass function *(pmf)* is defined for a **DISCRETE** random variable, X, only and is the *function*:

$$p(x) = \Pr(X = x), \quad \text{for all } x.$$

Note that:

- We use $p(x)$ here to emphasize that probabilities are being generated for the outcome x; e.g., $p(1) = \Pr(X = 1)$, etc.

- Note that $p(r) = 0$, if the number r is NOT a possible realisation of X. Thus, for the property crime random variable Y, with $p(y) = \Pr(Y = y)$, it must be that $p(0.5) = 0$ since a realisation of 0.5 is impossible for the random variable.

- If $p(x)$ is to be useful, then it follows from the *axioms of probability* that,

  $$p(x) \geq 0 \quad \textit{and} \quad \sum_x p(x) = 1$$

  where the sum is taken over all possible values that X can assume.

For example, when $X =$ '*the number of HEADs obtained when a fair coin is flipped* 3 times', we can write that $\sum_{j=0}^{3} p(j) = p(0)+p(1)+p(2)+p(3) = 1$ since the number of heads possible is either $0, 1, 2,$ or 3. Be clear about the notation being used here: $p(j)$ is being used to give the probability that j heads are obtained in 3 flips; i.e., $p(j) = \Pr(X = j)$, for values of j equal to $0, 1, 2, 3$.

The *pmf* tells us how probabilities are distributed across all possible outcomes of a discrete random variable X; it therefore generates a *probability distribution.*

4.2.2 A Bernoulli random variable

A *Bernoulli* random variable is a particularly simple (but very useful) discrete random variable. The 'property crime' random variable, Y, introduced above is a particular example. A Bernoulli random variable can only assume one of two possible values: 0 or 1; with probabilities $(1-\pi)$ and π, respectively. Often, the value 1 might be referred to as a success and the value 0 a failure. Here, π is any number satisfying $0 < \pi < 1$, since it is a probability. Note that, here, π is the Greek letter *pi* (lower case), with English equivalent p, and is *not* used here to denote the number $Pi = 3.14159...$. Clearly, different choices for π generate different probabilities for the outcomes of interest; it is an example of a very simple statistical model which can be written compactly as

$$p(y) = \pi^y (1-\pi)^{1-y}, \quad 0 < \pi < 1, \quad y = 0, 1.$$

- note that we have used $p(y)$ here rather than $p(x)$. This is absolutely inconsequential, since you should be able to satisfy yourself that $p(0) = 1 - \pi$ and $p(1) = \pi$, which is all that matters.

Another mathematical function of some importance, which also assigns probabilities (but in a rather different way) is the *cumulative (probability) distribution function*:

4.2.3 Cumulative distribution function

In the **DISCRETE** case the cumulative distribution function (*cdf*) is a function which *cumulates* (adds up) values of $p(x)$, the *pmf*. In general, it is defined as the function:

$$P(x) = \Pr(X \leq x);$$

e.g., $P(1) = \Pr(X \leq 1)$. Note the use of an upper case letter, $P(.)$, for the *cdf*, as opposed to the lower case letter, $p(.)$, for the *pmf*. In the case of a discrete random variable it is constructed as follows:

Suppose the discrete random variable, X, can take on possible values $x = a_1, a_2, a_3, ...$, etc, where the a_j are an increasing sequence of numbers $(a_1 < a_2 < \ldots)$. Then, for example, we can construct the following (cumulative) probability:

$$\Pr(X \leq a_4) = P(a_4) = p(a_1) + p(a_2) + p(a_3) + p(a_4) = \sum_{j=1}^{4} p(a_j),$$

i.e., we take all the probabilities assigned to possible values of X, up to the value under consideration (in this case a_4), and then add them up. It follows from the axioms of probability that $\sum_j p(a_j) = 1$, all the probabilities assigned must sum to unity, as noted before. Therefore,

$$\begin{aligned}\Pr(X \geq a_4) &= p(a_4) + p(a_5) + p(a_6) + \ldots \\ &= \left\{\sum_j p(a_j)\right\} - \{p(a_1) + p(a_2) + p(a_3)\} \\ &= 1 - \Pr(X \leq a_3),\end{aligned}$$

and, similarly,

$$\Pr(X > a_4) = 1 - \Pr(X \leq a_4)$$

which is always useful to remember.

As a more concrete example, consider the case where $X =$ *the number of HEADs obtained from* 3 *flips of a fair coin.* In this case, there are just four possible values of X with $a_1 = 0$, $a_2 = 1$, $a_3 = 2$ and $a_4 = 3$ in the notation used above. Furthermore, due to independence, we can write that

- $p(0) = \Pr(X = 0) = \Pr(\text{T,T,T}) = \Pr(\text{T}) \times \Pr(\text{T}) \times \Pr(\text{T}) = (1/2)^3 = 1/8$
- $p(1) = \Pr(X = 1) = \Pr(\text{H,T,T}) + \Pr(\text{T,H,T}) + \Pr(\text{T,T,H})$
  $= (1/2)^3 + (1/2)^3 + (1/2)^3 = 1/8 + 1/8 + 1/8 = 3/8$
- $p(2) = \Pr(X = 2) = \Pr(\text{H,H,T}) + \Pr(\text{H,T,H}) + \Pr(\text{T,H,H})$
  $= (1/2)^3 + (1/2)^3 + (1/2)^3 = 1/8 + 1/8 + 1/8 = 3/8$
- $p(3) = \Pr(X = 3) = \Pr(\text{H,H,H}) = (1/2)^3 = 1/8$

whilst

- $P(2) = \Pr(X \leq 2) = p(0) + p(1) + p(2) = 1/8 + 3/8 + 3/8 = 7/8$
- $\Pr(X > 1) = 1 - \Pr(X \leq 1) = 1 - (1/8 + 3/8) = 1/2.$
- Note also that $P(2.5) = P(X \leq 2.5)$ must be identical to $P(2) = \Pr(X \leq 2)$ - *think about it*!

In fact, this is a very simple example of a **Binomial distribution.**

4.2.4 The Binomial random variable

Any experiment which can result in only one of two possible outcomes (*success* with a probability denoted by π $(0 < \pi < 1)$ and *failure* with probability denoted by $1 - \pi$) is called a *Bernoulli* experiment and gives rise to a so-called *Bernoulli* random variable, as discussed above; e.g., (a) flipping a coin once: does it land *head* (1) or *tail* (0); (b) Opinion polls: should the UK join EMU? Individuals answer *yes* (1) or *no* (0).

A random variable, X, is said to have a **BINOMIAL** distribution (and is called a Binomial random variable) if it is defined to be the total number of successes in n *independent and identical Bernoulli* experiments; e.g. (i) the total number of heads (1's) obtained when a coin is flipped n times; (ii) of 20 people randomly selected, how many were in favour of the UK joining EMU; (ii) if you were to ask 10 students from this University's large student population if they are vegetarian then, of the 10, the total number of vegetarians would be (approximately) a binomial random variable.

Note that the possible realisations of X are $x = 0, 1, 2, ..., n$. The Binomial *probability mass function* is:

$$p(x) = P(X = x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}, \quad x = 0, 1, 2, ..., n; \quad 0 < \pi < 1.$$

where

$$\binom{n}{x} = \frac{n!}{x!(n-x)!},$$

and $n!$ denotes n *factorial*: $n! = n(n-1)(n-2)...2 \times 1$; e.g., $3! = 6$. In the above formula, we **define** $0! = 1$. Note that $\binom{n}{x}$ is called a *combinatorial coefficient* and always gives an integer; it simply counts the total number of different ways that we can arrange exactly x "ones" and $(n - x)$ "zeros" together.. For example, consider in how may different ways we can arrange (or combine) 2 "ones" and 1 "zero". The possibilities are

$$(1, 1, 0)\,; (1, 0, 1)\,; (0, 1, 1)$$

and that's it. There are only three ways we can do it. Using $\binom{n}{x}$ we need to substitute $x = 2$ and $n = 3$, since the total number of "ones" and "zeros" is 3. This gives $\binom{3}{2} = \frac{3!}{2!1!} = 3$; as we discovered above.

Also, it is worth observing that the transformed random variable, $Z = X/n$, defines the random variable which describes the proportion of successes.

4.3 Examples of Probability Mass Functions for Discrete Random Variables

4.3.1 Binomial Random Variable

As discussed in the main text,

$$p(x) = P(X = x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}, \quad x = 0, 1, 2, ..., n; \quad 0 < \pi < 1,$$

and if a discrete random variable has such a probability mass function then we say that is has a Binomial distribution with *parameters* n and π.

Consider, then, a Binomial random variable, with parameters $n = 5$ and $\pi = 0.3$.

$$\begin{aligned} p(2) &= \Pr(X = 2) = \binom{5}{2} (0.3)^2 (0.7)^3 \\ &= \frac{5!}{2!3!} (0.09) (0.343) \\ &= \frac{5 \times 4}{2} (0.09) (0.343) \\ &= 0.3087. \end{aligned}$$

A cumulative probability is worked as follows:

$$\begin{aligned} P(2) &= \Pr(X \leq 2) = \Pr(X = 0) + \Pr(X = 1) + \Pr(X = 2) \\ &= p(0) + p(1) + p(2) \\ &= \binom{5}{0} (0.3)^0 (0.7)^5 + \binom{5}{1} (0.3)^1 (0.7)^4 + \binom{5}{2} (0.3)^2 (0.7)^3 \\ &= 0.16807 + 0.36015 + 0.3087 \\ &= 0.83692. \end{aligned}$$

4.3.2 Geometric Random Variable

As in the Binomial case, consider repeating independent and identical Bernoulli experiments (each of which results in a success, with probability π, or a failure, with probability $1-\pi$). Define the random variable X to be the number of Bernoulli experiments performed in order to achieve the first success. This is a *Geometric* random variable.

The probability mass function is

$$p(x) = \Pr(X = x) = \pi (1 - \pi)^{x-1}, \quad x = 1, 2, 3, ...; \quad 0 < \pi < 1,$$

and if a discrete random variable has suvh a probability mass function then we say that is a Geometric distribution with *parameter* π.

Suppose the probability of success is $\pi = 0.3$. What is the probability that the first success is achieved on the second experiment? We require

$$p(2) = \Pr(X = 2) = (0.3)(0.7) = 0.21.$$

The probability that the first success is achieved on or before the third experiment is

$$\begin{aligned} P(3) &= \Pr(X \leq 3) \\ &= p(1) + p(2) + p(3) \\ &= (0.3)(0.7)^0 + (0.3)(0.7)^1 + (0.3)(0.7)^2 \\ &= (0.3)\left(1 + 0.7 + 0.7^2\right) \\ &= 0.657. \end{aligned}$$

4.3.3 Poisson Random Variable

The Poisson random variable is widely-used to count the number of events occurring in a given interval of time. Examples include (i) the number of cars passing an observation point, located on a long stretch of straight road, over a 10 minute interval; (ii) the number of calls received at a telephone exchange over a 10 second interval.

The probability mass function is

$$p(x) = \Pr(X = x) = \frac{\lambda^x}{x!}\exp(-\lambda), \quad x = 0, 1, 2, ...; \quad \lambda > 0$$

in which, exp(.) is the exponential function ($\exp(a) = e^a$) and λ is a positive real number (a parameter). We say that X has a Poisson distribution with parameter λ (note that λ will often be refered to as the *mean* which will be discussed in Section 9).

Suppose that the number of calls, X, arriving at a telephone exchange in any 10 second interval follows a Poisson distribution with $\lambda = 5$. What is $\Pr(X = 2)$?

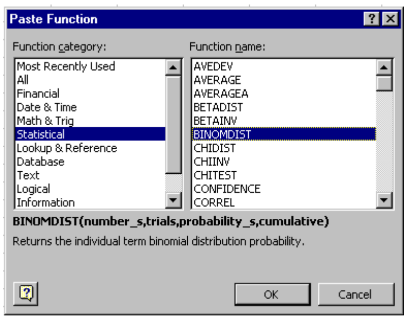
$$\begin{aligned} p(2) &= \Pr(X = 2) = \frac{5^2}{2!}\exp(-5) \\ &= 0.0842. \end{aligned}$$

The probability of more than 2 calls is

$$\begin{aligned} \Pr(X > 2) &= 1 - \Pr(X \leq 2) \\ &= 1 - \{\Pr(X = 0) + \Pr(X = 1) + \Pr(X = 2)\} \\ &= 1 - \left\{e^{-5} + 5 \times e^{-5} + 12.5 \times e^{-5}\right\} \\ &= 1 - 0.1247 \\ &= 0.8753. \end{aligned}$$

4.4 Using EXCEL to obtain probabilities: Binomial & Poisson

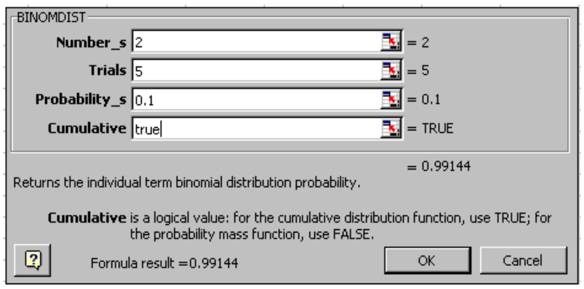
Probabilities for various statistical models can be obtained using **Paste Function** in EXCEL. To obtain **BINOMIAL** probabilities: click **Paste Function** - the **Paste Function** Dialogue Box appears. Select **Statistical** from the **Function Category** box**.** In **Function Name** box select **BINOMDIST** and then click **OK** - the **BINOMDIST** Dialogue Box appears. These two Dialogue Boxes are displayed below:



BINOMDIST(number_s,trials,probability_s,cumulative)

Returns the individual term binomial distribution probability.

Paste Function Dialogue Box



Number_s 2 = 2

Trials 5 = 5

Probability_s 0.1 = 0.1

Cumulative true = TRUE

= 0.99144

Returns the individual term binomial distribution probability.

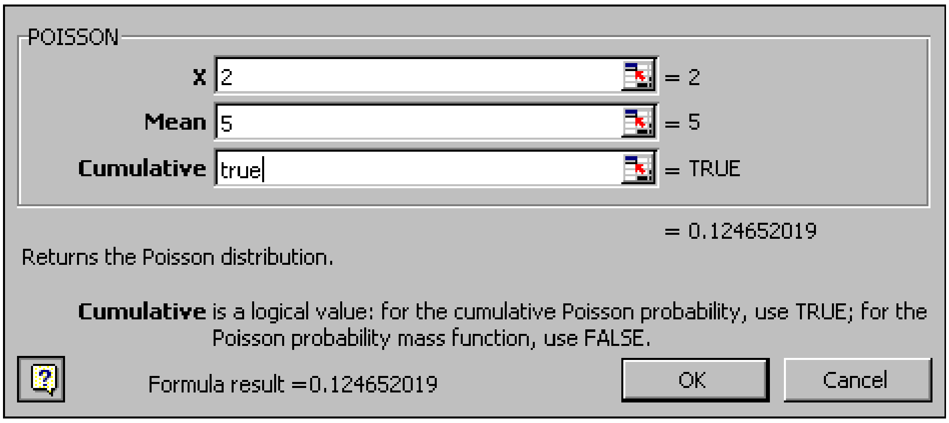
Cumulative is a logical value: for the cumulative distribution function, use TRUE; for the probability mass function, use FALSE.

Formula result =0.99144

The BINOMDIST Dialogue Box

You have to fill in **number_s, trials, probability_s** and **cumulative.** Once this is done, the required probability will appear in the **Formula Result** line at the bottom. This is illustrated above, showing you how to obtain $\Pr(X \leq 2) = 0.99144$, where X is a Binomial random variable which records the number of successes in 5 trials, with the probability of success equal to 0.1 (i.e., a Binomial random variable with $n = 5$ and $\pi = 0.1$). If the ‘value’ in the cumulative box is changed to **false** then the computed probability will be $\Pr(X = 2)$.

You can also use this feature to calculate Poisson probabilities. The Poisson Dialogue box looks like:



POISSON

X 2 = 2

Mean 5 = 5

Cumulative true = TRUE

= 0.124652019

Returns the Poisson distribution.

Cumulative is a logical value: for the cumulative Poisson probability, use TRUE; for the Poisson probability mass function, use FALSE.

Formula result =0.124652019

The POISSON Dialogue Box

where the above illustration shows how to calculate $\Pr(X \leq 2)$, for a Poisson random variable with $\lambda = 5$ (referred to as the mean in EXCEL).

Chapter 5

RANDOM VARIABLES & PROBABILITY DISTRIBUTIONS II

[HEALTH WARNING: Before reading this section you MUST revise your undertanding of integration.]

In the previous section, we introduced the notion of a *random variable* and, in particular a *discrete* random variable. It was then discussed how to use mathematical functions in order to assign probabilities to the various possible numerical values of such a random variable. A probability distribution is a method by which such probabilities can be assigned and in the discrete case this can be achieved via a probability mass function (*pmf*). Figure 6.1 illustrates the *pmf* for a particular *binomial random variable.*

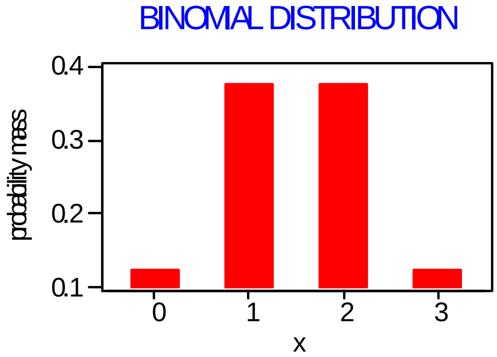


Figure 5.1: X = number of HEADS in three flips of a fair coin

Notice how *masses* of probability are dropped onto the possible discrete (isolated) outcomes. We now develop mathematical functions which can used to describe probability distributions associated with a *continuous random variable.*

5.1 Continuous random variables

Recall that a random variable is a function applied on a sample space, by which we mean that physical attributes of a sample space are mapped (by this function) into a number. When a *continuous* random variable is applied on a sample space, a range of possible numbers is implied (not just isolated numbers as with a discrete random variable).

As an example, let $X =$ '*the contents of a reservoir*', where the appropriate sample space under consideration allows for the reservoir being just about *empty*, just about *full* or *somewhere in between.* Here we might usefully define the range of possible values for X as $0 < x < 1$, with 0 signifying 'empty' and 1 signifying 'full'. As noted before when talking about the characteristics of continuous variables, theoretically, we can't even begin to list possible numerical outcomes for X; any value in the interval is possible.

How do we thus distribute probability in this case? Well, presumably the probability must be distributed only over the range of possible values for X; which, in the reservoir example, is over the unit interval $(0, 1)$. However, unlike the discrete case where a specific *mass* of probability is dropped on each of the discrete outcomes, for continuous random variables probability is distributed more smoothly, rather like brushing paint on a wall, over the *whole interval* of defined possible outcomes. Therefore in some areas the distribution of probability is quite thick and others it can be relatively thin. This is depicted in Figure 6.2.

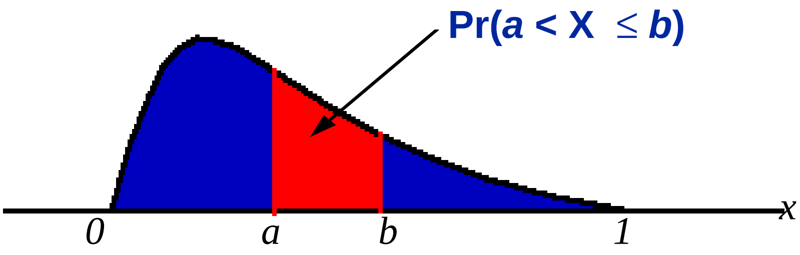


Figure 5.2: Distribution of probability

Some thought should convince you that for a *continuous* random variable, X, it must be the case that $\Pr(X = c) = 0$ for all real numbers c contained in the range of possible outcomes of X. If this were not the case, then the axioms of probability would be violated. However, there should be a positive probability of X being close, or in the *neighbourhood*, of c. (A neighbourhood of c might be $c \pm 0.01$, say.) For example, although the probability that the reservoir is *exactly* 90% full must be *zero* (who can measure 90% exactly?), the axioms of probability require that there is non-zero probability of the reservoir being between, say, 85% and 95% full. (Indeed, being able to make judgements like this is an eminently reasonable requirement if we wish to investigate, say, the likelihood of water shortages.) This emphasizes that we can not think of probability being assigned at specific points (numbers), rather probability is *distributed* over intervals of numbers.

We therefore must confine our attention to assigning probabilities of the form $\Pr(a < X \leq b)$, for some real numbers $a < b$; i.e., what is the probability that X takes on values between a and b. For an appropriately defined mathematical function describing how probability is distributed, as depicted in Figure 6.2, this would be the area under that function between the values a and b. Such functions can be constructed and are called *probability density functions*. Let us emphasize the role of *area* again: it is the *area* under the probability density function which provides probability, *not* the probability density function itself. Thus, by the axioms of probability, if a and b are in the range of possible values for the continuous random variable, X, then $\Pr(a < X \leq b)$ must always return a positive number, lying between 0 and 1, *no matter how close* b is to a (provided only that $b > a$).

What sorts of mathematical functions can usefully serve as probability density functions? To develop the answer to this question, we begin by considering another question: *what mathematical functions would be appropriate as cumulative distribution functions*?

5.2 Cumulative distribution function (*cdf*)

For a continuous random variable, X, the *cdf* is a *smooth* continuous function defined as $F(x) = \Pr(X \leq x)$, for *all* real numbers x; e.g., $F(0.75) = \Pr(X \leq 0.75)$. The following should be observed:

- such a function is defined for all real numbers x; not just those which are possible realisations of the random variable X;
- we use $F(.)$, rather than $P(.)$, to distinguish the cases of *continuous* and *discrete* random variables, respectively.

Let us now establish the *mathematical properties* of such a function. We can do this quite simply by making $F(.)$ adhere to the axioms of probability.

Firstly, since $F(x)$ is to be used to return probabilities, it must be that

$$0 \leq F(x) \leq 1, \quad \text{for all } x.$$

Secondly, it must be a smooth, increasing function of x (over intervals where possible values of X can occur). To see this, consider again the reservoir example and any arbitrary numbers a and b, satisfying $0 < a < b < 1$. Notice that $a < b$; b can be as a close as you like to a, but it must always be strictly greater than a. Therefore, the axioms of probability imply that $\Pr(a < X \leq b) > 0$, since the event '$a < X \leq b$' is possible. Now divide the real line interval $(0, b]$ into two mutually exclusive intervals, $(0, a]$ and $(a, b]$. Then we can write the event '$X \leq b$' as

$$(X \leq b) = (X \leq a) \cup (a < X \leq b).$$

Assigning probabilities on the left and right, and using the axiom of probability concerning the allocation of probability to mutually exclusive events, yields

$$\Pr(X \leq b) = \Pr(X \leq a) + \Pr(a < X \leq b)$$

or

$$\Pr(X \leq b) - \Pr(X \leq a) = \Pr(a < X \leq b).$$

Now, since $F(b) = \Pr(X \leq b)$ and $F(a) = \Pr(X \leq a)$, we can write

$$F(b) - F(a) = \Pr(a < X \leq b) > 0.$$

Thus $F(b) - F(a) > 0$, for all real numbers a and b such that $b > a$, no matter how close. This tells us that $F(x)$ must be an increasing function and a little more delicate mathematics shows that it must be a *smoothly* increasing function. All in all then, $F(x)$ appears to be smoothly increasing from 0 to 1 over the range of possible values for X. In the reservoir example, the very simple function $F(x) = x$ would appear to satisfy these requirements, provided $0 < x < 1$.

More generally, we now formally state the properties of a *cdf.* For complete generality, $F(x)$ must be defined over the whole real line even though in any given application the random variable under consideration may only be defined on an interval of that real line.

5.2.1 Properties of a cdf

A cumulative distribution function is a mathematical function, $F(x)$, satisfying the following properties:

1. $0 \leq F(x) \leq 1$.

2. If $b > a$ then $F(b) \geq F(a)$;

   i.e., F is increasing.

   In addition, over all intervals of possible outcomes for a continuous random variable, $F(x)$ is smoothly increasing; i.e., it has no *sudden jumps.*

3. $F(x) \to 0$ as $x \to -\infty$;

   $F(x) \to 1$ as $x \to \infty$;

   i.e., $F(x)$ decreases to 0 as x falls, and increases to 1 as x rises.

Any function satisfying the above may be considered suitable for modelling cumulative probabilities, $\Pr(X \leq x)$, for a continuous random variable. Careful consideration of these properties reveals that $F(x)$ can be *flat* (i.e., non-increasing) over some regions. This is perfectly acceptable since the regions over which $F(x)$ is flat correspond to those where values of X can not occur and. therefore, zero probability is distributed over such regions. In the reservoir example, $F(x) = 0$, for all $x \leq 0$, and $F(x) = 1$, for all $x \geq 1$; it is therefore flat over these two regions of the real line. This particular examples also demonstrates that the last of the three properties can be viewed as completely general; for example, the fact that $F(x) = 0$, in this case, for all $x \leq 0$ can be thought of as simply a special case of the requirement that $F(x) \to 0$ as $x \to -\infty$.

Some possible examples of *cdf*s, in other situations, are depicted in Figure 6.3.

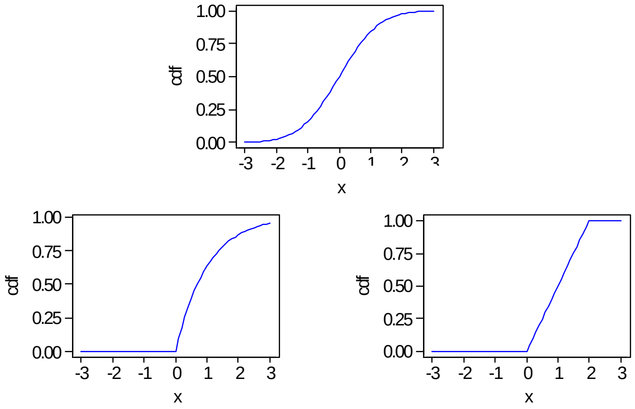


Figure 5.3: Some cdf's for a continuous random variable

The first of these is strictly increasing over the whole real line, indicating possible values of X can fall anywhere. The second is increasing, but only strictly over the interval $x > 0$; this indicates that the range of possible values for the random variable is $x > 0$ with the implication that $\Pr(X \leq 0) = 0$. The third is only strictly increasing over the interval $(0 < x < 2)$, which gives the range of possible values for X in this case; here $\Pr(X \leq 0) = 0$, whilst $\Pr(X \geq 2) = 0$.

Let us end this discussion by re-iterating the calculation of probabilities using a *cdf*:

- $\Pr(a < X \leq b) = F(b) - F(a)$, for any real numbers a and b;

as discussed above.

- $\Pr(X \leq a) = 1 - \Pr(X > a)$, since $\Pr(X \leq a) + \Pr(X > a) = 1$, for any real number a;

and, finally,

- $\Pr(X < a) = \Pr(X \leq a)$, since $\Pr(X = a) = 0$.

Our discussion of the *cdf* was introduced as a means of developing the idea of a *probability density function* (*pdf*). Visually the *pdf* illustrates how all of the probability is distributed over possible values of the continuous random variable; we used the analogy of paint being brushed over the surface of a wall. The *pdf* is also a mathematical function satisfying certain requirements in order that the axioms of probability are not violated. We also pointed out that it would be the *area* under that function which yielded probability. Note that this is in contrast to the *cdf*, $F(x)$, where the function itself gives probability. We shall now investigate how a *pdf* should be defined.

5.3 Probability Density Functions (*pdf*)

For a continuous random variable, it is well worth reminding ourselves of the following:

- There is no function which gives $Pr(X = x)$ for some number x, since all such probabilities are *identically zero*.

However, and as discussed in the previous section, there is a smooth, increasing, function $F(x)$, the *cdf*, which provides $\Pr(X \leq x)$. In particular

$$\Pr(a < X \leq b) = F(b) - F(a),$$

for real numbers $b > a$. Also, since $F(x)$ is smoothly continuous and differentiable over the range of possible values for X (see Figure 6.3), then there must exist a function $f(x) = dF(x)/dx$, the derivative of $F(x)$. Note that $f(x)$ must be positive over ranges where $F(x)$ is increasing; i.e., over ranges of possible values for X. On the other hand, $f(x) = 0$, over ranges where $F(x)$ is flat; i.e., over ranges where values of X can *not* occur.

Moreover, the *Fundamental Theorem of Calculus* (which simply states that *differentiation* is the opposite of *integration*) implies that if $f(x) = dF(x)/dx$, then

$$F(b) - F(a) = \int_a^b f(x)dx.$$

We therefore have constructed a function $f(x) = dF(x)/dx$, such that the area under it yields probability (recall that the integral of a function between two specified limits gives the area under that function). Such a function, $f\,(x)$, is the *probability density function.*

In general, $\lim_{a\to-\infty} F(a) = 0$, so by letting $a \to -\infty$ in the above we can define the fundamental relationship between the *cdf* and *pdf* as:

$$\begin{aligned} F(x) &= \Pr\left(X \le x\right) = \int_{-\infty}^{x} f(t)dt, \\ f(x) &= dF(x)/dx; \end{aligned}$$

i.e., $F(x)$ is the area under the curve $f(t)$ up to the point $t = x$. Now, letting $x \to \infty$, and remembering that $\lim_{x\to\infty} F(x) = 1$, we find that

$$\int_{-\infty}^{\infty} f(x)dx = 1;$$

i.e., *total area under* $f(x)$ *must equal* 1 (rather like the total area of the sample space as depicted by a Venn Diagram).

These definitions are all quite general so as to accommodate a variety of situations; however, as noted before, implicit in all of the above is that $f(x) = 0$ over intervals where no probability is distributed; i.e., where $F(x)$ is flat. Thus, in the reservoir example we could be more explicit with the limits and write

$$F\,(x) = \int_0^x f(t)dt, \quad 0 < x < 1,$$

for some suitable function $f(.)$, since $f(t) = 0$ for all $t \le 0$, and all $t \ge 1$. For example, suppose the contents of the reservoir can be modelled by the continuous random variable which has probability density function

$$f(x) = \begin{cases} 3(1-x)^2, & 0 \le x \le 1 \\ 0, & \text{otherwise.} \end{cases}$$

We can then calculate the probability that the reservoir will be over 75% full as:

$$\begin{aligned} \Pr\left(X > 0.75\right) &= \int_{0.75}^{1} 3(1-x)^2 dx \\ &= -\left[(1-x)^3\right]_{0.75}^{1} \\ &= \left(\frac{1}{4}\right)^3 \\ &= \frac{1}{64}. \end{aligned}$$

Figure 6.4 also gives a simple example of a *cdf* and *pdf*, where probability is distributed *uniformly* over a finite interval (in this case, it is the *unit* interval $[0, 1]$). Such a distribution is therefore said to be uniform.For example,

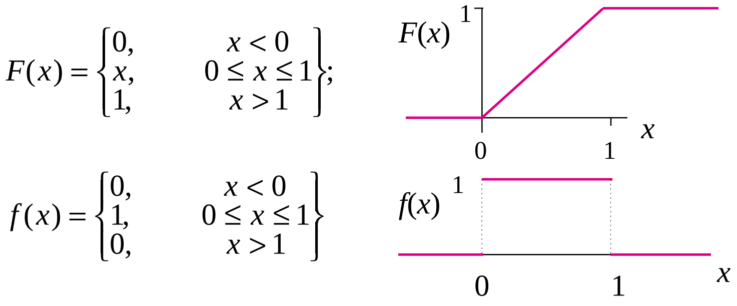


Figure 5.4: A simple *cdf* and *pdf*

$$\Pr(0.25 < X \le 0.5) = F(0.5) - F(0.25) = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}.$$

Alternatively, using the *pdf*,

$$\Pr(0.25 < X \le 0.5) = \int_{0.25}^{0.5} f(x)dx = 0.25.$$

Also note that the total area under $f(x)$, over the unit interval, is clearly equal to 1.

To recap, then, let us list the properties of the *pdf.*

5.3.1 Properties of the *pdf*

A *pdf* for a continuous random variable is a mathematical function which must satisfy,

1. $f(x) \ge 0$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$

Probabilities can be calculated as:

- $\Pr(a < X \le b) = \int_a^b f(x)dx$
  - i.e., it is the *area* under the *pdf* which gives probability

and the relationship with *cdf* is given by:

- - $f(x) = dF(x)/dx$
  - $F(x) = \int_{-\infty}^{x} f(t)dt.$

5.4 Exponential Distribution

Let the continuous random variable, denoted X, monitor the elapsed time, measured in minutes, between successive cars passing a junction on a particular road. Traffic along this road in general flows freely, so that vehicles can travel independently of one another, not restricted by the car in front. Occasionally, there are quite long intervals between successive vehicles, while more often there are smaller intervals. To accommodate this, the following *pdf* for X is defined:

$$f(x) = \begin{cases} \exp(-x), & x > 0, \\ 0, & x \leq 0. \end{cases}$$

The graph of $f(x)$ looks is depicted in Figure 6.5.

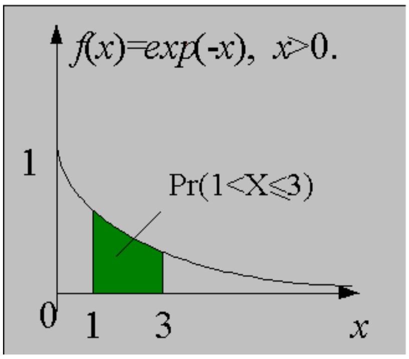


Figure 5.5: Exponential *pdf*

Now, you might care to verify that $\int_0^\infty \exp(-x)dx = 1$, and from the diagram it is clear that

$$\Pr(a < X \leq a+1) > \Pr(a+1 < X \leq a+2),$$

for any number $a > 0$. By setting $a = 1$, say, this implies that an elapsed time of between 1 and 2 minutes has greater probability than an elapsed time of between 2 and 3 minutes; and, in turn, this has a greater probability than an elapsed time of between 3 and 4 minutes, etc; i.e., smaller intervals between successive vehicles will occur more frequently than longer ones.

Suppose we are interested in the probability that $1 < X \leq 3$? i.e., the probability that the elapsed time between successive cars passing is somewhere between 1 and 3 minutes? To find this, we need

$$\begin{aligned} \Pr(1 < X \leq 3) &= \int_1^3 \exp(-x)dx \\ &= [-\exp(-x)]_1^3 \\ &= e^{-1} - e^{-3} \\ &= 0.318. \end{aligned}$$

One might interpret this as meaning that about 32% of the time, successive vehicles will be between 1 and 3 minutes apart.

The above distribution is known as the *unit* exponential distribution. In general, we say that the continuous random variable X has an *exponential distribution* if it has probablity density function given by

$$f(x) = \frac{1}{\theta}\exp\left(-\frac{x}{\theta}\right), \quad x > 0; \quad \theta > 0$$

where θ is called the parameter of the distribution (sometimes the *mean,* which will be discussed in Section 8). Note that when $\theta = 1$, we get back to the special case of the unit exponential distribution. Notice that the random variable, here, can only assume positive values.

Note that

$$\begin{aligned} \int_0^\infty f(x)dx &= \int_0^\infty \frac{1}{\theta}\exp\left(-\frac{x}{\theta}\right)dx \\ &= \left[-\exp\left(-\frac{x}{\theta}\right)\right]_0^\infty \\ &= 1 \end{aligned}$$

so that it is a proper probability density function (clearly, also, $f(x) > 0$ for all $x > 0$).

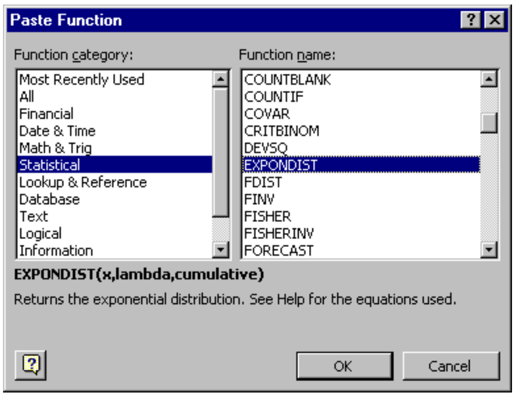
5.5 Revision: Properties of the Exponential Function

Refer to ES1070 Maths notes or ES2281 Maths notes. There is also additional material on the course website.

1. $y = e^x$ is a strictly positive and strictly increasing function of x
2. $\frac{dy}{dx} = e^x$ and $\frac{d^2y}{dx^2} = e^x$
3. $\ln y = x$
4. When $x = 0$, $y = 1$, and $y > 1$ when $x > 0$, $y < 1$ when $x < 0$
5. By the chain rule of differentiation, if $y = e^{-x}$ then $\frac{dy}{dx} = -e^{-x}$.

5.6 Using EXCEL to obtain probabilities: The Exponential Distribution

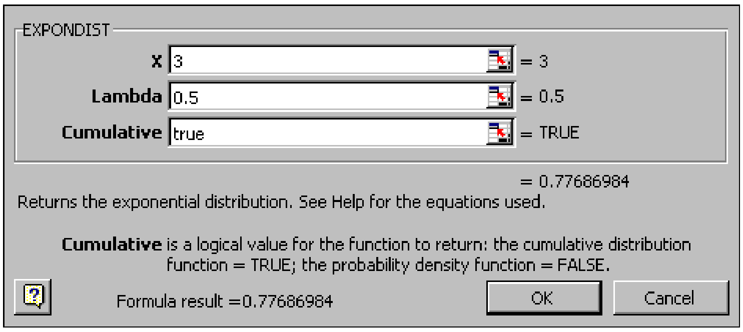
Probabilities for various statistical models can be obtained using **Paste Function** in EXCEL. To obtain **EXPONENTIAL** probabilities: click **Paste Function** - the **Paste Function** Dialogue Box appears. Select **Statistical** from the **Function Category** box. In **Function Name** box select **EXPONDIST** and then click **OK** - the **EXPONDIST** Dialogue Box appears. These two Dialogue Boxes are displayed below:



EXPONDIST(x,lambda,cumulative)

Returns the exponential distribution. See Help for the equations used.

Paste Function Dialogue Box



X 3 = 3

Lambda 0.5 = 0.5

Cumulative true = TRUE

= 0.77686984

Returns the exponential distribution. See Help for the equations used.

Cumulative is a logical value for the function to return: the cumulative distribution function = TRUE; the probability density function = FALSE.

Formula result =0.77686984

The EXPONDIST Dialogue Box

You have to fill in **X, Lambda,** and **cumulative.** Once this is done, the required probability will appear in the **Formula Result** line at the bottom. This is illustrated above, showing you how to obtain $\Pr(X \leq 3) = 0.77687$, where X is an Exponential random variable with parameter $\lambda = 0.5$. Note that the value for λ (lambda) required by EXCEL is the inverse value of the parameter θ defined in the notes above; i.e., $\lambda = \theta^{-1}$. Thus EXCEL parameterises the exponential probability density function as

$$f(x) = \lambda \exp(-\lambda x), \quad x > 0, \quad \lambda > 0.$$

If the 'value' in the cumulative box is changed from **true** to **false** then the formula result will simply give the value of the probability density function; i.e., in this case, $0.5\exp(-1.5) = 0.111565$, which is not a probability.

5.7 Exercise 3

1. In an experiment, if a mouse is administered dosage level A of a certain (harmless) hormone then there is a 0.2 probability that the mouse will show signs of aggression within one minute. For dosage levels B and C, the probabilities are 0.5 and 0.8, respectively. Ten mice are given exactly the same dosage level of the hormone and, of these, exactly 6 shows signs of aggression within one minute of receiving the dose.

   (a) Calculate the probability of this happening for each of the three dosage levels, A, B and C. (This is essentially a Binomial random variable problem, so you can check your answers using EXCEL.)

   (b) Assuming that each of the three dosage levels was equally likely to have been administered in the first place (with a probability of 1/3), use Bayes' Theorem to evaluate the likelihood of each of the dosage levels *given* that 6 out of the 10 mice were observed to react in this way.

2. Let X be the random variable indicating the number of incoming planes every k minutes at a large international airport, with probability mass function given by

   $p(x) = \Pr(X = x) = \frac{(0.9k)^x}{x!} \exp(-0.9k), \quad x = 0, 1, 2, 3, 4, \ldots$.

   Find the probabilities that there will be

   (a) exactly 9 incoming planes during a period of 5 minutes (i.e., find $\Pr(X = 9)$ when $k = 5$);

   (b) fewer than 5 incoming planes during a period of 4 minutes (i.e., find $\Pr(X < 5)$ when $k = 4$);

   (c) at least 4 incoming planes during an 2 minute period (i.e., find $\Pr(X \geq 4)$ when $k = 2$).

   Check all your answers using EXCEL.

3. The random variable Y is said to be *Geometric* if it has probability mass function given by

   $p(y) = \Pr(Y = y) = (1-\theta)\theta^{y-1}, \quad y = 1, 2, 3, ...; \quad 0 < \theta < 1;$

   where θ is an unknown 'parameter'.
   Show that the cumulative distribution function can be expressed as

   $$P(y) = \Pr(Y \leq y) = 1 - \theta^y, \quad y = 1, 2, 3, ...$$

   with $P(y) = 0$ for $y \leq 0$ and $P(y) \to 1$ as $y \to \infty$.

   (Note that $P(y) = p(1) + p(2) + ... + p(y) = \sum_{t=1}^{y} p(t)$ can be written in longhand as

   $$P(y) = (1-\theta)\left(1 + \theta + \theta^2 + \theta^3 + \ldots + \theta^{y-1}\right).$$

   The term in the second bracket on the right-hand side is the sum of a *Geometric Progression.*)

4. The weekly consumption of fuel for a certain machine is modelled by means of a continuous random variable, X, with probability density function

   $$g(x) = \begin{cases} 3(1-x)^2, & 0 \leq x \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

   Consumption, X, is measured in hundreds of gallons per week.

   (a) Verify that $\int_0^1 g(x)dx = 1$ and calculate $\Pr(X \leq 0.5)$.

   (b) How much fuel should be supplied each week if the machine is to run out fuel 10% of the time at most? (Note that if s denotes the supply of fuel, then the machine will run out if $X > s$.)

5. The lifetime of a electrical component is measured in 100s of hours by a random variable T having the following probability density function

   $$f(t) = \begin{cases} \exp(-t), & t > 0, \\ 0, & \text{otherwise.} \end{cases}$$

   (a) Show that the cumulative distribution function, $F(t) = \Pr(T \leq t)$ is given by

   $$F(t) = \begin{cases} 1 - \exp(-t), & t > 0 \\ 0 & t \leq 0. \end{cases}$$

   (b) Show the probability that a component will operate for at least 200 hours without failure is $\Pr(T \geq 2) \cong 0.135.$?

(c) Three of these electrical components operate independently of one another in a piece of equipment and the equipment fails if ANY ONE of the individual components fail. What is the probability that the equipment will operate for at least 200 hours without failure? (Use the result in (b) in a binomial context).

Chapter 6

THE NORMAL DISTRIBUTION

It could be argued that the most important probability distribution encountered thus far has been the *Binomial* distribution for a discrete random variable monitoring the total number of successes in n independent and identical Bernoulli experiments. Indeed, this distribution was proposed as such by Jacob Bernoulli (1654-1705) in about 1700. However as n becomes large, the Binomial distribution becomes difficult to work with and several mathematicians sought approximations to it using various limiting arguments. Following this line of enquiry two other important probability distributions emerged; one was the *Poisson* distribution, due to the French mathematician Poisson (1781-1840), and published in 1837. An exercise using the Poisson distribution is provided by Question 2, in Exercise 3. The other, is the *normal* distribution due to De Moivre (French, 1667-1754), but more commonly associated with the later German mathematician, Gauss (1777-1855), and French mathematician, Laplace (1749-1827). Physicists and engineers often refer to it as the *Gaussian* distribution. There a several pieces of evidence which suggest that the British mathematician/statistician, Karl Pearson (1857-1936) coined the phrase *normal distribution.*

Further statistical and mathematical investigation, since that time, has revealed that the normal distribution plays a unique role in the theory of statistics; it is without doubt the most important distribution. We introduce it here, and study its characteristics, but you will encounter it many more times in this, and other, statistical or econometric courses.

Briefly the motivation for wishing to study the normal distribution can be summarised in three main points:

- it can provide a good approximation to the binomial distribution
- it provides a natural representation for many *continuous* random variables that arise in the social (and other) sciences

- many functions of interest in statistics give random variables which have distributions closely approximated by the normal distribution.

We shall see shortly that the normal distribution is defined by a particular *probability density function*; it is therefore appropriate (in the strict sense) for modelling *continuous* random variables. Not withstanding this, it is often the case that it provide an adequate approximation to another distribution, even if the original distribution is *discrete* in nature, as we shall now see in the case of a binomial random variable.

6.1 The Normal distribution as an approximation to the Binomial distribution

Consider a Binomial distribution which assigns probabilities to the total number of successes in n identical applications of the same Bernoulli experiment. For the present purpose we shall use the example of flipping a coin a number of times (n). Let the random variable of interest be the *proportion* of times that a *HEAD* appears and let us consider how this distribution changes as n increases:

- If $n = 3$, the possible proportions could be $0,\ 1/3,\ 2/3\ \textit{or}\ 1$
- If $n = 5$, the possible proportions could be $0,\ 1/5,\ 2/5,\ 3/5,\ 4/5\ \textit{or}\ 1$
- If $n = 10$, the possible proportions could be $0,\ 1/10,\ 2/10$, etc ...

The probability distributions, over such proportions, for $n = 3$, 5, 10 and 50, are depicted in Figure 7.1.

Notice that the ‘bars’, indicating where masses of probability are dropped, get closer and closer together until, in the limit, all the space between them is squeezed out and a bell shaped mass appears, by joining up the tops of every bar: this characterises the *probability density function* of the *NORMAL DISTRIBUTION.*

Having motivated the normal distribution via this limiting argument, let us now investigate the fundamental mathematical properties of this *bell-shape.*

6.2 The Normal distribution

The normal distribution is characterised by a particular probability density function $f(x)$, the precise definition of which we shall divulge later. For the moment the important things to know about this function are:

- it is bell-shaped

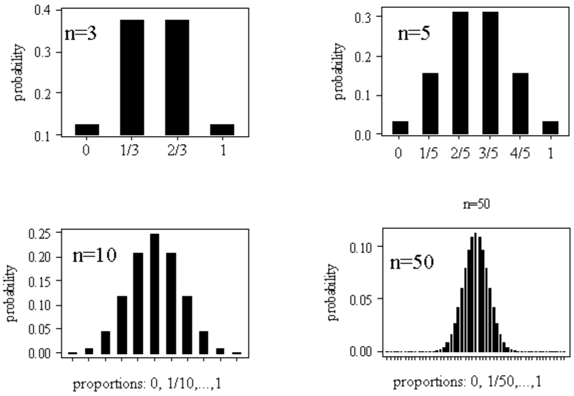


Figure 6.1: The changing shape of the Binomial distribution

- it tails off to zero as $x \to \pm\infty$

- area under $f(x)$ gives probability; i.e., $\Pr(a < X \leq b) = \int_a^b f(x)dx$.

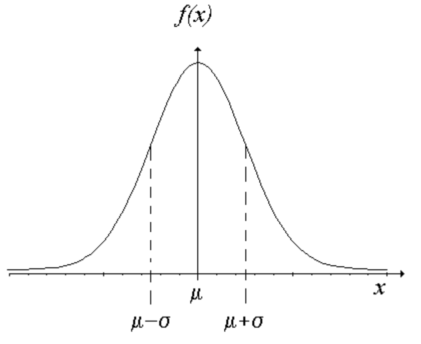


Figure 7.2: The Normal density function (the classic *bell* shape)

The specific *location* and *scale* of the bell depend upon two *parameters* (real numbers) denoted μ and σ (with $\sigma > 0$), as depicted in Figures 7.2. μ is the Greek letter *mu* (with English equivalent m) and σ is the Greek letter *sigma* with (English equivalent s). Changing μ relocates the density (shifting it to the left or right) but leaving it's scale and shape unaltered. Increasing σ makes the density 'fatter' with a lower peak; such changes are illustrated in Figure 7.3.

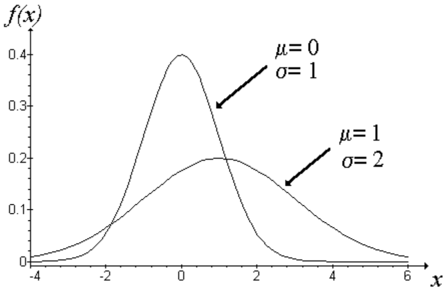


Figure 7.3: Location (μ) and scale (σ) changes

Furthermore:

- $f(x)$ is symmetric about the value $x = \mu$; i.e., $f(\mu + c) = f(\mu - c)$, for any real number c.
- $f(x)$ has points of inflection at $x = \mu \pm \sigma$; i.e., $d^2 f(x)/dx^2$ is zero at the values $x = \mu + \sigma$ and $x = \mu - \sigma$.

The above describes all the salient mathematical characteristics of the normal *pdf*. For what it's worth, although you will not be expected to remember this, the density is actually defined by:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad -\infty < x < \infty; \quad -\infty < \mu < \infty, \quad \sigma > 0,$$

and we say that a continuous random variable X has a normal distribution if and only if it has *pdf* defined by $f(x)$, above. Here, π is the number $Pi = 3.14159...$. In shorthand, we write $X \sim N\left(\mu, \sigma^2\right)$, meaning '$X$ is normally distributed with location μ and scale σ'. However, a perfectly acceptable alternative is to say 'X is normally distributed with *mean* μ and *variance* σ^2', for reasons which shall become clear in the next section.

An important special case of this distribution arises when $\mu = 0$ and $\sigma = 1$, yielding the standard normal density:

6.2.1 The standard normal density.

If $Z \sim N(0, 1)$, then the *pdf* for Z is written

$$\phi(z) = \frac{1}{\sqrt{2\pi}} exp(-z^2/2), \quad -\infty < z < \infty,$$

where ϕ is the Greek letter *phi,* equivalent to the English f. The *pdf,* $\phi(z)$, is given a special symbol because it is used so often and merits distinction. Indeed, the standard normal density is used to calculate probabilities associated with a normal distribution, even when $\mu \neq 0$ and/or $\sigma \neq 1$.

6.3 The normal distribution as a model for data

Apart from its existence via various mathematical limiting arguments, the normal distribution offers a way of approximating the distribution of many variables of interest in the social (and other) sciences. For example, in Exercise 2, some statistics were provided from The Survey of British Births which recorded the birth-weight of babies born to mothers who smoked and those who didn't. Figure 7.4, depicts the histogram of birth-weights for babies born to mothers who had never smoked. Superimposed on top of that is normal density curve with parameters set at $\mu = 3353.8$ and $\sigma = 572.6$.

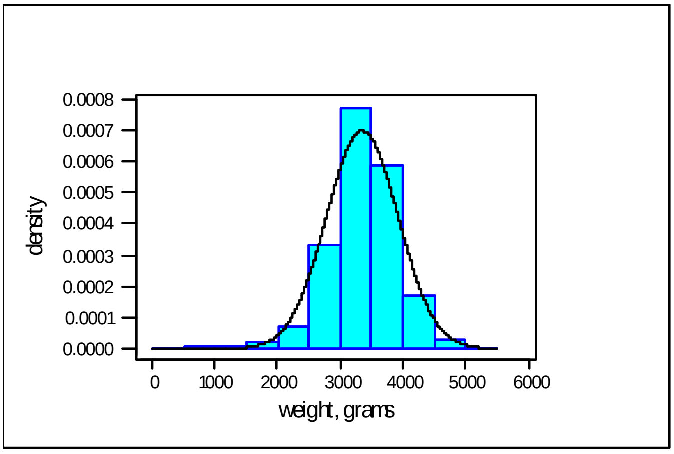


Figure 7.4: Histogram of birth weights and normal density

As can be seen, the fitted normal density does a reasonable job at tracing out the shape of the histogram, as constructed from the data. (I will leave it as a matter of conjecture as to whether the birth-weights of babies born to mothers who smoked are *normal.*) The nature of the approximation here is that areas under the histogram record the relative frequency, or *proportion* in the sample, of birth-weights lying in a given interval, whereas the area under the normal density, over the same interval, gives the *probability.*

Let us now turn the question of calculating such probabilities associated with a normal distribution.

6.4 Calculating probabilities

Since $f(x)$ is a *pdf*, to obtain probabilities we need to think *area* which means we have to *integrate.* Unfortunately, there is no easy way to integrate $\phi(z)$, let alone $f(x)$. To help us, however,

- special (statistical) tables (or computer packages such as EXCEL) provide probabilities about the standard normal random variable $Z \sim N(0,1)$, and

- from this initial calculation, probability statements about $X \sim N(\mu, \sigma^2)$ are easily obtained.

To develop how this works in practice, we require some elementary properties of $Z \sim N(0, 1)$

6.4.1 A few elementary properties: $Z \sim N(0, 1)$

Firstly, we introduce the *cdf* for Z, This functions is denoted $\Phi(z)$, where Φ is the upper case Greek F, and is defined as follows:

$$\Phi(z) = \Pr(Z \leq z) = \int_{-\infty}^{z} \phi(t)dt,$$

the area under $\phi(.)$ up to the point z, with $\phi(z) = d\Phi(z)/dz$.

Now, due to symmetry of $\phi(z)$ about the value $z = 0$, it follows that:

$$\Phi(0) = \Pr(X \leq 0) = 1/2$$

and, in general,

$$\begin{aligned} \Phi(-z) &= \Pr(Z \leq -z) \\ &= \Pr(Z > z) \\ &= 1 - \Pr(Z \leq z) \\ &= 1 - \Phi(z). \end{aligned}$$

The role of symmetry and calculation of probabilities as areas under $\phi(z)$ is illustrated in Figure 7.5. In this diagram, the area under $\phi(z)$ is divided up into 2 parts: the area to the left of a which is $\Phi(a)$; and the area to the right of a which is $1 - \Phi(a)$. These areas add up to 1.

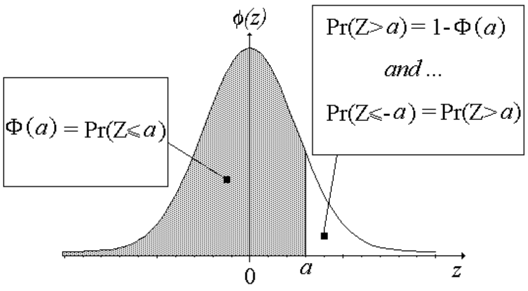


Figure 7.5: Area gives probability and symmetry is useful

Armed with these properties we can now use the “standard normal” table of probabilities. Such a table is given in the Appendix to these course notes. These are exactly the tables which you will be expected to use in the examination.

The probabilities provided by this table are of the form $\Pr(Z \leq z) = \Phi(z)$, for values of $0 < z < \infty$. ($\Pr(Z \leq z)$ for values of $z < 0$ can be deduced using symmetry.) For example, you should satisfy yourself that you understand the use of the table by verifying that,

$$\begin{array}{cc} \Pr(Z \leq 0] = 0.5; & \Pr(Z \leq 0.5) = 0.691; \\ \Pr(Z \leq 1.96) = 0.975; & \Pr(Z \geq 1) = \Pr(Z \leq -1) = 0.159, \ \textit{etc.} \end{array}$$

The calculation of the probability

$$\begin{aligned} \Pr(-1.62 &< Z \leq 2.2) = \Pr(Z \leq 2.2) - \Pr(Z \leq -1.62) \\ &= 0.986 - 0.053 \\ &= 0.933 \end{aligned}$$

is illustrated in Figure 7.6.

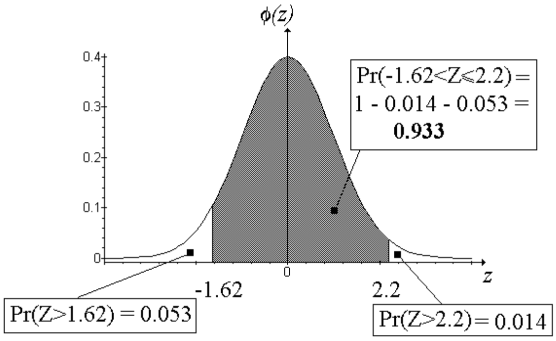


Figure 7.6: $\Pr(-1.62 < Z \leq 2.2) = 0.933$

In this diagram, the areas are divided into 3 mutually exclusive parts: the area to the left of $z = -1.62$, which equals 0.053; the area to the right of $z = 2.2$, which is equal to 0.014; and the area in between, which is equal to 0.933 the required probability.

6.4.2 Calculating probabilities when $X \sim N(\mu, \sigma^2)$.

We can calculate probabilities associated with the random variable $X \sim N(\mu, \sigma^2)$, by employing the following results which shall be stated without proof:

- If $Z \sim N(0,1)$, then $X = \sigma Z + \mu \ \ \sim \ \ N(\mu, \sigma^2)$.
- If $X \sim N\left(\mu, \sigma^2\right)$, then $Z = \frac{X-\mu}{\sigma} \sim N\left(0,1\right).$

For example, if $Z \sim N\left(0,1\right)$, then $X = 3Z + 6 \sim N\left(6,9\right)$; and, if $X \sim N\left(4,25\right)$, then $Z = \frac{X-4}{5} \sim N\left(0,1\right).$

It therefore transpires that if $X \sim N(\mu, \sigma^2)$, probabilities about X can be obtained from probabilities about Z via the relationship $X = \sigma Z + \mu$, since we can then write $Z = \frac{X-\mu}{\sigma}$.

Let $X \sim N(\mu, \sigma^2)$, with $Z = \frac{X-\mu}{\sigma} \sim N\left(0,1\right)$, then

$$\begin{aligned}
\Pr(a < X \leq b) &= \Pr\left(a-\mu < X-\mu < b-\mu\right), \quad \text{subtract } \mu \text{ throughout,} \\
&= \Pr\left(\tfrac{a-\mu}{\sigma} < \tfrac{X-\mu}{\sigma} \leq \tfrac{b-\mu}{\sigma}\right), \quad \text{divide through by } \sigma > 0 \text{ throughout,} \\
&= \Pr\left(\tfrac{a-\mu}{\sigma} < Z \leq \tfrac{b-\mu}{\sigma}\right), \quad \text{where } Z \sim N\left(0,1\right), \\
&= \Pr\left(Z \leq \tfrac{b-\mu}{\sigma}\right) - \Pr\left(Z \leq \tfrac{a-\mu}{\sigma}\right), \\
&= \Phi\left(\tfrac{b-\mu}{\sigma}\right) - \Phi\left(\tfrac{a-\mu}{\sigma}\right).
\end{aligned}$$

We thus find that $\Pr(a < X \leq b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$, and the probabilities on the right hand side are easily determined from Standard Normal Tables. The following example illustrates the procedure in practice:

- *Example*: Let $X \sim N(10, 16)$, what is $\Pr(0 < X \leq 14)$?

  Here, $\mu = 10, \sigma = 4, a = 0, b = 14$; so, $\frac{a-\mu}{\sigma} = -2.5$ and $\frac{b-\mu}{\sigma} = 1$.

  Therefore, the required probability is:

$$\begin{aligned}
Pr(-2.5 &< Z \leq 1) = \Pr\left(Z \leq 1\right) - \Pr\left(Z \leq -2.5\right) \\
&= 0.841 - 0.006 = 0.835.
\end{aligned}$$

- *Example*: A fuel is to contain $X\%$ of a particular compound. Specifications call for X to be between 30 and 35. The manufacturer makes a profit of Y pence per gallon where

$$Y = \begin{cases} 10, & \text{if} \quad 30 \leq x \leq 35 \\ 5, & \text{if} \quad 25 \leq x < 30 \quad \textit{or} \quad 35 < x \leq 40 \\ -10, & \text{otherwise.} \end{cases}$$

  If $X \sim N(33, 9)$, evaluate $\Pr\left(Y = 10,\right)$ $\Pr\left(Y = -10\right)$ and, hence, $\Pr\left(Y = 5\right).$

Here, $X \sim N(33,9)$; i.e., $\mu = 33$ and $\sigma = 3$. Now, since $\frac{X-33}{3} \sim N(0,1)$:

$$\begin{aligned}
\Pr(Y=10) &= \Pr(30 \leq X \leq 35) \\
&= \Pr\left(\frac{30-33}{3} \leq \frac{X-33}{3} \leq \frac{35-33}{3}\right) \\
&= \Pr(Z \leq 2/3) - \Pr(Z \leq -1), \qquad \text{where } Z \sim N(0,1) \\
&= \Phi(2/3) - \Phi(-1) \\
&= 0.74857 - 0.15866 \\
&= 0.58991.
\end{aligned}$$

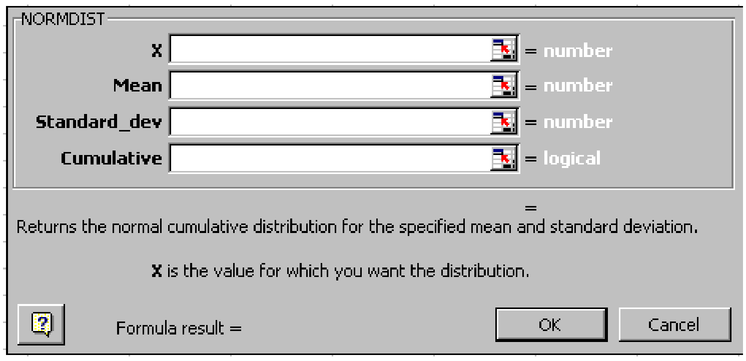
Similar calculations show that

$$\begin{aligned}
\Pr(Y=-10) &= \Pr(\{X<25\} \cup \{X>40\}) \\
&= 1 - \Pr(25 \leq X \leq 40) \\
&= 1 - \Pr\left(\frac{25-33}{3} \leq \frac{X-33}{3} \leq \frac{40-33}{3}\right) \\
&= 1 - \{\Phi(7/3) - \Phi(-8/3)\} \\
&= 1 - 0.99010 + 0.00379 \\
&= 0.01369.
\end{aligned}$$

Thus, $\Pr(Y=5) = 1 - 0.58991 - 0.01369 = 0.3964$.

6.5 Using EXCEL to obtain Probabilities for a Normal random variable

To obtain probabilities associated with the normal distribution: click **Paste Function** - the **Paste Function** Dialogue Box appears. Select **Statistical** from the **Function Category** box**.** In **Function Name** box select **NORMDIST** and then click **OK** - the **NORMDIST** Dialogue Box appears:



NORMDIST

X = number

Mean = number

Standard_dev = number

Cumulative = logical

=

Returns the normal cumulative distribution for the specified mean and standard deviation.

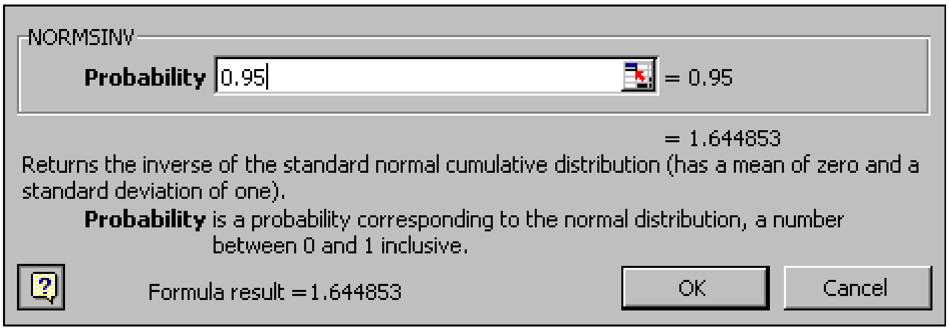
X is the value for which you want the distribution.

Formula result =

NORMDIST Dialogue Box

This allows you to obtain $\Pr(X \leq x)$, where $X \sim N(\mu, \sigma^2)$. You need to supply x, μ (the **mean**), σ (the **standard deviation**) and a value for **cumulative**, which should be **true** in order to obtain $\Pr(X \leq x)$.

Two other useful functions are **NORMINV** and **NORMSINV.** NORMSINV returns the inverse standard normal distribution function. That is if you supply a probability, say p, then you obtain the number z such that $\Pr(Z \leq z) = p$, where $Z \sim N(0, 1)$. The NORMSINV Dialogue Box looks like



NORMSINV

Probability 0.95 = 0.95

= 1.644853

Returns the inverse of the standard normal cumulative distribution (has a mean of zero and a standard deviation of one).

Probability is a probability corresponding to the normal distribution, a number between 0 and 1 inclusive.

Formula result =1.644853

The NORMSINV Dialogue Box

where, here, the supplied probability is 0.95 and the result is 1.644853; i.e., $\Pr(Z \leq 1.644853) = 0.95$.

NORMINV operates in a similar fashion, but for a general $N(\mu, \sigma^2)$ distribution (so you also need to supply μ and σ).

Chapter 7

MOMENTS AND EXPECTATION

7.1 Introduction

We have, thus far, developed the fundamental properties of probability distributions for both continuous and discrete random variables. These distributions are mathematical models which assign probabilities to various numerical values of the random variable under consideration. The *probability mass function* drops masses of probability onto *specific* values of the *discrete* random variable whereas the *probability density function* distributes the probability smoothly over the *range* of possible values of a *continuous random variable.* However, further theoretical properties (characteristics) of these distributions tell us how, exactly, these probabilities are assigned. For example, if $X \sim N\left(\mu, \sigma^2\right)$, consider the various possibilities for $\Pr\left(X \leq 2\right)$? For $\mu = 0$, $\sigma = 1$, the probability is 0.977; whereas for $\mu = 1$, $\sigma = 4$, the probability is 0.599. Thus, changing the values of μ and σ changes the allocation of probability.

The properties of interest which can affect the allocation of probability are called **MOMENTS**, and we shall be concerned with those *moments* which affect the location and scale of a probability distribution. Of particular importance are the **theoretical** (or *population*) **mean** and **theoretical variance.** These are quantities derived from theoretical mathematical models, and should not be confused with the sample mean and sample variance which are calculated from the observed sample data. We use the term population here because a mathematical/statistical model is often put forward as an appropriate *description* of the population from which a sample is drawn.

The relationship between theoretical and sample mean is motivated as follows.

7.2 Motivation for the DISCRETE case

Let X be a discrete random variable, which when applied can yield any one of r possible values denoted $a_1, a_2, ..., a_r$. Suppose that a suitable experiment is performed which can give rise to n observed values of this random variable. Let f_1 denote the number of times a_1 occurs, f_2 the number of times a_2 occurs, etc, once the experiment has been performed, with $\sum_{j=1}^{r} f_j = n$.

On the basis of observed sample data the sample mean would be

$$\bar{X} = \frac{1}{n}(f_1 a_1 + f_2 a_2 + \ldots + f_r a_r) = \sum_{j=1}^{r} a_j w_j,$$

where $w_j = f_j/n$ is the relative frequency of outcome a_j, with $\sum_{j=1}^{n} w_j = 1$. In this from $\bar{X}$ is a weighted average of the a_j's with the weights being the relative frequencies.

Now, the *relative frequency interpretation* of probability says that, if n is large, the relative frequency of observing outcome a_j should settle down to the probability of so doing; i.e., as n increases, $w_j \cong p(a_j) = \Pr(X = a_j)$. Then, substituting $p(a_j)$ for w_j in the expression for the sample mean, we have in theory that $\bar{X}$ should settle down to what is known as the **THEORETICAL MEAN.** This quantity is denoted by μ or $E[X]$ and defined as:

$$\mu = E[X] = \sum_{j=1}^{r} a_j p(a_j),$$

and is also referred to as the **EXPECTED VALUE of** X.

The following points are worth remembering:

- the mean, $E[X]$, has the interpretation of being the balancing point of the distribution;
- "*the mean of X*" and "*the mean of the distribution*" are equivalent expressions;
- in the discrete case, $E[X] = \mu$ need not necessarily be one of the possible outcomes of X;
- on occasion, we may need to write μ_X as the mean of X, to distinguish it from μ_Y the mean of another random variable Y.

Consider Figure 8.1, which depicts the probability mass function for, X, the number of dots on the upturned face after rolling a fair die. From visual inspection (and elementary physics) is seems clear that the balancing point of the probability distribution is located at $x = 3.5$, which is not one of the possible outcomes of X. This can be verified using the above formula, where outcomes, a_j, are simply the integers 1 to 6 and all probabilities are equal to $\frac{1}{6}$. Then

$$E[X] = \{1 \times (1/6)\} + \{2 \times (1/6)\} + ... + \{6 \times (1/6)\} = 3.5$$

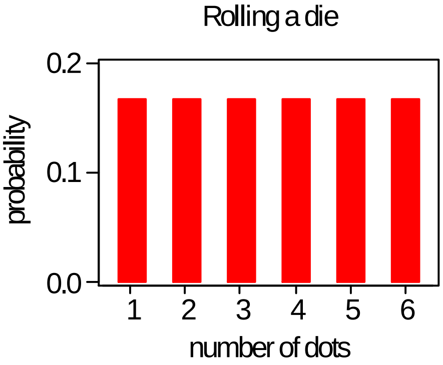


Figure 8.1: Discrete uniform: mean = balancing point= $3\frac{1}{2}$

In general, for a discrete random variable X, the theoretical mean, or expectation, is defined as

$$E[X] = \mu = \sum_x x \Pr(X = x),$$

where the sum is taken over all possible values of x for which $\Pr(X = x) \neq 0$.

- *Example*: Consider again the example given at the end of Section 7 where a fuel is to contain $X\%$ of a particular compound. Specifications call for X to be between 30 and 35. The manufacturer makes a profit of Y pence per gallon where

  $$Y = \begin{cases} 10, & \text{if} \quad 30 \leq x \leq 35 \\ 5, & \text{if} \quad 25 \leq x < 30 \quad \textit{or} \quad 35 < x \leq 40 \\ -10, & \text{otherwise.} \end{cases}$$

  If $X \sim N(33, 9)$, evaluate $E[Y]$. What profit per gallon must be gained on those loads for which $30 \leq x \leq 35$ to increase *expected* profit by 50%?

  Since Y is a discrete random variable its mean is given by (the required probabilities were obtained previously)

  $$\begin{aligned} E[Y] &= 10 \times \Pr(Y = 10) + 5 \times \Pr(Y = 5) - 10 \times \Pr(Y = -10) \\ &= 10 \times 0.58991 + 5 \times 0.3964 - 10 \times 0.01369 \\ &= 7.74. \end{aligned}$$

To increase expected profit by 50%, we require $E\left[Y\right] = 11.6$, $(= 7.74 \times 1.5)$. We therefore need to solve the following equation for p^* :

$$\begin{aligned} 11.6 &= p^* \times \Pr\left(Y = 10\right) + 5 \times \Pr\left(Y = 5\right) - 10 \times \Pr\left(Y = -10\right) \\ &= p^* \times 0.58991 + 1.7101 \end{aligned}$$

$$\Leftrightarrow p^* = 16.7651 \cong 17.$$

The theoretical mean, $E\left[X\right]$, is referred to as the first moment about the origin; other moments are also of interest as follows.

7.2.1 Higher order moments

Consider the random variable X^s, for some positive integer s (e.g., $s = 2$). If X can assume values x, then X^s assumes values x^s. We can also take the expectation of the random variable X^s:

- $E[X^s] = \sum_x x^s \Pr(X = x)$
- *Example*: Rolling a fair die.

$$E[X^2] = [1 + 4 + 9 + 16 + 25 + 36]/6 = 15\frac{1}{6}.$$

More generally, we can take the expectation of any function of X, provided the function itself is mathematically defined. Suppose that X can assume values denoted by x, and that $g(x)$ exists; then

$$E[g(X)] = \sum_x g(x) \Pr(X = x)$$

A particular application of this is when $g(X) = (X - \mu)^2$, where $\mu = E\left[X\right]$, in which case we get what is called the theoretical variance as $E\left[(X - \mu)^2\right]$.

7.2.2 The variance

The variance of a random variable (distribution), denoted $var[X]$, characterises the spread, or scale, of the distribution and is a *positive real number*. In terms of expectation, the variance is defined as the expected squared distance of X from its mean (location): i.e.,

$$var[X] = E[(X - \mu)^2] > 0,$$

where $\mu = E[X]$.

- $var[X]$ is sometimes called the *second moment about the mean.*

In the case of the discrete random variable, it is calculated as:

$$var[X] = \sigma^2 = \sum_x (x-\mu)^2 \Pr(X=x).$$

In the same way that the sample mean of observations settles down to the theoretical mean of the distribution (population) from which the sample was drawn, the sample variance can be expected to settle down to the corresponding theoretical variance. As with the mean, we may on occasion write σ_X^2 to denote the variance of X, with σ_Y^2 denoting the variance of another random variable, Y. Rather, than the variance (which is measured in squared units of the original variable), we often report the *standard deviation* (which is measured in the same units as the variable):

- *The standard deviation* is the positive square root of the variance and is denoted by σ. Thus, for example, $\sigma_X = +\sqrt{var\,[X]}$, and may on occasion be written $sd\,[X]$.

For example, the fuel consumption of a car may have a mean of 33 mpg and a standard deviation of 9 mpg; the variance in this case would be 81 mpg^2.

- *Example*: Rolling fair die.

$$\begin{aligned} var[X] &= [(2.5)^2 + (1.5)^2 + (0.5)^2 + (0.5)^2 + (1.5)^2 + (2.5)^2]/6 = 35/12. \\ sd\,[X] &= +\sqrt{\frac{35}{12}} \cong 1.7. \end{aligned}$$

As you will see later, understanding linear combinations (or linear transformations) of random variables plays a crucial role in the development of statistical inference. We therefore introduce some important preliminary results now.

7.3 The effect of linear transformations

The idea of a linear transformation is extremely simple. If X is some random variable and a and b are known constants (i.e., not random variables themselves), then $aX + b$ represents a linear transformation X. We have the following results:

1. $E[aX+b] = aE[X] + b$

   Proof:

In the discrete case, let $p(x) = \Pr(X = x)$; then

$$\begin{aligned} E[aX+b] &= \sum_x (ax+b)p(x) \\ &= \sum_x (axp(x) + bp(x)) \\ &= a\sum_x xp(x) + b\sum_x p(x) \\ &= aE[X] + b, \end{aligned}$$

since $\sum_x p(x) = 1$ and $\sum_x xp(x) = E[X]$, by definition. This result has great applicability; for example, if $Y = X^2$, then $E[aY+b] = aE[Y]+b = aE[X^2]+b$, etc. In particular any multiplicative constants, like a here, can be moved outside of the expectations operator, $E[.]$.

- *Example*: Rolling a fair die.

$$E[4X+1] = 4 \times (3.5) \ + \ 1 = 15.$$

The next result concerns the variance of a linear transformation.

2. $var[aX+b] = a^2 var[X]$

Proof:

Define the new random variable, $W = aX + b$. Then it is clear that $var[aX+b] = var[W]$. Thus, we need $var[W]$. Since W is a random variable, we have by definition that

$$var[W] = E\left[(W - E[W])^2\right],$$

it is the expected squared distance of W from its mean. But from above we know that

$$\begin{aligned} E[W] &= E[aX+b] \\ &= aE[X] + b. \end{aligned}$$

Therefore, since $W = aX + b$

$$\begin{aligned} var[W] &= E\left[(aX + b - aE[X] - b)^2\right] \\ &= E\left[(aX - aE[X])^2\right] \\ &= E\left[\{a(X - E[X])\}^2\right] \\ &= E\left[a^2(X - E[X])^2\right]. \end{aligned}$$

Then, noting that a^2 is a multiplicative constant we can move it outside of the expectations operator to give

$$\begin{aligned} var\left[W\right] &= a^2 E\left[\left(X - E\left[X\right]\right)^2\right] \\ &= a^2 var\left[X\right], \end{aligned}$$

by definition of $var\left[X\right]$.

- *Example*: Rolling a fair die.

$$var[4X + 1] = 16 \times 35/12 = 46\frac{2}{3}$$

Moments, means and variances, can also be defined for continuous random variables as well. Here, we simply substitute the *pdf* for the *pmf*, and change the *sum* to a smooth *integral*.

7.4 The continuous case

Let X be a continuous random variable with density function $f(x)$. Means, variances etc are defined as follows, where integration is effectively over the range of x, for which $f(x) > 0$:

1. The mean of X, $E\left[X\right]$,

$$E[X] = \mu = \int_{-\infty}^{\infty} xf(x)dx$$

2. The variance of X, $var\left[X\right]$,

$$var[X] = \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx$$

3. For admissible functions $g(X)$,

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx.$$

- *Example:* Suppose X has a continuous uniform density function

$$\begin{aligned} f(x) &= 1/6, \quad 0 < x < 6 \\ &= 0, \quad otherwise. \end{aligned}$$

  - $E[X] = 3$; clearly, it is the balancing point (you can check that $\int_0^6 \frac{x}{6}dx = 3$.)

– $var[X] = \frac{1}{6}\int_0^6 (x-3)^2 dx = 3.$

The linear transformation results apply in the continuous case as well:

$$E[aX + b] = aE[X] + b$$

whilst

$$var[aX + b] = a^2 var[X].$$

The interpretation for $E[X]$ and $var\,[X]$ are the same as in the discrete case. $E\,[X]$ is informative about where on the real line the centre of the probability distribution is located and $var\,[X]$ tells us something about how dispersed the probability distribution is about that central value; i.e., the spread or scale of the distribution.

7.5 Calculating a variance

In some situations, the calculation of $var[X]$ can be eased using the above results:

$$\begin{aligned} var[X] &= E[(X-\mu)^2], \quad where \quad \mu = E[X], \\ &= E[X^2 - 2\mu X + \mu^2] \\ &= E[X^2] - \mu^2 \end{aligned}$$

i.e., $var[X] = E[X^2] - \{E[X]\}^2$.

In the continuous uniform example above, we have that

$$E[X^2] = \frac{1}{6}\int_0^6 x^2 dx = 12$$

and $\{E[X]\}^2 = 9$, so $var\,[X] = 12 - 9 = 3$, which agrees with the previous calculation.

- *Example*: Consider a continuous random variable, X, having the follwoing probability density function

$$f(x) = 3(1-x)^2, \quad 0 < x < 1.$$

We can establish its mean and variance as follows:

$$\begin{aligned}
E[X] &= 3\int_0^1 x(1-x)^2\,dx \\
&= 3\int_0^1 \left\{x - 2x^2 + x^3\right\} dx \\
&= 3\left[\frac{x^2}{2} - \frac{2x^3}{3} + \frac{x^4}{4}\right]_0^1 \\
&= 3\left(\frac{1}{2} - \frac{2}{3} + \frac{1}{4}\right) \\
&= \frac{1}{4}
\end{aligned}$$

and

$$\begin{aligned}
E\left[X^2\right] &= 3\int_0^1 x^2(1-x)^2\,dx \\
&= 3\int_0^1 \left\{x^2 - 2x^3 + x^4\right\} dx \\
&= 3\left[\frac{x^3}{3} - \frac{2x^4}{4} + \frac{x^5}{5}\right]_0^1 \\
&= 3\left(\frac{1}{3} - \frac{1}{2} + \frac{1}{5}\right) \\
&= \frac{1}{10}.
\end{aligned}$$

So that

$$\begin{aligned}
var[X] &= E\left[X^2\right] - \{E[X]\}^2 \\
&= \frac{1}{10} - \frac{1}{16} \\
&= \frac{3}{80.}
\end{aligned}$$

7.6 Means and Variances for some Discrete Distributions

The following results are stated without proof. For the interested reader, proofs are available on the course website.

Binomial distribution

$$\begin{aligned} p(x) &= \binom{n}{x}\pi^x(1-\pi)^{n-x}, \quad x = 0,1,2,...,n; \quad 0<\pi<1 \\ E\left[X\right] &= n\pi \\ var\left[X\right] &= n\pi\left(1-\pi\right) \end{aligned}$$

Geometric distribution

$$\begin{aligned} p\left(x\right) &= \pi\left(1-\pi\right)^{x-1}, \quad x = 1,2,3,...; \quad 0<\pi<1 \\ E\left[X\right] &= \frac{1}{\pi} \\ var\left[X\right] &= \frac{1-\pi}{\pi^2} \end{aligned}$$

Poisson distribution

$$\begin{aligned} p\left(x\right) &= \Pr\left(X=x\right) = \frac{\lambda^x}{x!}\exp\left(-\lambda\right), \quad x = 0,1,2,...; \quad \lambda>0 \\ E\left[X\right] &= \lambda \\ var\left[X\right] &= \lambda \end{aligned}$$

7.7 Means and Variances for some Continuous Distributions

The following results are stated without proof. For the interested reader, proofs are available on most texts

7.7.1 Uniform Distribution

$$\begin{aligned} f\left(x\right) &= \frac{1}{b-a}, \quad a<x<b \\ E\left[X\right] &= \frac{b+a}{2} \\ var\left[X\right] &= \frac{1}{12}\left(b-a\right)^2 \end{aligned}$$

7.7.2 Exponential distribution

$$\begin{aligned} f(x) &= \frac{1}{\beta}\exp\left(-\frac{x}{\beta}\right), \quad x>0; \quad \beta>0 \\ E\left[X\right] &= \beta \\ var\left[X\right] &= \beta^2 \end{aligned}$$

7.7.3 Gamma Distribution

The Gamma function is defined as $\Gamma(x) = \int_0^\infty t^{x-1} \exp(-t)\, dt$, with the property that $\Gamma(x) = (x-1)\Gamma(x-2)$, so that for any positive integer, a, we have $\Gamma(a) = (a-1)! = (a-1)(a-2)...2.1$. The density of a Gamma random variable is

$$\begin{aligned} f(x) &= \frac{x^{\alpha-1}}{\beta^\alpha \Gamma(\alpha)} \exp\left(-\frac{x}{\beta}\right), \quad x > 0; \quad \alpha > 0,\ \beta > 0 \\ E[X] &= \alpha\beta \\ var[X] &= \alpha\beta^2. \end{aligned}$$

With $\alpha = 1$, we get back to the Exponential distribution.

7.7.4 Chi-squared Distribution

$$\begin{aligned} f(x) &= \frac{x^{(v-2)/2}}{2^{v/2}\Gamma(v/2)} \exp\left(-\frac{x}{2}\right), \quad x > 0; \quad v > 0 \\ E[X] &= v \\ var[X] &= 2v. \end{aligned}$$

where v is called the shape parameter of more commonly the “degrees of freedom”. Note that a Gamma distribution with $\alpha = v/2$ and $\beta = 2$ is a Chi-squared distribution. If a random variable, X, has such a density we usually write that $X \sim \chi_v^2$.

7.7.5 The Normal distribution

If $Z \sim N(0,1)$, then it can be shown that $E[Z] = 0$ and $var[Z] = 1$. Using this result and the fact that if $X = \sigma Z + \mu$, then $X \sim N(\mu, \sigma^2)$, it follows that

$$E[X] = E[\sigma Z + \mu] = \sigma E[Z] + \mu = \mu$$

and

$$var[X] = var[\sigma Z + \mu] = \sigma^2 var[Z] = \sigma^2.$$

Hence we often say that “*X is normally distributed with mean μ and variance σ^2*”, rather than location μ and scale σ.

Finally, it can be shown that the square of a standard normal random variable is a chi-squared random variable with $v = 1$ degree of freedom as follows:

Proposition 1 *Let Z be a standard normal random variable with distribution function $\Phi(z)$, then $X = Z^2$ has a chi-squared distribution with* 1 *degree of freedom: $X \sim \chi_1^2$.*

Proof. *Let $F(x) = \Pr(X \leq x)$, $x > 0$, be the distribution function of X, so that it's density is given by $f(x) = dF(x)/dx$. We first find $F(x)$, and then differentiate this to get $f(x)$. Thus, since $X = Z^2$, we have*

$$\begin{aligned} F(x) &= \Pr\left(Z^2 \leq x\right) \\ &= \Pr\left(-\sqrt{x} \leq Z \leq \sqrt{x}\right) \\ &= \Phi\left(\sqrt{x}\right) - \Phi\left(-\sqrt{x}\right) \\ &= 2\Phi\left(\sqrt{x}\right) - 1 \end{aligned}$$

and differentiating yields (where $\phi(x) = d\Phi(x)/dx$, the standard normal density)

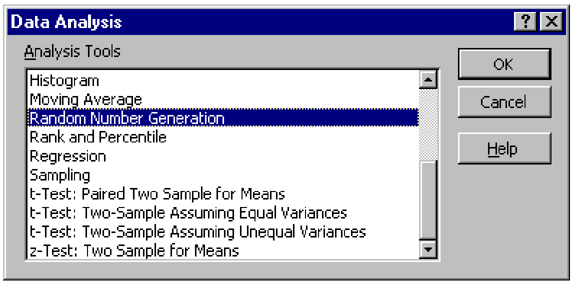
$$\begin{aligned} f(x) &= x^{-1/2}\phi(\sqrt{x}) \\ &= \frac{x^{-1/2}}{2^{1/2}\sqrt{\pi}} \exp\left(-\frac{x}{2}\right) \end{aligned}$$

and this is the density of a χ_1^2 random variable, noting that $\Gamma(1/2) = \sqrt{\pi}$. ■

7.8 Using EXCEL to create random numbers

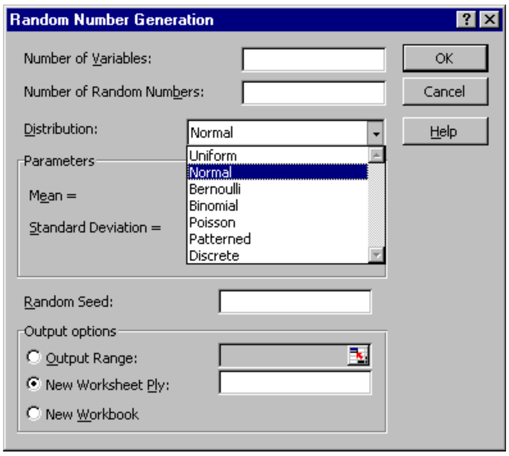
Here, we illustrate how to obtain 100 typical observations from a $N(4, 16)$ distribution.

To do this, select **Data Analysis** from the **Tools** menu, then select **Random Number Generator** from the **Data Analisis Dialogue Box**


Data Analysis
?
Analysis Tools
Histogram
Moving Average
Random Number Generation
Rank and Percentile
Regression
Sampling
t-Test: Paired Two Sample for Means
t-Test: Two-Sample Assuming Equal Variances
t-Test: Two-Sample Assuming Unequal Variances
z-Test: Two Sample for Means
OK
Cancel
Help

Data Analysis Dialogue box

Choose the distribution from the **Random Number Generation Dialogue Box**



Random Number Generation

Number of Variables:

Number of Random Numbers:

Distribution: Normal

Parameters

Mean =

Standard Deviation =

Random Seed:

Output options

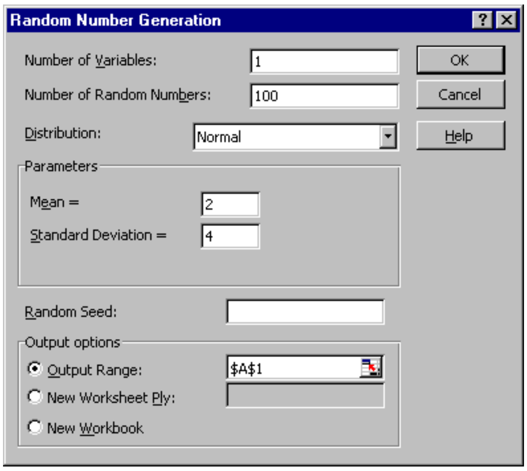
○ Output Range:

● New Worksheet Ply:

○ New Workbook

Random Number Generator Dialogue Box

(in this case **Normal**) then choose the number of (random) variables, number of random numbers (observations) for each variable, and parameters for the distribution which will be used to generate the random numbers. In the following example, we will generate 100 random numbers for a single random variable which has a normal distribution with parameters $\mu = 2$ (mean) and $\sigma = 4$ (standard deviation). If you select the cell of the first number in the **Output Range** (here **A1**) then the 100 random numbers will be placed in cells **A1:A100**. You can leave the **Random Seed** field blank, or use you registration number.



Random Number Generation

Number of Variables: 1

Number of Random Numbers: 100

Distribution: Normal

Parameters

Mean = 2

Standard Deviation = 4

Random Seed:

Output options

● Output Range: A1

○ New Worksheet Ply:

○ New Workbook

Setting the options

The effect of this is to place 100 typical observations, from a $N(2, 16)$ distribution in the cells **A1:A100.**

7.9 Exercise 4

1. Find the number z_0 such that if $Z \sim N(0,1)$

   (a) $\Pr(Z \geq z_0) = 0.05$

   (b) $\Pr(Z < -z_0) = 0.025$

   (c) $\Pr(-z_0 < Z \leq z_0) = 0.95$

   and check your answers using EXCEL.

2. If $X \sim N(4, 0.16)$ evaluate

   (a) $\Pr(X \geq 4.2)$

   (b) $\Pr(3.9 < X \leq 4.3)$

   (c) $\Pr\left((X \leq 3.8) \cup (X \geq 4.2)\right)$

   and check your answers using EXCEL. (Note for part (c), define the "events" $A = (X \leq 3.8)$ and $B = (X \geq 4.2)$ and calculate $\Pr\left(A \cup B\right)$.

3. Suppose that X is a Binomial random variable with parameters $n = 3$, $\pi = 0.5$, show by direct calculation that $E\left[X\right] = 1.5$ and $var\left[X\right] = 0.75$.

4. The continuous random variable X has probability density function given by

$$f(x) = \begin{cases} 0.1 + kx, & 0 \leq x \leq 5, \\ 0, & \text{otherwise.} \end{cases}$$

   (a) Find the value of the constant, k, which ensures that this is a proper density function.

   (b) Evaluate $E[X]$ and $var[X]$.

5. Use the random number generator in EXCEL to obtain 100 observations from a $N\left(2,1\right)$ distribution. When doing so, enter the *last four digits* from your registration number in the **Random Seed** field.

   Use EXCEL to calculate the following:

   (a) the simple average and variance of these 100 observations

   (b) the proportion of observations which are less than 1.

   Now compare these with

   (a) the theoretical mean and variance of a $N\left(2,1\right)$ distribution

(b) the probability that a random variable, with a $N(2,1)$ distribution, is less than 1.

What do you think would might happen to these comparisons if you were to generate 1000 obervations, rather than just 100?

Part II

Statistical Inference

Chapter 8

JOINT PROBABILITY DISTRIBUTIONS

The objective of statistics is to learn about population characteristics: this was first mentioned in Section 1. An **EXPERIMENT** is then any process which generates data. It is easy to imagine circumstances where an experiment generates two pieces of information, for example the weekly income of husbands and the weekly income of wives in a particular population of husbands and wives. One possible use of such data is to investigate the relationship between the observed values of the two variables. Section 2 discussed the use of correlation and regression to summarise the extent and nature of any *linear relationship* between these observed values. It has to be said that the discussion of relationships between variables in this section does but scratch the surface of a very large topic. Subsequent courses in Econometrics take the analysis of relationships between two or more variables much further.

If these two pieces of information generated by the experiment are considered to be the values of two random variables defined on the **SAMPLE SPACE** of an experiment (cf. Section 3), then the discussion of random variables and probability distributions in Sections 5 and 6 needs to be extended.

8.1 Joint Probability Distributions

Let X and Y be the two random variables: for simplicity, they are considered to be discrete random variables. The outcome of the experiment is a pair of values (x, y). The probability of this outcome is a **joint** probability which can be denoted

$$\Pr(X = x \cap Y = y),$$

emphasising the analogy with the probability of a joint event $\Pr(A \cap B)$, or, more usually, by

$$\Pr(X = x, Y = y).$$

- The collection of these probabilities, for all possible combinations of x and y, is the **joint probability distribution** of X and Y, denoted

$$p(x, y) = \Pr(X = x, Y = y).$$

- The **Axioms of Probability** in Section 3.2 carry over to imply

$$0 \leqslant p(x, y) \leqslant 1,$$

$$\sum_x \sum_y p(x, y) = 1,$$

- where the sum is over all (x, y) values.

8.1.1 Examples

Example 2 *Let H and W be the random variables representing the population of weekly incomes of husbands and wives, respectively, in some country. There are only three possible weekly incomes, £0, £100 or £ 200. The joint probability distribution of H and W is represented as a table:*

		Values of H :		
Probabilities		0	1	2
Values of W :	0	0.05	0.15	0.10
	1	0.10	0.10	0.30
	2	0.05	0.05	0.10

Then we can read off, for example, that

$$\Pr(H = 0, W = 0) = 0.05,$$

or that in this population, 5% *of husbands and wives have each a zero weekly income.*

In this example, the nature of the experiment underlying the population data is not explicitly stated. However, in the next example, the experiment is described, the random variables defined in relation to the experiment, and their probability distribution deduced directly.

Example 3 *Consider the following simple version of a lottery. Players in the lottery choose one number between* 1 *and* 5*, whilst a machine selects the lottery winners by randomly selecting one of five balls (numbered* 1 *to* 5*). Any player whose chosen number coincides with the number on the ball* *is a winner. Whilst the machine selects one ball at random (so that each ball has an* 0.2 *chance of selection), players of the lottery have "lucky" and "unlucky" numbers, with the probabilities of choosing each specific number as follows:*

Number chosen by player	*Probability of being chosen*
1	0.40
2	0.20
3	0.05
4	0.10
5	0.25
	1.0

Let X denote the number chosen by a player and Y the number chosen by the machine. If they are assumed to be independent events, then for each possible value of X and Y, we will have

$$\Pr(X \cap Y) = \Pr(X) \Pr(Y).$$

The table above gives the probabilities for X, and $\Pr(Y) = 0.2$, *so that a table can be drawn up displaying the joint distribution $p(x, y)$:*

Probabilities	*Y :*					
X :	*Selected by machine*					
Chosen by player	1	2	3	4	5	*Row Total*
1	0.08	0.08	0.08	0.08	0.08	0.40
2	0.04	0.04	0.04	0.04	0.04	0.20
3	0.01	0.01	0.01	0.01	0.01	0.05
4	0.02	0.02	0.02	0.02	0.02	0.10
5	0.05	0.05	0.05	0.05	0.05	0.25
Column Total	0.20	0.20	0.20	0.20	0.20	1.00

The general question of independence in joint probability distributions will be discussed later in the section.

8.2 Marginal Probabilities

Given a joint probability distribution

$$p(x, y) = \Pr(X = x, Y = y)$$

for the random variables X and Y, a probability of the form $\Pr(X = x)$ or $\Pr(Y = y)$ is called a **marginal** probability.

The collection of these probabilities for all values of X is the **marginal probability distribution** for X,

$$p_X(x) = \Pr(X = x).$$

If it is clear from the context, write $p_X(x)$ as $p(x)$.

Suppose that Y takes on values $0, 1, 2$. Then

$$\Pr(X = x) = \Pr(X = x, Y = 0) + \Pr(X = x, Y = 1) + \Pr(X = x, Y = 2),$$

the sum of all the joint probabilities favourable to $X = x$. So, marginal probability distributions are found by summing over all the values of the other variable:

$$p_X(x) = \sum_y p(x, y), \qquad p_Y(y) = \sum_x p(x, y).$$

This can be illustrated using Example 1 of Section 8.1.1 again:

$$\begin{aligned} \Pr(W = 0) &= \Pr(W = 0, H = 0) + \Pr(W = 0, H = 1) + \Pr(W = 0, H = 2) \\ &= 0.05 + 0.15 + 0.10 \\ &= 0.30. \end{aligned}$$

There is a simple recipe for finding the marginal distributions in the table of joint probabilities: find the row sums and column sums. From Example 1 in Section 8.1.1,

Probabilities		Values of H : 0	1	2	Row Sums $p_W(w)$
Values of W :	0	0.05	0.15	0.10	0.30
	1	0.10	0.10	0.30	0.50
	2	0.05	0.05	0.10	0.20
Column Sums: $p_H(h)$		0.20	0.30	0.50	1.00

from which the marginal distributions should be written out explicitly as

Values of W	$p_W(w)$	Values of H	$p_H(h)$
0	0.3	0	0.2
1	0.5	1	0.3
2	0.2	2	0.5
	1.0		1.0

By calculation, we can find the expected values and variances of W and H as

$$\begin{aligned} E[W] &= 0.9, \quad \operatorname{var}[W] = 0.49, \\ E[H] &= 1.3, \quad \operatorname{var}[H] = 0.61. \end{aligned}$$

Notice that a marginal probability distribution has to satisfy the usual properties expected of a probability distribution (for a discrete random variable):

$$\begin{aligned} 0 &\le p_X(x) \le 1, \quad \sum_x p_X(x) = 1, \\ 0 &\le p_Y(y) \le 1, \quad \sum_y p_Y(y) = 1. \end{aligned}$$

8.3 Functions of Two Random Variables

Given the experiment of Example 1 of section 8.1.1, one can imagine defining further random variables on the sample space of this experiment. One example is the random variable T representing total household income:

$$T = H + W.$$

This new random variable is a (linear) **function** of H and W, and we can deduce the probability distribution of T from the joint distribution of H and W. For example,

$$\begin{aligned} \Pr(T = 0) &= \Pr(H = 0, W = 0), \\ \Pr(T = 1) &= \Pr(H = 0, W = 1) + \Pr(H = 1, W = 0). \end{aligned}$$

The complete probability distribution of T is

Values of T	$\Pr(T = t)$	$t \times \Pr(T = t)$
0	0.05	0
1	0.25	0.25
2	0.25	0.5
3	0.35	1.05
4	0.10	0.40
	1.00	2.20

from which we note that $E[T] = 2.2$, indicating that the population mean income for married couples in the specific country is £220.

Now we consider a more formal approach. Let X and Y be two discrete random variables with joint probability distribution $p(x, y)$. Let V be a random variable defined as a function of X and Y :

$$V = g(X, Y).$$

Here, $g(X, Y)$ is not necessarily a linear function: it could be any function of two variables. In principle, we can deduce the probability distribution of V from $p(x, y)$ and thus deduce the mean of V, $E[V]$, just as we did for T in Example 1.

However, there is a second method that works directly with the joint probability distribution $p(x, y)$: the expected value of V is

$$E[V] = E[g(X, Y)] = \sum_x \sum_y g(x, y) p(x, y).$$

The point about this approach is that it avoids the calculation of the probability distribution of V.

To apply this argument to find $E[T]$ in Example 1, it is helpful to modify the table of joint probabilities to display the value of T associated with each pair of values for H and W :

		h		
	(t)	0	1	2
w	0	(0) 0.05	(1) 0.15	(2) 0.10
	1	(1) 0.10	(2) 0.10	(3) 0.30
	2	(2) 0.05	(3) 0.05	(4) 0.10

.

Then, the double summation required for the calculation of $E[T]$ can be performed along each row in turn:

$$\begin{aligned} E[T] &= (0) \times 0.05 + (1)(0.15) + (2) \times 0.10 \\ &\quad + (1) \times 0.10 + (2)(0.10) + (3) \times 0.30 \\ &\quad + (2) \times 0.05 + (3)(0.05) + (4) \times 0.10 \\ &= 2.20. \end{aligned}$$

So, the recipe is to multiply, for each cell, the implied value of T in that by the probability in that cell, and add up the calculated values over all the cells.

8.4 Independence, Covariance and Correlation

8.4.1 Independence

If the random variables X and Y have a joint probability distribution

$$p(x, y) = \Pr(X = x, Y = y),$$

then it is possible that for some combinations of x and y, the events $(X = x)$ and $(Y = y)$ are independent events:

$$\Pr(X = x, Y = y) = \Pr(X = x) \Pr(Y = y),$$

or, using the notation of joint and marginal probability distributions,

$$p(x, y) = p_X(x) \, p_Y(y).$$

If this relationship holds for all values of x and y, the random variables X and Y are said to be **independent:**

- X and Y are **independent** random variables if and only if

$$p(x, y) = p_X(x) \, p_Y(y) \qquad \text{for } \textbf{all } x, y.$$

Each joint probability is the product of the corresponding marginal probabilities. Independence also means that $\Pr(Y = y)$ would not be affected by knowing that $X = x$: knowing the value taken on by one random variable does not affect the probabilities of the outcomes of the other random variable. A corollary of this is that if two random variables X and Y are independent, then there can be no relationship of any kind, linear or non-linear, between them.

- The joint probabilities and marginal probabilities for Example 1 in section 8.1.1 are

Probabilities		Values of H : 0	 1	 2	Row Sums $p_W(w)$
Values of W :	0	0.05	0.15	0.10	0.30
	1	0.10	0.10	0.30	0.50
	2	0.05	0.05	0.10	0.20
Column Sums: $p_H(h)$		0.20	0.30	0.50	1.00

Here $p(0,0) = 0.05$, whilst $p_W(0) = 0.30$, $p_H(0) = 0.20$, with

$$p(0,0) \neq p_W(0)\,p_H(0).$$

So, H and W cannot be independent.

For X and Y to be independent, $p(x,y) = p_X(x)\,p_Y(y)$ has to hold for all x, y. Finding one pair of values x, y for which this fails is sufficient to conclude that X and Y are not independent. However, one may also have to check every possible pair of values to confirm independence: think what would be required in Example 2 of Section 8.1.1, if one did not know that the joint probability distribution had been constructed using an independence property.

8.4.2 Covariance

A popular measure of association for random variables X and Y is the (population) correlation coefficient. It is the population characteristic analogous to the (sample) correlation coefficient introduced in Section 2.3.3. It will be seen that this (population) correlation coefficient is really only a measure of strength of any linear relationship between the random variables.

The first step is to define the (population) covariance as a characteristic of the joint probability distribution of X and Y. Let

$$E[X] = \mu_X, \qquad E[Y] = \mu_Y.$$

- the (population) **covariance** is defined as

$$\begin{aligned} \operatorname{cov}[X,Y] &= E[(X-\mu_X)(Y-\mu_Y)] \\ &= \sigma_{XY}. \end{aligned}$$

Notice that by this definition, $\operatorname{cov}[X,Y] = \operatorname{cov}[Y,X]$.

- There are a number of alternative expressions for the covariance. The first follows from seeing

  $$(X-\mu_X)(Y-\mu_Y)$$

  as a function $g(X,Y)$ of X and Y :

  $$\operatorname{cov}[X,Y] = \sum_x \sum_y (x-\mu_X)(y-\mu_Y)\, p(x,y)$$

  We can see from this expression that if enough (x,y) pairs have $x-\mu_X$ and $y-\mu_Y$ values with the same sign, $\operatorname{cov}[X,Y] > 0$, so that large (small) values of $x-\mu_X$ tend to occur with large (small) values of $y-\mu_Y$. Similarly, if enough (x,y) pairs have $x-\mu_X$ and $y-\mu_Y$ values with different signs, $\operatorname{cov}[X,Y] < 0$. Here, large (small) values of $x-\mu_X$ tend to occur with **small** (large) values of $y-\mu_Y$.

- $\operatorname{cov}[X,Y] > 0$ gives a "positive" relationship between X and Y, $\operatorname{cov}[X,Y] < 0$ a "negative" relationship.

- There is a shorthand calculation for covariance, analogous to that given for the variance in Section 8.5:

  $$\begin{aligned} \operatorname{cov}[X,Y] &= E[(X-\mu_X)(Y-\mu_Y)] \\ &= E[XY - X\mu_Y - \mu_X Y + \mu_X\mu_Y] \\ &= E[XY] - E[X]\mu_Y - \mu_X E[Y] + \mu_X\mu_Y \\ &= E[XY] - \mu_X\mu_Y - \mu_X\mu_Y + \mu_X\mu_Y \\ &= E[XY] - \mu_X\mu_Y. \end{aligned}$$

  Here, the linear function rule of Section 8.3 has been used to make the expected value of a sum of terms equal to the sum of expected values, and then to make, for example,

  $$E[X\mu_Y] = E[X]\,\mu_Y.$$

Even with this shorthand method, the calculation of the covariance is rather tedious. To calculate $\operatorname{cov}[W,H]$ in Example 1, the best approach is to imitate the way in which $E[T]$ was calculated in Section 8.3. Rather than display the values of T, here we display the values of $W \times H$ in order to first calculate $E[WH]$:

		h		
	$(w \times h)$	0	1	2
w	0	(0) 0.05	(0) 0.15	(0) 0.10
	1	(0) 0.10	(1) 0.10	(2) 0.30
	2	(0) 0.05	(2) 0.05	(4) 0.10

.

Using the same strategy of multiplication within cells, and adding up along each row in turn, we find

$$\begin{aligned} E[WH] &= (0)\times 0.05 + (0)\times 0.15 + (0)\times 0.10 \\ &\quad + (0)\times 0.10 + (1)\times 0.10 + (2)\times 0.30 \\ &\quad + (0)\times 0.05 + (2)\times 0.05 + (4)\times 0.10 \\ &= 0.1 + 0.6 + 0.1 + 0.4 \\ &= 1.2. \end{aligned}$$

We found in Section 8.2 that $E[W] = 0.9, E[H] = 1.3$, so that

$$\begin{aligned} \operatorname{cov}[W,H] &= E[WH] - E[W]E[H] \\ &= 1.2 - (0.9)(1.3) \\ &= 0.03. \end{aligned}$$

Strength of Association and Units of Measurement

How does covariance measure the **strength** of a relationship between X and Y? Not well is the answer, because the value of the covariance is dependent on the units of measurement. Suppose in Example 1 of section 8.1.1 the units of measurement of W are changed to pounds rather than the original hundreds of pounds. Let V represent W in the new units:

$$V = 100W:$$

where W had values $0, 1, 2$, V now has values $0, 100, 200$. Notice from section 8 that

$$E[V] = \mu_V = 100E[W] = 100(0.9) = 90,$$

and in turn

$$V - \mu_V = 100W - 100\mu_W = 100(W - \mu_W).$$

Then,

$$\begin{aligned} \operatorname{cov}[V,H] &= E[(V-\mu_V)(H-\mu_H)] \\ &= E[100(W-\mu_W)(H-\mu_H)] \\ &= 100E[(W-\mu_W)(H-\mu_H)] \\ &= 100\operatorname{cov}[W,H] \\ &= 3. \end{aligned}$$

Simply by changing units of measurement, we can make the strength of the association between wife's income and husband's income bigger. If in addition, we also measured husband's income in pounds rather than hundreds of pounds, the covariance would increase to 300 - even better, it seems! This is easily confirmed by replacing $(H - \mu_H)$ by $100(H - \mu_H)$ in the covariance calculation.

Correlation

This cannot be sensible: what is required is a measure of the strength of association which is **invariant** to changes in units of measurement. Generalising what we have just seen, if the units of measurement of two random variables X and Y are changed to produce new random variables αX and βY, then the covariance in the new units of measurement is related to the covariance in the original units of measurement by

$$\operatorname{cov}[\alpha X, \beta Y] = \alpha\beta \operatorname{cov}[X, Y].$$

What are the variances of αX and βY in terms of $\operatorname{var}[X]$ and $\operatorname{var}[Y]$? By Section 8.3, they are

$$\operatorname{var}[\alpha X] = \alpha^2 \operatorname{var}[X], \quad \operatorname{var}[\beta Y] = \beta^2 \operatorname{var}[Y].$$

The (population) **correlation coefficient** between X and Y is defined by

$$\rho_{XY} = \frac{\operatorname{cov}[X, Y]}{\sqrt{\operatorname{var}[X] \operatorname{var}[Y]}}.$$

This is also the correlation between αX and βY :

$$\begin{aligned} \rho_{\alpha X, \beta Y} &= \frac{\operatorname{cov}[\alpha X, \beta Y]}{\sqrt{\operatorname{var}[\alpha X] \operatorname{var}[\beta Y]}} \\ &= \frac{\alpha\beta \operatorname{cov}[X, Y]}{\sqrt{\alpha^2\beta^2 \operatorname{var}[X] \operatorname{var}[Y]}} \\ &= \rho_{XY}, \end{aligned}$$

so that the correlation coefficient does **not** depend on the units of measurement.

In Section 8.2, we found for Example 1 of section 8.1.1 that $\operatorname{var}[W] = 0.49, \operatorname{var}[H] = 0.61$, so that

$$\begin{aligned} \rho_{WH} &= \frac{0.03}{\sqrt{(0.49)(0.61)}} \\ &= 0.0549. \end{aligned}$$

Is this indicative of a strong relationship? Just like the sample correlation coefficient of Section 2.3.3, it can be shown that

- the correlation coefficient ρ_{XY} always satisfies $-1 \leqslant \rho_{XY} \leqslant 1$.

- The closer ρ is to 1 or -1, the stronger the relationship.

So, $\rho_{WH} = 0.0549$ is indicative of a very weak relationship.

It can shown that if X and Y are exactly linearly related by

$$Y = a + bX \quad \text{with} \quad b > 0$$

then $\rho_{XY} = 1$ - that is, X and Y are perfectly correlated. X and Y are also perfectly correlated if they are exactly linearly related by

$$Y = a + bX \quad \text{with} \quad b < 0,$$

but $\rho_{XY} = -1$. Thus,

- correlation measures only the strength of a **linear** relationship between X and Y;
- correlation does **not** imply **causation.**

Other notations for the correlation coefficient are

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

which uses covariance and standard deviation notation, and

$$\rho_{XY} = \frac{E\left[(X - \mu_X)(Y - \mu_Y)\right]}{\sqrt{E\left[(X - \mu_X)^2\right] E\left[(Y - \mu_Y)^2\right]}}.$$

Correlation, Covariance and Independence

Non-zero correlation and covariance between random variables X and Y indicate some linear association between them, whilst (Section 8.4.1) independence of X and Y implies no relationship or association of any kind between them. So, it is not surprising that

- **independence** of X and Y implies **zero covariance:** $\text{cov}[X, Y] = 0$;
- **independence** of X and Y implies **zero correlation:** $\rho_{XY} = 0$.

The converse is not true, in general:

- **zero covariance or correlation** does not imply independence.

The reason is that there may be a relationship between X and Y which is not linear.

8.5 Conditional Distributions

In Sections 8.1 and 8.2, we looked at the jointe and marginal distributions for a pair of discrete random variables. Continuing the previous discussion of relationships between variables, we are often interested in econometrics in how random variable X affects random variable Y. This information is contained in something called the *conditional distribution* of "Y given X". For discrete random variables, X and Y, this distribution is defined by the following probabilities

$$\begin{aligned} p_{Y|X}\left(y|x\right) &= \Pr\left(Y=y|X=x\right) \\ &= \frac{p\left(x,y\right)}{p_X\left(x\right)}, \end{aligned}$$

where reads as "the probability that Y takes the value y given that (conditional on) X takes the value x". As with the discussion of conditional probability in Chapter 3, these conditional probabilities are defined on a restricted sample space of $X = x$ (hence the rescaling by $p_X(x)$) and they are calculated on a sequence on restricted sample spaces; one for each possible value of x (in the discrete case).

As an illustration of the calculations, consider again Example 8.1.1 and the construction of the conditional distribution of W given H for which we had the following joint distribution:

		Values of H :		
Probabilities		0	1	2
Values of W :	0	0.05	0.15	0.10
	1	0.10	0.10	0.30
	2	0.05	0.05	0.10

We consider, in turn, conditional probablities for the values of W given, first $H = 0$, then $H = 1$ and finally $H = 2$. Intuitively, think of the probabilities in the cells as indicating sub-areas of the entire sample space, with the latter having and area of 1 and the former (therefore) summing to 1. With this interpretation, the restriction $H = 0$ "occupies" 20% of the entire sample space (recall the marginal probability, $\Pr(H = 0)$, from Section 8.2). The three cells corresponding to $H = 0$ now correspond to the restricted sample space of $H = 0$, and the outcome $W = 0$ takes up $0.05/0.2 = 0.25$ of this restricte sample space; thus $\Pr(W = 0|H = 0) = \Pr(W = 0.H = 0) / \Pr(H = 0) = 0.25$. Similarly, $\Pr(W = 1|H = 0) = 0.10/0.2 = 0.5$ and $\Pr(W = 2|H = 0) = 0.05/0.2 = 0.25$. Notice that $\sum_{j=0}^{2} \Pr(W = j|H = 0) = 1$, as it should do for the restricted sample space of $H = 0$. For all possible restrictions imposed by H we get the following conditional distributions for W (we get three conditional distributions, one for each of $h = 0$, $h = 1$, $h = 2$) :

		Values of H :			$Pr\left(W=w \mid H=h\right)$		
Probabilities		0	1	2	$h=0$	$h=1$	$h=2$
Values of W :	0	0.05	0.15	0.10	1/4	1/2	1/5
	1	0.10	0.10	0.30	1/2	1/3	3/5
	2	0.05	0.05	0.10	1/4	1/6	1/5

Notice how the probabilities for particular values of W change according to the restriction imposed by H; for example, $\Pr\left(W=0|H=0\right) \neq \Pr\left(W=0|H=1\right)$, say. Thus knowledge of, or information about, H changes probabilities concerning W. Because of this, and as ascertained previously, W and H are NOT independent.

In general,.X and Y are independent, if and only if knowledge of the value taken by X does not tell us anything about the probability that Y takes any particular value. Indeed, from the definition of $p_{Y|X}\left(y|x\right)$, we see that X and Y are independent if and only if $p_{Y|X}\left(y|x\right) = p_Y\left(y\right)$, for **all** x, y.

There is a similar treatment for conditional distributions for continuous random variables.

8.5.1 Conditional Expectation

While correlation is a useful summary of the relationship between two random variables, in econometrics we often want to go further and explain one random variable Y as a function of some other random variable X. One way of doing this is to look at the properties of the distribution of Y conditional on X, as intruduced above. In general these properties, such as expectation and variance, will depend on the value of X, thus we can think of them as being functions of X. The conditional expectation of Y is denoted $E(Y|X=x)$ and tells us the expectation of Y given that X has taken the particular value x. Since this will vary with the particular value taken by X we can think of $E(Y|X=x)=m(x)$, as a function of x.

As an example think of the population of all working individuals and let X be *years of education* and Y be *hourly wages.* $E(Y|X=12)$ is the expected hourly wage for all those people who have 12 years of education while $E(Y|X=16)$ tells us the expected hourly wage for all those who have 16 years of education. *Tracing out* the values of $E(Y|X=x)$ for all values of X tells us a lot about how education and wages are related.

In econometrics we typically summarise the relationship represented by $E(Y|X)=m(X)$ in the form of a simple function. For example we could use a simple linear function:

$$E(WAGE|EDUC) = 1.05 + 0.45 * EDUC$$

or a non-linear function:

$$E(QUANTITY|PRICE) = 10/PRICE,$$

with the latter example demonstrating the deficiencies of correlation as a measure of association (since it confines itself to the consideration of linear relationships only).

Properties of Conditional Expectation

The following properties hold for both discrete and continuous random variables.

- $E[c(X)|X] = c(X)$ for any function $c(X)$.
  Functions of X behave as constants when we compute expectations conditional on X. (If we know the value of X then we know the value of $c(X)$ so this is effectively a constant.)

For functions $a(X)$ and $b(X)$

- $E[a(X)Y + b(X)|X] = a(X)E(Y|X) + b(X)$
  This is an extension of the previous rule's logic and says that since we are conditioning on X, we can treat X, and any function of X, as a constant when we take the expectation.

- If X and Y are independent, then $E(Y|X) = E(Y)$.
  This follows immediately from the earlier discussion of conditional probability distributions. If the two random variables are independent then knowledge of the value of X should not change our view of the likelihood of any value of Y. It should therefore not change our view of the expected value of Y.
  A special case is where U and X are independent and $E(U) = 0$. It is then clear that $E(U|X) = 0$.

- $E[E(Y|X)] = E(Y)$
  This result is known as the "iterative expectations" rule. We can think of $E(Y|X)$ as being a function of X. Since X is a random variable then $E(Y|X) = m(X)$ is a random variable and it makes sense to think about its distribution and hence its expected value. Think about the following example: suppose $E(WAGE|EDUC) = 4 + 0.6 * EDUC$. Suppose $E(EDUC) = 11.5$. Then according to the iterative expectation rule

  $$E(WAGE) = E(4 + 0.6 * EDUC) = 4 + 0.6(11.5) = 10.9.$$

- If $E(Y|X) = E(Y)$ then $Cov(X, Y) = 0$.

The last two properties have immediate applications in econometric modelling: if U and X are random variables with $E(U|X) = 0$, then $E(U) = 0$ and $Cov(U, X) = 0$.

Finally, $E(Y|X)$ is often called the "*regression*" of Y on X. We can always write

$$Y = E(Y|X) + U$$

where, by the above properties, $E(U|X) = 0$. Now consider $E(U^2|X)$, which is

$$E\left(U^2|X\right) = E\left[\left(Y - E\left(Y|X\right)\right)^2|X\right] = var\left(Y|X\right)$$

the conditional variance of Y given X.

In general, it can be shown that

$$var\left(Y\right) = E\left[var\left(Y|X\right)\right] + var\left[E\left(Y|X\right)\right].$$

Chapter 9

LINEAR COMBINATIONS OF RANDOM VARIABLES

In this section, some properties of linear functions of random variables X and Y are considered. In Section 8.3, a new random variable V was defined as a function of X and Y,

$$V = g(X, Y),$$

with no restriction on the nature of the function or transformation g. In this section, the function g is restricted to be a **linear** function of X and Y :

$$V = aX + bY + c,$$

where a, b and c are constants. V is also called a **linear combination** of X and Y.

The properties developed in this section are specific to linear functions: they do not hold in general for nonlinear functions or transformations.

9.1 The Expected Value of a Linear Combination

> This result is easy to remember: it amounts to saying that
>
> *the expected value of a linear combination is the linear combination of the expected values.*
>
> Even more simply, *the expected value of a sum is a sum of expected values.*

If $V = aX + bY + c$, where a, b, c are constants, then

$$E[V] = E[aX + bY + c] = aE[X] + bE[Y] + c.$$

This result is a natural generalisation of that given in Section 8.3.

- **Proof** (discrete random variables case only). Using the result in Section 8.3,

$$\begin{aligned} E[V] &= E[g(X,Y)] \\ &= E[aX+bY+c] \\ &= \sum_x \sum_y (ax+by+c)\, p(x,y). \end{aligned}$$

  From this point on, the proof just involves manipulation of the summation signs:

$$\begin{aligned} \sum_x \sum_y (ax+by+c)\, p(x,y) &= a\sum_x \sum_y xp(x,y) + b\sum_x \sum_y yp(x,y) + c\sum_x \sum_y p(x,y) \\ &= a\sum_x \left[x\left(\sum_y p(x,y) \right) \right] + b\sum_y \left[y\left(\sum_x p(x,y) \right) \right] + c \\ &= a\sum_x \left[xp_X(x) \right] + b\sum_y \left[yp_Y(y) \right] + c \\ &= aE[X] + bE[Y] + c. \end{aligned}$$

Notice the steps used:

- $\sum_y xp(x,y) = x\sum_y p(x,y) = xp_X(x)$, $\sum_x yp(x,y) = y\sum_x p(x,y) = xp_X(x)$,
- because x is constant with respect to y summation and y is constant with respect to x summation;
- $\sum_x \sum_y p(x,y) = 1$,
- the definitions of marginal distributions from Section 8.2;
- the definitions of expected value for discrete random variables.

Notice that nothing need be known about the joint probability distribution $p(x,y)$ of X and Y. The result is also valid for continuous random variables, nothing need be known about $p(x,y)$.

9.1.1 Examples

1. Example 1 from Section 8.1.1: we defined

$$T = W + H,$$

and had $E[W] = 1.3$, $E[H] = 0.9$, giving

$$\begin{aligned} E[T] &= E[W] + E[H] \\ &= 2.2 \end{aligned}$$

confirming the result obtained earlier.

2. Suppose that the random variables X and Y have $E[X] = 0.5$ and $E[Y] = 3.5$, and let

$$V = 5X - Y.$$

Then,

$$\begin{aligned} E[V] &= 5E[X] - E[Y] \\ &= (5)(0.5) - 3.5 \\ &= -1. \end{aligned}$$

9.1.2 Generalisation

Let $X_1, ..., X_n$ be random variables and $a_1, ..., a_n, c$ be constants, and define the random variable W by

$$\begin{aligned} W &= a_1X_1 + ... + a_nX_n + c \\ &= \sum_{i=1}^{n} a_iX_i + c. \end{aligned}$$

Then,

$$E[W] = \sum_{i=1}^{n} a_iE[X_i] + c.$$

The proof uses the linear combination result for two variables **repeatedly**:

$$\begin{aligned} E[W] &= a_1E[X_1] + E[a_2X_2 + ... + a_nX_n + c] \\ &= a_1E[X_1] + a_2E[X_2] + E[a_3X_3 + ... + a_nX_n + c] \\ &= ... \\ &= a_1E[X_1] + a_2E[X_2] + ... + a_nE[X_n] + c. \end{aligned}$$

Example: let $E[X_1] = 2, E[X_2] = -1, E[X_3] = 3$, $W = 2X_1 + 5X_2 - 3X_3 + 4$ and then

$$\begin{aligned} E[W] &= E[2X_1 + 5X_2 - 3X_3 + 4] \\ &= 2E[X_1] + 5E[X_2] - 3E[X_3] + 4 \\ &= (2)(2) + (5)(-1) - (3)(3) + 4 \\ &= -6. \end{aligned}$$

9.2 The Variance of a Linear Combination

9.2.1 Two Variable Case

Let V be the random variable defined in Section 9.1:

$$V = aX + bY + c.$$

What is $\operatorname{var}[V]$? To find this, it is helpful to use notation that will simplify the proof. By definition,

$$\operatorname{var}[V] = E\left[(V - E[V])^2\right].$$

Put

$$\tilde{V} = V - E[V]$$

so that

$$\operatorname{var}[V] = E\left[\tilde{V}^2\right].$$

We saw that

$$E[V] = aE[X] + bE[Y] + c,$$

so that

$$\begin{aligned}\tilde{V} &= (aX + bY + c) - (aE[X] + bE[Y] + c) \\ &= a(X - E[X]) + b(Y - E[Y]) \\ &= a\tilde{X} + b\tilde{Y}\end{aligned}$$

and then

$$\operatorname{var}[V] = E\left[\tilde{V}^2\right] = E\left[\left(a\tilde{X} + b\tilde{Y}\right)^2\right].$$

Notice that this does not depend on the constant c.

To make further progress, recall that in the current notation,

$$\begin{aligned}\operatorname{var}[X] &= E\left[\tilde{X}^2\right], \quad \operatorname{var}[Y] = E\left[\tilde{Y}^2\right], \\ \operatorname{cov}[X, Y] &= E[(X - E[X])(Y - E[Y])] \\ &= E\left[\tilde{X}\tilde{Y}\right].\end{aligned}$$

Then,

$$\begin{aligned}\operatorname{var}[V] &= E\left[\left(a\tilde{X} + b\tilde{Y}\right)^2\right] \\ &= E\left[a^2\tilde{X}^2 + 2ab\tilde{X}\tilde{Y} + b^2\tilde{Y}^2\right] \\ &= a^2E\left[\tilde{X}^2\right] + 2abE\left[\tilde{X}\tilde{Y}\right] + b^2E\left[\tilde{Y}^2\right] \\ &= a^2\operatorname{var}[X] + 2ab\operatorname{cov}[X, Y] + b^2\operatorname{var}[Y],\end{aligned}$$

using the linear combination result for expected values.

Summarising,

- if $V = aX + bY + c$, then
$$\text{var}[V] = a^2 \text{var}[X] + 2ab \,\text{cov}[X, Y] + b^2 \text{var}[Y].$$
- If X and Y are **uncorrelated**, so that $\text{cov}[X, Y] = 0$,
$$\text{var}[V] = a^2 \text{var}[X] + b^2 \text{var}[Y].$$
This special case can be nicely summarised as

  the variance of a (weighted) sum is a (weighted) sum of variances.

  Notice that the weights get squared, as is usual for variances.
- If X and Y are independent, the same result holds.

Examples

1. Suppose that X and Y are independent random variables with $\text{var}[X] = 0.25$, $\text{var}[Y] = 2.5$. If
$$V = X + Y,$$
then
$$\begin{aligned}\text{var}[V] &= \text{var}[X] + \text{var}[Y] \\ &= 0.25 + 2.5 \\ &= 2.75.\end{aligned}$$
2. This uses Example 1 of Section 8.1.1, with random variables W and H, and T defined by
$$T = W + H.$$
In Section 8.2, we found that $\text{var}[W] = 0.49, \text{var}[H] = 0.61$, whilst in Section 8.4.2 we found $\text{cov}[W, H] = 0.03$. Then,
$$\begin{aligned}\text{var}[T] &= \text{var}[W + H] \\ &= \text{var}[W] + 2\,\text{cov}[W, H] + \text{var}[H]\end{aligned}$$
(since this is a case with $a = b = 1$). So,
$$\text{var}[T] = 0.49 + (2)(0.03) + 0.61 = 1.16.$$
3. For the same joint distribution, the difference between the income of husbands and wives is
$$D = H - W.$$
This case has $a = 1$ and $b = -1$, so that
$$\begin{aligned}\text{var}[D] &= (1)^2 \text{var}[H] + 2(1)(-1)\,\text{cov}[W, H] + (-1)^2 \text{var}[W] \\ &= 0.61 - (2)(0.03) + 0.49 \\ &= 1.04.\end{aligned}$$

Generalisation

To extend the result to the case of a linear combination of n random variables $X_1, ..., X_n$ is messy because of the large number of covariance terms involved. So, we simplify by supposing that $X_1, ..., X_n$ are uncorrelated random variables, with all covariances equal to zero: $\text{cov}[X_i, X_j] = 0, i \neq j$. Then,

- for $X_1, ..., X_n$ uncorrelated,

$$\text{var}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i^2 \,\text{var}[X_i].$$

- This also applies when $X_1, ..., X_n$ are **independent** random variables.

Standard Deviations

None of these results apply **directly** to standard deviations. Consider the simple case where X and Y are independent random variables and

$$W = X + Y.$$

Then,

$$\begin{aligned}\text{var}[W] &= \sigma_W^2 \\ &= \text{var}[X] + \text{var}[Y] \\ &= \sigma_X^2 + \sigma_Y^2\end{aligned}$$

and then

$$\sigma_W = \sqrt{\sigma_X^2 + \sigma_Y^2}.$$

In general it is true that

$$\sigma_W \neq \sigma_X + \sigma_Y.$$

To illustrate, if X_1, X_2 and X_3 are independent random variables with $\text{var}[X_1] = 3, \text{var}[X_2] = 1$ and $\text{var}[X_3] = 5$, and if

$$P = 2X_1 + 5X_2 - 3X_3,$$

then

$$\begin{aligned}\text{var}[P] &= 2^2\,\text{var}[X_1] + 5^2\,\text{var}[X_2] + (-3)^2\,\text{var}[X_3] \\ &= (4)(3) + (25)(1) + (9)(5) \\ &= 82, \\ \sigma_P &= \sqrt{82} = 9.06.\end{aligned}$$

9.3 Linear Combinations of Normal Random Variables

The results in this section so far relate to characteristics of the probability distribution of a linear combination like

$$V = aX + bY,$$

not to the probability distribution itself. Indeed, part of the attraction of these results is that they can be obtained without having to find the probability distribution of V.

However, if we knew that X and Y had normal distributions then it would follow that V is also normally distributed. This innocuous sounding result is **EXTREMELY IMPORTANT!** It is also rather unusual: there are not many distributions for which this type of result holds.

More specifically,

- if $X \sim N\left(\mu_X, \sigma_X^2\right)$ and $Y \sim N\left(\mu_Y, \sigma_Y^2\right)$ and $W = aX + bY + c$, then

  $$W \sim N\left(\mu_W, \sigma_W^2\right)$$

  with

  $$\begin{aligned} \mu_W &= a\mu_X + b\mu_Y + c, \\ \sigma_W^2 &= a^2\sigma_X^2 + 2ab\sigma_{XY} + b^2\sigma_Y^2 \end{aligned}$$

  Note that **independence** of X and Y has **not** been assumed.

- If $X_1, ...X_n$ are uncorrelated random variables with $X_i \sim N\left(\mu_i, \sigma_i^2\right)$ and $W = \sum\limits_{i=1}^{n} a_i X_i$, where $a_1, ..., a_n$ are constants, then

  $$W \sim N\left(\sum_{i=1}^{n} a_i\mu_i, \sum_{i=1}^{n} a_i^2\sigma_i^2\right).$$

Of course, standard normal distribution tables can be used in the usual way to compute probabilities of events involving W. This is illustrated in the following

- Example. If $X \sim N(20, 5), Y \sim N(30, 11)$, X and Y independent, and

  $$D = X - Y,$$

  then

  $$D \sim N(-10, 16)$$

and

$$\begin{aligned}\Pr(D > 0) &= \Pr\left(Z > \frac{0-(-10)}{4}\right) \\ &= \Pr(Z > 2.5) \\ &= 0.00621,\end{aligned}$$

where $Z \sim N(0,1)$.

9.4 Exercise 5

1. Consider the joint probability distributions

Probabilities		Values of Y: -1	0	1
Values of X:	-1	0.0	0.1	0.0
	0	0.2	0.3	0.3
	1	0.0	0.0	0.1

,

Probabilities		Values of Y: -1	0	1
Values of X:	-1	0.02	0.10	0.08
	0	0.06	0.30	0.24
	1	0.02	0.10	0.08

.

In each case,

(a) find the marginal probability distributions of X and Y;

(b) find out whether X and Y are independent.

2. You are an investment consultant employed by an investor who intends to invest in the stock market or in a deposit account with a building society. The percentage annual rate of return for the stock market is denoted by the random variable S. For simplicity we assume that this rate of return will be one of four values: -10%, 0%,10% or 20%. The annual rate of interest on the deposit account (denoted R) will be 4%, 6% or 8%. From previous experience, you believe that the joint probability distribution for these variables is:

Probabilities		Values of S: -10	0	10	20
Values of R:	4	0	0	0.1	0.1
	6	0	0.1	0.3	0.1
	8	0.1	0.1	0.1	0

(a) i. Find the marginal probability distributions for R and S. What is the overall probability that the rate of return for the stock market will be positive? What is the probability that the rate of return for the stock market will exceed that from the building society deposit account?

   ii. Calculate the mean and variance using each of these marginal distributions. What does this information imply about the relative merits of the two types of investment?

(b) Calculate the (population) covariance and (population) correlation between R and S. How would you interpret the value of the correlation?

(c) i. One proposal you make to the investor is to split her savings equally between the stock market and the building society account. Find (using any appropriate method) the mean and variance of the random variable

$$A = 0.5R + 0.5S.$$

   ii. Why might the investor prefer the proposed new 50/50 strategy to the simple strategies of investing all her money in the building society or in the stock market?

3. The random variables X and Y have $\mu_X = 10, \sigma_X = 3, \mu_Y = -1, \sigma_Y = 4$.

   (a) Find the mean and standard deviation of $V = X + Y$ when

      i. X and Y are independent;

      ii. $\sigma_{XY} = -8$.

   (b) Find the mean and standard deviation of

$$W = 3X - 2Y + 8$$

   when X and Y are independent.

4. In the manufacture of disposable syringes, the manufacturing process produces cylinders of diameter $X_1 \sim N(20.2, 0.04)$ and plungers of diameter $X_2 \sim N(19.7, 0.0225)$.

   (a) If the components are combined so that X_1 is independent of X_2, what proportion of plungers will not fit?

   (b) Suppose now that the components are not independently combined, but that larger plungers tend to be combined with larger cylinders, leading to $\text{cov}[X_1, X_2] = 0.02$. What proportion of plungers will not fit now?

9.5 Exercises in EXCEL

EXCEL can be used to perform the routine calculations involved in the material of Sections 9 and 10. For example, summation in EXCEL can be used to obtain the marginal probability distributions obtained in Section 9. Similarly, the tedious calculations involved in obtaining the separate means, variances, the covariance and the correlation from a given joint probability distribution can be carried out in EXCEL. Note, however, that there are no functions available within EXCEL specifically designed to compute the expected value and standard deviation for a random variable.

EXCEL can be used to obtain probabilities for a Normal random variable, as discussed in Section 8.10. Thus, having obtained the mean and variance of a linear combination of normal random variables, the NORMDIST function can be used to obtain probabilities for values of this linear combination. This yields more accurate probabilities than those which can be obtained using statistical tables.

Chapter 10

POPULATIONS, SAMPLES & SAMPLING DISTRIBUTIONS

10.1 Experiments and Populations

Section 1 provided some basic definitions and concepts of statistics - **data**, **experiment**, **sampling**, **population**, **sample**, **sample space** and **statistic**. In this section, the links from these ideas to probability distributions and random variables are made explicit.

There are two aspects to the idea of an **experiment**. It is any process which generates data, and, at least in some cases, generates a sample of data, in which case, this is also considered to be sampling from a population. Here the population is defined as the totality of items that we are interested in. On the other hand, the sample space of the experiment lists all the possible outcomes of the experiment. Much effort was devoted in Sections 5 - 7 to establishing the properties of random variables which can be defined on the sample space of the experiment.

From this perspective, then, an experiment generates the values of a random variable, and possibly even several random variables. The values of a random variable are assumed to occur with a known probability for a given experiment, and the collection of these probabilities constitute the probability distribution of the random variable.

Pooling the two aspects, we see that data generated by experiments can be considered both as values of a random variable and as a sample of data from a population. More succinctly, we can argue that values that occur in a population are generated by an experiment. Pursuing this argument one stage further, we can conclude that values occurring in a population can be considered as the values of a random variable.

This is the crucial idea, but it can be extended still further. The relative frequencies with which values occur in the population must equal the probability of these values as values of a random variable. So, we could argue that a **population** is "equivalent" (in this sense) to a random variable. This reasoning permits us to use the language of probability and probability distributions alongside that of populations and population relative frequencies.

10.1.1 Examples

1. The population of January examination results for ES1070 are the values of some random variable.
2. The number of cars passing an observation point on the M60 in a short interval of time is the value of some random variable.
3. Whether or not a household in the UK owns a DVD player is the value of a random variable.

The skill of the statistician lies in deciding which is the appropriate random variable to describe the underlying populations. This is not always easy, and usually one tries to use a well known random variable and probability distribution, at least as a first approximation. Hence the need to discuss Binomial, Geometric, Poisson, Uniform, Exponential and Normal random variables in this course.

10.2 Populations and random variables

The discussion above helps us to see why it is that the **expected value**, $E[X]$, of some random variable X is simultaneously

- the **theoretical** mean
- the **mean** of the probability distribution of X
- the **population** mean.

The same applies to the variance $\text{var}[X]$ of X : it is

- the theoretical variance
- the variance of the probability distribution of X
- the population variance.

Population characteristics like mean and variance are usually called population **parameters,** but they are also characteristics of the probability distribution of X. There are other sorts of parameters that we may be interested in - for example, the population relative frequency π in Example 3 of section

10.1.1. If we define X to be a random variable taking on the value 1 if a household owns a DVD player, and 0 otherwise, the population proportion becomes the parameter of a probability distribution as $\pi = \Pr(X = 1)$.

10.2.1 Objective of Statistics

As far as this course is concerned, the objective of statistics is to learn about population characteristics or parameters. It is important to remember that the values of these parameters are unknown to us, and generally, we will never discover their values exactly. The idea is to get as close to the truth as possible, even though the truth may never be revealed to us. All we can do in practice is to make reasonable judgments based on the evidence (data) and analysis of that data: this process is called **statistical inference.** So,

- the objective of statistics is statistical inference on (unknown) population parameters.

10.3 Samples and Sampling

Put simply, we use a sample of data from a population to draw inferences about the unknown population parameters. However, the argument in Section 10.1 makes it clear that this idea of sampling from a population is equivalent to **sampling from the probability distribution of a random variable.**

It is important to note that the way in which a sample is obtained will influence inference about population parameters. Indeed, badly drawn samples will **bias** such inference.

10.3.1 Example

Suppose that an investigator is interested in the amount of debt held by students when they graduate from a UK university. If the investigator samples only graduating students from the University of Manchester, there can be no presumption that the sample is representative of all graduating UK students.

10.3.2 Sampling from a population

It is easier to start by discussing appropriate sampling methods from a population, and then discuss the equivalence with sampling from probability distributions. Our objective of avoiding biased inferences is generally considered to be met if the sampling procedure satisfies two conditions:

1. Each element of the population has an equal chance of being drawn for inclusion in the sample.

2. Each draw from the population is independent of the preceding and succeeding draws.

A sample meeting these conditions is called a **simple random sample**, although frequently this is abbreviated to **random sample.**

How can these conditions be physically realised? Drawing names from a hat is one possible method, although not very practical for large populations. The electronic machine used to make draws for the National Lottery is apparently considered a fair way of drawing 6 or 7 numbers from 50. The use of computers to draw "pseudo-random" numbers is also a standard method.

There are some technical complications associated with this description of random sampling. One is that with a population of finite size, the conditions can be met only if an item drawn from the population is "replaced" after drawing - this is **sampling with replacement.** So, **sampling without replacement** makes the chance of being drawn from the population different at each draw. However, if we have a "large" population, and a "small" sample size relative to the population size, there is little practical difference between sampling with and sampling without replacement.

These distinctions are ignored in what follows.

10.3.3 Sampling from a probability distribution

It is helpful to see an example of sampling from a population: the example is simple enough to make the link to sampling from a probability distribution transparent.

The population contains 1000 elements, but only three distinct values occur in this population, $0, 1, 2$, with population relative frequencies p_0, p_1, p_2 respectively. We can consider this population as being equivalent to a random variable X taking on the values $0, 1, 2$ with probabilities p_0, p_1, p_2. The probability distribution of X can be represented as the table

Values of X	Probability
0	p_0
1	p_1
2	p_2

.

In this population, 0 occurs $1000p_0$ times, 1 occurs $1000p_1$ times and 2 occurs $1000p_2$ times. If we select an element from the population at random, we don't know in advance which element will be drawn, but every element has the same chance, $\frac{1}{1000}$, of being selected. What is the chance that a 0 value is selected? Presumably this is ratio of the number of $0's$ to the population size:

$$\frac{1000p_0}{1000} = p_0.$$

Exactly the same argument applies to selecting a 1 or a 2, producing selection probabilities p_1 and p_2 respectively.

It is clear that the probability distribution of what might be drawn from the population is that of the random variable X which "describes" this population. So, it is appropriate to define a random variable X_1, say, which describes what might be obtained on the first draw from this population. The possible values of X_1 are the three distinct values in the population, and their probabilities are equal to the probabilities of drawing these values from the population:

- X_1 has the **same** probability distribution as X.

By the principle of (simple) random sampling, what might be drawn at the second draw is **independent** of the first draw. The same values are available to draw, with the same probabilities. What might be drawn at the second drawing is also described by a random variable, X_2, say, which is independent of X_1 but has the same probability distribution as X_1.

- X_2 is independent of X_1, and has the same distribution as X.

We can continue in this way until n drawings have been made, resulting in a **random sample of size** n. Each of the n random variables $X_1, ..., X_n$ describing what might be drawn are independent random variables with the same probability distribution as the random variable X describing the population. To use a jargon phrase, these **sample** random variables are **independently and identically distributed.**

We have to translate this process of sampling from a population to sampling from the probability distribution of X. All we have to do is to say that what one might get in a random sample of size 1 from the probability distribution of X are the values of a random variable X_1 with the same probability distribution as X. For a random sample of size 2, what we might get are the values of a pair of independent random variables X_1, X_2, each having the same probability distribution as X. For a random sample of size n, what we might get are the values of n independent random variables $X_1, ..., X_n$, each having the same probability distribution as X.

Although a particular population was used as an example, one can see that the description of sampling from the corresponding probability distribution yields properties that apply generally. Specifically, they apply even when the random variable used to describe a population is a **continuous** random variable.

To summarise, using the language of sampling from a probability distribution,

- a **random sample of size** n from the probability distribution of a random variable X

- consists of sample random variables $X_1, ..., X_n$
- that are mutually independent
- and have the same probability distribution as X;
- $X_1, ..., X_n$ are **independently and identically distributed** random variables
- $X_1, ..., X_n$ are **i.i.d.**

It is important to note that this discussion relates to what **might** be obtained in a random sample. The sample of **data** consists of the **values** $x_1, ..., x_n$ of the sample random variables $X_1, ..., X_n$.

10.3.4 Examples

1. A random sample of size 3 is drawn from the population above. It consists of three i.i.d random variables X_1, X_2, X_3. Suppose that the values in the sample of data are 0,0,1: that is,

$$x_1 = 0, x_2 = 0, x_3 = 1.$$

2. Pursuing the graduate debt example of Section 10.3.1, suppose that graduate debt is described by a random variable X with a normal distribution, so that X is distributed as $N\left(5000, 1000^2\right)$:

$$X \sim N\left(5000, 1000^2\right).$$

If 10 students are drawn at random from the population of students (i.e. using a random sample), the debt at each drawing also has this distribution. The random sample consists of 10 random variables $X_1, ..., X_{10}$, mutually independent, and each X_i is normally distributed:

$$X_i \sim N\left(5000, 1000^2\right), \quad i = 1, ..., 10.$$

The sample of data is the values $x_1, ..., x_n$, for example,

$$\begin{aligned} x_1 &= 5754.0, \quad x_2 = 6088.0, \quad x_3 = 5997.5, \quad x_4 = 5572.3, \quad x_5 = 4791.9, \\ x_6 &= 4406.9, \quad x_7 = 5366.1, \quad x_8 = 6083.3, \quad x_9 = 6507.9, \quad x_{10} = 4510.7. \end{aligned}$$

3. An alternative version of Example 2. Suppose that $X \sim N\left(\mu, \sigma^2\right)$, with μ and σ^2 unknown: then the random sample of size 10 consists of $X_1, ..., X_{10}$, mutually independent, and

$$X_i \sim N\left(\mu, \sigma^2\right), \quad i = 1, ..., 10.$$

If we suppose that the sample values are as shown above, it is tempting to use this data to make a guess at μ and σ^2.

Chapter 11

STATISTICS & SAMPLING DISTRIBUTIONS

11.1 Statistics

Suppose that a random sample of size n has been obtained from the probability distribution of a random variable X. This gives n sample random variables $X_1, ..., X_n$, and the **sample of data** consists of the values $x_1, ..., x_n$ of these random variables. In Section 2, *Basic Descriptive Statistics*, the sample mean,

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

was considered an appropriate measure of location, and the sample variance

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

an appropriate measure of dispersion for a data set. There is a notation change here, compared with Section 2, a notation change that is **important.**

The sample mean and the sample variance are both examples of a **statistic,** a quantity which is a function of the data. We now consider them as functions of the **values** $x_1, ..., x_n$ of the sample random variables $X_1, ..., X_n$. There are expressions corresponding to $\bar{x}$ and s^2 in terms of these sample random variables $X_1, ..., X_n$:

$$\begin{aligned} \bar{X} &= \frac{1}{n} \sum_{i=1}^{n} X_i, \\ S^2 &= \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 . \end{aligned}$$

These expressions show that $\bar{X}$ and S^2 are **functions** of the sample random variables $X_1, ..., X_n$, and are therefore random variables themselves, having probability distributions, expected values etc.

- The probability distributions of $\bar{X}$ and S^2 are called **sampling distributions** because they depend on a random sampling procedure.

Notice that with this new perspective, the **statistic** $\bar{x}$ is a sample value of the **sample statistic** $\bar{X}$, and the statistic s^2 is a sample value of the **sample statistic** S^2.

- It is important to distinguish a statistic and a sample statistic: one is a numerical value, the other a random variable, with a probability distribution referred to as a **sampling distribution**.

11.2 Sampling Distributions

How are sampling distributions found? The following example shows that it can be quite laborious to find them from first principles.

11.2.1 Example

Suppose that a random sample of size 2 is to be drawn from the probability distribution of the random variable X, where this is given in the table

Values of X	0	1	2	$E[X]$
Probability	0.2	0.3	0.5	1.3

.

The random sample will consist of the independent random variables X_1, X_2, each with this probability distribution. So, for example, the probability of obtaining the sample $x_1 = 0, x_2 = 1$ (for example), is, by independence,

$$\begin{aligned} \Pr(X_1 = 0, X_2 = 1) &= \Pr(X_1 = 0)\Pr(X_2 = 1) \\ &= 0.06. \end{aligned}$$

Here, the sample mean is

$$\bar{X} = \frac{1}{2}(X_1 + X_2).$$

What is its probability distribution?

The strategy is to find out what possible samples can be drawn, what their probability of occurrence is, and the value of $\bar{X}$ implied by that sample.

From this information we can deduce the probability distribution of $\bar{X}$. All of these pieces of information are displayed in the table below:

Samples	Probability	Value of $\bar{X}$
$(0,0)$	0.04	0
$(0,1)$	0.06	0.5
$(0,2)$	0.1	1
$(1,0)$	0.06	0.5
$(1,1)$	0.09	1
$(1,2)$	0.15	1.5
$(2,0)$	0.1	1
$(2,1)$	0.15	1.5
$(2,2)$	0.25	2

We can see what the possible values for $\bar{X}$ are, and the probabilities of the samples which are favourable to each value of $\bar{X}$. This leads to a table displaying the probability distribution of $\bar{X}$:

Value of $\bar{X}$	0	0.5	1	1.5	2	$E\left[\bar{X}\right]$
Probability	0.04	0.12	0.29	0.3	0.25	1.3

.

It is easily checked that the probabilities add up to 1, and that the expected value calculation for $\bar{X}$ is correct.

An important aspect of this example is that the expected value of $\bar{X}$ is equal to the expected value of X, which could here be described as the population mean.

Another Example: the Binomial Distribution

This distribution is described in Section 5.2.4. Suppose that

- a (large) population consists of values 0 or 1,
- 1 indicates (for example) a household in the UK owning a DVD player,
- the (unknown) population relative frequency of $1's$ is π.

A random variable which can describe this population is one which takes on values 0 and 1 with probabilities $1-\pi$ and π respectively. Denote this random variable by X, a *Bernoulli* random variable, discussed in Section 5.2.2.

Imagine that the experiment generating the value of a Bernoulli random variable is repeated n times under identical conditions, in such a way that the potential outcome of one experiment is independent of the other experiments. Then, these repetitions are equivalent to drawing a random sample of size n from this Bernoulli distribution. If the outcome of an experiment is a value 1, call it a "success".

Each experiment generates the value of a Bernoulli random variable X_i, having the same distribution as X. Let the random variable T be the total number of successes,

$$T = \sum_{i=1}^{n} X_i.$$

Section 5.2.4 explains that in this situation, the probability distribution of T is a binomial distribution:

$$\Pr(T = t) = \binom{n}{t} \pi^t (1-\pi)^{n-t}, \quad t = 0, ..., n, \quad 0 < \pi < 1.$$

How does this relate to the previous example?

We can use this to deduce the sampling distribution of the sample mean $\bar{X}$, since it is related very simply to T :

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

We can deduce that if T can take on values $0, 1, 2, ..., n-1, n$ then $\bar{X}$ can take on values

$$0, \frac{1}{n}, \frac{2}{n}, ..., \frac{n-1}{n}, 1$$

and that

$$\Pr\left(\bar{X} = \frac{t}{n}\right) = \Pr(T = t), \qquad t = 0, ..., n.$$

In principle, we could now try to show that the expected value of $\bar{X}$ is equal to the expected value of X, (which is π), the population mean. This will follow from the discussion in the next section.

11.3 Sampling Distribution of $\bar{X}$

Assuming random sampling, we can find the mean and variance of the sampling distribution of $\bar{X}$, without actually knowing what the sampling distribution is. This is a very useful and important result.

Suppose that a random sample of size n is drawn from a population with mean μ and variance σ^2, or equivalently, from the probability distribution of a random variable X with $E[X] = \mu$, $\text{var}[X] = \sigma^2$.

Since

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

$\bar{X}$ is a linear combination of the sample random variables $X_1, ..., X_n$, so that the results in Sections 9.1.2 and 9.2.1 can be used to find $E\left[\bar{X}\right]$ and $\operatorname{var}\left[\bar{X}\right]$. The weights in the linear combination are all the same:

$$a_i = \frac{1}{n},$$

in the notation of those sections.

In turn, from the properties of random sampling, we know that

$$\begin{aligned} E\left[X_i\right] &= E\left[X\right] = \mu, \\ \operatorname{var}\left[X_i\right] &= \operatorname{var}\left[X\right] = \sigma^2. \end{aligned}$$

11.3.1 The mean of the sample mean

This heading is deliberately misleading: what is meant precisely is the expected value or expectation of the sampling distribution of the sample mean.

- In random sampling, the expected value of $\bar{X}$ is equal to the population mean, μ.

Proof: From Section 9.1.2,

$$\begin{aligned} E\left[\bar{X}\right] &= E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} E\left[X_i\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu \\ &= \mu. \end{aligned}$$

11.3.2 The variance of the sample mean

Random sampling makes the sample random variables $X_1, ..., X_n$ independent and therefore uncorrelated. We can then use the slogan of Section 9.2: *the variance of a weighted sum is a weighted sum of variances* to prove that

- in random sampling, the variance of the sample mean is $\dfrac{\sigma^2}{n}$.

Proof:

$$\begin{aligned}
\operatorname{var}\left[\bar{X}\right] &= \operatorname{var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\
&= \sum_{i=1}^{n} \operatorname{var}\left[\frac{1}{n} X_i\right] \\
&= \sum_{i=1}^{n} \left(\frac{1}{n}\right)^2 \operatorname{var}\left[X_i\right] \\
&= \sum_{i=1}^{n} \left(\frac{1}{n}\right)^2 \sigma^2 \\
&= \frac{n\sigma^2}{n^2} \\
&= \frac{\sigma^2}{n}.
\end{aligned}$$

The square root of $\operatorname{var}\left[\bar{X}\right]$ is the **standard deviation** of $\bar{X}$, and this is usually given the specific name of **standard error**. So far, we have identified population parameters with the parameters of the distribution of the corresponding random variable X. We can extend this to cover characteristics or parameters of the probability distributions of sample statistics like $\bar{X}$. So, $\operatorname{var}\left[\bar{X}\right]$ is a **parameter** of the probability distribution of $\bar{X}$, and so is the standard error,

$$\mathrm{SE}\left[\bar{X}\right] = \frac{\sigma}{\sqrt{n}}.$$

11.3.3 Example

In the Example of Section 11.2.1, we found $E\left[X\right] = 1.3$, and it is easy to find that

$$E\left[X^2\right] = (0)^2\,(0.2) + (1)^2\,(0.3) + (2)^2\,(0.5) = 2.3,$$

so that the population variance is

$$\operatorname{var}\left[X\right] = 2.3 - (1.3)^2 = 0.61.$$

We also found that $E\left[\bar{X}\right] = 1.3$, and we can calculate from the probability distribution of $\bar{X}$ that

$$\begin{aligned}
E\left[\bar{X}^2\right] &= (0)^2\,(0.04) + (0.5)^2\,(0.12) + (1)^2\,(0.29) + (1.5)^2\,(0.3) + (2)^2\,(0.25) \\
&= 1.995.
\end{aligned}$$

This gives

$$\operatorname{var}\left[\bar{X}\right] = 1.995 - (1.3)^2 = 0.305$$

which is precisely

$$\frac{\sigma^2}{2} = \frac{0.61}{2},$$

matching the theoretical result exactly, since $n = 2$ here.

Here, the standard error is

$$\mathrm{SE}\left[\bar{X}\right] = \sqrt{0.305} = 0.552.$$

11.3.4 Summary

The results presented above are so important that they need to be stated compactly.

- If a random sample of size n is drawn from a population with mean μ and variance σ^2,
- the expected value of $\bar{X}$ is equal to the population mean, $\mu : E\left[\bar{X}\right] = \mu$,
- the variance of the sample mean is $\dfrac{\sigma^2}{n}$: $\mathrm{var}\left[\bar{X}\right] = \dfrac{\sigma^2}{n}$.

Notice that the variance of $\bar{X}$ declines, relative to the population variance, as the sample size n increases. This is due explicitly to the "averaging" effect contained in the sample mean. Clearly, the same applies to the standard error $\mathrm{SE}\left[\bar{X}\right]$ as well.

11.3.5 The Binomial case

Here, $\bar{X}$ is obtained by random sampling from a Bernoulli distribution, with success probability π - see Section 11.2.1. So, if X has a Bernoulli distribution, the population mean and population variance are

$$\begin{aligned}
E[X] &= (0)(1-\pi) + (1)(\pi) = \pi, \\
E\left[X^2\right] &= (0)^2(1-\pi) + (1)^2(\pi) = \pi, \\
\mathrm{var}[X] &= E\left[X^2\right] - (E[X])^2 = \pi - \pi^2 = \pi(1-\pi).
\end{aligned}$$

Using the general properties from Sections 11.3.1 and 11.3.2, it will follow that

$$E\left[\bar{X}\right] = \pi, \qquad \mathrm{var}\left[\bar{X}\right] = \frac{\pi(1-\pi)}{n}.$$

In turn, we can use the relationship

$$\bar{X} = \frac{T}{n}$$

to deduce that if T has a Binomial distribution,

$$E[T] = n\pi, \qquad \mathrm{var}[T] = n^2 \,\mathrm{var}\left[\bar{X}\right] = n\pi(1-\pi).$$

This confirms the results given in Section 8.6.1.

11.3.6 The Sampling Distribution

It is worth emphasising that the results above have been deduced without knowing the nature of the distribution which has been sampled. Without such information we cannot calculate, for any value x,

$$\Pr\left(\bar{X} \leqslant x\right).$$

In the example of Section 11.2.1, it was easy to find the probability distribution of $\bar{X}$ from first principles. In the next example, in Section 11.2.1, we saw that the nature of the population and its corresponding Bernoulli random variable generated a sampling distribution which was a Binomial probability distribution. So, the sampling distribution of $\bar{X}$ changes as we change the population probability distribution.

The classical example of the sampling distribution of $\bar{X}$ is where the population probability distribution is a **normal distribution.** In Section 9.3, it was stated that a linear combination of normal random variables also has a normal distribution. Since $\bar{X}$ is such a linear combination - see the beginning of Section 11.3, it follows that $\bar{X}$ also has a normal distribution.

This result is **very important**. Almost all of the rest of this course depends on this result.

- If a random sample of size n is drawn from the distribution $X \sim N\left(\mu, \sigma^2\right)$, then

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

  Note that the mean and variance of this distribution is exactly that deduced above without using knowledge of the population probability distribution.

- Example. IQ tests are designed to behave as if they are drawings from a normal distribution with mean 100 and variance 400 : $X \sim N(100, 400)$. Suppose that a random sample of 25 individuals is obtained. Then,

  $$\bar{X} \sim N\left(100, \frac{400}{25}\right), \quad \text{or,} \quad \bar{X} \sim N(100, 16).$$

  We can then calculate, for example,

  $$\begin{aligned} \Pr\left(\bar{X} < 90\right) &= \Pr\left(\frac{\bar{X} - 100}{4} < \frac{90 - 100}{4}\right) \\ &= \Pr(Z < -2.5) \\ &= 0.0062. \end{aligned}$$

11.3.7 Sampling from Non-Normal distributions

In this case, the sampling distribution of $\bar{X}$ will not be normal. However, if we imagine that the sample size n is allowed to increase without bound, so that $n \to \infty$, we can appeal to a famous theorem (more accurately, a collection of theorems) in probability theory called the **Central Limit Theorem.** This states that

- if $\bar{X}$ is obtained from a random sample of size n from a population with mean μ and variance σ^2, then, irrespective of the distribution sampled,

  $$\frac{\bar{X}-\mu}{\mathrm{SE}\left[\bar{X}\right]} = \frac{\bar{X}-\mu}{(\sigma/\sqrt{n})} \to N\left(0,1\right) \quad \text{as } n \to \infty.$$

  That is, the probability distribution of $\dfrac{\bar{X}-\mu}{\mathrm{SE}\left[\bar{X}\right]}$ approaches the standard normal distribution as $n \to \infty$.

We **interpret** this as saying that

$$\frac{\bar{X}-\mu}{\mathrm{SE}\left[\bar{X}\right]} = \frac{\bar{X}-\mu}{(\sigma/\sqrt{n})} \sim N\left(0,1\right), \quad \textbf{approximately}$$

for finite n.

- An alternative is to say that

  $$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \textbf{approximately}$$

  for finite n.

The rate at which the standard normal distribution is approached influences the quality of the approximation. This is expected to improve as n increases, and textbooks usually claim that the approximation is good enough if

$$n \geqslant 20 \quad \text{or} \quad n \geqslant 30.$$

The idea that $\bar{X}$ has an approximate normal distribution as $n \to \infty$ is often described as the **large sample normality** of the sample mean. The textbook claim here is that a “large” sample is at least 20. This is not really reasonable, but is adequate for use in a course like this.

So, in the IQ example above, we can argue that $\Pr\left(\bar{X} < 90\right) = 0.0062$ approximately if in fact IQ’s are not normally distributed, but do have population mean 100 and population variance 400.

Chapter 12

POINT ESTIMATION

It was stated in Section 10.2.1 that the objective of statistics is to learn about population characteristics or parameters. The values of these parameters are unknown to us, and generally, we will never discover their values exactly. The objective can then be restated as making (statistical) inferences on (unknown) population parameters, using data from a random sample from the population. We now know that this is equivalent to drawing a random sample from the probability distribution of a random variable X, producing sample random variables $X_1, ..., X_n$ which are mutually independent and have the same probability distribution as X. We can construct sample statistics like $\bar{X}$, and conceptually find their sampling distributions, and the characteristics or parameters of these sampling distributions.

12.1 Estimates and Estimators

12.1.1 Example

The basic principle of estimation of population parameters is very simple, and is best motivated by an example. Suppose that a random sample of size 3 is drawn from a population with mean μ and variance σ^2, both parameters being unknown. The values in the sample are

$$x_1 = 5, \quad x_2 = 20, \quad x_3 = 11$$

with

$$\bar{x} = \frac{5 + 20 + 11}{3} = 12.$$

Then, the (point) estimate of μ is $\bar{x} = 12$.

Similarly, we can calculate the sample variance s^2 from

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right). \end{aligned}$$

Here,

$$\sum_{i=1}^{n} x_i^2 = 25 + 400 + 121 = 546,$$

so that

$$s^2 = \frac{1}{2}(546 - (3)\,144) = \frac{114}{2} = 57.$$

Then, the (point) estimate of σ^2 is $s^2 = 57$.

12.1.2 Principle of Estimation

The principle appears to be that of estimating a population characteristic by its corresponding sample version. However, there is another important principle here. We are calculating numerical estimates of the population parameters using the sample values $\bar{x}$ and s^2 of what have previously been called **sample statistics** like $\bar{X}$ and S^2. The latter are random variables, and have probability distributions; $\bar{x}$ and s^2 are values of these random variables. Additional terminology is required to make sure that this distinction is preserved when using the language of estimation of population parameters:

- an **estimator** is the **sample statistic**;
- an **estimator** is the random variable which is a function of the sample random variables $X_1, ..., X_n$;
- an **estimate** is the **value** of the **sample statistic**;
- an **estimate** is the **statistic** or number calculated from the sample values $x_1, ..., x_n$.

Another aspect is that

- an **estimator** is a random variable and hence has a probability distribution
- an **estimate** is a **value** of this random variable.

In this course, we have encountered a number of population parameters: mean, variance, proportion, covariance, correlation. The following table displays the parameters, their estimators and their estimates.

Population Parameter	Estimator	Estimate
mean: μ	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
variance: σ^2	$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$
covariance: σ_{XY}	$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)$	$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$
correlation: ρ	$R = \frac{S_{XY}}{S_X S_Y}$	$r = \frac{s_{XY}}{s_X s_Y}$
proportion: π	P	p

Most of the quantities in this table have appeared in the discussion of descriptive statistics in Section 2.3. The (sample) correlation coefficient R is shown explicitly as a function of the sample covariance and the sample variances to emphasise that the estimator of the population correlation coefficient ρ is derived from other estimators. Out of the quantities in this table, we shall be concerned almost exclusively with the behaviour of the sample mean $\bar{X}$. The sample variance S^2 will also appear in a minor role, but none of the other quantities will concern us further.

One other small detail to explain in this table is the absence of an expression for the estimator or estimate of the population proportion. This is because the sample mean is the required quantity, when sampling from a population of zeros and ones or equivalently from the distribution of a Bernoulli random variable - see Section 11.2.1 above. The possible sample values for $\bar{x}$ are

$$0, \frac{1}{n}, \frac{2}{n}, ..., \frac{n-1}{n}, 1:$$

these are precisely the proportions of $1's$ in the sample, and are thus values of the sample proportion.

12.1.3 Point Estimation

The adjective “point” seems to play no role in the discussion. It serves mainly to distinguish the ideas from another concept of estimation called *interval estimation*, which will be discussed shortly. The distinction is simply that a **point estimate** is a single number, whilst an **interval estimate** is an interval of numbers.

12.2 Properties of Estimators

The discussion of the sampling distribution of the sample mean $\bar{X}$ in Section 11.3 can be carried over to the case where $\bar{X}$ is considered as an estimator of μ. More generally, if θ is some population parameter which we estimate by a sample statistic U, then we expect that U will have a sampling distribution. We can use this sampling distribution to obtain information on how good U is as an estimator of θ. After all, there may be a number of possible estimators of θ and it is natural to want to use the estimator that is *best* in some sense, or at least, avoid using estimators that have undesirable properties. This raises the issue of what desirable properties an estimator should possess.

The relevance of the sampling distribution of an estimator for this issue can be motivated in the following way. Different samples of data will generate different numerical values for u - that is, different values for the estimator U. A value of u will only equal the population parameter θ by chance. Because population parameters are generally unknown, we will not know when this happens anyway. But, the sampling distribution of U represents, intuitively, the "chance" of such an occurrence, and we can calculate from it, in principle, the appropriate probabilities.

12.2.1 Unbiased Estimators

Following this rather intuitive argument, if we cannot detect whether the estimate is actually correct, we could resort to demanding that the estimate be correct "on average". Here, the appropriate concept of averaging is that embodied in finding the expected value of U. We can then say that

- if $E[U] = \theta$, then U is an **unbiased estimator** of θ;
- an unbiased estimator is *correct on average;*
- if $E[U] \neq \theta$, then U is a **biased** estimator of θ.
- a biased estimator is *incorrect on average.*

It is clear that *unbiasedness* is a desirable property for an estimator, whilst *bias* is an undesirable property.

So, to show that an estimator is unbiased, we have to find its expected value, and show that this is equal to the population parameter being estimated.

12.2.2 Examples

In Section 11.3.1, we showed that in sampling from a population with mean μ and variance σ^2,

$$E\left[\bar{X}\right] = \mu.$$

So, without any further conditions, the sample mean is unbiased for μ.

It is also true that the sample variance S^2 is unbiased for σ^2 :

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

One can guess from this that the use of the divisor $n-1$ rather than n is important in obtaining this property. Given this result, we can see that

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

so that this alternative definition of sample variance produces a **biased** estimator of σ^2. On the other hand, the systematic underestimation of σ^2 implied by the nature of the bias will disappear as the sample size n increases. The estimator S^2 is used simply because it is unbiased.

It is possible to show, but not in this course, that the sample covariance S_{XY} is unbiased for the population covariance σ_{XY}.

The sample correlation coefficient R is in general **biased** for the population correlation coefficient ρ, because it is a ratio of random variables. This does not seem to prevent its widespread practical use, however.

12.2.3 Biased Sampling Procedures

All of these results are based on drawing a random sample from the population. One would expect that a sampling procedure not based on the principles behind random sampling (see Section 10.3) would automatically generate biased estimators of population parameters. This is an important issue for the design of official sample surveys and market research surveys. The consequences of inappropriate design are illustrated in the following example.

To estimate the mean income of households in the UK, a researcher uses a sample of households selected at random from the telephone directory. The sample mean income will be taken as an estimate of the population mean income. To illustrate the possible consequences of this sample design, suppose that it is known that the population of households with telephones have a mean income of £20000, whilst households without telephones have mean income £10000. Out of the total population, 80% of households have a telephone.

One can show that the overall UK population mean income is

$$(0.8)\,(\pounds 20000) + (0.2)\,(\pounds 10000) = \pounds 18000.$$

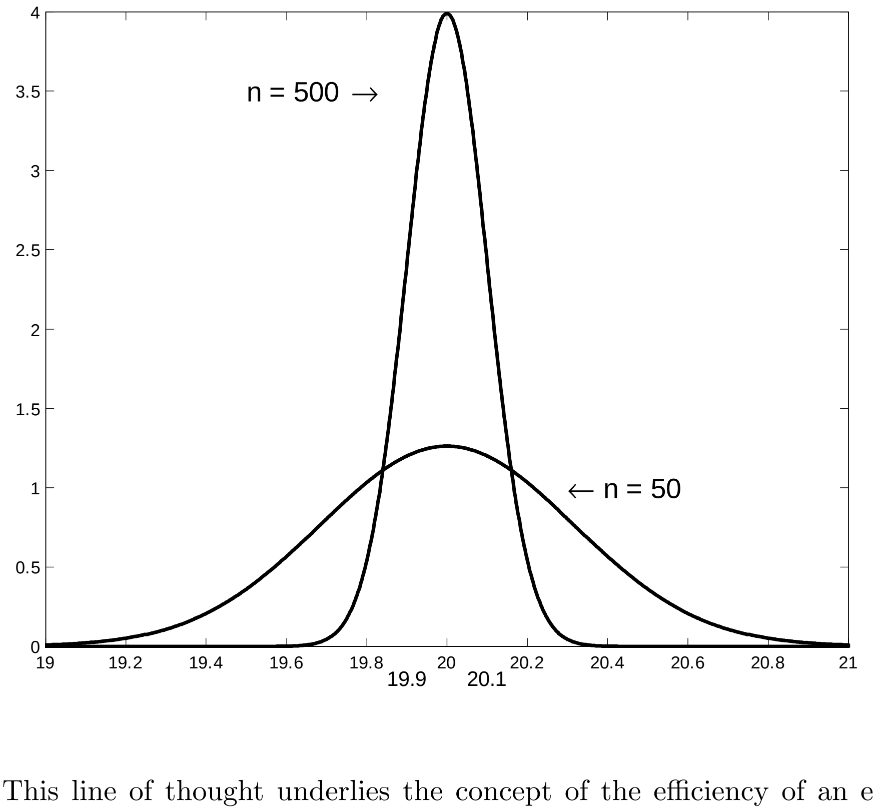
By construction, the sample design here gives an unbiased estimator of the mean of the population of households with a telephone, £20000, but a biased estimator of the overall population mean of £18000.

12.2.4 Efficiency

Unbiasedness is a weak property of an estimator: even a random sample of size 1 from a population with mean μ yields a trivial sample mean which is unbiased. The sample random variable is X_1, and by the random sampling principle, has a distribution which is the same as the population being sampled. Hence

$$E\left[X_1\right] = \mu.$$

It was seen in Section 11.3.2 that var $\left[\bar{X}\right] = \dfrac{\sigma^2}{n}$, which diminishes as n increases. If we suppose temporarily that sampling is from a normal distribution $N\left(\mu, \sigma^2\right)$, so that $\bar{X} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$, we can see graphically the effect of increasing the sample size. Figure 12.2.4 shows that a sample size of $n = 500$ yields a sampling distribution for $\bar{X}$ which is much more concentrated around the population mean μ than for the case $n = 50$. Intuitively, this means that "large" deviations away from μ are much less likely for $n = 500$ than for $n = 50$.



This line of thought underlies the concept of the efficiency of an estimator. Suppose that U and V are unbiased estimators of a population parameter θ. Then,

- the efficiency of U relative to V is defined as

$$\operatorname{eff}(U, V) = \frac{\operatorname{var}[V]}{\operatorname{var}[U]};$$

- if $\operatorname{var}[U] < \operatorname{var}[V]$, U is **more** efficient than V, with

$$\operatorname{eff}(U, V) > 1;$$

- if $\operatorname{var}[U] > \operatorname{var}[V]$, U is **less** efficient than V, with

$$\operatorname{eff}(U, V) < 1;$$

- efficiency does not depend on the nature of the sampling distribution.

12.2.5 Example

For the sake of illustration, suppose that it is known that household income in £'000 is described by a random variable $X \sim N(20, 5)$: one would usually say household income in £'000 is **distributed as** $N(20, 5)$

From earlier results, if $\bar{X}_{50}$ is the sample mean of a sample of size 50, and $\bar{X}_{500}$ the sample mean of a sample of size 500, then

$$\bar{X}_{50} \sim N\left(20, \frac{5}{50}\right), \quad \bar{X}_{500} \sim N\left(20, \frac{5}{500}\right).$$

We can easily calculate that

$$\operatorname{eff}\left(\bar{X}_{500}, \bar{X}_{50}\right) = \frac{(5/50)}{(5/500)} = \frac{500}{50} = 10.$$

There are other ways of representing efficiency, **if**, as here, the nature of the sampling distribution is known.

12.2.6 How close is $\bar{X}$ to μ?

This discussion follows on from the previous example. If we sample from $N\left(\mu, \sigma^2\right)$, so that $\bar{X} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$, we can measure the closeness of $\bar{X}$ to μ by the probability

$$\Pr\left(-\varepsilon < \bar{X} - \mu < \varepsilon\right),$$

where ε is an arbitrary positive number. If we recall that for any number z, the event

$$|z| < \varepsilon$$

is equivalent to

$$(z < \varepsilon), \quad \text{and} \quad (-z < \varepsilon),$$

or

$$(z < \varepsilon), \quad \text{and} \quad (z > -\varepsilon),$$

we see that

$$\begin{aligned} \Pr\left(-\varepsilon < \bar{X} - \mu < \varepsilon\right) &= \Pr\left(\left|\bar{X} - \mu\right| < \varepsilon\right) \\ &= \Pr\left(\frac{\left|\bar{X} - \mu\right|}{\sigma/\sqrt{n}} < \frac{\varepsilon}{\sigma/\sqrt{n}}\right). \end{aligned}$$

This probability can be expressed as the probability that $\bar{X}$ *lies within* $\pm\varepsilon$ *of* μ. The point is that this probability will increase with n, since the factor $\dfrac{\varepsilon}{\sigma/\sqrt{n}}$ is equal to

$$\frac{\varepsilon}{\sigma/\sqrt{n}} = \sqrt{n}\frac{\varepsilon}{\sigma}.$$

For the example in Section 12.2.5, we arbitrarily choose $\varepsilon = 0.1$. The usual normal probability calculations give, for $\mu = 20$ and $\sigma^2 = 5$,

$$\begin{aligned} \Pr\left(-0.1 < \bar{X}_{50} - \mu < 0.1\right) &= \Pr\left(-0.32 < Z < 0.32\right) = 0.252, \\ \Pr\left(-0.1 < \bar{X}_{500} - \mu < 0.1\right) &= \Pr\left(-1 < Z < 1\right) = 0.682. \end{aligned}$$

This comparison is more obvious in Figure 12.2.6. This shows that more efficient estimators have distributions that are much more concentrated around the mean μ. Calculation of the probabilities for a given ε is a way of quantifying this relative concentration.

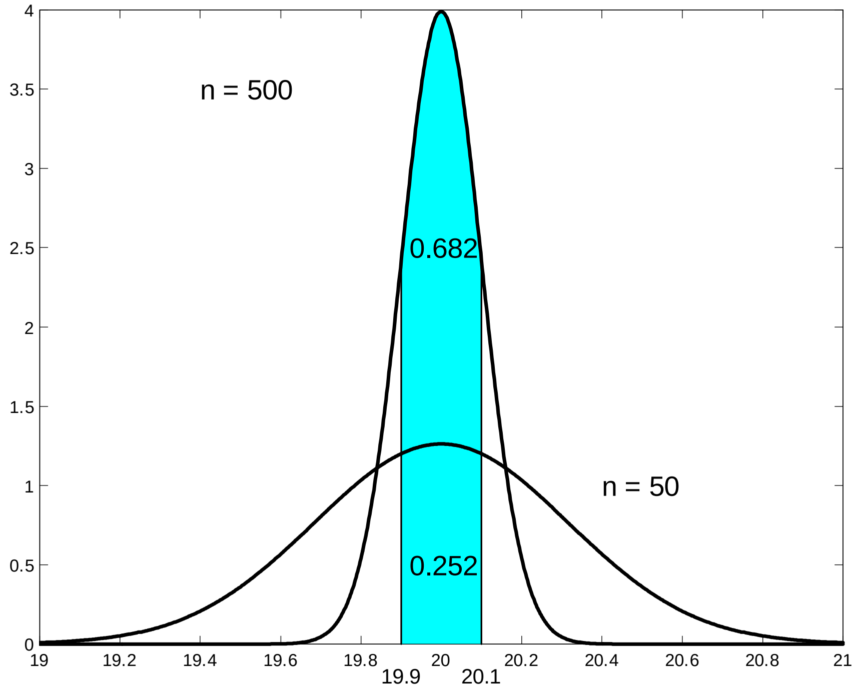
12.3 Estimating the population proportion

Section 11.2.1 showed how a Bernoulli random variable X, taking on the value 0 with probability $1 - \pi$ and the value 1 with probability π, could represent a population of zeros and ones, with 1 representing the fact that a household in the population owned a DVD player, and a 0 that it doesn't. Here, π is the unknown population proportion of successes. In Section 12.1.2, it was shown that the sample mean $\bar{X}$ of a random sample of size n from a Bernoulli distribution takes on values

$$0, \frac{1}{n}, \frac{2}{n}, ..., \frac{n-1}{n}, 1$$

and is the sample proportion P of successes. Recall that in Section 11.2.1, the total number of successes in n Bernoulli trials, T, was shown to have a Binomial distribution, and that $\bar{X}$ was represented as

$$\bar{X} = \frac{T}{n}.$$



As a result, we can also define P in terms of T :

$$P = \frac{T}{n},$$

with sample value

$$p = \frac{t}{n}.$$

Here, t is the total number of successes in a given sample.

So, we know (from Section 11.3.5) that

$$\begin{aligned} E\left[\bar{X}\right] &= E\left[P\right] = \pi, \\ \text{var}\left[\bar{X}\right] &= \text{var}\left[P\right] = \frac{\pi\left(1-\pi\right)}{n}, \end{aligned}$$

so that P is an unbiased estimator of π.

From the results in Section 11.2.1, we know that the sampling distribution of P is related to that of the Binomial random variable T. Suppose that we draw a random sample of size 50 from the distribution of a Bernoulli random variable with success probability $\pi = 0.6$, say. What is

$$\Pr\left(P \leqslant 0.4\right)?$$

Since

$$P = \frac{T}{n},$$

it must be true that

$$P \leqslant 0.4$$

is equivalent to

$$\frac{T}{50} \leqslant 0.4 \quad \text{or} \quad T \leqslant 20.$$

Then,

$$\begin{aligned} \Pr(P \leqslant 0.4) &= \Pr(T \leqslant 20) \\ &= \binom{50}{0}(0.6)^0(0.4)^{50} + ... + \binom{50}{20}(0.6)^{20}(0.4)^{30}. \end{aligned}$$

This is impossible to calculate by hand, but not a problem for a computer.

EXCEL and other statistical packages can handle these calculations with ease: instructions for doing this are given Section 12.4. A simple use of the appropriate function in EXCEL yields, with no further effort,

$$\Pr(T \leqslant 20) = 0.003360382.$$

The traditional response to this practical difficulty in calculation is to argue that since the sample proportion P is a sample mean, and since the Central Limit Theorem of Section 11.3.7 can be applied to the sample mean, we have

$$\frac{P - \pi}{\sqrt{\pi(1-\pi)/n}} \sim N(0,1) \quad \text{approximately}$$

or, equivalently,

$$P \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right) \qquad \text{approximately.}$$

Further, if $P = \dfrac{T}{n}$, it should follow that the Binomial random variable T has an approximate normal distribution,

$$T \sim N(n\pi, n\pi(1-\pi)) \qquad \text{approximately.}$$

12.3.1 An Example

Under the conditions above,

$$P \sim N\left(0.6, \frac{(0.6)(0.4)}{50}\right) \quad \text{approximately.}$$

The variance here is

$$\frac{(0.6)(0.4)}{50} = \frac{0.24}{50} = 0.0048.$$

Applying the usual arguments,

$$\begin{aligned}\Pr(P \leqslant 0.4) &= \Pr\left(\frac{P-0.6}{\sqrt{0.0048}} \leqslant \frac{0.4-0.6}{\sqrt{0.0048}}\right) \\ &= \Pr(Z \leqslant -2.8868) \\ &= 0.001945974.\end{aligned}$$

This latter calculation was also performed in EXCEL. This is not a particularly good approximation to the exact value given in the previous section, and illustrates one of the drawbacks to the sweeping statements that the approximation is good for $n \geqslant 20$, irrespective of any other conditions.

12.3.2 The normal approximation for T

Consider the claim that

$$T \sim N(n\pi, n\pi(1-\pi)) \qquad \text{approximately.}$$

Suppose that Y is a random variable such that

$$Y \sim N(n\pi, n\pi(1-\pi)) \qquad \text{exactly.}$$

The Central Limit Theorem asserts that T has approximately the same probability distribution as Y, if n is large. So, we might try to approximate $\Pr(T = t)$ by $\Pr(Y = t)$. But, Y is a continuous random variable, so that

$$\Pr(Y = y) = 0$$

for any value y - see Section 6.1. Approximating $\Pr(T = t)$ in this way is clearly not a good idea.

The solution is to approximate $\Pr(T = t)$ by an area under the density of Y as

$$\Pr(T = t) \cong \Pr\left(t - \frac{1}{2} \leqslant Y \leqslant t + \frac{1}{2}\right)$$

for $t = 1, ..., n-1$. For $t = 0$ and $t = n$, use

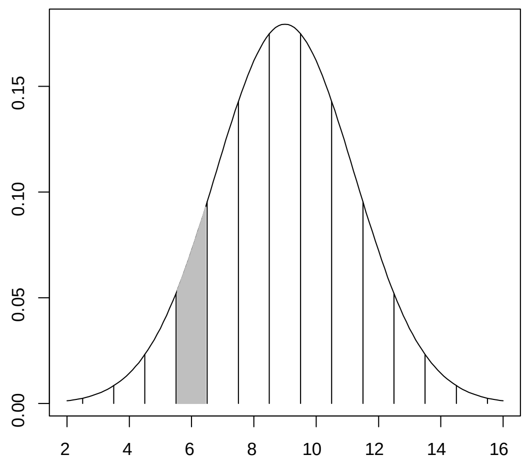
$$\Pr(T = 0) \cong \Pr\left(Y \leqslant \frac{1}{2}\right), \qquad \Pr(T = n) \cong 1 - \Pr\left(Y \leqslant n - \frac{1}{2}\right):$$

see Figure 12.3.2.

This approximation has the advantage that the approximate probabilities for T add up to 1, as they should.

We can also see that (for $t = 1, ..., n-1$)

$$\Pr(T \leqslant t) = \Pr\left(P \leqslant \frac{t}{n}\right) \cong \Pr\left(Y \leqslant t + \frac{1}{2}\right),$$



so that it may provide a better approximation for P probabilities. This is easily adapted to the cases $t = 0$ and $t = n$.

In the example of the previous Section, where $n = 50$ and $\pi = 0.6$,

$$Y \sim N\left(30, 12\right)$$

with

$$t = 20, \qquad \frac{t}{n} = \frac{20}{50} = 0.4.$$

So,

$$\begin{aligned} \Pr\left(P \leqslant 0.4\right) &\cong \Pr\left(Y \leqslant 20 + 0.5\right) \\ &= \Pr\left(Z \leqslant \frac{20.5 - 30}{\sqrt{12}}\right) \\ &= \Pr\left(Z \leqslant -2.74\right) \\ &= 0.00307, \end{aligned}$$

which is closer to the correct value

$$\Pr\left(T \leqslant 20\right) = 0.003360382$$

than the previous calculation.

12.3.3 Another Example

For another example, suppose that it is known that a certain parliamentary constituency contains 45% of Tory voters. A random sample of 20 electors is drawn, yielding 6 Tory voters, or a sample value for the Binomial random variable T of 6. What are the exact and approximate values of

$$\Pr(T=6)?$$

The exact value is

$$\Pr(T=6)=\binom{20}{6}(0.45)^6(0.55)^{14}=0.0745996.$$

Since

$$\begin{aligned} E[T] &= (20)(0.45)=9, \\ \operatorname{var}[T] &= (20)(0.45)(0.55)=4.95, \end{aligned}$$

the distribution of T is approximated by

$$Y \sim N(9, 4.95).$$

Then,

$$\Pr(T=6) \cong \Pr(5.5 \leqslant Y \leqslant 6.5)=0.07273327,$$

which is reasonable, compared with the exact probability. Figure 12.3.2 displays the approximate probability.

Why then is this type of approximation, called the *continuity correction*, not always widely used? Consider the approximation formula

$$\Pr\left(P \leqslant \frac{t}{n}\right) \cong \Pr\left(Z \leqslant \frac{t+0.5-n\pi}{\sqrt{n\pi(1-\pi)}}\right)$$

which underlies the calculation of $\Pr(P \leqslant 0.4)$ in the previous section. One has to convert the P probability into a T probability, apply the approximation with continuity correction, and then standardise - quite difficult to remember in an examination! In addition, the factor

$$\frac{0.5}{\sqrt{n\pi(1-\pi)}}$$

approaches 0 as $n \to \infty$, so that the correction is irrelevent for large enough n.

12.3.4 Exercise 6

1. Suppose that $Y \sim N(6, 2)$, and that $\bar{Y}$ is the sample mean of a (simple) random sample of size n. Find:

   (a) $\Pr(Y > 8)$;

(b) $\Pr\left(\bar{Y} > 8\right)$ when $n = 1$;

(c) $\Pr\left(\bar{Y} > 8\right)$ when $n = 2$;

(d) $\Pr\left(\bar{Y} > 8\right)$ when $n = 5$;

Sketch, on the same axes, the sampling distribution of $\bar{Y}$ for $n = 1, 2, 5$.

2. In a certain population, 60% of all adults own a car. If a simple random sample of 100 adults is taken, what is the probability that at least 70% of the sample will be car owners? (Optional: use EXCEL to find the exact probability.)

3. When set correctly, a machine produces hamburgers of mean weight $100g$ each and standard deviation $5g$ each. The weight of hamburgers is known to be normally distributed. The hamburgers are sold in packets of four.

   (a) What is the sampling distribution of the total weight of hamburgers in a packet? In stating this sampling distribution, state carefully what results you using and any assumptions you have to make.

   (b) A customer claims that packets of hamburgers are underweight. A trading standards officer is sent to investigate. He selects one packet of four hamburgers and finds that the weight of hamburgers in it is $390g$. What is the probability of a packet weighing as little as $390g$ if the machine is set correctly? Do you consider that this finding constitutes evidence that the machine has been set to deliver underweight hamburgers?

4. A discrete random variable, Y, has the following probability distribution:

y	0	1	2
$p(y)$	0.3	0.4	0.3

   (a) What are $E[Y]$ and $y_{\min}$, where $y_{\min}$ is the smallest possible value of Y?

   (b) Simple random samples of two observations are to be drawn with replacement from this population. Write down all possible samples, and the probability of each sample. Use this to obtain the sampling distribution of each of the following statistics:

      i. the sample mean, $\bar{Y}$;

      ii. the minimum of the two observations, M.

   (c) Calculate $E\left[\bar{Y}\right]$ and $E[M]$. State whether each is an unbiased estimator of the corresponding population parameter.

5. A random sample of size three is drawn from the distribution of a Bernoulli random variable X, where

$$\Pr(X = 0) = 0.3, \quad \Pr(X = 1) = 0.7.$$

   (a) Enumerate all the possible samples, and find their probabilities of being drawn. You should have eight possible samples.

   (b) Find the sampling distribution of the random variable T, the total number of ones in each sample.

   (c) Check that the probability distribution of T is the Binomial distribution for $n = 3$ and $\pi = 0.7$, by calculating

$$\Pr(T = t) = \binom{3}{t} (0.7)^t (0.3)^{3-t}$$

   for $t = 0, 1, 2, 3$.

   (d) Find the probability distribution of P, the sample proportion of ones. How is this probability distribution related to that of T?

   (e) Is P an unbiased estimator of $\Pr(X = 1)$?

6. A simple random sample of three observations is taken from a population with mean μ and variance σ^2. The three sample random variables are denoted Y_1, Y_2, Y_3. A sample statistic is being sought to estimate μ. The statistics being considered are

   (a) i. $A_1 = \frac{1}{3}(Y_1 + Y_2 + Y_3)$;

   ii. $A_2 = \frac{1}{2}(Y_1 + Y_2)$;

   iii. $A_3 = \frac{1}{2}(Y_1 + Y_2 + Y_3)$;

   iv. $A_4 = 0.75Y_1 + 0.75Y_2 - 0.5Y_3$.

   (b) Which of these statistics yields an unbiased estimator of μ?

   (c) Of those that are unbiased, which is the most efficient?

   (d) Of those that are unbiased, find the efficiency with respect to A_1.

12.4 Using EXCEL

The calculation of Binomial probabilities in EXCEL uses the statistical function

```
Binomdist(Number_s,Trials,Probability_s,TRUE)
```

for

$$\Pr(T \leqslant \texttt{Number\_s})$$

where T has the Binomial distribution with $n =$ `Trials` trials, and success probability $\pi =$ `Probability_s`. The component `TRUE` indicates the calculation of a cumulative probability. If this is replaced by `FALSE`, the probability

$$\Pr(T = \texttt{Number\_s})$$

is calculated.

Chapter 13

CONFIDENCE INTERVALS

13.1 Point and Interval Estimation

In Section 12.1.3, it was noted that an estimate of a population parameter is a single number. Bearing in mind the fact that in general, the values of population parameters are unknown, it is easy to fall into the trap of treating the estimated value as if it were actually the "true" population value. After all, the estimate is derived from a single sample out of all the possible samples that might be drawn. Different samples will yield different estimates of the population parameter: this is the idea of **sampling variability**. One can see the usefulness of obtaining from a **single sample,** some idea of the range of values of the estimate that might be obtained in different samples. This is the purpose of an *interval estimate.* There is an alternative and more popular name, *confidence interval.* Initially this name is not used because it does not permit the distinction between an **interval estimator** and an **interval estimate.**

13.1.1 Sampling Variability

Consider the simplest case of sampling from a normal distribution, $X \sim N\left(\mu, \sigma^2\right)$ with the intention of estimating μ. The obvious estimator is the sample mean, $\bar{X}$, with sampling distribution $\bar{X} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$. Here, μ is unknown: one issue will be whether the population parameter σ^2 is also unknown. From the general principle that population parameters are unknown, the answer should be "yes", but it will be convenient to assume, for simplicity, that σ^2 is actually known.

The variance of the distribution of $\bar{X}$ measures the dispersion of this distribution, and thus gives some idea of the range of values of $\bar{X}$ that might be obtained in drawing different samples of size n. The question is, though, what is an "extreme" or "untypical" value of $\bar{X}$? The conventional way to define this is by using a multiple of the **standard error** of $\bar{X}$, $\mathrm{SE}\left(\bar{X}\right)$, as a measure of sampling variability. Then, "extreme" values do not belong to the interval

$$\bar{x} \pm k \, \mathrm{SE}\left(\bar{X}\right),$$

with the value k chosen suitably. Then, the factor $\pm k \, \mathrm{SE}\left(\bar{X}\right)$ is the measure of sampling variability around $\bar{x}$. Clearly, the parameter $\mathrm{SE}\left(\bar{X}\right)$ has to be known in order for the measure of 'sampling variability to be computed.

This measure partially reflects the inherent variability in the population, as represented by the population variance σ^2. Usually, it has to be estimated by using the sample variance s^2 instead of σ^2. It is via the use of s^2 that the measure of sampling variability is being calculated from a single sample.

How should k be chosen? A conventional value is $k = 2$: why this might be popular will be seen shortly.

To illustrate the reasoning here, we use an example that also appears later in Section 13.2.3. Suppose that a random sample of size 50 is drawn from the distribution of household incomes, where the latter is supposed to be $N(\mu, 5)$, and that the mean of the sample is $\bar{x} = 18$. If we choose $k = 2$, the measure of sampling variability is

$$\pm 2 \, \mathrm{SE}\left(\bar{X}\right) = \pm (2) \sqrt{\frac{5}{50}} = \pm 0.6325,$$

which could reasonably be said to be rather small.

13.2 Interval Estimators

It is simplest to see how an interval estimator is constructed within the context of estimating the mean μ of a normal distribution $N\left(\mu, \sigma^2\right)$ using the sample mean $\bar{X}$ of a random sample of size n. An interval has to have two endpoints, and an estimator is a random variable, so we seek two random variables C_L and C_U such that the closed interval $[C_L, C_U]$ contains the parameter μ with a pre-specified probability. Rather obviously, this is a **random interval** because its endpoints are random variables.

This interval estimator $[C_L, C_U]$ is also called a **confidence interval**, and the prespecified probability is called the **confidence coefficient** or **confidence level**. The corresponding interval estimate is then the sample value of this random interval: $[c_L, c_U]$. The sample values c_L, c_U are called the **lower** and **upper confidence bounds** or **limits**.

13.2.1 Construction of the Interval Estimator

From the sampling distribution of $\bar{X}$, for any given value k we can find the probability that

$$\Pr\left(-k \leqslant \frac{\bar{X} - \mu}{\mathrm{SE}\left(\bar{X}\right)} \leqslant k\right)$$

as

$$\Pr\left(-k \leqslant \frac{\bar{X}-\mu}{\mathrm{SE}\left(\bar{X}\right)} \leqslant k\right) = \Pr\left(Z \leqslant k\right) - \Pr\left(Z \leqslant -k\right)$$

for $Z \sim N\left(0,1\right)$, just as in Section 12.2.6. So, if we choose $k = 1.96$,

$$\Pr\left(-1.96 \leqslant \frac{\bar{X}-\mu}{\mathrm{SE}\left(\bar{X}\right)} \leqslant 1.96\right) = 0.95.$$

By manipulating the **two** inequalities inside the brackets, but **without** changing the truth content of the inequalities, we can rewrite this so that the centre of the inequalities is the unknown parameter μ. The sequence is to

1. multiply across by $\mathrm{SE}\left(\bar{X}\right)$ to give

$$\Pr\left\{-1.96\,\mathrm{SE}\left(\bar{X}\right) \leqslant \bar{X}-\mu \leqslant 1.96\,\mathrm{SE}\left(\bar{X}\right)\right\} = 0.95;$$

2. move $\bar{X}$ from the centre:

$$\Pr\left\{-\bar{X}-1.96\,\mathrm{SE}\left(\bar{X}\right) \leqslant -\mu \leqslant -\bar{X}+1.96\,\mathrm{SE}\left(\bar{X}\right)\right\} = 0.95;$$

3. multiply through by -1 :

$$\Pr\left\{\bar{X}+1.96\,\mathrm{SE}\left(\bar{X}\right) \geqslant \mu \geqslant \bar{X}-1.96\,\mathrm{SE}\left(\bar{X}\right)\right\} = 0.95;$$

4. tidy up:

$$\Pr\left\{\bar{X}-1.96\,\mathrm{SE}\left(\bar{X}\right) \leqslant \mu \leqslant \bar{X}+1.96\,\mathrm{SE}\left(\bar{X}\right)\right\} = 0.95.$$

Notice that because the manipulations do not change the truth content of the inequalities, the probability of the $\bar{X}$ event defined by the inequalities is not changed.

If we identify the endpoints C_L, C_U of the interval estimator with the endpoints of the interval in part (4),

$$C_L = \bar{X}-1.96\,\mathrm{SE}\left(\bar{X}\right), \qquad C_U = \bar{X}+1.96\,\mathrm{SE}\left(\bar{X}\right),$$

we will have constructed a random interval with the desired properties:

$$\Pr\left(C_L \leqslant \mu \leqslant C_U\right) = 0.95.$$

An alternative expression for this uses membership of the interval:

$$\Pr\left(\mu \in \left[C_L, C_U\right]\right) = 0.95.$$

In both expressions, μ is fixed and unknown. It is the random variables C_L and C_U which supply the chance behaviour, giving the possibility that $\mu \notin [C_L, C_U]$ with some non-zero probability. Indeed, **by construction,** the interval estimator **fails** to contain (more strictly, **cover**) the unknown value μ with probability 0.05 :

$$\Pr\left(\mu \notin [C_L, C_U]\right) = 0.05.$$

To summarise, in the standard jargon:

- a 95% confidence interval for μ is given by the random interval or interval estimator

$$\left[\bar{X} - 1.96\,\mathrm{SE}\left(\bar{X}\right),\ \bar{X} + 1.96\,\mathrm{SE}\left(\bar{X}\right)\right]$$

13.2.2 The Interval Estimate

This interval is defined by the sample values of C_L and C_U : these are the lower and upper confidence bounds

$$c_L = \bar{x} - 1.96\,\mathrm{SE}\left(\bar{X}\right), \quad c_U = \bar{x} + 1.96\,\mathrm{SE}\left(\bar{X}\right).$$

The interval estimate

$$\bar{x} \pm 1.96\,\mathrm{SE}\left(\bar{X}\right)$$

contains the measure of sampling variability discussed in Section 13.1.1. The choice of k as $k = 1.96$ is now determined by the desired confidence level. Why choose the latter to be 0.95 or 95%? This is really a matter of convention.

It is a common abuse of language to call the interval

$$\bar{x} \pm 1.96\,\mathrm{SE}\left(\bar{X}\right)$$

the confidence interval - indeed it is so common that this abuse will be allowed. Strictly, this is an interval estimate, which is now seen to be a combination of a point estimate, $\bar{x}$, of μ, and a measure of sampling variability determined by the constant k, which sets the confidence coefficient or level, in this case, 0.95.

There is a common misinterpretation of a confidence interval, based on this abuse of language, which says that *the confidence interval*

$$\bar{x} \pm 1.96\,\mathrm{SE}\left(\bar{X}\right)$$

contains μ with 95% *confidence.* Why is this a misinterpretation? For the following reasons:

- μ is unknown;

- so this “confidence interval” may or may not contain μ;
- since μ is unknown, we will **never** know which is true;
- the “confidence level” is either 0 or 1, not 0.95.

A better interpretation is based on the relative frequency interpretation of probability. If samples are repeatedly drawn from the population, say $X \sim N\left(\mu, \sigma^2\right)$, and the interval estimates (“confidence interval”) for a 95% confidence level calculated for each sample, about 95% of them will contain μ. However, this doesn’t help when only a single sample is drawn. In any case, this interpretation is only a relative frequency restatement of the principle behind the construction of the interval estimator.

Ultimately, we have to abandon interpretations like this and return to the idea of obtaining from a single sample, a point estimate of a population parameter and a measure of sampling variability. An interval estimate (“confidence interval”) does precisely this, in a specific way.

13.2.3 An Example

As in Section 13.1.1, suppose that a random sample of size 50 is drawn from the distribution of household incomes, where the latter is supposed to be $N(\mu, 5)$. Notice that σ^2 here is supposed to be **known** to equal 5. Suppose that the mean of the sample is $\bar{x} = 18$. Then, the 95% confidence interval for μ is (allowing the abuse of language)

$$\begin{aligned} \bar{x} \pm 1.96\,\mathrm{SE}\left(\bar{X}\right) &= 18 \pm 1.96\sqrt{\frac{5}{50}} \\ &= 18 \pm 0.62 \\ &= [17.38, 18.62]\,. \end{aligned}$$

Here the measure of sampling variability around $\bar{x} = 18$ is ± 0.62. One might reasonably conclude that since this measure of sampling variability is small compared to $\bar{x}$, $\bar{x} = 18$ is a relatively precise estimate of μ. To refer to *precision* here is fair, since we are utilising the variance of the sampling distribution of $\bar{X}$.

13.2.4 Other Confidence Levels

Instead of choosing k so that

$$\Pr\left(-k \leqslant \frac{\bar{X} - \mu}{\mathrm{SE}\left(\bar{X}\right)} \leqslant k\right) = 0.95,$$

we choose it to deliver the desired probability, usually expressed as $1 - \alpha$:

$$\Pr\left(-k \leqslant \frac{\bar{X} - \mu}{\mathrm{SE}\left(\bar{X}\right)} \leqslant k\right) = 1 - \alpha.$$

The reason for the use of $1-\alpha$ is explained later. Since

$$Z = \frac{\bar{X}-\mu}{\mathrm{SE}\left(\bar{X}\right)} \sim N\left(0,1\right),$$

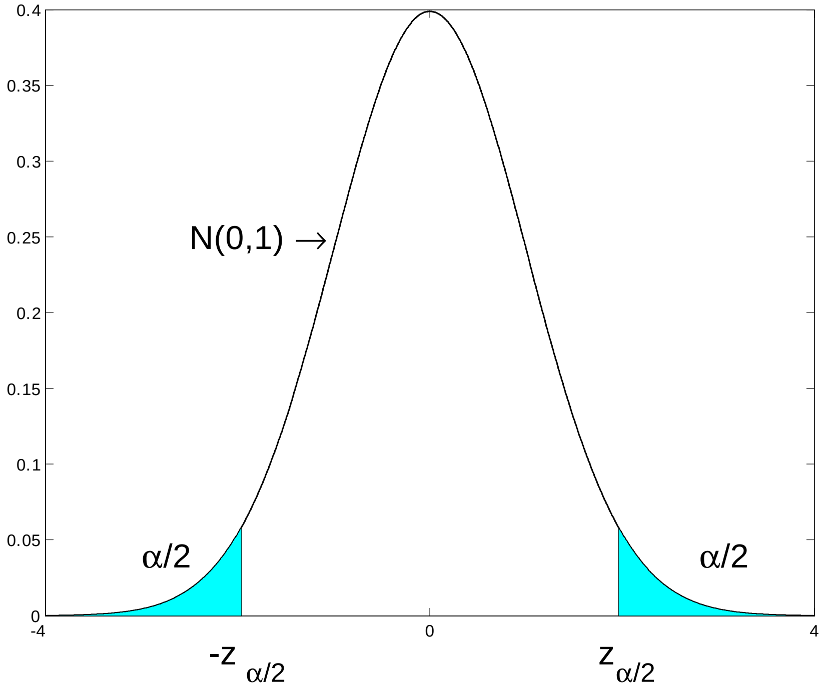
we can find from the tables of the standard normal distribution the value (*percentage point*) $z_{\alpha/2}$ such that

$$\Pr\left(Z > z_{\alpha/2}\right) = \frac{\alpha}{2}.$$

This implies that

$$\Pr\left(-z_{\alpha/2} \leqslant Z \leqslant z_{\alpha/2}\right) = 1-\alpha.$$

This is clear from the familiar picture in Figure 13.2.4.



To find a confidence interval for μ with confidence level $1-\alpha$, or equivalently, $100\left(1-\alpha\right)\%$, we can follow through the derivation in section 13.2.1, replacing 1.96 by $z_{\alpha/2}$ to give

$$\Pr\left\{\bar{X} - z_{\alpha/2}\,\mathrm{SE}\left(\bar{X}\right) \leqslant \mu \leqslant \bar{X} + z_{\alpha/2}\,\mathrm{SE}\left(\bar{X}\right)\right\} = 1-\alpha.$$

That is, the random variables C_L and C_U defining the interval estimator are

$$C_L = \bar{X} - z_{\alpha/2}\,\mathrm{SE}\left(\bar{X}\right), \quad C_U = \bar{X} + z_{\alpha/2}\,\mathrm{SE}\left(\bar{X}\right)$$

The sample value of this confidence interval ("the" confidence interval) is then

$$\bar{x} \pm z_{\alpha/2} \, \mathrm{SE}\left(\bar{X}\right).$$

Using the example of the previous section, we can calculate, for example, a 99% confidence interval for μ. Here,

$$1 - \alpha = 0.99, \quad \alpha = 0.01, \quad \alpha/2 = 0.005,$$

and then from tables,

$$z_{\alpha/2} = 2.58.$$

This gives the lower and upper confidence bounds as

$$\begin{aligned} [c_L, c_U] &= \bar{x} \pm z_{\alpha/2} \, \mathrm{SE}\left(\bar{X}\right) \\ &= 18 \pm 2.58 \sqrt{\frac{5}{50}} \\ &= 18 \pm 0.82 \\ &= [17.18, 18.82]. \end{aligned}$$

Notice that the measure of sampling variability has increased from 0.62 for a 95% confidence interval to 0.82 for a 99% confidence interval. This illustrates the general proposition that the confidence interval gets wider as the confidence coefficient is increased. There has been no change in the precision of estimation here.

Why the use of $1 - \alpha$ in the probability statement underlying the confidence interval? The random variables C_L and C_U are designed here to make

$$\Pr\left(\mu \in [C_L, C_U]\right) = 1 - \alpha$$

and therefore

$$\Pr\left(\mu \notin [C_L, C_U]\right) = \alpha.$$

This probability that μ does **not** belong to the confidence interval turns out to be very important for the topic of *hypothesis testing* which will be discussed in Section 14. As a result, α is considered to be "important", and the confidence coefficient is then stated in terms of α. Again, this is largely due to convention.

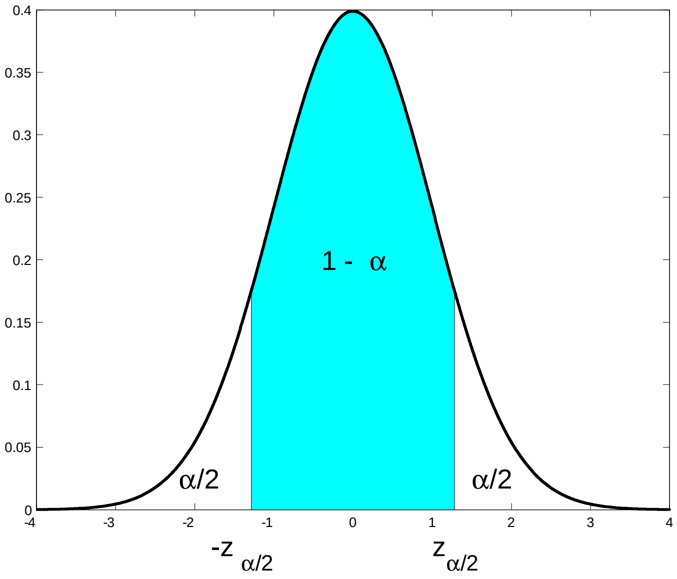
13.2.5 A small table of percentage points

For the $N(0, 1)$ distribution, it is possible in principle to find the appropriate z_α or $z_{\alpha/2}$ from the table of the standard normal distribution in the Appendix. But, this soon becomes tiresome. The table below gives $z_{\alpha/2}$

to four decimal places for a range of common confidence levels. With some care, it can be used for z_α as well - this will be useful for later work.

$1-\alpha$	α	$\alpha/2$	$z_{\alpha/2}$
0.80	0.2	0.10	1.2816
0.90	0.1	0.05	1.6449
0.95	0.05	0.025	1.9600
0.98	0.02	0.01	2.3263
0.99	0.01	0.005	2.5758

Figure 13.2.5 shows the notation graphically.



13.2.6 A small but important point

We have assumed that a random sample of size n has been drawn from a normal population, where $X \sim N\left(\mu, \sigma^2\right)$, and it is clear that an important role in a confidence interval for μ is played by

$$\mathrm{SE}\left(\bar{X}\right) = \sqrt{\frac{\sigma^2}{n}}.$$

This standard error has to be known in order to calculate the interval estimate. But, it has frequently been emphasised that population parameters are in general unknown. So, **assuming** that σ^2 **is known** has to be seen as an unrealistic but simplifying assumption. This assumption allows us to see the principles behind the construction of an interval estimator or confidence interval without other complications. We shall have to investigate the consequences of relaxing this assumption.

13.3 The t distribution

What happens if the parameter σ^2 is unknown, as is likely to be the case usually? Underlying the construction of a 95% confidence interval for μ is the sampling distribution for $\bar{X}, \bar{X} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$ and a true probability statement,

$$\Pr\left(-1.96 \leqslant \frac{\bar{X}-\mu}{\sqrt{\dfrac{\sigma^2}{n}}} \leqslant 1.96\right) = 0.95.$$

This is still true when σ^2 is unknown, but it is of no help, since we cannot construct the confidence interval (i.e. interval estimate)

$$\bar{x} \pm 1.96\sqrt{\frac{\sigma^2}{n}}.$$

In the light of the discussion starting in Section 12 on the role of *estimation* in statistics, there is what seems to be an obvious solution. This is to replace the unknown σ^2 by an estimate, s^2, derived from the same sample as $\bar{x}$. However, one has to be a little careful. First, s^2 is the **estimate** of the (population) variance , and is the sample value of the **estimator** S^2. The probability statement above is based on the fact that

$$Z = \frac{\bar{X}-\mu}{\sqrt{\dfrac{\sigma^2}{n}}} \sim N\left(0,1\right),$$

and in this one has to replace σ^2 by S^2, not by s^2. In effect, we are talking about the estimator and the estimate of SE $\left(\bar{X}\right)$:

- the estimator of SE $\left(\bar{X}\right)$ is $\sqrt{\dfrac{S^2}{n}}$;
- the estimate of SE $\left(\bar{X}\right)$ is $\sqrt{\dfrac{s^2}{n}}$.

Sometimes the estimator of SE $\left(\bar{X}\right)$ is denoted $\widehat{\text{SE}}\left(\bar{X}\right)$, with estimate $\widehat{\text{ese}}\left(\bar{X}\right)$, but these are a bit clumsy to use in general.

13.4 Using $\widehat{\text{SE}}\left(\bar{X}\right)$

So, instead of using

$$Z = \frac{\bar{X}-\mu}{\text{SE}\left(\bar{X}\right)} \sim N\left(0,1\right),$$

we should use

$$T = \frac{\bar{X} - \mu}{\widehat{\mathrm{SE}}\left(\bar{X}\right)}.$$

This seems like a simple solution, but unfortunately it is not an innocuous solution. This is because the distribution of the random variable T combines two sources of randomness, $\bar{X}$ and S^2. As a result, the distribution of T is **NOT** $N(0,1)$.

The distribution of T was discovered by a statistician called W.S. Gossett who worked at the Guinness Brewery in Dublin, and wrote under the pen name 'Student'. The distribution of T is called *Student's t distribution*, or more commonly, just the *t distribution.* This distribution depends on a parameter, just like other distributions: here, the parameter is called the *degrees of freedom.* Before discussing the properties of this distribution, we summarise the distribution statement:

- in random sampling from $N\left(\mu, \sigma^2\right)$,

$$T = \frac{\bar{X} - \mu}{\widehat{\mathrm{SE}}\left(\bar{X}\right)} \sim t_{n-1},$$

- the t distribution with $n-1$ degrees of freedom.

The presence of $n-1$ degrees of freedom can be explained in a number of ways. One explanation is based on the expression for the estimator S^2 :

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

Here, the divisor $n-1$ in this expression leads to the degrees of freedom for T. This is actually the main justification for using a divisor $n-1$ in a sample variance S^2 rather than n, although using $n-1$ also leads to an unbiased estimator.

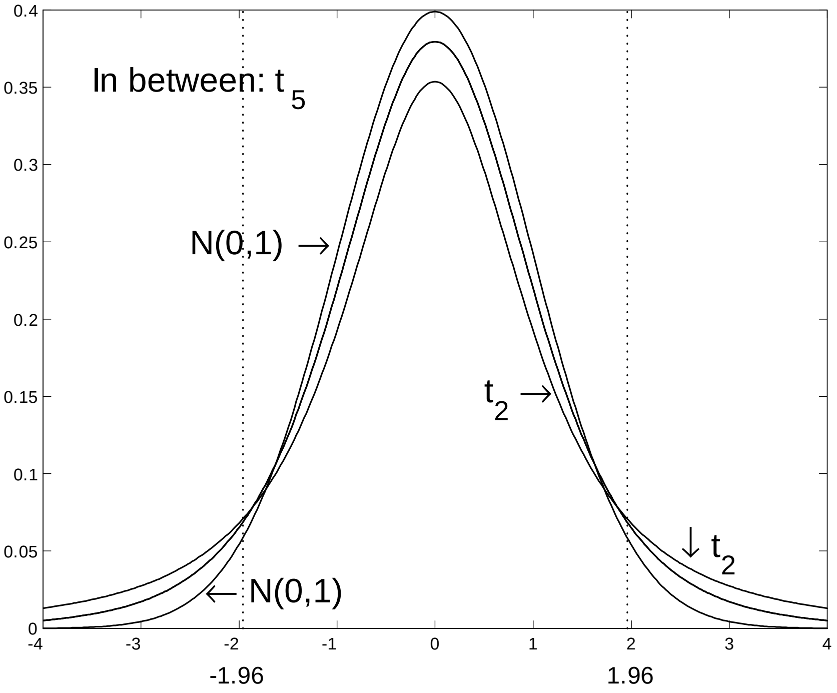
13.4.1 Properties of the t distribution

In general, the parameter ν of the t distribution is a positive real number, although in most applications, it is an integer, as here:

$$\nu = n - 1.$$

Unlike many (population) parameters, this one has a known value once the sample size is known. Figure 13.4.1 shows a plot of the $N(0,1)$ distribution, a t distribution with $\nu = 2$ and a t distribution with $\nu = 5$ degrees of freedom.

It can be seen from Figure 13.4.1 that the t distribution is



- symmetric about zero, like the $N\left(0,1\right)$ distribution
- more dispersed than $N\left(0,1\right)$;
- as ν increases, approaches the $N\left(0,1\right)$ distribution.

One can show that if

$$T \sim t_\nu,$$

then

$$E\left[T\right] = 0, \qquad \mathrm{var}\left[T\right] = \frac{\nu}{\nu - 2}.$$

So, the variance is only defined for $\nu \geqslant 2$, and

$$\mathrm{var}\left[T\right] \geqslant 1,$$

which explains the extra dispersion of the t distribution relative to $N\left(0,1\right)$.

13.4.2 Comparing the t and $N\left(0,1\right)$ distributions

One way of doing this is to compare some "typical" probabilities. The difficulty with this is that one cannot produce a table of t distribution probabilities

$$\Pr\left(T \leqslant t\right) \quad \text{for} \quad T \sim t_\nu$$

to compare with those for

$$\Pr(Z \leqslant z) \quad \text{for} \quad Z \sim N(0,1):$$

there would have to be a table for each value of ν.

The alternative is to use a program like EXCEL. The implication is that one has to understand the appropriate EXCEL commands. One soon discovers that EXCEL works differently for the $N(0,1)$ distribution and the t_ν distributions. First,

$$\Pr(Z \leqslant 1.96) = \texttt{normsdist(1.96)}.$$

Next, the EXCEL commands for computing t distribution probabilities are

$$\texttt{tdist(1.96,df,1)} \quad \text{or} \quad \texttt{tdist(1.96,df,2)}.$$

However, the first of these is

$$\Pr(T > 1.96) = \texttt{tdist(1.96,df,1)},$$

whilst the second is

$$\begin{aligned}\Pr(|T| > 1.96) &= \Pr(T < -1.96) + \Pr(T > 1.96) \\ &= \texttt{tdist}(1.96, \texttt{df}, 2).\end{aligned}$$

In effect, EXCEL gives "lower tail" probabilities for $N(0,1)$, but upper tail probabilities for t_ν, which is a little confusing.

There is another point to bear in mind when using EXCEL, and which becomes relevant later in the course. In trying to compute

$$\Pr(T > t) = \texttt{tdist}(\texttt{t}, \texttt{df}, 1),$$

EXCEL only allows **positive** values of t. So, to calculate, for example,

$$\Pr(T > -1.96),$$

an appeal to symmetry is required first to give

$$\Pr(T > -1.96) = \Pr(T \leqslant 1.96) = 1 - \Pr(T > 1.96),$$

and the latter probability is obtained from EXCEL.

The following table gives some numerical values:

	$\Pr(T \leqslant 1.96)$	$\Pr(T > 1.96)$	$\Pr(\lvert T\rvert > 1.96)$
t_2	0.9055	0.0945	0.1891
t_4	0.9464	0.0536	0.1073
t_{40}	0.9715	0.0285	0.0570
t_{100}	0.9736	0.0264	0.0528
$N(0,1)$	0.9750	0.0250	0.0500

One can see that the tail probabilities for the t_ν distribution actually approach those of the $N\left(0,1\right)$ distribution as ν increases, although the rate of convergence is quite slow in fact. Conventionally, one treats $N\left(0,1\right)$ as if it were t_∞, as in the tables in the Appendix to this book.

An alternative comparison is in terms of *percentage points* - values t and z such that, for example,

$$\begin{aligned}\Pr\left(T \leqslant t\right) &= 0.975 \quad \text{for} \quad T \sim t_\nu, \\ \Pr\left(Z \leqslant z\right) &= 0.975 \quad \text{for} \quad Z \sim N\left(0,1\right).\end{aligned}$$

More generally, z_α is the percentage point such that

$$\Pr\left(Z > z_\alpha\right) = \alpha,$$

and, for the t distribution, $t_{\nu,\alpha}$ is the percentage point such that

$$\Pr\left(T > t_{\nu,\alpha}\right) = \alpha.$$

The Appendix to this book has a table giving values of $t_{\nu,\alpha}$ for various combinations of ν and α such that

$$\Pr\left(T \leqslant t_{\nu,\alpha}\right) = 1 - \alpha.$$

Other texts have similar tables, but may employ (for example)

$$\Pr\left(T > t_{\nu,\alpha}\right) = \alpha.$$

As usual, EXCEL is different: its function

$$\texttt{tinv}\left(\alpha, \texttt{df}\right)$$

gives $t_{\nu,\alpha}$ such that

$$\Pr\left(\left|T\right| > t_{\nu,\alpha}\right) = \alpha.$$

So, EXCEL always defines its percentage point by the probability in **both** tails.

The table below shows some of these values:

	Appendix	*Other Texts*	`tinv(0.05,df)`
	$\Pr\left(T \leqslant t_{\nu,0.025}\right)$	$\Pr\left(T > t_{\nu,0.025}\right)$	$\Pr\left(\left|T\right| > t_{\nu,0.025}\right)$
t_2	4.303	4.3	4.303
t_5	2.571	2.57	2.571
t_{40}	2.021	2.02	2.021
t_{100}	1.984	1.98	1.984
$N\left(0,1\right)$	1.96	1.96	1.96

One can see the same sort of effects: the value that puts 2.5% in the upper tail of a t_ν distribution approaches that for the $N\left(0,1\right)$ distribution. There is a conventional textbook presumption that for ν sufficiently large, one can use the $N\left(0,1\right)$ percentage points as good enough practical approximations to those from the t_ν distribution. The figure $\nu = 40$ is often mentioned for this purpose.

13.4.3 Confidence intervals using the t distribution

Suppose that a $100\,(1-\alpha)\,\%$ interval estimator or confidence interval is wanted for the mean of a normal distribution, $N\left(\mu,\sigma^2\right)$, where σ^2 is unknown. A random sample of size n will be drawn from this distribution. In Section 13.2.1, the facts that

$$\begin{aligned} \bar{X} &\sim N\left(\mu,\frac{\sigma^2}{n}\right), \\ Z &= \frac{\bar{X}-\mu}{\sqrt{\dfrac{\sigma^2}{n}}} \sim N\left(0,1\right) \end{aligned}$$

were used to derive the interval estimator from the probability statement

$$\Pr\left(-z_{\alpha/2} \leqslant Z \leqslant z_{\alpha/2}\right) = 1-\alpha,$$

where

$$\Pr\left(Z > z_{\alpha/2}\right) = \alpha/2.$$

We cannot use this argument when σ^2 is unknown. Instead, σ^2 is replaced by its estimator S^2, and the random variable

$$T = \frac{\bar{X}-\mu}{\sqrt{\dfrac{S^2}{n}}} \sim t_{n-1}$$

used rather than Z. The replacement probability statement is

$$\Pr\left(-t_{n-1,\alpha/2} \leqslant T \leqslant t_{n-1,\alpha/2}\right) = 1-\alpha,$$

in the form

$$\Pr\left(-t_{n-1,\alpha/2} \leqslant \frac{\bar{X}-\mu}{S/\sqrt{n}} \leqslant t_{n-1,\alpha/2}\right) = 1-\alpha.$$

Exactly the same sequence of manipulations as described in Section 13.2.1 is used to generate the probability statement which defines the endpoints of the interval estimator as

$$\Pr\left(\bar{X} - t_{n-1,\alpha/2}\sqrt{\frac{S^2}{n}} \leqslant \mu \leqslant \bar{X} + t_{n-1,\alpha/2}\sqrt{\frac{S^2}{n}}\right) = 1-\alpha.$$

That is, the interval estimator or confidence interval is

$$\begin{aligned} \left[C_L, C_U\right] &= \left[\bar{X} - t_{n-1,\alpha/2}\sqrt{\frac{S^2}{n}}, \bar{X} + t_{n-1,\alpha/2}\sqrt{\frac{S^2}{n}}\right] \\ &= \bar{X} \pm t_{n-1,\alpha/2}\sqrt{\frac{S^2}{n}}. \end{aligned}$$

The sample value of this confidence interval ("the" confidence interval) is then

$$\bar{x} \pm t_{n-1,\alpha/2}\sqrt{\frac{s^2}{n}} :$$

notice the use of the sample value s^2 of S^2 in this expression. This can be usefully compared with the corresponding confidence interval for the case where σ^2 is known:

$$\bar{x} \pm z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}.$$

Two things are different in the t based confidence interval: the use of $t_{n-1,\alpha/2}$ rather than $z_{\alpha/2}$, and the use of s^2 rather than σ^2.

13.4.4 Example

This is the same as that from Section 13.2.3, but now assuming that the population variance σ^2 is unknown. Household income in £ '000 is $X \sim N\left(\mu, \sigma^2\right)$, where **both** μ and σ^2 are unknown. A random sample of size $n = 5$ (previously 50) yields $\bar{x} = 18$ (as before) and $s^2 = 4.5$. Here,

$$T = \frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}} \sim t_4.$$

For a 95% confidence interval for μ, we need the percentage point $t_{4,0.025}$ such that

$$\Pr\left(T \leqslant t_{4,0.025}\right) = 0.975.$$

From the tables in the Appendix this is found to be

$$t_{4,0.025} = 2.776.$$

The confidence interval for μ is then

$$\begin{aligned} \bar{x} \pm t_{n-1,\alpha/2}\sqrt{\frac{s^2}{n}} &= 18 \pm (2.776)\sqrt{\frac{4.5}{5}} \\ &= 18 \pm 2.634 \\ &= [15.366, 20.634]. \end{aligned}$$

For comparison with the original example, if we had used $\sigma^2 = 5$ with a sample of size 5, the resulting normal-based confidence interval would be

$$\begin{aligned} \bar{x} \pm z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} &= 18 \pm (1.96)\sqrt{\frac{5}{5}} \\ &= 18 \pm 1.96 \\ &= [16.04, 19.96]. \end{aligned}$$

This normal-based confidence interval is narrower than the t based one: this is the consequence of the extra dispersion of the t distribution compared with the $N(0,1)$ distribution.

13.5 Realationships between Normal, t and χ^2 distributions

As this point we offer a digression an report some well known properties that link normal, Student t and χ^2 distributions. Some of the following results have been discussed above, butball are included for completeness:

1. Let $X \sim N(\mu, \sigma^2)$; i.e., a normally distributed random variable with mean μ and variance σ^2. Then $Z = (X - \mu)/\sigma \sim N(0,1)$, standard normal, and $W = Z^2 \sim \chi_1^2$, chi-squared with 1 degree of freedom. Generally, χ_v^2 denotes a chi-squared distribution with v degrees of freedom.

2. Let X_i, $i = 1, \ldots, n$, be *iid* (independently and identically distributed) $N\left(\mu, \sigma^2\right)$ variates, then

$$\sum_{i=1}^{n} Z_i^2 \sim \chi_n^2,$$

where $Z_i = (X_i - \mu)/\sigma$, and

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \sim \chi_{n-1}^2,$$

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

Furthermore,

$$\sqrt{n}\left(\bar{X} - \mu\right)/\sigma \sim N(0,1)$$

and

$\sqrt{n}\left(\bar{X} - \mu\right)/s \sim t_{n-1}$, *Student-t distribution* with $n-1$ degrees of freedom,

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is distributed *independently* of $\bar{X}$.

3. Let $Z \sim N(0,1)$ independently of $Y \sim \chi_v^2$. Then, $S = \frac{Z}{\sqrt{Y/v}} \sim t_v$.

4. Let $W \sim \chi_m^2$ independently of $V \sim \chi_p^2$, then $U = W + V \sim \chi_{m+p}^2$ and $R = \frac{W/m}{V/p} \sim F_{m,p}$; i.e., R has an *F-distribution* with m and p degrees of freedom. Hence,

   (a) $R^{-1} \sim F_{p,m}$; and,

   (b) using previous results, if $S \sim t_q$ then $S^2 \sim F_{1,q}$.

13.6 Large Sample Confidence Intervals

The discussion so far has been based on the idea that we are sampling from a normal distribution, $N\left(\mu, \sigma^2\right)$, in which both μ and σ^2 may have to be estimated. In effect, since we have no way of knowing for sure, we **assume** that we are sampling from a normal distribution. Assumptions need not be true: what can be done if the **assumption of normality** is false?

13.6.1 Impact of Central Limit Theorem

The usual sampling distribution of the sample mean $\bar{X}$ also assumes sampling from a normal distribution. In Section 11.3.7, the effect of sampling from a non-normal distribution was discussed. Provided that one draws a random sample from a population with mean μ and variance σ^2, the Central Limit Theorem assures us that

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \qquad \text{approximately,}$$

or, equivalently,

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad \text{approximately.}$$

There is the presumption that the quality of the approximation improves as $n \to \infty$, that is, as the sample size increases. The larger the n, the better.

If σ^2 is known, then we are **approximately** back in the context of Section 13.2.1. That is, we proceed as if σ^2 is known, and simply qualify the confidence level as an **approximate** confidence level.

If σ^2 is unknown, we can use S^2 to estimate σ^2. However, unless we sample from a normal distribution, it will **not** be true that

$$T = \frac{\bar{X} - \mu}{\sqrt{\dfrac{S^2}{n}}} \sim t_{n-1}.$$

Rather,

$$T = \frac{\bar{X} - \mu}{\sqrt{\dfrac{S^2}{n}}} \sim N(0, 1) \qquad \textbf{approximately}.$$

That is, replacing σ^2 by an estimator still allows the **large sample normal approximation** to hold. Only an intuitive justification for this can be given here. This is simply that as $n \to \infty$,

$$S^2 \to \sigma^2 :$$

S^2 gets so close to σ^2 that its influence on the distribution of T disappears.

13.6.2 The large sample confidence interval for μ

From the initial probability statement

$$\Pr\left(-z_{\alpha/2} \leqslant \frac{\bar{X}-\mu}{\sqrt{\dfrac{S^2}{n}}} \leqslant z_{\alpha/2}\right) = 1-\alpha \qquad \text{approximately,}$$

we derive, as in Section 13.2.1, the interval estimator or confidence interval

$$[C_L, C_U] = \bar{X} \pm z_{\alpha/2}\sqrt{\frac{S^2}{n}}$$

having **approximate** confidence level $100\,(1-\alpha)\,\%$. The sample value of this confidence interval is

$$\bar{x} \pm z_{\alpha/2}\sqrt{\frac{s^2}{n}}.$$

For an example, consider that of Section 13.2.3, but now not assuming that sampling takes place from a normal distribution. We assume that the sample information is $n = 50$, $\bar{x} = 18$ and $s^2 = 4.5$. Then, an approximate 95% confidence interval is

$$\begin{aligned} \bar{x} \pm z_{\alpha/2}\sqrt{\frac{s^2}{n}} &= 18 \pm 1.96\sqrt{\frac{4.5}{50}} \\ &= 18 \pm 0.588 \\ &= [17.412, 18.588]\,. \end{aligned}$$

In general, how this compares with the exact confidence interval based on knowledge of σ^2 depends on how good s^2 is as an estimate of σ^2. Nothing can be said about this usually.

13.6.3 Confidence intervals for population proportions

In Section 12.3, estimation of a population proportion π was discussed. To revise this, a random sample is obtained from the distribution of a Bernoulli random variable, a random variable X taking on values 0 and 1 with probabilities $1-\pi$ and π respectively. The sample mean $\bar{X}$ here is the random variable representing the sample proportion of $1's$, and so is usually denoted P, “the” sample proportion. It was shown in Section 12.3 that the sampling distribution of P is related to a Binomial distribution, and that the Central Limit Theorem can be used to provide an approximate normal sampling distribution.

Since

$$E\,[P] = \pi, \qquad \operatorname{var}[P] = \frac{\pi\,(1-\pi)}{n},$$

$$\frac{P-\pi}{\sqrt{\operatorname{var}[P]}} \sim N(0,1) \qquad \text{approximately.}$$

So, one could hope to use this to provide an approximate confidence interval for π. There is a minor complication here in that $\operatorname{var}[P]$ depends on the unknown parameter π, but, there is an obvious estimator (P) available which could used to provide an estimator of $\operatorname{var}[P]$.

Following the previous reasoning, we argue that

$$\frac{P-\pi}{\sqrt{\dfrac{P(1-P)}{n}}} \sim N(0,1) \qquad \text{approximately.}$$

By analogy with the case of the population mean, an approximate $100(1-\alpha)\%$ confidence interval for π is

$$[C_L, C_U] = P \pm z_{\alpha/2}\sqrt{\frac{P(1-P)}{n}},$$

with sample value

$$p \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}.$$

13.6.4 Example

A random sample of 300 households is obtained, with 28% of the sample owning a DVD player. An approximate 95% confidence interval for the population proportion of households owning a DVD player is

$$\begin{aligned} p \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} &= 0.28 \pm (1.96)\sqrt{\frac{0.28(1-0.28)}{300}} \\ &= 0.28 \pm 0.0508 \\ &= [0.229, 0.331]. \end{aligned}$$

For such an apparently large sample size, this is quite a wide confidence interval. Better precision of estimation would require a larger sample size.

13.7 Exercise 7

1. You are interested in the mean duration of a spell of unemployment for currently unemployed women in a particular city. It is known that the unemployment duration of women is normally distributed with variance 129.6. The units of measurement for the variance are therefore *months squared*. You draw a random sample of 20 unemployed women, and they have an average unemployment duration of 14.7 months. Obtain a 98% confidence interval for the population mean unemployment duration for women.

2. A simple random sample of 15 pupils attending a certain school is found to have an average IQ of 107.3 with a sample variance of 32.5.

   (a) Calculate a 95% confidence interval for the unknown population mean IQ, stating any assumptions you need to make. Interpret this interval.

   (b) Explain whether you would be happy with a parent's claim that the average IQ at the school is 113.

3. An internet service provider is investigating the length of time its subscribers are connected to its site, at any one visit. Having no prior information about this, it obtains three random samples of these times, measured in minutes. The first has sample size 25, the second sample size 100 and the third 250. The sample information is given in the table below:

	n	$\bar{x}$	s^2
Sample 1	25	9.8607	2.1320
Sample 2	100	9.8270	2.1643
Sample 3	250	9.9778	2.0025

   (a) Calculate a 95% confidence interval for each sample: state any assumptions you make.

   (b) Do the confidence intervals get narrower as the sample size increases? Why would you expect this?

4. In an opinion poll based on 100 interviews, 34 people say they are not satisfied with the level of local Council services. Find a 99% confidence interval for the true proportion of people who are not satisfied with local Council services.

5. Explain the difference between

   (a) an interval estimator and an interval estimate;

   (b) an interval estimate and a confidence interval.

6. Which of these interpretations of a 95% confidence interval $[c_L, c_U]$ for a population mean are valid, and why?

   (a) μ lies in the interval $[c_L, c_U]$ with probability 0.95;

   (b) in repeated sampling, approximately 95% of confidence intervals will contain μ;

   (c) μ lies in the interval $[C_L, C_U]$ with probability 0.95;

   (d) the confidence interval $[c_L, c_U]$ contains a point estimate of μ and an allowance for sampling variability;

(e) $[c_L, c_U]$ displays the likely range of values of μ.

(f) $[c_L, c_U]$ shows how precise the estimator of μ is expected to be.

13.8 Using EXCEL for the t distribution

Some of the information in Section 13.4.2 is deliberately repeated here.

The table for percentage points of the t distribution in the Appendix refers only to some specific probabilities and degrees of freedom. EXCEL has no such restrictions, either in calculating probabilities or percentage points.

Use the **Paste Function** and select **Statistical** from the **Function Category** box.

- Select `tdist` in the **Function Name** box to obtain probabilities of either the one-tailed form $\Pr(T > t) = \alpha$ or of the two-tailed form $\Pr(|T| > t) = \alpha$. In the dialogue box you supply the value t (which must be nonnegative), the degrees of freedom ν, and also 1 for a one tailed probability *or* 2 for a two tailed probability. In each case, the probability α is returned. Symmetry of the t distribution allows probabilities for $t < 0$ to be obtained.

- Select `tinv` in the **Function Name** to obtain the value $t_{\nu,\alpha/2} \geqslant 0$ such that
$$\Pr\left(T > t_{\nu,\alpha/2}\right) = \alpha/2$$
or
$$\Pr\left(|T| > t_{\nu,\alpha/2}\right) = \alpha.$$
In the Dialogue Box you supply the probability (i.e. the value of α) and the degrees of freedom ν. Note that the function assumes a two-tailed form, so that the probability must be doubled if a one tailed probability is required. Once again, symmetry allows values $t_{\nu,\alpha/2}$ to be obtained.

Chapter 14

HYPOYHESIS TESTING

14.1 Introduction

Despite the fact that population parameters are usually unknown, we now know how to estimate population means, variances and proportions. The population mean is the most important in this course, and we have seen how to obtain both point and interval estimators and estimates of this parameter, whether or not sampling takes place from a normal distribution. It is very common for an investigator to have some sort of preconception or expectation (in the ordinary sense of the word) about the value of a parameter, say the mean, μ. Statisticians usually talk about having a *hypothesis* about the value of the parameter.

We have already seen a number of examples of this. In Section 11.3.6, there is an example which asserts that IQ tests are designed to produce test results which are distributed as $N(100, 400)$. In effect, this is equivalent to saying that IQ test results are drawings from $N(\mu, 400)$, and it is believed that $\mu = 100$. When applied to a specific population of individuals, this hypothesis may or may not be true, and an investigator may want to decide whether sample evidence is, in some sense to be discussed further, compatible with, or in favour of, this hypothesis.

Another example concerns the distribution of household incomes (in £ '000) used in Section 12.2.5, which is supposed be distributed as $N(20, 5)$. Here, household incomes are drawings from $N(\mu, 5)$, with the hypothesis that $\mu = 20$. Again, this may or may not be true, and sample evidence can again to used to "test" compatibility with this hypothesis.

It is important to notice that an investigator is "deciding" whether the sample evidence is compatible with or favourable to the hypothesis. As usual, since population parameters are unknown, the investigator may decide wrongly - and will never know this.

How could one "decide" whether a hypothesis is "true", in this sense? The first thing to do is to introduce notation that distinguishes the supposed value from the actual population parameter value μ. Call the hypothesised value μ_0 (the reason for this strange notation will be explained in Section 14.3.1).

14.2 Confidence intervals again

If we sample from $N\left(\mu, \sigma^2\right)$, with σ^2 again assumed known for simplicity, a $100\,(1-\alpha)\,\%$ confidence interval is a pair of random variables C_L, C_U such that

$$\Pr\left(\mu \in [C_L, C_U]\right) = 1-\alpha, \qquad \Pr\left(\mu \notin [C_L, C_U]\right) = \alpha,$$

where

$$C_L = \bar{X} - z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}, \qquad C_U = \bar{X} + z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}.$$

In Section 13.2.2, interpretations of the corresponding interval estimate ("the" confidence interval) $[c_L, c_U]$ are discussed. There it is emphasised that such a confidence interval is really augmenting the point estimate $\bar{x}$ of μ with a measure of sampling variability.

14.2.1 Rejection criteria

Another possible interpretation is that the interval estimate displays, in some sense, the "likely range of values of μ".

If we accept this interpretation uncritically for the moment, one way of making a decision that the population parameter μ is equal to the hypothesised value μ_0, so that $\mu = \mu_0$, is to ask if μ_0 is contained in the interval estimate. That is, decide "yes" if

$$\mu_0 \in [c_L, c_U]$$

and "no" if

$$\mu_0 \notin [c_L, c_U],$$

where

$$c_L = \bar{x} - z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}, \qquad c_U = \bar{x} + z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}.$$

The more usual language here is to *accept* or *reject* the hypothesis.

To reject the hypothesis $\mu = \mu_0$, we need μ_0 to satisfy

$$\text{either} \quad \mu_0 < c_L = \bar{x} - z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} \quad \text{or} \quad \mu_0 > c_U = \bar{x} + z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}},$$

These can be rearranged to give the rejection criterion in terms of $\bar{x}$ as

$$\text{either} \quad \bar{x} > \mu_0 + z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} \quad \text{or} \quad \bar{x} < \mu_0 - z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}$$

or even as

$$\text{either} \quad \frac{\bar{x}-\mu_0}{\sqrt{\sigma^2/n}} > z_{\alpha/2} \quad \text{or} \quad \frac{\bar{x}-\mu_0}{\sqrt{\sigma^2/n}} < -z_{\alpha/2}.$$

This last statement can be compressed even further: put

$$z = \frac{\bar{x}-\mu_0}{\sqrt{\sigma^2/n}}.$$

Then the rejection criterion

$$\text{either} \quad z > z_{\alpha/2} \quad \text{or} \quad z < -z_{\alpha/2}.$$

is actually equivalent to

$$|z| > z_{\alpha/2}.$$

14.2.2 Critical values

The values

$$\mu_0 - z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} \quad \text{or} \quad \mu_0 + z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}$$

against which $\bar{x}$ is compared are called **critical values**, and will be denoted $\bar{x}_L, \bar{x}_U$:

$$\bar{x}_L = \mu_0 - z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}, \qquad \bar{x}_U = \mu_0 + z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}.$$

The values $\bar{x}_L, \bar{x}_U$ are critical because they are the boundary between accepting and rejecting the hypothesis.

In the same way, the values $-z_{\alpha/2}$ and $z_{\alpha/2}$ are critical values: if $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$, the hypothesis is rejected. Notice that these critical values are the percentage points of the $N(0,1)$ distribution, whereas the critical values $\bar{x}_L, \bar{x}_U$ are derived from these percentage points.

To summarise: deciding to reject the hypothesis $\mu = \mu_0$ if

$$\mu_0 \notin [c_L, c_U],$$

is equivalent to rejecting if

- $\bar{x} > \bar{x}_U = \mu_0 + z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} \quad \text{or} \quad \bar{x} < \mu_0 - z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}$

  or if

- $|z| > z_{\alpha/2}$.

In words, we reject if $\bar{x}$ is too large relative to the critical value $\bar{x}_U = \mu_0 + z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}$ or too small relative to the critical value $\bar{x}_L = \mu_0 - z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}$, or if $|z|$ is too large relative to $z_{\alpha/2}$.

Using the example from Section 13.2.3, where a sample of size 50 is drawn from $N(\mu, 5)$ with $\bar{x} = 18$, we found that the 95% confidence interval was

$$[c_L, c_U] = [17.38, 18.62].$$

If the hypothesised value of μ is $\mu = \mu_0 = 20$, we can see immediately that this should be rejected. We can find the critical values as

$$\begin{aligned}
\bar{x}_L &= \mu_0 - z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} = 20 - (1.96)\sqrt{\frac{5}{50}} = 19.3802, \\
\bar{x}_U &= \mu_0 + z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} = 20 + (1.96)\sqrt{\frac{5}{50}} = 20.6198.
\end{aligned}$$

Since

$$\bar{x} = 18 < \bar{x}_L = 19.3802,$$

we confirm the rejection.

Similarly, we can find

$$z = \frac{\bar{x} - \mu_0}{\sqrt{\sigma^2/n}} = \frac{18 - 20}{\sqrt{5/50}} = -6.3246,$$

which clearly satisfies

$$z < -z_{\alpha/2} = -1.96.$$

14.2.3 Properties of the rejection criteria

To investigate these properties, we can use the repeated sampling interpretation of a confidence interval given in Section 13.2.2. If random samples are repeatedly drawn from $N(\mu_0, \sigma^2)$, then approximately $100(1-\alpha)\%$ of these interval estimates will contain μ_0. The use of the hypothesised value in the normal distribution here is deliberate: only under this condition will the repeated sampling statement be true. This is also equivalent to the truth of the probability statement underlying the interval estimator.

So, we can say that *if* we sample from $X \sim N(\mu_0, \sigma^2)$, so that $\bar{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right)$, we will have

$$\Pr(\mu_0 \in [C_L, C_U]) = 1 - \alpha, \qquad \Pr(\mu_0 \notin [C_L, C_U]) = \alpha,$$

with

$$C_L = \bar{X} - z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}, \qquad C_U = \bar{X} + z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}.$$

Notice the implication: even when the hypothesised value μ_0 is the "true" value of μ, there is some chance, α, that we will make an incorrect decision. This is an inherent feature of the use of an interval estimate argument to decide whether to accept or reject the hypothesis $\mu = \mu_0$.

In effect, we have chosen a procedure which will reject the hypothesis $\mu = \mu_0$ with probability α even when it is true. Ideally, one would like this probability to be small.

This probability of rejection can also be expressed as an $\bar{X}$ or Z probability. For, the probability

$$\Pr\left(\mu_0 \in [C_L, C_U]\right) = 1 - \alpha$$

is derived, as in Section 13.2.4, from the probability

$$\Pr\left(-z_{\alpha/2} \leqslant Z \leqslant z_{\alpha/2}\right) = 1 - \alpha,$$

with

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}},$$

and $Z \sim N(0,1)$ when $\mu = \mu_0$. The corresponding probability for rejecting $\mu = \mu_0$,

$$\Pr\left(\mu_0 \notin [C_L, C_U]\right) = \alpha,$$

is then

$$\Pr\left(|Z| > z_{\alpha/2}\right) = \alpha.$$

Another arrangement of the first Z probability is

$$\Pr\left(\mu_0 - z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} \leqslant \bar{X} \leqslant \mu_0 + z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right) = 1 - \alpha,$$

with rejection probability version using the critical values $\bar{x}_L, \bar{x}_U$,

$$\begin{aligned} \Pr\left\{\left(\bar{X} < \mu_0 - z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right) \cup \left(\bar{X} > \mu_0 + z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right)\right\} &= \Pr\left\{\left(\bar{X} < \bar{x}_L\right) \cup \left(\bar{X} > \bar{x}_U\right)\right\} \\ &= \alpha. \end{aligned}$$

What can be said about the value of μ if $\mu = \mu_0$ does not appear to be true (i.e. is rejected, using sample information)? At this point, we need to be a bit more formal.

14.3 Hypotheses

In the discussion of the previous two sections, certain words, *hypothesis, test, decide, accept, reject,* were used as if we were referring to their ordinary meaning. However, these words also belong to the jargon of that part of statistical inference called *hypothesis testing:* their specific meaning in this context will become clear.

14.3.1 Null hypotheses

A “hypothesis” appears to be a claim about the value of some population parameter: for example

$$\mu = 20, \qquad \pi = 0.45, \quad \sigma^2 = 1.$$

It can also be much more general than this, and refer to functions of population parameters. For example, we may obtain a random sample from a population with population mean μ_1, and then a random sample from another population with mean μ_2, and ask whether

$$\mu_1 = \mu_2, \qquad \text{equivalently}, \qquad \mu_1 - \mu_2 = 0.$$

However, these examples would usually be called **null** hypotheses: the last example shows the link between the adjective *null* and the numerical value 0. The earlier examples can be rephrased to emphasise this as

$$\mu - 20 = 0, \quad \pi - 0.45 = 0, \quad \sigma^2 - 1 = 0.$$

So, a **null hypothesis** expresses the preconception or “expectation” (compare the language of Section 14.1) about the value of the population parameter. There is a standard, compact, notation for this:

$$H_0 : \mu = \mu_0$$

expresses the null hypothesis that $\mu = \mu_0$.

Deciding whether or not $\mu = \mu_0$ on the basis of sample information is usually called *testing the hypothesis*. There is actually a question about **what** sample evidence should be used for this purpose. This issue is has to do with the question of how the true population parameter μ is related to the “null hypothesis value” μ_0 if μ is not equal to μ_0.

14.3.2 Alternative hypotheses

The **alternative hypothesis** defines this relationship. The possible alternatives to H_0 are (only one of)

$$\mu > \mu_0, \qquad \mu < \mu_0, \qquad \mu \neq \mu_0.$$

Formal statements of these as hypotheses are

$$H_A : \mu > \mu_0$$

or

$$H_A : \mu < \mu_0$$

or

$$H_A : \mu \neq \mu_0.$$

Here, H_A stands for the *alternative hypothesis.* In some textbooks, this is denoted H_1, but is still called the alternative hypothesis.

Notice that the null and alternative hypotheses are expected to be mutually exclusive: it would make no sense to reject H_0 if it was also contained in H_A!

If the null hypothesis is rejected, then we are apparently deciding to "accept" the truth of the alternative hypothesis H_A. Notice that H_A gives a range of values of μ : it does not point to any specific value. So, in rejecting the null hypothesis, we are not deciding in favour of another specific value of μ.

14.3.3 Example

Which sort of alternative hypothesis should be chosen? This all depends on context and perspective. Suppose that a random sample of jars of jam coming off a packing line is obtained. The jars are supposed to weigh, on average, 454gms. You have to decide, on the basis of the sample evidence, whether or not this is true. So, the null hypothesis here is

$$H_0 : \mu = 454.$$

As a jam manufacturer, you may be happy to get away with selling, on average, shortweight jars, since this gives more profit, but be unhappy at selling, on average, overweight jars, owing to the loss of profit. So, the manufacturer might choose the alternative hypothesis

$$H_A : \mu > 454.$$

A consumer, or a trading standards conscious jam manufacturer, might be more interested in underweight jars, so the alternative might be

$$H_A : \mu < 454.$$

A mechanic from the packing machine company might simply want evidence that the machine is not working to specification, and would choose

$$H_A : \mu \neq 454$$

as the alternative.

14.3.4 Sides and Tails

Alternative hypotheses are often called **one sided** or **one tailed, two sided** or **two tailed.**

- $H_A : \mu > \mu_0$ is an **upper one tailed** (or **sided**) hypothesis

- $H_A : \mu < \mu_0$ is a **lower one tailed** (or **sided**) hypothesis
- $H_A : \mu \neq \mu_0$ is a **two tailed** (or **sided**) hypothesis.

  One can see that *upper* and *lower* come from the direction in which the "arrow" of the inequality points.

14.4 Types of Error

The rejection criteria based on a confidence interval argument in Section 14.2.1 are really designed for testing the hypotheses

$$H_0 : \mu = \mu_0$$

against

$$H_A : \mu \neq \mu_0,$$

as we shall shortly see. The point to be derived from the discussion in Section 14.2.3 is that even when **the null hypothesis** H_0 **is true**, i.e. $\mu = \mu_0$, there is some chance of rejecting it, that is, of denying its truth. This is an "error". It is also true, although not apparent from the discussion in Section 14.2.3, that when H_0 is not true, and H_A is therefore presumed to be true, there is some chance that one will **accept** (more precisely, **not reject**) H_0. This too is an error, of a different sort. These points are also true in the case of one-tailed alternative hypotheses.

These errors are called **Type I** and **Type II** errors, and can be summarised as:

- a Type I error is incorrectly rejecting the null hypothesis;
- a Type II error is incorrectly accepting the null hypothesis.

It is also possible to make correct decisions, and the possible combinations of correct and incorrect decisions are laid out in the following table:

	Decision	
Truth	H_0 accepted	H_0 rejected
H_0 true	correct decision	Type I error
H_0 false	Type II error	correct decision

The objective in designing a procedure to test an H_0 against an H_A - i.e. decide whether to accept H_0 or to reject H_0, is to ensure that a Type I error does not occur to often. More precisely, the objective is to design a procedure which fixes the probability of a Type I error occurring at a **prespecified** level, α, which is small and therefore presumably tolerable. In the new terminology, the procedure based on a confidence interval described in Section 14.2.3 has, by construction, a Type I error probability of 1 minus the confidence level, $1 - \alpha$, used: i.e. a Type I error probability of α.

Conventionally, the Type I error probability α is set to 5%, or sometimes 1%. For the moment, we shall simply accept the convention. More will be said about this in Section 14.8.

Once a procedure is designed in this way, one can determine in principle the probability of the Type II error for each value of μ that could be true under the alternative hypothesis. For a given sample size, n, this probability cannot be controlled, although it can be reduced by increasing n. It is also true that for a given value of μ under H_A, as the Type I error probability α is made smaller, the corresponding Type II error probability increases.

14.4.1 Example

This classic example is intended to sharpen the contrast between the two sorts of error. In a court of law, the accused is either innocent or guilty, and is usually presumed innocent until proven guilty. In terms of hypotheses, this suggests that the null hypothesis is

$$H_0 : \text{accused is innocent}$$

with the implied alternative

$$H_A : \text{the accused is guilty.}$$

The court makes a decision as to whether the accused is innocent or guilty, but may make a mistake:

- a Type I error, rejecting H_0 when it is true, consists of convicting an innocent person;
- a Type II error, accepting H_0 when false, consists of acquitting a guilty person.

Which of these two errors is the most serious? The standard approach to hypothesis testing **presumes** that the Type I error is the most important, as noted above. The legal system also appears to take this view.

14.5 The traditional test procedure

Suppose that we wish to test the null hypothesis

$$H_0 : \mu = \mu_0.$$

The traditional test procedure (sometimes called the *classical* test procedure) is based on a simple principle: is the value of $\bar{x}$ too extreme relative to the value μ_0 to be compatible with the truth of this null hypothesis? If the answer is yes, reject the null hypothesis. Two questions arise naturally here. What is extreme? What is too extreme?

The first question is equivalent to asking in which direction we should look for extreme values. This information is provided to us by the alternative hypothesis we specify. So, in the case of a

- **lower one tailed** alternative, $H_A : \mu < \mu_0$, "extreme" is "$\bar{x}$ too small relative to μ_0";
- **upper one tailed** alternative, $H_A : \mu > \mu_0$, "extreme" is "$\bar{x}$ too large relative to μ_0";
- **two tailed** alternative, $H_A : \mu \neq \mu_{0,}$ "extreme" is "$\bar{x}$ too small **or** too large relative to μ_0".

Recall that the confidence interval procedure of Section 14.2.1 had the latter as the rejection criteria, and so can be seen to be appropriate for a test against a two-tailed alternative:

$$H_0 : \mu = \mu_0 \quad \text{against} \quad H_A : \mu \neq \mu_0.$$

The rejection criterion here uses the critical values $\bar{x}_L$ and $\bar{x}_U$ and critical region

$$\bar{x} < \bar{x}_L = \mu_0 - z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} \qquad \text{or} \qquad \bar{x} > \bar{x}_U = \mu_0 + z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}.$$

This rejection criterion has the property that

$$\Pr\left\{\left(\bar{X} < \bar{x}_L\right) \cup \left(\bar{X} > \bar{x}_U\right)\right\} = \alpha,$$

the probability of a Type I error occurring.

It is the **choice** of Type I error probability which defines the concept of "how extreme" for the hypothesis testing procedure. In fact, the critical values $\bar{x}_L$ and $\bar{x}_U$ are designed to generate the specific Type I error probability α.

14.5.1 An upper one-tailed test

In this light, and without reference to the confidence interval case, we construct a test procedure for an upper one tailed test,

$$H_0 : \mu = \mu_0 \quad \text{against} \quad H_A : \mu > \mu_0$$

for the case of sampling from $N\left(\mu, \sigma^2\right)$ with σ^2 known. According to the arguments above, we will reject this null hypothesis if the sample value $\bar{x}$ is too large, where this is determined by a **critical value**, $\bar{x}_U$**.**

This critical value $\bar{x}_U$ defines an **acceptance region** and a **critical** or **rejection region,** with the obvious idea that if $\bar{x}$ falls in the acceptance region, H_0 is accepted, whilst if it falls in the critical or rejection region, H_0 is rejected. Clearly, the acceptance region is

$$(-\infty, \bar{x}_U]$$

and the rejection region is

$$(\bar{x}_U, \infty).$$

The critical value $\bar{x}_U$ is determined by the requirement that

$$\Pr\left(\bar{X} > \bar{x}_U \,|H_0 \text{ true}\right) = \alpha,$$

i.e. that the Type I error probability, or **level of significance** for short, be set to α. Notice the importance of the requirement "H_0 true": it means that in computing this probability, or more accurately determining $\bar{x}_U$ from it, we use the sampling distribution of $\bar{X}$ **under the null hypothesis,**

$$\bar{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right).$$

Standardising in the usual way, we get

$$\Pr\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} > \frac{\bar{x}_U-\mu_0}{\sigma/\sqrt{n}} \,|H_0 \text{ true}\right) = \alpha,$$

or

$$\Pr\left(Z > \frac{\bar{x}_U-\mu_0}{\sigma/\sqrt{n}} \,|H_0 \text{ true}\right) = \alpha$$

for

$$Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}},$$

Since $Z \sim N(0,1)$, this implies that

$$\frac{\bar{x}_U-\mu_0}{\sigma/\sqrt{n}} = z_\alpha$$

where $\Pr(Z > z_\alpha) = \alpha$, which gives

$$\bar{x}_U = \mu_0 + \frac{\sigma}{\sqrt{n}} z_\alpha.$$

One can also see in this argument an alternative approach, which is to compute the sample value of Z : if

$$z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} > z_\alpha,$$

reject the null hypothesis. This works simply because the inequalities

$$\bar{x} > \mu_0 + \frac{\sigma}{\sqrt{n}} z_\alpha$$

and

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} > z_\alpha$$

are equivalent.

To summarise: in sampling from $N\left(\mu, \sigma^2\right)$ with σ^2 known, a test of the hypotheses

$$H_0 : \mu = \mu_0 \quad \text{against} \quad H_A : \mu > \mu_0$$

at level of significance α is given by rejecting H_0 if

- $\bar{x} > \bar{x}_U = \mu_0 + \dfrac{\sigma}{\sqrt{n}} z_\alpha$

  or

- $z = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} > z_\alpha.$

  where $\Pr\left(Z > z_\alpha\right) = \alpha.$

Of the two approaches, the first fits better the intuitive idea of asking whether $\bar{x}$ is too large to be compatible with $\mu = \mu_0$, but the second is better for performing the test quickly.

An equivalent approach using a suitable confidence interval will be described in Section 14.9.

14.5.2 Example

This example is deliberately free from a real world context. A random sample of size 50 is drawn from $N\left(\mu, 5\right)$, to test the hypotheses

$$H_0 : \mu = 20 \quad \text{against} \quad H_A : \mu > 20.$$

Suppose that the sample mean is $\bar{x} = 20.7$. We choose the conventional Type I error probability or level of significance of 5%. We need the upper 5% percentage point of the $N\left(0, 1\right)$ distribution:

$$\Pr\left(Z > z_\alpha\right) = \alpha = 0.05$$

implies $z_\alpha = 1.645$. The critical value $\bar{x}_U$ is then

$$\bar{x}_U = \mu_0 + \frac{\sigma}{\sqrt{n}} z_\alpha = 20 + \sqrt{\frac{5}{50}}\left(1.645\right) = 20.52.$$

We reject H_0 if

$$\bar{x} > \bar{x}_U$$

and $\bar{x} = 20.7 > \bar{x}_U$ here, so H_0 is rejected.

The more direct approach starts by finding the critical value z_α from the tables, and computes

$$z = \frac{\bar{x} - \mu_0}{\sqrt{\sigma^2/n}} = \frac{20.7 - 20}{\sqrt{5/50}} = 2.2136.$$

Since $z > z_\alpha$, here, we reject H_0.

14.5.3 A lower one tailed test

This follows the lines one might anticipate from the discussion above. To test

$$H_0 : \mu = \mu_0 \quad \text{against} \quad H_A : \mu < \mu_0$$

using a random sample from $N\left(\mu, \sigma^2\right)$ with σ^2 known, we know that small values of $\bar{x}$ are evidence against the null hypothesis. How small is determined by the critical value $\bar{x}_L$ designed to set the level of significance to a value $\alpha :$

$$\Pr\left(\bar{X} < \bar{x}_L \,|H_0 \text{ true}\right) = \alpha.$$

Following through the arguments of the preceding case, we get

$$\Pr\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < \frac{\bar{x}_L - \mu_0}{\sigma/\sqrt{n}} \,|H_0 \text{ true}\right) = \alpha,$$

or

$$\Pr\left(Z < \frac{\bar{x}_L - \mu_0}{\sigma/\sqrt{n}} \,|H_0 \text{ true}\right) = \alpha$$

for

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}},$$

Since $Z \sim N(0,1)$, this implies that

$$\frac{\bar{x}_L - \mu_0}{\sigma/\sqrt{n}} = -z_\alpha$$

for $\Pr(Z < -z_\alpha) = \alpha$, and then

$$\bar{x}_L = \mu_0 - \frac{\sigma}{\sqrt{n}} z_\alpha.$$

Suppose that in the context of the previous example (in Section 14.5.2) the sample information is now $\bar{x} = 19.7$, and the hypotheses to be tested are

$$H_0 : \mu = 20 \quad \text{against} \quad H_A : \mu < 20.$$

This time, the level of significance or Type I error probability will be set at $\alpha = 0.01$, leading to

$$\Pr(Z < -z_\alpha) = 0.01$$

or

$$-z_\alpha = -2.33.$$

The critical value is then

$$\bar{x}_L = \mu_0 - \frac{\sigma}{\sqrt{n}} z_\alpha = 20 - \sqrt{\frac{5}{50}}\,(2.33) = 20 - 0.74 = 19.26.$$

Here,

$$\bar{x} = 19.7 > \bar{x}_L = 19.26,$$

so that the null hypothesis is **not** rejected: $\bar{x}$ is not sufficiently extreme in the "right" direction.

The direct approach calculates the sample value of Z as

$$z = \frac{\bar{x} - 20}{\sqrt{5/50}} = \frac{19.7 - 20}{\sqrt{5/50}} = -0.95,$$

to be compared with the lower tail percentage point $-z_\alpha = -2.33$. Again, the null hypothesis is not rejected at the 1% level.

14.5.4 A two tailed test

This will be described independently of the motivation via confidence intervals in Section 14.2, and the observation in Section 14.5 that the confidence interval reasoning was suited to two tailed tests.

Suppose that a sample is drawn from $N\left(\mu, \sigma^2\right)$, with σ^2 assumed known, with the intention of testing the hypotheses

$$H_0 : \mu = \mu_0 \quad \text{against} \quad H_A : \mu \neq \mu_0.$$

Following the reasoning of the upper and lower one-tailed tests, we can argue that values of $\bar{x}$ which are **either** too small **or** too large relative to μ_0 count as evidence against H_0. So, we need to find two critical values $\bar{x}_L$, $\bar{x}_U$ such that

$$\Pr\left\{\left(\bar{X} < \bar{x}_L\right) \text{ or } \left(\bar{X} > \bar{x}_U\right)\right\} = \alpha.$$

This is cumbersome to standardise directly, so we use the probability of the complementary event,

$$\Pr\left(\bar{x}_L \leqslant \bar{X} \leqslant \bar{x}_U\right) = 1 - \alpha$$

and standardise using

$$\bar{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right)$$

to give

$$\Pr\left(\frac{\bar{x}_L - \mu_0}{\sqrt{\sigma^2/n}} \leqslant Z \leqslant \frac{\bar{x}_U - \mu_0}{\sqrt{\sigma^2/n}}\right) = 1 - \alpha.$$

This will be true if

$$\frac{\bar{x}_L - \mu_0}{\sqrt{\sigma^2/n}} = -z_{\alpha/2}, \qquad \frac{\bar{x}_U - \mu_0}{\sqrt{\sigma^2/n}} = z_{\alpha/2}$$

or

$$\bar{x}_L = \mu_0 - z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}, \qquad \bar{x}_U = \mu_0 + z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}.$$

If

$$\bar{x} < \bar{x}_L \quad \text{or} \quad \bar{x} > \bar{x}_U,$$

reject the null hypothesis.

The direct version again computes the sample value of

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} :$$

if

$$z < -z_{\alpha/2} \quad \text{or} \quad z > z_{\alpha/2}$$

reject the null hypothesis.

For an example, consider the case of packing jars of jam, as in Section 14.3.3. Here, the hypotheses are

$$H_0 : \mu = 454 \quad \text{against} \quad H_A : \mu \neq 454.$$

Suppose that jam jar weights are distributed as $X \sim N(\mu, 16)$ and that a random sample of 25 jars gives a sample mean weight of $\bar{x} = 452.41$ grams. Then, using a 5% level of significance, $z_{\alpha/2} = 1.96$, so that

$$\begin{aligned}\bar{x}_L &= \mu_0 - z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} = 454 - 1.96\sqrt{\frac{16}{25}} = 452.432 \\ \bar{x}_U &= \mu_0 + z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} = 454 + 1.96\sqrt{\frac{16}{25}} = 455.568.\end{aligned}$$

Here, the null hypothesis is just rejected.

Using the direct method we calculate

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{452.41 - 454}{0.8} = -1.9875,$$

with the same conclusion.

14.6 Hypothesis tests: other cases

We know from the discussion of confidence intervals leading up to Section 13.2.6 that the assumption that the population variance σ^2 is known in sampling from $N\left(\mu, \sigma^2\right)$ has to be seen as a convenient, simplifying but unrealistic assumption. Indeed, this can also be said of the assumption that we sample from a normal distribution.

We have seen that the first issue leads to confidence intervals based on the t distribution in Section 13.4.3, and the second issue leads to large sample or approximate confidence intervals in Section 13.6. The discussion of hypothesis testing was motivated via an argument based on confidence intervals, so that it is reasonable to expect that the hypothesis tests described in Sections 14.5.1, 14.5.3 and 14.5.4 can be adapted to deal with these cases.

14.6.1 Test statistics

One can see that in the case that σ^2 is known, the random variable

$$Z = \frac{\bar{X} - \mu_0}{\sqrt{\sigma^2/n}}$$

drives the hypothesis testing procedure. It is convenient to call Z a **test statistic**, to distinguish it from other sample statistics, and it has a known distribution, when the null hypothesis $\mu = \mu_0$ is true, namely $N\left(0,1\right)$.

The distribution of this test statistic, **under the null hypothesis**, provides the critical values (directly or indirectly) required for the test. The sample value of Z, z, will be called the **value of the test statistic.** The "direct" method of carrying out the hypothesis test involves a comparison of the value of the test statistic with these critical values.

14.6.2 The t case

If we wish to test the hypotheses

$$H_0 : \mu = \mu_0 \quad \text{against} \quad H_A : \mu \neq \mu_0$$

(for example) about the mean of $N\left(\mu, \sigma^2\right)$, with σ^2 unknown, the analogy with confidence intervals suggests that we should base the test procedure on another **test statistic,**

$$T = \frac{\bar{X} - \mu_0}{\sqrt{S^2/n}},$$

since we know that under the null hypothesis,

$$T \sim t_{n-1}.$$

A critical value approach based on $\bar{x}_L$ and $\bar{x}_U$ here is a little clumsy, so we use the "direct method", arguing intuitively that if $\bar{x}$ is too small or too large relative to μ_0, the sample value of T will also be too small or too large relative to critical values which are percentage points of the t distribution. These are denoted $t_{n-1,\alpha}$:

$$\Pr\left(T > t_{n-1,\alpha}\right) = \alpha, \qquad \Pr\left(T \leqslant t_{n-1,\alpha}\right) = 1 - \alpha.$$

For this test, the acceptance region is defined by

$$\Pr\left(-t_{n-1,\alpha/2} \leqslant T \leqslant t_{n-1,\alpha/2}\right) = 1 - \alpha$$

and then the rejection region is

$$\Pr\left\{\left(T < -t_{n-1,\alpha/2}\right) \cup \left(T > t_{n-1,\alpha/2}\right)\right\} = \alpha.$$

But, this is also equivalent to

$$\Pr\left(|T| > t_{n-1,\alpha/2}\right) = \alpha :$$

see Section 12.2.6 for a revision of the argument.

So, to test the hypotheses

$$H_0 : \mu = \mu_0 \quad \text{against} \quad H_A : \mu \neq \mu_0$$

we

- calculate the sample value of T,

$$t = \frac{\bar{x} - \mu_0}{\sqrt{s^2/n}};$$

- compare $|t|$ with $t_{n-1,\alpha/2}$;
- if $|t| > t_{n-1,\alpha/2}$, reject the null hypothesis.

One can immediately see how the one sided cases would work. Again compute the sample value of T, and use as critical values $-t_{n-1,\alpha}$ for the lower one tailed case, and $t_{n-1,\alpha}$ for the upper one tailed case.

Summarising the procedure for each case,

1. to test the hypotheses

$$H_0 : \mu = \mu_0 \quad \text{against} \quad H_A : \mu \neq \mu_0$$

at level of significance α, we

- calculate the sample value of T,
$$t = \frac{\bar{x} - \mu_0}{\sqrt{s^2/n}};$$
- compare $|t|$ with $t_{n-1,\alpha/2}$;
- if $|t| > t_{n-1,\alpha/2}$, reject the null hypothesis.

2. to test the hypotheses
$$H_0 : \mu = \mu_0 \quad \text{against} \quad H_A : \mu < \mu_0$$
at level of significance α we
   - calculate the sample value of T,
   $$t = \frac{\bar{x} - \mu_0}{\sqrt{s^2/n}};$$
   - compare t with $-t_{n-1,\alpha}$;
   - if $t < -t_{n-1,\alpha}$, reject the null hypothesis.

3. to test the hypotheses
$$H_0 : \mu = \mu_0 \quad \text{against} \quad H_A : \mu > \mu_0$$
at level of significance α we
   - calculate the sample value of T,
   $$t = \frac{\bar{x} - \mu_0}{\sqrt{s^2/n}};$$
   - compare t with $t_{n-1,\alpha}$;
   - if $t > t_{n-1,\alpha}$, reject the null hypothesis.

For an example, we adapt the jam example in Section 14.5.4. Here the objective is to test the hypotheses
$$H_0 : \mu = 454 \quad \text{against} \quad H_A : \mu \neq 454$$
using the results of a random sample of size 25 from $N\left(\mu, \sigma^2\right)$, yielding $\bar{x} = 452.41$ and $s^2 = 12.992$. We choose the conventional level of significance, 5%, so we need the value $t_{24,0,025}$ which gives
$$\Pr\left(|T| > t_{24,0,025}\right) = 0.05.$$

From the tables in the Appendix, we need the column headed 0.975, giving

$$t_{24,0,025} = 2.064.$$

The value t of the test statistic T is

$$t = \frac{\bar{x} - \mu_0}{\sqrt{s^2/n}} = \frac{452.41 - 454}{\sqrt{12.992/25}} = -2.206,$$

so that

$$|t| = 2.206 > t_{24,0,025} = 2.064,$$

leading in this case to rejection of the null hypothesis.

14.6.3 Tests on μ using large sample normality

Consider the example used in Section 13.2.3 and again in Section 14.5.2. There, a random sample of size 50 was drawn from a distribution of household incomes in £'000, which was assumed to be normally distributed. There is plenty of empirical evidence that household incomes are **not** normally distributed, so if $n = 50$ is considered sufficiently large, a **large sample test** can be used, without assuming normality.

This is based on the argument that

$$T = \frac{\bar{X} - \mu_0}{\sqrt{S^2/n}} \sim N(0,1) \qquad \text{approximately.}$$

So, test procedures for the hypothesis

$$H_0 : \mu = \mu_0$$

against the usual alternatives use the rejection regions in Section 14.6.2, using the sample value t of T, and percentage points from $N(0,1)$ instead of the t distribution used there.

For the example suppose we wish to test

$$H_0 : \mu = 20 \qquad \text{against} \qquad H_A : \mu > 20$$

at the conventional 5% level of significance. The random sample of size 50 yields

$$\bar{x} = 22.301, \qquad s^2 = 12.174.$$

First, we need the value z_α such that

$$\Pr(Z > z_\alpha) = 0.05,$$

which is

$$z_\alpha = 1.645.$$

Next, we need the sample value of T :

$$t = \frac{\bar{x} - \mu_0}{\sqrt{s^2/n}} = \frac{22.301 - 20}{\sqrt{12.174/50}} = 4.66.$$

This clearly exceeds $z_\alpha = 1.645$, so the null hypothesis is rejected $-$ **approximately**.

The idea that we can only perform an approximate test does lead to difficulties. If, as in Section 14.5.4, a null hypothesis is only marginally rejected (or even accepted), there is always the question as to whether this is "correct" or simply a consequence of a poor approximation given the sample size available. This comment points to a possible solution: get a bigger sample and try again.

14.6.4 Tests on a proportion π

In random sampling from a Bernoulli distribution with success probability π, it is known from Sections 12.3 and 13.6.3 that the sample proportion P is approximately normally distributed in large samples:

$$P \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right) \qquad \text{approximately.}$$

This was used to construct a large sample confidence interval for π in Section 13.6.3. There, the fact that the variance of P is unknown is handled by working with the **approximate** distribution of the random variable T^* :

$$T^* = \frac{P - \pi}{\sqrt{\dfrac{P(1-P)}{n}}} \sim N(0,1), \qquad \text{approximately.}$$

Suppose that we wished to test the hypotheses that

$$H_0 : \pi = \pi_0 \qquad \text{against} \qquad H_A : \pi \neq \pi_0$$

for example. This null hypothesis specifies the value of the mean of the approximate distribution of P, and also the variance. So, it would also be justifiable to use as a test statistic,

$$T = \frac{P - \pi_0}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}}$$

with

$$T \sim N(0,1) \qquad \text{approximately,}$$

under the null hypothesis.

Textbooks differ on whether one should use T or T^* as a test statistic. In this course, we will be agnostic, and allow the use of either. However, there is the caveat that it is possible to reject a hypothesis using the test statistic T, and accept it with T^* – the converse is also possible.

Apart from this difficulty, carrying out an approximate large sample test for π is just the same as the large sample normal case for the μ.

To illustrate the ideas, we use the example of households owning a DVD player used in Section 13.6.4. A market researcher suspects that the ownership of DVD players has risen above the previously assumed level of 25%. He collects information from the sample of 300 households, and finds $p = 0.28$.

Here we test

$$H_0 : \pi = 0.25 \quad \text{against} \quad H_A : \pi > 0.25,$$

using a level of significance of 1%. So, we need the value z_α such that

$$\Pr(Z > z_\alpha) = 0.01 \quad \text{for} \quad Z \sim N(0,1),$$

which is

$$z_\alpha = 2.3263$$

from the table in Section 13.2.5. Here we use T as the test statistic, with sample value

$$t = \frac{p - \pi_0}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}} = \frac{0.28 - 0.25}{\sqrt{\dfrac{0.25(1-0.25)}{300}}} = 1.2.$$

This is an upper one tailed test, and t does not exceed the critical value $z_\alpha = 2.3263$, so we do not reject (i.e. accept) the null hypothesis.

It should be clear how to extend this reasoning to examples with lower one-tailed or two tailed alternative hypotheses.

14.7 p values: the modern approach

Consider the example of Section 14.5.2 again. Here, a random sample of size 50 was drawn from $N(\mu, 5)$, yielding $\bar{x} = 20.7$. The hypotheses

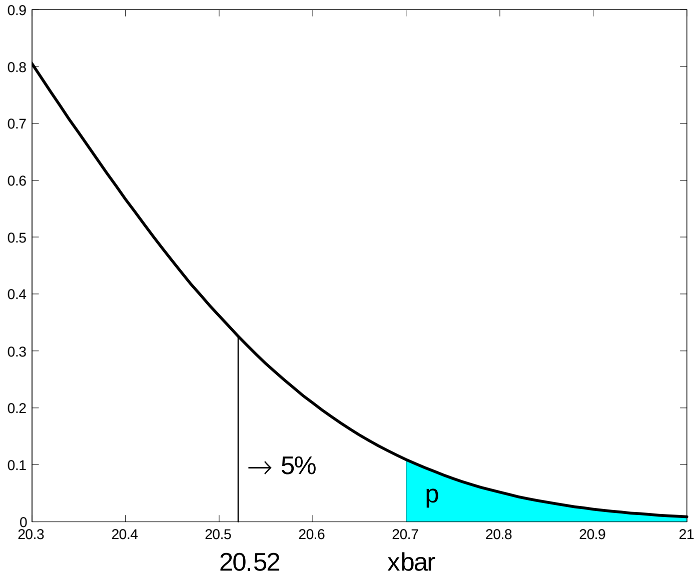
$$H_0 : \mu = 20 \quad \text{against} \quad H_A : \mu > 20$$

were tested at a 5% level of significance. The critical value for $\bar{x}$ was found to be

$$\bar{x}_U = 20.52,$$

and, for the direct method,

$$z_\alpha = 1.645.$$



The z value derived from $\bar{x}$ was

$$z = 2.2136,$$

so that the null hypothesis was rejected.

Examine Figure 14.7 carefully: it shows a fragment of the sampling distribution of $\bar{X}$ under the null hypothesis, more specifically, the upper tail. The critical value $\bar{x}_U = 20.52$ is shown, and is such that

$$\Pr\left(\bar{X} > 20.52\,|H_0 \text{ true}\right) = 0.05.$$

Also, the position of the sample mean, $\bar{x} = 20.7$, is shown. It should be **obvious** from the picture that

$$\Pr\left(\bar{X} > 20.7\,|H_0 \text{ true}\right) < 0.05:$$

just in case it isn't, the corresponding area is shaded. Denote this probability by p :

$$\Pr\left(\bar{X} > 20.7\,|H_0 \text{ true}\right) = p.$$

What should be deduced from Figure 14.7 is that

$$\bar{x} > \bar{x}_U \quad \textbf{if and only if} \quad p < 0.05.$$

In other words, comparing the value of the probability p with the level of significance is equivalent to comparing the value of $\bar{x}$ with its critical value, which we also know is equivalent to comparing the value z with $z_\alpha = 1.645$.

The probability p is a **p-value,** and is sometimes called an **observed significance level**.

14.7.1 p values

This initial definition is only appropriate for the case where we sample from $N\left(\mu,\sigma^2\right)$, with σ^2 known. Extensions to other cases will follow. For one tailed hypothesis tests, a p value is the probability of a value **at least as extreme** as the **sample value** $\bar{x}$, given that the null hypothesis is true. So,

- for a lower one tailed hypothesis, the p value is $\Pr\left(\bar{X} < \bar{x}\,|H_0 \text{ true}\right)$;
- for an upper one tailed hypothesis, the p value is $\Pr\left(\bar{X} > \bar{x}\,|H_0 \text{ true}\right)$.

Computing the p value follows the usual method of standardisation: e.g. when $\bar{X} \sim N\left(\mu_0, \dfrac{\sigma^2}{n}\right)$,

$$\begin{aligned}
\Pr\left(\bar{X} > \bar{x}\,|H_0 \text{ true}\right) &= \Pr\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} > \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\,|H_0 \text{ true}\right) \\
&= \Pr\left(Z > \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\,|H_0 \text{ true}\right) \\
&= \Pr\left(Z > z\,|H_0 \text{ true}\right),
\end{aligned}$$

where z is the "observed value" of Z.

For the example above, we find

$$\begin{aligned}
\Pr\left(\bar{X} > 20.7\,|H_0 \text{ true}\right) &= \Pr\left(Z > 2.2136\,|H_0 \text{ true}\right) \\
&= 0.01343 \\
&= p
\end{aligned}$$

which is clearly less than 0.05. Exactly the same principle can be used for lower tail critical values.

It should be clear from this argument that a p value could be defined directly in terms of the value of the test statistic Z, rather than $\bar{X}$, as $p = \Pr\left(Z > z\,|H_0 \text{ true}\right)$. This is very much in the spirit of the direct approach to performing a hypothesis test.

14.7.2 Upper and lower p values

How can this idea be adapted to the case of two tailed tests? Using the direct approach, to test

$$H_0 : \mu = \mu_0 \quad \text{against} \quad H_A : \mu \neq \mu_0,$$

we calculate the value of

$$Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$$

and compare with critical values $-z_{\alpha/2}$ and $z_{\alpha/2}$. It should be clear, by inspecting Figure 13.2.5 in Section 13.2.5 and using the arguments above that

$$z < -z_{\alpha/2} \quad \text{if and only if} \quad \Pr\left(Z < z\,|H_0 \text{ true}\right) < \alpha/2$$

and

$$z > z_{\alpha/2} \quad \text{if and only if} \quad \Pr\left(Z > z\,|H_0 \text{ true}\right) < \alpha/2.$$

To follow the logic of p values for one tailed hypotheses, we have to accept that there are two p values for the two tailed case, a **lower** p value denoted p_L, and an **upper** p value denoted p_U :

$$p_L = \Pr\left(Z < z\,|H_0 \text{ true}\right), \qquad p_U = \Pr\left(Z > z\,|H_0 \text{ true}\right).$$

How would these be used to reject a null hypothesis? Only one of the two inequalities

$$z < -z_{\alpha/2} \quad \text{and} \quad z > z_{\alpha/2}$$

defining the rejection region can be true, so only one of the lower and upper p values can be less that $\alpha/2$. It is of course possible that neither p_L nor p_U are less than $\alpha/2$. So, a rejection rule

- reject H_0 if $\min\left(p_L, p_U\right) < \alpha/2$

would match the rule based on the critical values $-z_{\alpha/2}$ and $z_{\alpha/2}$. However, it is conventional to modify this so that the comparison is with the level of significance:

- reject H_0 if $2\min\left(p_L, p_U\right) < \alpha$.

This need to calculate upper and lower p values, and then double the smallest one may seem complicated, but in practice it is easy to use. For, if $z < 0$, then the lower p value p_L must be the smallest, whereas if $z > 0$, the upper p value p_U is the smallest. So, one only needs a single calculation.

Consider the example in Section 14.5.4. Here the observed value of Z is

$$z = -1.9875.$$

The lower p value is (using EXCEL)

$$\begin{aligned} p_L &= \Pr\left(Z < -1.9875\,|H_0 \text{ true}\right) \\ &= 0.023433 \end{aligned}$$

and the upper p value is therefore

$$p_U = 1 - p_L = 0.976567.$$

Comparing the smaller of the two, p_L, with $\alpha/2 = 0.025$, or comparing twice the smaller,

$$p = 0.046866,$$

with $\alpha = 0.05$ leads to the same conclusion as before, reject the null hypothesis - but in a sense, only just. One can also see confirmation of the idea that if $z < 0$, p_L will be smaller than p_U.

14.7.3 p values for the t distribution case

Exactly analogous arguments apply to the case where we sample from $N\left(\mu, \sigma^2\right)$, with σ^2 unknown, in order to test hypotheses like

$$H_0 : \mu = \mu_0 \quad \text{against} \quad H_A : \mu > \mu_0.$$

Instead of using the test statistic Z, we use the test statistic

$$T = \frac{\bar{X} - \mu_0}{\sqrt{S^2/n}} \sim t_{n-1},$$

where this distribution statement is true under the null hypothesis. So, given the sample value of T, t, we compute (for this upper tailed case)

$$p = \Pr\left(T > t \,|H_0 \text{ true}\right).$$

Equally, for a lower one tailed alternative,

$$H_0 : \mu = \mu_0 \quad \text{against} \quad H_A : \mu < \mu_0.$$

we compute

$$p = \Pr\left(T < t \,|H_0 \text{ true}\right).$$

For the two-tailed case,

$$H_0 : \mu = \mu_0 \quad \text{against} \quad H_A : \mu \neq \mu_0,$$

we would calculate both of these probabilities and call them the **upper** and **lower** p values.

The rejection rules are just the same as in the cases discussed in the previous few sections. However, there seems to be a major practical difficulty: there are no tables of probabilities for the t distribution. One therefore has to use EXCEL: how this can be done is outlined in Section 13.4.2.

The reasoning is illustrated by using the jam example of Sections 14.3.3, 14.5.4 and 14.6.2. The t test for a two tailed hypothesis performed in Section 14.6.2 gave a value of the test statistic T as

$$t = -2.206,$$

with 24 degrees of freedom. Since t is negative, we know that the lower p value

$$\Pr\left(T < t \,|H_0 \text{ true}\right)$$

will be the smaller of the lower and upper p values. Recall from Section 13.4.2 that EXCEL calculates

$$\Pr\left(T > t\right) = \texttt{tdist}\left(t, \texttt{df}, \texttt{1}\right),$$

for $t > 0$. Here, we exploit symmetry of the t distribution to give

$$\begin{aligned}\Pr(T < -2.206 \,|H_0 \text{ true}) &= \Pr(T > 2.206 \,|H_0 \text{ true}) \\ &= 0.018601.\end{aligned}$$

Here, then

$$p_L = 0.018601 < \alpha/2 = 0.025$$

so that

$$p = 2\min(p_L, p_U) = 0.037203 < \alpha = 0.05,$$

leading to rejection of the null hypothesis, as in Section 14.6.2.

14.7.4 Large sample p values

Whether for population means or proportions, hypothesis tests for these are based on the large sample normality of the test statistic

$$T = \frac{\bar{X} - \mu_0}{\sqrt{S^2/n}} :$$

$$T \sim N(0,1) \qquad \text{approximately}$$

under the null hypothesis. Once the sample value of T has been obtained, the arguments laid in Sections 14.7.1 and 14.7.2 can be applied directly.

To illustrate, the example of a test on a proportion in Section 14.6.4 is used. The alternative hypothesis is upper one tailed, and the value of the test statistic T is

$$t = 1.2,$$

so that the p value is

$$\Pr(T > 1.2 \,|H_0 \text{ true}) = 0.11507.$$

Since this exceeds the level of significance $\alpha = 0.01$ originally chosen for the example, we accept the null hypothesis, as before.

14.8 Why use p values?

As presented, the calculation of p values is yet another way to carry out a hypothesis test. In the case of tests based on the t distribution, it is more difficult for a student to carry out the test using a p value compared with the classical approach. Why then are they used?

Hypothesis tests have been motivated by the question: *is the sample evidence too extreme compared with the hypothesised value of the population parameter?* The construction of hypothesis tests relies on finding a measure of "extremeness", but these measures are different for different sampling distributions. The logic of the arguments presented above show that p values are also a measure of extremeness, but no matter what sampling distribution is used, p values appear on a common scale, the interval $(0,1)$. This common scale, independent of context, is one reason for the use and importance of p values, but there are other reasons.

We noted in constructing hypothesis tests that the rejection region is designed to deliver a pre-specified Type I error probability or level of significance. This is *conventionally* chosen to be 5% or perhaps 1%. Why? There is no simple answer to this. In principle, one can attempt to evaluate "losses" which might be incurred if a Type I error or a Type II error occurs, and then choose the level of significance to reflect the relative importance of these losses. In practice this is never done, so this argument does not help to explain the essentially arbitrary choice of $\alpha = 0.05$.

The advantage of a p value is that it can be interpreted as a measure of the strength of the sample evidence for or against a null hypothesis:

- the smaller is the p value, the stronger the sample evidence against the null hypothesis;
- the larger the p value, the stronger the sample evidence in favour of the null hypothesis.

In effect, one steps away from the hypothesis testing context towards evaluating the *weight of evidence* about the null hypothesis: **decisions** about its truth need not be made. In practice, such decisions are still made, and the p value used to refine these decisions. The point here is that the p values (for example)

$$p = 0.045, \qquad p = 0.00001$$

both lead to a rejection at a 5% significance level, but one of them conveys much stronger information about the weight of evidence against the null hypothesis than the other.

There is another, perhaps more powerful, reason to explain the importance of p values in statistical practice. One has to accept that all of the calculations required for statistical inference discussed in this course are routinely performed using a statistical package like SPSS, SAS, R, etc. etc., and these packages always produce p values. Rather than looking at the value of a test statistic, then finding a critical value from tables, and then making a decision, one simply looks at the p value produced by the computer package.

14.9 Using confidence intervals for one sided tests

In Sections 14.2 and 14.5, it was argued that the confidence interval procedure was really designed for two-sided alternative hypotheses, i.e. for testing that

$$H_0 : \mu = \mu_0 \quad \text{against} \quad H_A : \mu \neq \mu_0.$$

So, the null hypothesis is rejected if

$$\text{either} \quad \mu_0 < c_L = \bar{x} - z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} \quad \text{or} \quad \mu_0 > c_U = \bar{x} + z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}},$$

Many students find it economical to use this **specific** rejection criterion to handle the case of one-sided alternative hypotheses as well. It is **not** correct to do this. Fortunately, it is easy to modify the confidence interval procedure to produce appropriate rejection criteria for testing one-sided alternatives.

Recall from Section 14.2.1 that the rejection criterion above is equivalent to a rejection criterion using critical values :

$$\text{either} \quad \bar{x} > \bar{x}_U = \mu_0 + z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} \quad \text{or} \quad \bar{x} < \bar{x}_L = \mu_0 - z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}.$$

The link between the two sets of rejection criteria is clear:

$$\begin{aligned} \bar{x} &> \bar{x}_U \Longleftrightarrow \mu_0 < c_L, \\ \bar{x} &< \bar{x}_L \Longleftrightarrow \mu_0 > c_U. \end{aligned}$$

With this link in mind, we can use the critical value rejection rules for one sided alternatives to generate corresponding rules using the appropriate confidence bound.

For upper one sided alternative hypotheses, the rejection rule is (see Section 14.5.1)

$$\bar{x} > \bar{x}_U = \mu_0 + z_{\alpha}\sqrt{\frac{\sigma^2}{n}}$$

and for lower one sided alternative hypotheses, the rejection rule is (see Section 14.5.3)

$$\bar{x} < \bar{x}_L = \mu_0 - z_{\alpha}\sqrt{\frac{\sigma^2}{n}}.$$

Each of these is designed to set the Type I error probability or level of significance at α.

Translating into corresponding confidence bounds, we would have to replace $z_{\alpha/2}$ in c_L and c_U by z_α :

$$c_L = \bar{x} - z_{\alpha}\sqrt{\frac{\sigma^2}{n}}, \qquad c_U = \bar{x} + z_{\alpha}\sqrt{\frac{\sigma^2}{n}}.$$

These confidence bounds do **NOT** correspond to a $100\,(1-\alpha)\,\%$ confidence interval. Rather, they correspond to a $100\,(1-2\alpha)\,\%$ confidence interval.

This is the key point. So, for example, to carry out a 5% one sided test, one has to construct first a 90% confidence interval, **NOT** a 95% confidence interval.

Once the correct $100\,(1-2\alpha)\,\%$ confidence interval is calculated, use the **upper** confidence bound c_U, in the case of a **lower** one tailed alternative, to reject the null hypothesis if

$$\mu_0 > c_U,$$

since this corresponds to

$$\bar{x} < \bar{x}_L.$$

Use the **lower** confidence bound c_L, in the case of an **upper** one tailed alternative, to reject the null hypothesis if

$$\mu_0 < c_L,$$

since this corresponds to

$$\bar{x} > \bar{x}_U.$$

Although this discussion uses the original simplified case of sampling from $N\left(\mu, \sigma^2\right)$ with σ^2 known, it is clear that the reasoning carries over to all the other cases discussed above, simply by changing the percentage points as necessary.

Consider the example of Section 14.5.2. The aim is to test

$$H_0 : \mu = 20 \quad \text{against} \quad H_A : \mu > 20$$

in sampling from $N\,(\mu, 5)$ using a 5% level of significance. To use the appropriate confidence interval procedure, we need the 90% confidence interval

$$[c_L, c_U] = \bar{x} \pm z_{\alpha/2}\,\mathrm{SE}\left(\bar{X}\right),$$

where

$$\alpha = 0.1, \qquad z_{\alpha/2} = 1.6449.$$

Since $\bar{x} = 20.7, n = 50$, we have

$$\begin{aligned} [c_L, c_U] &= 20.7 \pm (1.6449)\sqrt{\frac{5}{50}} \\ &= 20.7 \pm 0.52 \\ &= [20.18, 21.12]. \end{aligned}$$

If $\mu_0 < c_L$, we should reject the null hypothesis. Since $\mu_0 = 20 < c_L = 20.18$, we reject the null, as in Section 14.5.2.

14.9.1 Exercise 8

1. Imagine that you are a member of a team of scientific advisors considering whether genetic modification of crops has any health consequences for the population at large. Having some knowledge of statistics, you set up the issue as one of hypothesis testing.

(a) What would your null and alternative hypotheses be?

(b) Explain the interpretation of a Type I error and a Type II error in this context.

(c) What sort of costs would arise as a consequence of each type of error?

(d) What sort of sample evidence would be needed to enable a statistical conclusion to be reached?

(e) If suitable sample evidence were available, what advice would you give about the strength of the evidence that would be required to reject your null hypothesis?

2. Weekly wages in a particular industry are known to be normally distributed with a **standard deviation** of £2.10. An economist claims that the mean weekly income in the industry is £72.40. A random sample of 35 workers yields a mean income of £73.20.

   (a) What null hypothesis would you specify?

   (b) Without any further information, explain the justification for choosing

      i. a two tailed alternative hypothesis;

      ii. an upper one tailed alternative hypothesis.

   (c) Perform the tests for each of these alternative hypotheses in turn, using

      i. p values

      ii. classical hypothesis tests

      iii. a suitable confidence interval procedure

      at a 5% level of significance.

3. This question is a version of Question 2.

   Weekly wages in a particular industry are known to be normally distributed, with an unknown variance. An economist claims that the mean weekly income in the industry is £72.40. A random sample of 15 workers gives a sample mean of £73.20 and a sample standard deviation of £ 2.50. Redo part (c) of Question 2 with this new information. You will need to use EXCEL to compute the p values required.

4. A motoring organisation is examining the reliability of imported and domestically produced vans. Service histories for 500 domestically made and 500 imported vans were examined. Of these, 159 domestically produced vans and 121 imported vans had repairs for breakdowns. Test the hypothesis that the true proportion of breakdowns to be expected in the two populations of vans is 0.5,

(a) using an upper one sided alternative hypothesis for domestically produced vans;

(b) using a two-sided alternative hypothesis for imported vans.

Chapter 15

EXTENSIONS TO THE CATALOGUE

It should have become apparent that statistical inference appears to deal with a catalogue of cases:

- inference on a population mean μ in sampling from $N\left(\mu, \sigma^2\right)$ when σ^2 known;
- inference on a population mean μ in sampling from $N\left(\mu, \sigma^2\right)$ when σ^2 unknown;
- inference on a population mean μ in sampling from a population with mean μ and variance σ^2 when the sample is large;
- inference on a population proportion when the sample is large.

In reality, the catalogue is extremely extensive. Introductory statistics courses typically discuss more entries in the catalogue than have been covered in this course. This section covers some of these additional cases.

Please note that NONE of the topics in this Section are EXAMINABLE. They may be ignored with impunity.

The reason for these extensions is that they help to show how the principles developed in earlier sections carry over easily to other situations. In addition, these extensions are practically useful.

15.1 Differences of population means

Many types of investigation involve comparisons of population means. Is population mean income for men and women the same? Does school A have the same average reading score as school B? Other examples refer to a *treatment effect.* Some kind of treatment - a new reading scheme, a training scheme, a new drug treatment, is conceptually applied to a population. Is the post-treatment population mean an improvement (i.e. in whatever the appropriate direction) over the pre-treatment population mean?

To formalise this, let $X \sim N\left(\mu_X, \sigma_X^2\right)$ describe one population, and $Y \sim N\left(\mu_Y, \sigma_Y^2\right)$ describe the other. The questions above are most naturally phrased in terms of the difference of the two population means, $\mu_X - \mu_Y$. The first, most basic question is how can one estimate this parameter. For, once this is dealt with, the sampling distribution of the estimator can be used to construct a confidence interval for $\mu_X - \mu_Y$, or tests of hypotheses about $\mu_X - \mu_Y$.

Estimating μ_X is easy: use the sample mean $\bar{X}$ of a random sample from the distribution of X. The same is true for μ_Y : use the sample mean $\bar{Y}$. There is no reason why the same size of sample has to be drawn from each distribution, so suppose that a sample of size n_1 is drawn from X, and n_2 from Y. Each of the population mean estimators is unbiased, so it should follow that

$$\bar{X} - \bar{Y}$$

is an unbiased estimator of $\mu_X - \mu_Y$.

15.1.1 The sampling distribution of $\bar{X} - \bar{Y}$

In deriving this distribution, we have to know how sampling from the X distribution and sampling from the Y distribution are related. Does making a drawing from the X distribution influence what is drawn from the Y distribution? Usually, we would like the answer to this question to be *no,* in which case, we can assume that the samples are **independent** of each other, thus making the sample random variables

$$X_1, ..., X_{n_1} \qquad \text{and} \qquad Y_1, ..., Y_{n_2}$$

independent of each other. In turn, this makes $\bar{X}$ and $\bar{Y}$ independent random variables.

Given independence, we can deduce from

$$\bar{X} \sim N\left(\mu_X, \frac{\sigma_X^2}{n_1}\right), \qquad \bar{Y} \sim N\left(\mu_Y, \frac{\sigma_Y^2}{n_2}\right)$$

that

$$\bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}\right).$$

Although the parameters are different, this has **exactly** the same **structure** as the sampling distribution of $\bar{X}$. If the parameters σ_X^2 and σ_Y^2 are **known,** then we can use **exactly** the same arguments as in the case of a single population mean μ to construct confidence intervals and hypothesis tests using the standardised random variable

$$Z = \frac{\left(\bar{X} - \bar{Y}\right) - \left(\mu_X - \mu_Y\right)}{\sqrt{\dfrac{\sigma_X^2}{n_1} + \dfrac{\sigma_Y^2}{n_2}}} \sim N\left(0, 1\right).$$

So, details are not discussed here.

15.1.2 Unknown variances

What happens if we abandon the simplifying assumption that σ_X^2 and σ_Y^2 are known? Arguing by analogy with Section 13.4, one might anticipate replacing these variance parameters by their unbiased estimators S_X^2 and S_Y^2. However, the random variable

$$\frac{\left(\bar{X} - \bar{Y}\right) - \left(\mu_X - \mu_Y\right)}{\sqrt{\dfrac{S_X^2}{n_1} + \dfrac{S_Y^2}{n_2}}}$$

does **not** have a t distribution. Finding the sampling distribution here is a "well-known", old problem in statistics called the *Behrens-Fisher* problem, which we need not investigate.

In the spirit of making assumptions to simplify a difficult problem, we make the conventional assumption here. This is to **assume that**

$$\sigma_X^2 = \sigma_Y^2 = \sigma^2 :$$

i.e. assume that the population variances are the same. There is no reason why this should be true in practice: indeed, this is the chief objection to the sampling distribution to be outlined and used below.

Under this *common variance* assumption,

$$\operatorname{var}\left[\bar{X} - \bar{Y}\right] = \frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2} = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right).$$

The common variance is still unknown, and will have to be estimated. Both S_X^2 and S_Y^2 are still unbiased for σ^2, but there is no reason why the estimates from each sample will be equal. It is therefore reasonable to combine or **pool** the X and Y samples to construct an estimator of σ^2.

However, the ordinary sample variance estimator based on the pooled sample cannot be used because each sample comes from a population with a different population mean, and an estimator of σ^2 which reflects this has to be used. This is constructed as a weighted average of S_X^2 and S_Y^2. The weights are

$$\frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)}, \qquad \frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)},$$

being almost equal to the relative size of each sample to the pooled sample. The **pooled variance estimator** is

$$S_p^2 = \frac{(n_1 - 1)\, S_X^2 + (n_2 - 1)\, S_Y^2}{(n_1 - 1) + (n_2 - 1)}$$

Why is this complication helpful? Because it can be shown that the random variable

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{S_p^2 \left(\dfrac{1}{n_1} + \dfrac{1}{n_2} \right)}} \sim t_{n_1 + n_2 - 2}.$$

Notice that the degrees of freedom in the t distribution is the value in the denominator of S_p^2.

So, again by re-defining variables names, the logic for confidence intervals or hypothesis tests for $\mu_X - \mu_Y$ on hypothesis tests using the t distribution again carries over straightforwardly. However, since there are some important detail changes, two examples will be given

The interval estimator or confidence interval formula for $\mu_X - \mu_Y$ follows that given in Section 13.4.3 as

$$[C_L, C_U] = \bar{X} - \bar{Y} \pm t_{n_1 + n_2 - 2, \alpha/2} \sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)},$$

with sample confidence bounds ("confidence interval")

$$\bar{x} - \bar{y} \pm t_{n_1 + n_2 - 2, \alpha/2} \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}.$$

Suppose that independent random samples of size 6 and 5 respectively are drawn independently from two normal distributions with unknown means and variances, μ_X, σ_X^2, and μ_Y, σ_Y^2. The sample information is

$$\begin{aligned} n_1 &= 6, \quad \bar{x} = 19, \quad s_X^2 = 100; \\ n_2 &= 5, \quad \bar{y} = 25, \quad s_Y^2 = 64. \end{aligned}$$

We construct a 90% confidence interval for $\mu_X - \mu_Y$. First, assume a common population variance and calculate the pooled variance estimate:

$$\begin{aligned} s_p^2 &= \frac{(n_1 - 1)\, s_X^2 + (n_2 - 1)\, s_Y^2}{(n_1 - 1) + (n_2 - 1)} \\ &= \frac{(5)\,(100) + (4)\,(64)}{9} \\ &= 84. \end{aligned}$$

For a 90% confidence interval we need

$$t_{9,0.05} = 1.833.$$

Then, the confidence interval is

$$\begin{aligned}
\bar{x} - \bar{y} \pm t_{n_1+n_2-2,\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} &= -6 \pm (1.833)\sqrt{84\left(\frac{1}{6}+\frac{1}{5}\right)} \\
&= -6 \pm (1.833)(5.550) \\
&= -6 \pm 10.173 \\
&= [-16.173, 4.173].
\end{aligned}$$

Next, a hypothesis test using different data, but involving the same ideas. This question is whether mean household income in city A is the same as in city B. Independent random samples are drawn from

$$X_A \sim N\left(\mu_A, \sigma_A^2\right), \qquad X_B \sim N\left(\mu_B, \sigma_B^2\right)$$

to test the hypotheses

$$H_0 : \mu_A = \mu_B \quad \text{against} \quad H_A : \mu_A \neq \mu_B,$$

at a 1% level of significance. The sample information is

	n	$\bar{x}$	s^2
A	10	20.8	8.65
B	9	15.7	5.82

Again it is necessary to make the common population variance assumption,

$$\sigma_A^2 = \sigma_B^2 = \sigma^2.$$

The hypothesis test will be based on large or small values of the test statistic

$$T = \frac{\bar{X}_A - \bar{X}_B - (\mu_A - \mu_B)}{\sqrt{S_p^2\left(\frac{1}{n_A}+\frac{1}{n_B}\right)}} \sim t_{n_A+n_B-2}$$

where

$$S_p^2 = \frac{(n_A-1)S_A^2 + (n_B-1)S_B^2}{(n_A-1)+(n_B-1)}.$$

Note that under the null hypothesis,

$$\mu_A - \mu_B = 0$$

and so can be dropped from T.

The calculations for s_p^2 are

$$\begin{aligned} s_p^2 &= \frac{(n_A - 1)\, s_A^2 + (n_B - 1)\, s_B^2}{(n_A - 1) + (n_B - 1)} \\ &= \frac{(9)\,(8.65)^2 + (8)\,(5.82)^2}{9 + 8} \\ &= 55.55186, \end{aligned}$$

and the sample value of T is

$$\begin{aligned} t &= \frac{\bar{x}_A - \bar{x}_B}{\sqrt{s_p^2 \left(\dfrac{1}{n_A} + \dfrac{1}{n_B} \right)}} \\ &= \frac{20.8 - 15.7}{\sqrt{(55.55186)\left(\frac{1}{10} + \frac{1}{9}\right)}} \\ &= 1.4892. \end{aligned}$$

Under H_0, $T \sim t_{n_A + n_B - 2}$ i.e. 17 df, and for a 1% test, we need the value $t_{17,\alpha/2}$ such that

$$\Pr\big(T > t_{17,\alpha/2}\,|H_0 \text{ true}\big) = \alpha = 0.01.$$

Using the Appendix, the column headed 0.995 has to be used, giving

$$t_{17,\alpha/2} = 2.898.$$

For the p value, we need only calculate the upper p value, since $t > 0$, and this is

$$\begin{aligned} \Pr\left(T > 1.4892\,|H_0 \text{ true}\right) &= \texttt{tdist(1.4892, 17, 1)} \\ &= 0.077376, \end{aligned}$$

making the p value

$$p = 0.154752.$$

Both pieces of information lead to the same conclusion - do not reject the null hypothesis.

15.1.3 Large sample ideas

The t test arguments for $\mu_X - \mu_Y$ tend to get labelled as a *small sample* procedure, to be used when the population variances are unknown. However, one sometimes gets the impression from textbooks that in testing the difference of two means, with variances unknown, that one *must* make the common variances assumption. If this is not actually true, then this is not a good assumption to make. One has to accept that little can be done with small samples, **unless** the common variances assumption is reasonable. In any case, as has been argued before, sampling from normal distributions is typically an **assumption**, and therefore not necessarily true.

Provided that the two sample sizes are sufficiently large, the random variable

$$T = \frac{\left(\bar{X} - \bar{Y}\right) - \left(\mu_X - \mu_Y\right)}{\sqrt{\dfrac{S_X^2}{n_1} + \dfrac{S_Y^2}{n_2}}}$$

does have an approximate standard normal distribution,

$$T \sim N\left(0, 1\right) \quad \text{approximately,}$$

under the null hypothesis. Again, this puts us back into the context of confidence intervals and tests for normal distributions, as discussed in Section 14.6.3. There is therefore no need to give more details.

15.1.4 Paired Samples

The introduction to this section mentioned situations in which shifts in population means might occur because some "treatment" has been administered. At a practical level, the same treatment administered to two different subjects may produce wildly different effects basically due to the difference in natures of the two subjects. Clearly it would be desirable to use a procedure which eliminated as many of these subject-specific effects as possible, so that any observed differences can be ascribed solely to the treatment.

One possible strategy here is to measure a subjects' response (whatever this is) before the treatment is administered, and then again after the treatment. The treatment effect for the individual is then the difference in responses. The point is that differences in the **level** of response due to differences in subjects are eliminated in this method. So, the responses are "paired".

A simple example will help.

Concerned with the obesity of some of its citizens, a small town wants to instigate a fitness campaign. Before spending large amounts of money on the fitness campaign, it carries out a trial using a sample of 6 volunteers. The "before" and "after" body weights (in kg) are

Individual	1	2	3	4	5	6
Before	76	89	70	83	77	74
After	71	88	67	80	73	75
Differences	-5	-1	-3	-3	-4	1

There is some slight evidence from the sample that weight loss for the larger weights is smaller than that for the smaller weights. In other words, there *is* some evidence of subject-specific effects.

If we suppose that for each individual, the *before* and *after* weights are normal random variables, it will follow that the difference D_i is also normal, with some mean and variance. Treating the mean **and** the variance as unknown avoids any difficulties in finding out exactly what the parameters of this normal distribution are. The only other required assumption is that the sample of differences

$$D_1, ..., D_6$$

form a random sample from

$$D \sim N\left(\mu, \sigma^2\right).$$

Detecting an effect of the treatment can then be done by performing a t test of the hypotheses

$$H_0 : \mu = 0 \quad \text{against} \quad H_A : \mu < 0.$$

The usual sample information is

$$\bar{x} = -2.5, s^2 = 4.7,$$

so that the value of the test statistic is

$$t = \frac{-2.5}{\sqrt{4.7/6}} = -2.8247.$$

Performing the test at a 5% level, the critical value is

$$-t_{5,0.05} = -2.015,$$

with the obvious conclusion of rejecting the null hypothesis. The p value is

$$\Pr\left(T < -2.8247 \,|H_0 \text{ true}\right) = 0.018452,$$

confirming the conclusion.

15.2 Differences of proportions

Exactly the same sort of ideas as in Section 15.1 apply here. Two populations may have different population proportions, even if they are both unknown. Can one estimate the difference in proportions, either point or interval, and test hypotheses about the difference? The answer is of course yes, using the large sample normal approximation for each sample proportion, as outlined in Sections 12.3, 13.6.3, and 14.6.4.

So, imagine that independent random samples are drawn from distributions of two independent Bernoulli random variables X and Y, producing sample proportions P_1 and P_2. Each of these, as estimators of the population proportions π_1, π_2, is approximately normal in large samples:

$$P_i \sim N\left(\pi_i, \frac{\pi_i(1-\pi_i)}{n_i}\right), \qquad \text{approximately,}$$

so that

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

The fact that π_1 and π_2 are unknown requires them to be estimated to provide another approximate distribution:

$$T = \frac{(P_1 - P_2) - (\pi_1 - \pi_2)}{\sqrt{\dfrac{P_1(1-P_1)}{n_1} + \dfrac{P_2(1-P_2)}{n_2}}} \sim N(0,1) \qquad \text{approximately.}$$

An approximate confidence interval is easily deduced from this as

$$[C_L, C_U] = P_1 - P_2 \pm z_{\alpha/2}\sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}$$

with sample value

$$p_1 - p_2 \pm z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}.$$

For hypothesis tests, the same sort of arguments apply: we might wish to test

$$H_0 : \pi_1 = \pi_2 \qquad \text{against} \qquad H_A : \pi_1 \neq \pi_2.$$

The variance of $P_1 - P_2$ involves both π_1 and π_2, which are assumed equal under the null hypothesis. This type of issue also arose in the case of the difference of two means in Section 15.1.2, where the population variances are unknown. It is therefore natural to ask if we should use a **pooled** estimator of

$$\pi = \pi_1 = \pi_2$$

to compute an estimator of the variance of $P_1 - P_2$. The answer here is *perhaps.* Whether or not a pooled estimator of π is used has no effect on the sampling distribution of the test statistic: it is still $N(0,1)$ under the null hypothesis. In the case of the difference of two means the use of the pooled variance estimator was **essential** to ensure that we obtained a test statistic with a t distribution under the null hypothesis.

So, for the sake of simplicity, we will use separate estimators of π_1 and π_2 in such hypothesis tests, and therefore use the test statistic T above. This means that hypothesis tests for the difference of two proportions follows much the same lines as tests on a single proportion.

15.3 Overview

One can see from the discussion in this section that there are some basic principles being used in statistical inference, no matter what the specific context nor the specific sampling distribution. But it is also unfortunately true that there are cases where there are serious detail differences that have to dealt with. It is in this sense that we referred to a *catalogue* of cases.

15.4 Exercise 9

1. A simple random sample of 15 pupils attending a certain school is found to have an average IQ of 107.3, whilst a random sample of 12 pupils attending another school has an average IQ of 104.1. Obtain the 95% confidence interval for the difference between the mean IQ's of the pupils at the two schools when

   (a) the true variances of the IQ's for the children at the two schools are 39 and 58 respectively;

   (b) the true variances are unknown, but the sample variances are 32.5 and 56.5 respectively.

   In both cases, state any assumptions you make. For both parts of the question, do you consider that there is strong evidence that pupils attending the first school have higher mean IQ than those attending the second?

2. A drug manufacturer has two products which should contain the same amount of "Super E" additive. As part of the quality control process, regular tests are to be made to check this. Because of the costs involved in performing the analysis, these checks are based on small samples. It has been suspected for some time that Product 2 contains less of the additive than Product 1. A sample from each product yields the following results:

Product	n	$\bar{x}$	s^2
1	8	147	0.088
2	7	142	0.035

   Conduct a hypothesis test at a significance level of 10%, clearly stating any assumptions you have to make. Do you think these assumptions are reasonable?

3. The scores of males who play ASTRO (a video game) after drinking a pint of water are normally distributed with mean μ_1, whilst the scores

of males who play this game after drinking a pint of beer are normally distributed with mean μ_2.

Four male students play ASTRO after drinking a pint of water, and then again after drinking a pint of beer. The scores are

Name:	Chris	Martin	Alex	Nick
After Water	120	125	135	100
After Beer	115	115	95	120

(a) Compute a 95% confidence interval for the difference in means $\mu_1 - \mu_2$. Hint: you will have to calculate a suitable sample variance.

(b) Briefly interpret this confidence interval.

(c) Does it contain the value 0? What would be the interpretation of this?

(d) Have you had to make any assumptions in computing this confidence interval?

4. In a survey of 60 women and 100 men, 60% of women favour a ban on smoking in restaurants, whilst 45% of men favour such a ban. Find a 95% confidence interval for the difference between the proportions of all men and all women in favour of a ban. State any assumptions you make in computing this confidence interval.

5. Using the information in Question 4 of Exercise 6, test the hypothesis that the proportions of breakdowns to be expected for domestically produced and imported vans are equal.

Appendix A

Statistical Tables

Statistical tables for Standard Normal and Student t distributions.

Standard Normal Distribution Function

The table provides values of p where $\Pr(Z \leq z) = p$.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361
2.5	.99379	.99396	.99413	.99430	.99446	.99461	.99477	.99492	.99506	.99520
2.6	.99534	.99547	.99560	.99573	.99585	.99598	.99609	.99621	.99632	.99643
2.7	.99653	.99664	.99674	.99683	.99693	.99702	.99711	.99720	.99728	.99736
2.8	.99744	.99752	.99760	.99767	.99774	.99781	.99788	.99795	.99801	.99807
2.9	.99813	.99819	.99825	.99831	.99836	.99841	.99846	.99851	.99856	.99861
3.0	.99865	.99869	.99874	.99878	.99882	.99886	.99889	.99893	.99896	.99900
3.1	.99903	.99906	.99910	.99913	.99916	.99918	.99921	.99924	.99926	.99929
3.2	.99931	.99934	.99936	.99938	.99940	.99942	.99944	.99946	.99948	.99950
3.3	.99952	.99953	.99955	.99957	.99958	.99960	.99961	.99962	.99964	.99965
3.4	.99966	.99968	.99969	.99970	.99971	.99972	.99973	.99974	.99975	.99976
3.5	.99977	.99978	.99978	.99979	.99980	.99981	.99981	.99982	.99983	.99983
3.6	.99984	.99985	.99985	.99986	.99986	.99987	.99987	.99988	.99988	.99989
3.7	.99989	.99990	.99990	.99990	.99991	.99991	.99992	.99992	.99992	.99992
3.8	.99993	.99993	.99993	.99994	.99994	.99994	.99994	.99995	.99995	.99995
3.9	.99995	.99995	.99996	.99996	.99996	.99996	.99996	.99996	.99997	.99997
4.0	.99997	.99997	.99997	.99997	.99997	.99997	.99998	.99998	.99998	.99998

Student's t Distribution Function for Selected Probabilities

The table provides values of c where $\Pr(t_\nu \leq c) = p$.

p	0.750	0.800	0.900	0.950	0.975	0.990	0.995	0.9975	0.999	0.9995
ν	Values of c									
1	1.000	1.376	3.078	6.314	12.706	31.821	63.657			
2	0.816	1.061	1.886	2.920	4.303	6.965	9.925			
3	0.765	0.978	1.638	2.353	3.182	4.541	5.841			
4	0.741	0.941	1.533	2.132	2.776	3.747	4.604			
5	0.727	0.920	1.476	2.015	2.571	3.365	4.032	4.773		
6	0.718	0.906	1.440	1.943	2.447	3.143	3.707	4.317	5.208	
7	0.711	0.896	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.688	0.861	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
70	0.678	0.847	1.294	1.667	1.994	2.381	2.648	2.899	3.211	3.435
80	0.678	0.846	1.292	1.664	1.990	2.374	2.639	2.887	3.195	3.416
90	0.677	0.846	1.291	1.662	1.987	2.368	2.632	2.878	3.183	3.402
100	0.677	0.845	1.290	1.660	1.984	2.364	2.626	2.871	3.174	3.390
110	0.677	0.845	1.289	1.659	1.982	2.361	2.621	2.865	3.166	3.381
120	0.677	0.845	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	0.674	0.842	1.282	1.645	1.960	2.326	2.576	2.808	3.090	3.297