University of the Witwatersrand

STAT2012A

Introduction to Mathematical Statistics II

UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

Lecturer: Anna Kaduma Gumbie

Contents

Preface

Whether you are doing this course by choice, or as a filler or because you are forced to; the aim of this course is to introduce you to Mathematical Statistics and to realise its applications. The notes will attempt to clarify links between the techniques taught and the interpretation of the results obtained. It will not be assumed that you are familiar with any statistical techniques at the onset of this course.

The best way for you to study this book is to work in systematic manner through all the chapters. After understanding a chapter, you should be able to answer the questions at the end of each chapter, to move onto the next chapter.

These notes have been revised from the original notes compiled by Dr. Charles Chimedza and Ms. Nothabo Ndebele, and a revision by Dr. Raeesa Ganey. The chapter on Probability has been adapted from the Advanced Level Mathematics Statistics 1 written by Steve Dobbs and Jane Miller, published by Cambridge University Press 2002.

Chapter 1

What is Statistics?

1.1 Introduction

Statistics a field of study is not primarily the adding of numbers to come to the conclusion that there are 33 346 registered students at Wits and 30% of them are Science students in the year 2014, or that the demographic profile of First Year students consists of more females (54%) than males (46%).

Statistics, rather can be described as the science of **decision making in the face of uncertainty**. The emphasis here is not placed so much on the collection of data, but rather **drawing conclusions** from the data.

Mathematical Statistics is the application of mathematical concepts to decision making in Statistics. In this course, the focus will be placed on parts from Linear Algebra and Mathematical Analysis.

On the other hand, a *statistic* (plural: *statistics*), is an estimate of a *parameter*. A parameter is a characteristic of a population.

1.2 Population and Samples

A **population** is the collection of items under investigation. It may be finite, or infinite. However, when doing a statistical analysis, it is not always practical to collect all the data of a population, since a population - even if finite - could be very large. This could then mean that collection of data is not done in a timely manner, and/or it could be very costly. Examples could include full information on every birth in South Africa in a particular year, or the details of all the stars in the sky.

Given the limitations of collecting data of a population, a **sample** is rather drawn to make a conclusion. A sample is a subset of the population and it is selected

using **approved scientific methods** such that it can represent the population as accurately as possible. This means that any conclusions made from or using the sample, can be generalized to the entire population. Therefore, great care should be taken when selecting a sample and collecting the data. **Sampling** is a statistical technique used to select a sample from a population.

Descriptive Statistics as a statistical method, is a method used to summarise patterns (if any) observed from the sample, so that conclusions can be drawn without scrutinising each and every observation. The data is summarized in a meaningful way, so patterns or trends can be seen. Some examples of descriptive statistics are graphs, tables or even **statistics** such as the mean, median and mode.

Inferential Statistics on the other hand, is a method used to make inferences and predictions about a population based on a sample collected. The *estimation of parameters* and *testing of statistical hypotheses* are the primary methods of inferential statistics.

1.3 Types of variables

1.3.1 Qualitative Variables

A **qualitative variable** is a variable whose values assigned in measurements are usually a representation of a category of the characteristic being measured. Examples here would include colour, satisfaction level, gender, etc.

1.3.2 Quantitative Variables

Discrete Variables:

A **quantitative discrete variable** is a quantity that varies from item to item under observation that can assume any of a prescribed set of whole values/numbers. This set of values is called the **domain** of the variable. Think of a throwing dice, it can only take the values $1, 2, 3, 4, 5$ and 6. The values a discrete variable can take are integer values. The type of data which results from observations of this type of variable is *quantitative discrete data*.

Continuous Variables:

A **quantitative continuous variable** can take any value in a given *interval*. Example can be the height of a student that could can be 158.7cm, 164.2cm or 168.9cm.

The values lie on a continuous scale. The type of data which results from this measurement is *quantitative continuous data*.

1.4 Measurement Scales

This section looks at how data is measured, which refers to the different scales of measurement.

Measurement is the assignment of names or numerals to objects or events according to certain rules. The four common types of measurement scales are:

- Nominal,
- Ordinal,
- Interval and
- Ratio.

Nominal

This is the weakest/most general of the four measurement scales of data. It distinguishes one object or event from another on the basis of a 'name'. An example of this, is classifying items coming off an assembly line as defective or non-defective, or classifying a bank account as open or closed. The 'naming' can be coded, e.g. if the bank account is open, then use the value 1, and closed use the value 2. Data of this type are typically refereed to as: count data, frequency data or categorical data. Notes that order or rank of values/names in this case does not matter.

Ordinal

Objects or events are distinguished on the basis of the relative amounts of some characteristic they possess. These measurements enable observations to be ranked (ordered). An example is ranking different sized jerseys from smallest to largest by assigning the smallest as rank = 1 and increasing the rank by 1 up to the largest jersey with a rank = 4, say. Note that the magnitude of the difference between measurements is not reflected in the rank.

Interval Scale

This scale is applied when objects, or events can be distinguished one from another and ranked, and when the differences between measurements have meaning. Suppose that four objects A, B, C and D are assigned scores of 20, 30, 60 and 70 respectively. If the interval scale is used then we can say that the difference between A and B is equal to the difference between C and D, i.e. there are equal differences in the amount of trait or characteristic being measured. However the ratios of the scores cannot be used. The score of 60 for C does not mean that C has twice as much of trait as B which has a score of 30. The values 20, 30, 60 and 70 are scores assigned, and not measurements as it were.

Ratio scale

This kind of scale applies to all scales above and has the additional property that the ratios are meaningful. This scale includes the familiar measurements of height, weight, etc. that is quantitative data. The difference in magnitude and the ratio can all be used for analysis as they have meaning attached to it.

Exercise 1.1

1. The banks in South Africa are assigned positions according to their reported profits. The bank with the highest profit is given position 1, the bank with the second highest profit is given position 2, etc. What type of data is this?

2. If the actual profit for each bank is recorded, what type of data would it be?

1.5 Uses of Statistics

One could possibly use this entire book to write on the uses of statistics and application of it in different kind of fields. From forecasting future trends of the exchange rate, to modelling rainfall patterns to provide food security, to understanding astronomical data, the uses and applications are endless. However, some of you might question why you have to do a course in Statistics and how will it actually help in your career. Although this course is offered as an introductory course, there are techniques that will be taught that provide to be very useful and can be easily applied to daily life.

1.6 Conclusion

Statistics plays an important role in almost every field as it helps in decision making. Before any decision can be made about any business or society or problem it is often necessary to gather enough data or information to support a decision being made. Statistics helps in the collection of information in a scientific and systematic manner, and to make decisions based on the descriptive and inferential statistics.

Chapter 2

Descriptive Statistics

This chapter of descriptive statistics is split into two parts : **graphical summary**, which includes presenting information summarised in tables and graphs and **numerical summary** which involves presenting information summarised after numerical calculations.

2.1 Graphical Summary

A picture is worth a thousand words, or numbers, and there is no better way of getting a 'feel' for the data than to display them in a figure or graph. The general principle should be to convey as much information as possible in the figure, with the constraint that it is not overwhelmed by too much detail. Graphical techniques are therefore representations of data such that the main features of the data are captured.

2.1.1 Tabulation

Data is typically presented in a tabular form. However, the data can be summarised in a simple and easier way to understand and further analyse. Suppose data is recorded on the gender of each lecturer in a school, and the results are presented in the following way:

male	male	female	male	female	female	male
female	male	female	male	male	female	male

By looking at this information, the gender distribution is not immediately clear. It becomes more difficult if there were more data, say 1000 lecturers. A better way of presenting this information without losing any of the detail, is *tabulation*.

Gender	Number of lecturers
Male	8
Female	6
Total	14

Table 2.1: Number of lecturers by gender

This is especially useful if the number of observations is large and the distinct categories are few. This leads to the idea of **contingency tables**. A contingency table is a convenient way of summarising data with more than one variable. It consists of row(s) and column(s) of data, that represent the variables. Suppose more information on the lecturers are recorded like in the table below.

Lecturer	Gender	School
1	male	Maths
2	male	Statistics
3	female	Maths
4	male	Statistics
5	female	Maths
6	female	CSAM
7	male	CSAM
8	female	CSAM
9	male	Statistics
10	female	Statistics
11	male	CSAM
12	male	Maths
13	female	Maths
14	male	Maths

Table 2.2: Number of lecturers by gender and school

This information can be presented in a contingency table as shown in Table 2.3.

School / Gender	Male	Female	Total
Maths	3	3	6
Statistics	3	1	4
CSAM	2	2	4
Total	8	6	14

Table 2.3: Contingency table of lecturers by gender and school

The contingency table in this example is a 3 x 2 table as there are 3 rows and 2 columns. The number of rows and columns is determined by the number of categories in each variable. For gender, there are 2 and for school, there are 3. Contingency tables are sometimes referred to as cross tabulations and are used for data that has two variables.

2.1.2 Pie Chart

A pie chart is a circle that is divided into segments like a pie cut into pieces from the center outwards. Each segment represents one of more values taken by a variable and is used to illustrate **proportion**. Pie charts are a useful way to organise data in order to see the size of components relative to the whole.

Table 2.4 lists the number of 91 staff members working at a company tabulated by their qualification. Each category of the qualification can be expressed by a proportion and percentage. The proportion is calculated by the number in each category divided by the total number of staff members. E.g. Engineering qualifications gives a proportion $\frac{38}{91}$, and Science $\frac{24}{91}$ and so on. The percentage is the proportion multiplied by 100.

Qualification	Frequency	Proportion	Percentage
Engineering	38	$\frac{38}{91}$	41.75 %
Science	24	$\frac{24}{91}$	26.37 %
Arts	13	$\frac{13}{91}$	14.29 %
Commerce	8	$\frac{8}{91}$	8.79 %
Medicine	5	$\frac{5}{91}$	5.49 %
Other	3	$\frac{3}{91}$	3.29 %
Total	91	1	100 %

Table 2.4: Table of staff members working in a certain company tabulated by their qualification.

A pie chart is constructed by using the proportions in Table 2.4. As the pie is a circle, the calculation of the angle in each category is the **proportion x** 360°. The pie chart is shown in Figure 2.1, with the angles calculated in Table 2.5.

Qualification	Angle
Engineering	$\frac{38}{91}$ x 360 = 150.3
Science	$\frac{24}{91}$ x 360 = 94.9
Arts	$\frac{13}{91}$ x 360 = 51.4
Commerce	$\frac{8}{91}$ x 360 = 31.6
Medicine	$\frac{5}{91}$ x 360 = 19.8
Other	$\frac{3}{91}$ x 360 = 11.8

Table 2.5: Calculation of angles in the pie chart

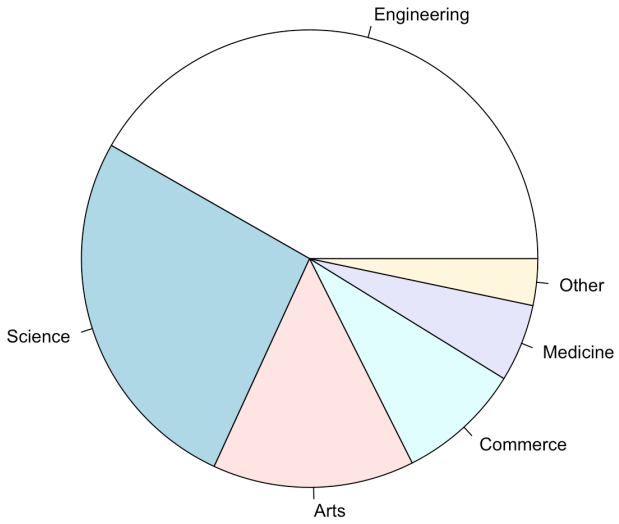


Figure 2.1: Pie chart of qualifications

2.1.3 Bar Chart

A bar chart or a bar graph, is a visual representation of data by means of bars or blocks put side by side. Each bar represents a count of the different categories of the data. Bar charts can use the actual frequency or proportions unlike the pie chart that only uses proportions.

Consider the same example in Table 2.4, a bar chart is constructed and given in Figure 2.2 that uses the frequency to construct the bars.

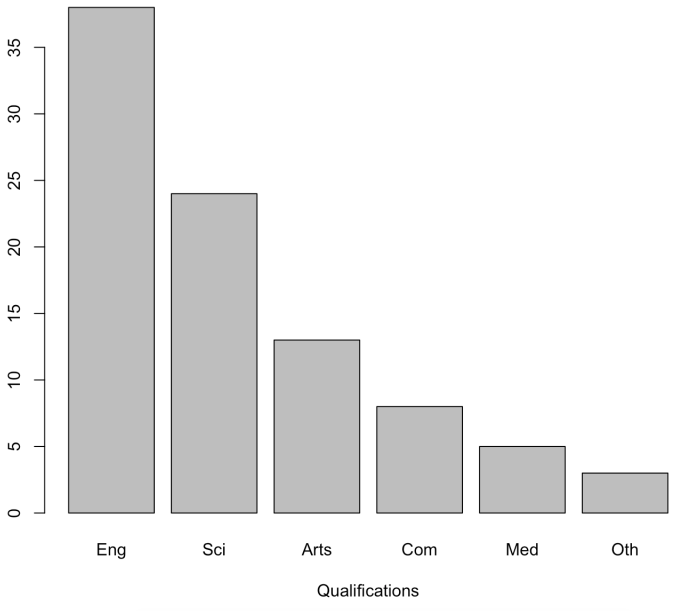


Figure 2.2: Bar chart of qualifications

2.1.4 Multiple Bar Graph

A multiple bar graph is a bar graph with multiple bars in each category. Consider the contingency table in Table 2.3, by using the school as the variable, and splitting the frequency by the gender. The multiple bar graph is shown in Figure 2.3. When the gender is taken as the variable and is split by the school, then the multiple bar graph will look like the one in Figure 2.4. The multiple graphs in Figure 2.3 and 2.4 are the same, but are displayed differently.

2.1.5 Component Bar Graph

Charts help answer and explain the main characteristics of data, without the need to browse through all the observations. Another variation of the bar chart family is the component bar graph. The component bar graph also known as a stacked bar graph, allows to examine the composition of several variables over time or some other entity. It is important to note that the variables you compare should have the same units of measurement. In this graph, the bars are stacked on top of each other. Have a look at Figure 2.5 which are stacked bar graphs of Table 2.3.

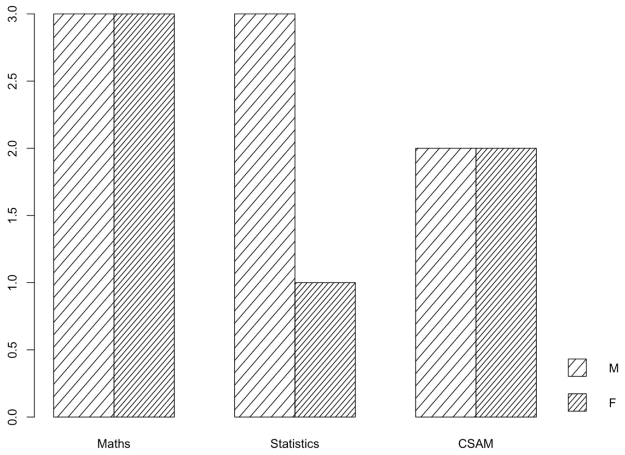


Figure 2.3: Multiple bar graph of number of lecturers by school, split by gender.

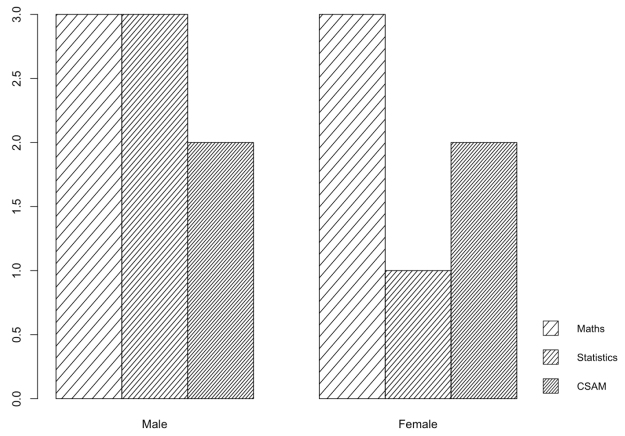


Figure 2.4: Multiple bar graph of number of lecturers by gender, split by school.

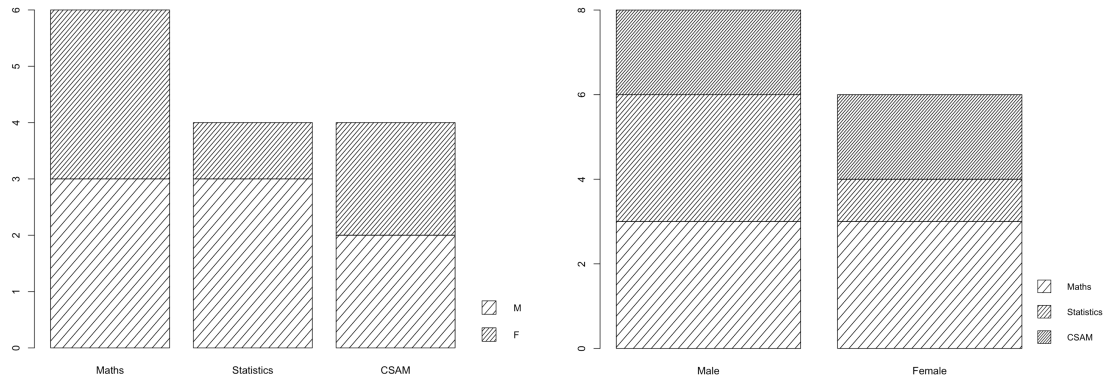


Figure 2.5: Stacked bar graphs

2.1.6 Percentage Component Graph

The percentage component graph is a component graph constructed using the percentages instead of the observed values or frequencies like in Figure 2.5. Using the same data in Table 2.3, a percentage bar is constructed and is seen in Figure 2.6. The two graphs can be constructed differently, either as percentage of lecturers per gender group (Table 2.6) or as a percentage of males and females in each school (Table 2.7).

School / Gender	Male	Female
Maths	$\frac{3}{8} = 37.5\%$	$\frac{3}{6} = 50\%$
Statistics	$\frac{3}{8} = 37.5\%$	$\frac{1}{6} = 17\%$
CSAM	$\frac{2}{8} = 25\%$	$\frac{2}{6} = 33\%$

Table 2.6: Calculating the percentage of lecturers by gender

School / Gender	Male	Female
Maths	$\frac{3}{6} = 50\%$	$\frac{3}{6} = 50\%$
Statistics	$\frac{3}{4} = 75\%$	$\frac{1}{4} = 25\%$
CSAM	$\frac{2}{4} = 50\%$	$\frac{2}{4} = 50\%$

Table 2.7: Calculating the percentage of males and females in each school

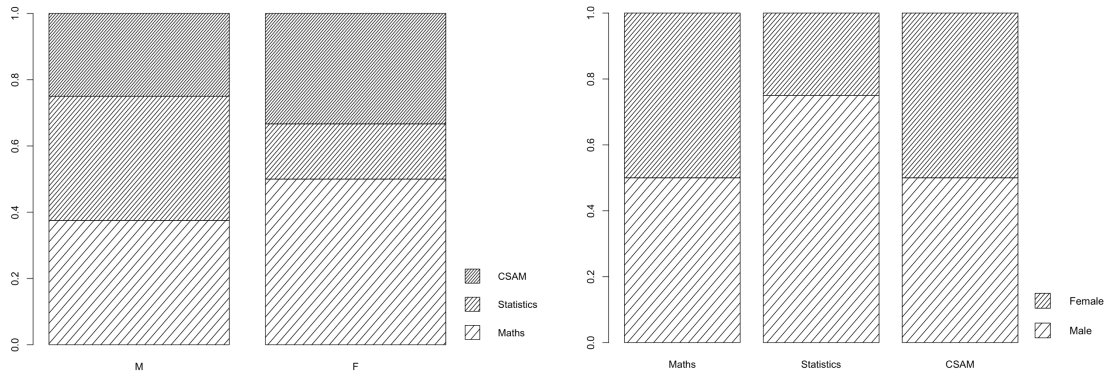


Figure 2.6: Percentage stacked bar graphs

2.1.7 Line graphs

A line graph is a graphical display of information that changes continuously over time. Within a line graph, there are points connecting the data to show a continuous change. The lines in a line graph can descend and ascend based on the data. A line graph can be used to compare different events, situations, and information.

Consider data of average temperatures (in °C) in Cape Town in each month of the year in a specific year. To plot a line graph, use the variable Month on the x-axis, and the Average Temperature on the y-axis in a x-y plot. Join the points to form a line to produce a line graph, like Figure 2.7. The line graph is useful in detecting trends or patterns over time.

Month	Jan	Feb	Mar	Apr	May	June	July	Aug	Sep	Oct	Nov	Dec
Average Temp	22	23	21	18	16	13	13	13	14	16	18	20

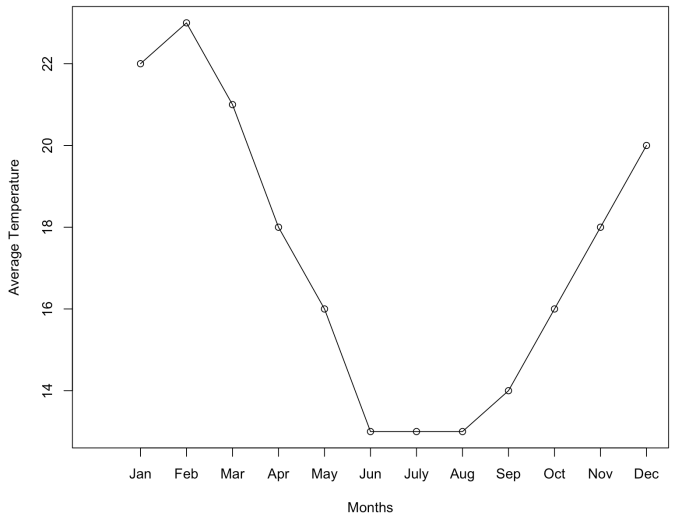


Figure 2.7: Line graph of average temperatures during a year in Cape Town

2.1.8 Frequency Distribution

Frequency tells you how often something occurs. The frequency of an observation in Statistics tells you the number of times the observation occurs in the data. A frequency distribution is a listing of observations according to their frequencies or occurrences.

Consider test marks of 20 students in a first Statistics course:

6	4	7	10
5	6	7	8
7	8	8	9
7	5	6	6
9	4	7	8

A frequency table of this data will look like:

Mark	Frequency
4	2
5	2
6	4
7	5
8	4
9	2
10	1
Total	20

Table 2.8: Frequency table of marks

Sometimes the frequency distribution is not as simple to construct as the one above. Observations are not always easy to group as there may be too many unique values. Consider the following example: the number of calls from motorists per day for roadside service in a certain month.

28	122	217	130	120	86	80	90
120	140	70	40	145	187	113	90
68	174	194	170	100	75	104	97
75	123	100	82	109	120	81	

To construct a frequency distribution for this data, the following can be done:

1. Determine the smallest and largest value in the data.
   Minimum = 28 and maximum = 217.

2. Calculate the range of the data, i.e. the difference between the maximum and minimum.
   Range = 217 - 28 = 189.

3. Calculate $k = 1 + 3.322$ x $\log_{10}(n)$ to find the number of classes to have in the frequency table.
   $k = 1 + 3.322$ x $\log_{10}(31) \approx 6$.

4. Estimate the approximate class size by dividing the range by k.
   Class size = $\frac{189}{6} \approx 32$.

5. Determine the lower end of the first class making sure the smallest value is equal to or less than the lower end.
   Lower end = 28.

6. Determine the frequencies for each class by counting the number of observations falling in each class.

Number of calls	Frequency
28 - 59	2
60 - 91	10
92 - 123	11
124 - 155	3
156 - 187	3
188 - 219	2

Table 2.9: Frequency table of number of calls

The **class limits** are defined as the starting and ending point in each class. The **class boundary** is the average of the end of the current class limit and the starting of the next class limit. The **class midpoint** is the average of the current class starting and ending limit.

2.1.9 Histogram

A **histogram** is a picture of a frequency distribution. It is used to represent continuous quantitative data. It usually consists of adjacent rectangles that are not separated. The area of each rectangle is drawn in proportion to the frequency corresponding to that frequency class. When the class intervals are equal, the area of each rectangle is a constant multiple of the height and the histogram can be drawn like a bar chart, except the bars are not seperated. It is important to note that the class intervals need not be equal.

Consider an example where the results of a survey carried on 45 shipping companies. They were asked how long it took to ship heavy equipment (in days) from the Far East. The frequency distribution table is shown in Table 2.10.

Shipping time	Frequency
10 - 19	7
20 - 29	20
30 - 39	9
40 - 49	3
50 - 59	5
60 - 69	1

Table 2.10: Frequency of shipping time (in days).

In constructing a histogram, the class intervals should be continuous. In this frequency table, the classes are discontinuous. There are gaps between the classes, for example which class does the value 19.5 fit in? Whenever there are gaps between the classes, it is said to have **imaginary limits**. This happens when the class boundaries and the class limits are not the same.

One cannot use imaginary limits to construct a histogram. **Real limits** will need to be constructed in this case as shown in the table below. The real limits are used to construct a histogram shown in Figure 2.8.

Shipping time	Frequency	Lower limit	Upper limit	Lower real limit	Upper real limit
10 - 19	7	10	19	$\frac{10+9}{2} = 9.5$	$\frac{19+20}{2} = 19.5$
20 - 29	20	20	29	$\frac{20+19}{2} = 19.5$	$\frac{29+30}{2} = 29.5$
30 - 39	9	30	39	$\frac{30+29}{2} = 29.5$	$\frac{39+40}{2} = 39.5$
40 - 49	3	40	49	$\frac{40+39}{2} = 39.5$	$\frac{49+50}{2} = 49.5$
50 - 59	5	50	59	$\frac{50+49}{2} = 49.5$	$\frac{59+50}{2} = 59.5$
60 - 69	1	60	69	$\frac{60+59}{2} = 59.5$	$\frac{69+70}{2} = 69.5$

Table 2.11: Computing the real limits of the shipping time frequencies

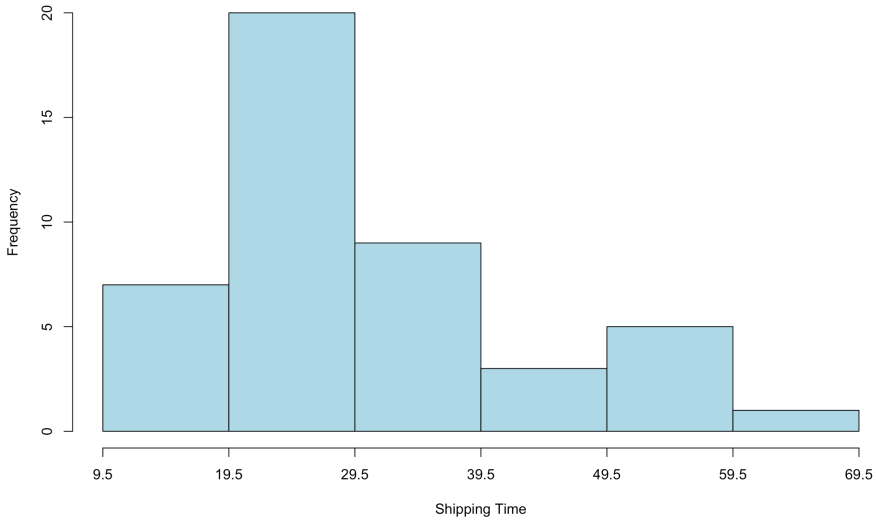


Figure 2.8: Histogram of shipping time.

2.1.10 Cumulative Frequency Curve

A cumulative frequency curve is drawn using the sum of all previous frequencies at each point. It is a plot of the real limits of the classes against the cumulative frequencies. If the frequency is divided by the total of all the available frequencies then the frequencies becomes a **relative frequency**. If the relative frequency is multiplied by 100, then a percentage cumulative frequency curve is formed.

Using the same example in Table 2.10, to find the values to plot in the cumulative frequency plot can be done in two different ways; the less than cumulative frequency which is accumulating the frequencies starting from the lowest class, and the greater than cumulative frequency which is accumulating the frequencies from the highest class. Computation of these curves are shown in Tables 2.12 and 2.13, repsectively.

The two curves can be drawn on the same plot, and where these curves meet or intersect is known as the **median**. The cumulative frequency curves is shown in Figure 2.9.

Shipping time	Frequency	Lower real limit	Upper real limit	< cumulative frequency
10 -19	7	9.5	19.5	7
20 -29	20	19.5	29.5	27
30 - 39	9	29.5	39.5	36
40 - 49	3	39.5	49.5	39
50 - 59	5	49.5	59.5	44
60 - 69	1	59.5	69.5	45

Table 2.12: Less than cumulative frequency computation

Shipping time	Frequency	Lower real limit	Upper real limit	> cumulative frequency
10 -19	7	9.5	19.5	45
20 -29	20	19.5	29.5	38
30 - 39	9	29.5	39.5	18
40 - 49	3	39.5	49.5	9
50 - 59	5	49.5	59.5	6
60 - 69	1	59.5	69.5	1

Table 2.13: Greater than cumulative frequency computation

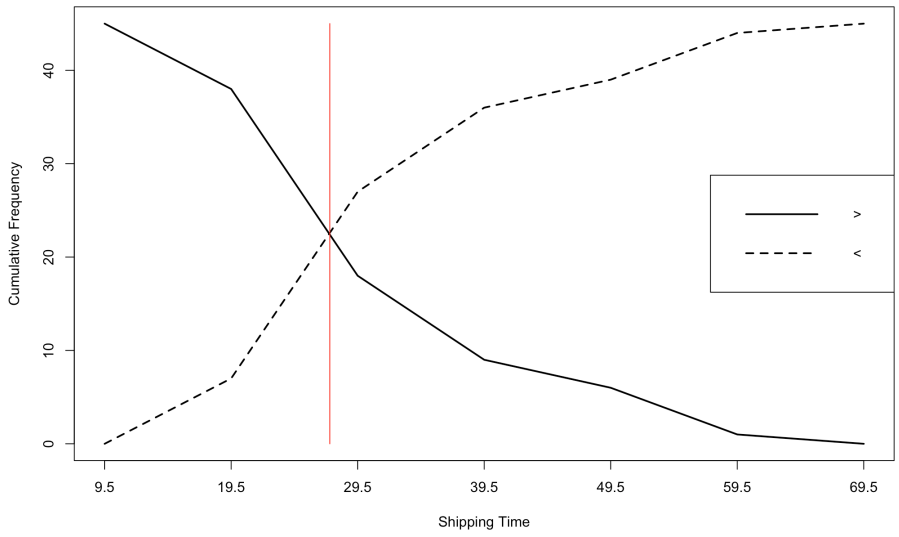


Figure 2.9: Cumulative frequency curves. Lower real limit vs. > cumulative frequency and upper real limit vs. < cumulative frequency.

Exercise 2.1

1. What is the difference between a histogram and a bar graph?

2. A company that produces timber is interested in the distribution of the heights of their pine trees. Compute the frequency distribution and then construct a histogram to display the heights (in metres) of the following sample of 30 trees:

18.3	19.1	17.3	19.4	17.6	20.1	19.9	20.0	19.5	19.3
17.7	19.1	17.4	19.3	18.7	18.2	20.0	17.7	20.0	17.5
18.5	17.8	20.1	19.4	20.5	16.8	18.8	19.7	18.4	20.4

3. A sample of 13 men working in a mine was collected during an investigation into allegations that the mine only employs men who are older than 25 years of age. Their ages are recorded as follows:

19.0	30.0	29.1	20.6	27.9
26.9	23.3	31.3	32.3	24.8
29.9	21.5	26.6		

   (a) Classify these ages into four classes; A: ages below 22.5, B: ages between 22.5 and 25 inclusive, C: ages between 25 and 27.5 inclusive, and D: above 27.5 to create a frequecny table.

   (b) Construst a pie chart from these classes.

   (c) Construct a bar chart from the data.

4. Given the following frequency distribution:

Classes	Frequency
5-9	1
10-14	9
15-19	20
20-24	12
25-29	5

   (a) What is the sample size?

   (b) What is the class size?

(c) What are the boundaries of the third class?

(d) What is the midpoint of the second class?

(e) Obtain the real limits.

5. The monthly sales (in millions of $) of a large business are given as:

149	148	189	167
380	170	216	155
280	655	250	235
221	950	750	912
510	565	215	842

(a) Construct a frequency distribution.

(b) Draw a histogram.

(c) Compute the greater and less than cumulative frequency values.

(d) Plot the cumulative frequency curves, and hence determine the median.

6. A company has been selling two types of cars A and B from 1992 to 1998. The number of sales obtained (in billions of $) is given as:

Year	A	B
1992	134	119
1993	126	96
1994	198	182
1995	144	98
1996	164	78
1997	200	197
1998	213	187

(a) Draw the line graphs of the sales of the two data sets on the same plot, and comment on the trend of the lines.

2.2 Numerical Summary

Numerical summaries are measures that characterise the distribution of the data. It can be described in terms of measures of location and its spread or variation. Here, the focus is on quantitative continuous variables.

2.2.1 Measures of Location

Arithmetic Mean

The arithmetic mean is defined as the sum of all observations, divided by the number of observations. Suppose there are n observations denoted as $x_1, x_2, \ldots, x_n$, then:

$$\text{sample mean} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.1}$$

for data that is not grouped/classified like frequency data (raw data). If the data is grouped/classified, with x_i occurring f_i times with a total of n observations, then

$$\text{sample mean} = \bar{x} = \frac{1}{n}\sum_{i=1}^{k} f_i x_i, \quad n = \sum_{i=1}^{k} f_i \tag{2.2}$$

where k is the number of classes.

NB: Note that the sample mean is also known as the **average**.

Example 2.1

Calculate the average age of the 45 people that attended a cultural movie on a specific day.

7	9	11	12	12	12	13	13	14	14
14	14	15	15	15	16	17	18	18	19
19	19	20	20	20	21	22	22	22	23
24	24	25	26	28	29	31	31	32	34
38	39	39	16	25					

The sample mean will be:

$$\bar{x} = \frac{1}{45}[7 + 9 + 11 + 12 \ldots 16 + 25] = \frac{1}{45}[927] = 20.6$$

Example 2.2

Calculate the sample mean of the following grouped/classisfied data of average income per hour on 2000 participants in a survey:

Income per hour	Frequency
R 0	1235
R 1 - R50	459
R 51 - R100	121
R 101 - R200	29

The sample mean is

$$\bar{x} = \frac{1}{1844}\sum_{i=1}^{k} f_i x_i = \frac{0\text{x}1235 + 25.5\text{x}459 + 75.5\text{x}121 + 150.5\text{x}29}{1844} = 13.67$$

Therefore the average income per hour is R 13.67 .

Median

Suppose sorting all observations into a numerical order ranging from lowest to highest. The **median** will be the middle value in the sorted list. Half of all the observations will be greater than the median and the other half will be less than the median. The median is also known as the **50th percentile** or the **second quartile**.

The median $\tilde{x}$ is calculated by first sorting the data in ascending order to get a new data array: $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, and then finding the central value. If n is odd, then median is the $\frac{n+1}{2}$th value,

$$\tilde{x} = x_{(\frac{n+1}{2})}, \tag{2.3}$$

and when n is even,

$$\tilde{x} = \frac{1}{2}[x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}] \tag{2.4}$$

Example 2.3

Calculate median of the data in Example 2.1.
Sort the data:

7	9	11	12	12	12	13	13	14	14
14	14	15	15	15	16	16	17	18	18
19	19	19	20	20	20	21	22	22	22
23	24	24	25	25	26	28	29	31	31
32	34	38	39	39					

Because n is odd, the median is given by:

$$\tilde{x} = x_{(\frac{45+1}{2})} = x_{(23)} = 19.$$

The computation of the median for grouped data is given by:

$$\text{Median} = L_m + \frac{c_m(\frac{n}{2} - F_{m-1})}{f_m}, \tag{2.5}$$

where
L_m is the lower limit of the class containing the median,
c_m is the difference between the upper end and lower end of the median class,
f_m is the frequency of the median class,
F_{m-1} is the cumulative frequency of the class just before the median class,
n is $\sum_{i=1}^{k} f_i$ the sum of frequencies and
k is the number of classes.

When calculating the median for grouped data it is important to remember that the real limits or class boundaries are used.

Example 2.4

The time it takes to build a three-roomed house is believed to be at most 12 weeks. The man in charge of time and service delivery for a building company took a random sample of three-roomed house constructions and inquired how long it took to build them. The data is given as:

Time in weeks	Frequency
5 - 7	5
8 - 10	20
11 - 13	45
14 - 16	10
17 - 19	6
20 - 22	4

The following steps are taken to calculate the median:

1. Calculate the real limits and the cumulative frequency.

Time in weeks	Frequency	Real limits	Cumulative Frequency
5 - 7	5	4.5 - 7.5	5
8 - 10	20	7.5 - 10.5	25
11 - 13	45	10.5 - 13.5	70
14 - 16	10	13.5 - 16.5	80
17 - 19	6	16.5 - 19.5	86
20 - 22	4	19.5 - 22.5	90

2. $L_m = 10.5$, since the median $x_{(45)}$ lies in the class 10.5 - 13.5.
3. $c_m = 13.5 - 10.5 = 3$.
4. $f_m = 45$.
5. $F_{(m-1)} = 25$.
6. Median $= 10.5 + \frac{3(\frac{90}{2} - 25)}{45} = 11.833$.
7. Important to note: check that the median calculated falls in the median class.

Mode

The mode is the observation with the largest frequency for ungrouped/unclassified data. The data is said to have no mode if all the observations are unique, as observations only occur once in the data. It is also possible to have more than one mode.

With grouped/classified data, there will not be a single most frequently occurring observation. However, it will be the class with the highest frequency. The mode will be found in the class with the highest frequency.

The mode for grouped/classified data is estimated as follows:

$$\text{Mode} = L_m + \frac{c_m(f_m - f_{m-1})}{2f_m - (f_{m-1} + f_{m+1})}, \tag{2.6}$$

where
L_m is the lower end of the modal class,
c_m is the upper end of the modal class - lower end of the modal class,
f_m is the frequency of the modal class,

f_{m-1} is the frequency of the class before the modal class and
f_{m+1} is the frequency of the class after the modal class.

Example 2.5

Using the data in Example 2.4, to calculate the mode is done in the following way:

1. The modal class is 10.5 - 13.5, as it has the highest frequency, therefore $L_m = 10.5$.
2. $c_m = 13.5 - 10.5 = 3$.
3. $f_m = 45$.
4. $f_{m-1} = 20$.
5. $f_{m+1} = 10$.
6. Mode $= 10.5 + \frac{3(45-20)}{2(45)-(20+10)} = 11.75$.

For any set of data; the mean, median and mode are likely to be different, thus it has to be decided which is the best one to use in given situation.

Quantiles/Quartiles/Percentiles

While the mean, median and mode describe the center of the data, it is sometimes useful to also summarise other specific points of location of the data. Suppose sorting or ranking data values in ascending order, the values can then be partitioned into equal size portions with dividing points called **quantiles**.

Quantiles are used to describe the percentage or proportion of observations lying above or below a certain level. The **quartiles** (when data is divided into four) are:

1. 1st quartile or the 25th percentile:
   The value where 25% of the observations lie below.
2. 2nd quartile or the 50th percentile, the median:
   The value where 50% of observations lie above and below.
3. 3rd quartile or the 75th percentile:
   The value where 75% observations lie below.

For ungrouped data, the 1st quartile is calculated by first ordering the data, and then computing,

$$\text{1st quartile} = x_{[\frac{n+1}{4}]} + \frac{1}{4}[x_{[\frac{n+1}{4}]+1} - x_{[\frac{n+1}{4}]}], \tag{2.7}$$

where $[\frac{n+1}{4}]$ and only takes the integer value. For example $[\frac{13}{4}] = 3$ and $[\frac{27}{4}] = 6$.

The 3rd quartile is calculated in a similar manner:

$$\text{3rd quartile} = x_{[\frac{3}{4}(n+1)]} + \frac{3}{4}[x_{[\frac{3}{4}(n+1)]+1} - x_{[\frac{3}{4}(n+1)]}]. \tag{2.8}$$

Example 2.6

Calculate the 1st and 3rd quartile of the following data:

0	0.20	10.00	20.12
20.20	23.90	122.13	200.00

The 1st quartile is given by:

$$\begin{aligned} \text{1st quartile} &= x_{[\frac{8+1}{4}]} + \frac{1}{4}[x_{[\frac{8+1}{4}]+1} - x_{[\frac{8+1}{4}]}] \\ &= x_{(2)} + \frac{1}{4}[x_{(3)} - x_{(2)}] \\ &= 0.20 + \frac{1}{4}[10.00 - 0.20] \\ &= 2.65. \end{aligned} \tag{2.9}$$

The 3rd quartile is given by:

$$\begin{aligned} \text{3rd quartile} &= x_{[\frac{3}{4}(8+1)]} + \frac{3}{4}[x_{[\frac{3}{4}(8+1)]+1} - x_{[\frac{3}{4}(8+1)]}] \\ &= x_{(6)} + \frac{3}{4}[x_{(7)} - x_{(6)}] \\ &= 23.90 + \frac{3}{4}[122.13 - 23.90] \\ &= 97.57. \end{aligned} \tag{2.10}$$

The quantiles for grouped data are calculated much like the grouped data median. The following calculation is for the qth percentile:

$$q\text{-th percentile} = L_q + \frac{c_q(\frac{qn}{100} - F_{q-1})}{f_q}, \tag{2.11}$$

where
L_q is the lower limit of the class containing the qth percentile,
c_q is the difference between the upper end and lower end of the qth percentile class,
f_q is the frequency of the qth percentile class,
F_{q-1} is the cumulative frequency of the class before the qth percentile class and
n is the sum of frequencies $\sum_{i=1}^{k} f_i$.

For example, to calculate the first quartile (25th percentile) the formula will look as follows:

$$\text{25-th percentile} = L_{25} + \frac{c_{25}(\frac{25n}{100} - F_{25-1})}{f_{25}},$$

where
L_{25} is the lower limit of the class containing the 25th percentile,
c_{25} is the difference between the upper end and lower end of the 25th percentile class,
f_{25} is the frequency of the 25th percentile class,
F_{q-1} is the cumulative frequency of the class before the 25th percentile class and
n is the sum of frequencies $\sum_{i=1}^{k} f_i$.

Example 2.7

Using the data in Example 2.4, the

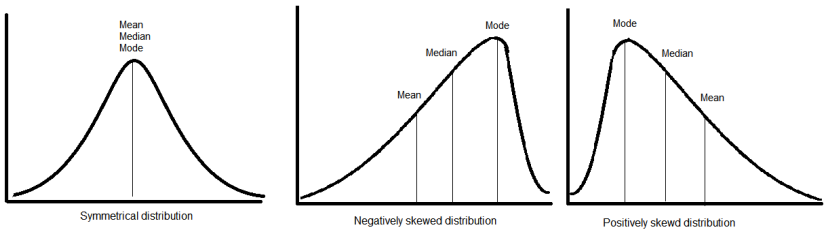
$$\begin{aligned}\text{25th percentile} &= 7.5 + \frac{3(\frac{90}{4} - 5)}{20} \\ &= 10.125\end{aligned}$$

and the

$$\begin{aligned}\text{75th percentile} &= 10.5 + \frac{3(\frac{75\text{x}90}{100} - 25)}{45} \\ &= 13.33.\end{aligned}$$

Skewness

Skewness is a measure of symmetry of a distribution. A distribution can have a positive or a negative skew, depending on where the mean, median and mode are situated. The following figure presents the three different scenarios.



Karl Pearson suggested two calculations as a measure of skewness:

$$P_{s1} = \frac{\text{(mean - mode)}}{\text{standard deviation}} \tag{2.12}$$

$$P_{s2} = \frac{3\text{(mean - median)}}{\text{standard deviation}} \tag{2.13}$$

Values of P_{s1} and P_{s2} less than 0 indicates negative skewness in the data, while values greater than 0 indicates positive skewness.

Kurtosis

Kurtosis measures how peaked the distribution of the data is. If the data set has a high kurtosis, then the histogram of the data will have a high peak. It also means that there is a great number of observations around the mode or the modal class has a high frequency.

The kurtosis is measured by:

$$m_4 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^4. \tag{2.14}$$

The value of m_4 can be standardised in order to get a scale free value which can be used in comparisons,

$$m_4' = \frac{m_4}{s^4}. \tag{2.15}$$

Data with a kurtosis equal to 3 is said to have a **mesokurtic distribution**, that is, it is peaked like the normal distribution. Data with a kurtosis greater than 3 is said to have a **leptokurtic distribution**, that is, it is more peaked than a normal distribution with values concentrated around the mean. When the data has a kurtosis less than 3 then it is said to have a **platykurtic distribution**, which means it is flatter than a normal distribution with a wider peak. The values are wider spread around the mean.

2.2.2 Measures of Dispersion

Variability is an important feature of data. It is a way of checking how the data fluctuates between observations.

The Range

The range (R) is the difference between the minimum and maximum value in the data set, it is given by:

$$\text{R} = \text{Maximum value} - \text{Minimum value} \tag{2.16}$$

The range has the limitation that is depends only on the extreme values in the data, which means it is sensitive to outliers.

Quartile deviation (IQR)

This measure is the difference between the third and first quartile. It is given by:

$$\text{IQR} = \text{3rd quartile} - \text{1st quartile} \tag{2.17}$$

This measure is not greatly affected by outliers - which has the property of **robustness**.

Variance

The variance is the most commonly used measure of dispersion or variability in statistical analysis. This measurements takes into account all observations in the data set. It is denoted by s^2, and the greater the value, the greater the variability. If all observations are close or almost equal then the variance will be low.

The variance is calculated by taking the average of the sum of squared deviations of each observation from the mean, it is given by:

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right] \\ &= \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{[\sum_{i=1}^{n} x_i]^2}{n}\right]. \end{aligned} \tag{2.18}$$

For grouped data, then the variance is calculated by:

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{k} x_i^2 f_i - \frac{[\sum_{i=1}^{k} x_i f_i]^2}{n}\right], \tag{2.19}$$

where f_i are the k frequencies corresponding to each x_i. If the data is categorical data or classified data, the midpoint m_i of a class is used instead of x_i.

The positive square root of the variance s^2 is called the **standard deviation**, and is denoted by s.

2.2.3 Box and whisker plot

The box and whisker plot is constructed using the summary statistics, as opposed to the other plots that uses the raw data. The diagram shows the

1. quartiles,
2. median,
3. maximum and minimum values.

Outliers can also be indicated in the box and whisker plot. The features of the box and whisker plot is that the box contains 50% of the observations and the whiskers are lines which extend from the box to the maximum and minimum values.

Example 2.8

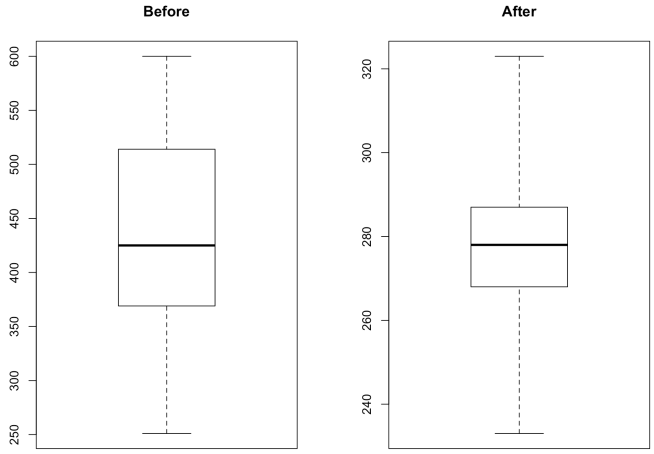
The rate at which accidents occurred at a road junction controlled by yield signs were recorded. Road accidents were also recorded after traffic lights were erected in place of the yield signs. The results are given as:

Year	Before	After
1	395	280
2	425	284
3	415	271
4	369	308
5	514	268
6	548	287
7	498	323
8	548	275
9	355	278
10	251	296
11	600	233
12	358	251
13	478	244

The summary statistics are calculated and are given as:

Period	n	Mean	Median	s	Min	Max	Q_1	Q_3
Before	13	442.615	425	98.619	251	600	369	514
After	13	276.769	278	24.863	233	323	268	287

The box and whisker plots of the accidents before and after the traffic lights were erected is given as:



Draw the boxplots to scale in one plot, and comment on the distribution of the accidents before and after the traffic lights were erected.

2.3 Conclusion

In this chapter, a variety of diagrammatic representations of data was looked at. The diagrams and summary statistics help one understand the data better and to deduce the structure of the data. A step above descriptive statistics is inferential statistics which will be looked at in further chapters.

Exercise 2.2

1. A glass manufacturing company recorded the following profits (in millions). The profits are recorded every four months from January 1985 to December 1987):

   12 18 10 13 20 11 12 19 10

   (a) Calculate the mean, median, mode and variance of the data.

   (b) Comment of the median and mean, in terms of skewness.

   (c) Construct a bar chart and comment on the distribution.

   (d) Construct a box-and-whisker plot for the data.

Chapter 3

Probability

Probability is the measure of the likelihood that an event will occur. Probability quantifies events into a number which can be used to make decisions. It is measured on a scale from zero representing impossibility to one representing certainty.

3.1 Assigning probabilities

In many situations, you may be unsure of the outcome of some activity or experiment, but are sure of the possible outcomes. For example, when you roll a dice, the possible outcomes are getting a 1,2,3,4,5 or 6. But you do not know which number you will get. If you toss a coin twice, then the possible outcomes could be (H,H), (H,T), (T,H) and (T,T). The list of all possible outcomes is called a **sample space,** Ω.

Each of the possible outcomes has an assigned probability to it. For example the sample space for throwing a dice is {1,2,3,4,5,6}, and assigning a probability to it each of the outcomes will be $\frac{1}{6}$, in belief that the dice is fair, and that is each outcome is equally likely. When probabilities are assigned to possible outcomes,

1. each probability must lie between 0 and 1, inclusively, and
2. the sum of all probabilities assigned must equal to 1.

Example 3.1

Assign probabilities to the following experiments:

1. Choosing a card from a standard pack of playing cards.
   The sample space consists of 52 playing cards {Ace of Clubs, 2 of Clubs, 3 of Clubs, ..., King of Spades}. The probability assigned to each item will be $\frac{1}{52}$, assuming all the cards are equally likely to be picked.

2. The combined experiment of tossing a coin and rolling a dice.
   The sample space is $\{(H,1),(H,2),(H,3),(H,4),(H,5),(H,6),$
   $(T,1),(T,2),(T,3),(T,4),(T,5),(T,6)\}$, and each of the outcomes will have an assigned probability of $\frac{1}{12}$.

3.2 Probability of events

Sometimes it may be of interest, not in one particular outcome, but in two or three or more of them. For example, suppose tossing a coin twice. You might be interested in whether the result is the same both times. The list of outcomes in which you are interested in is called an **event**. The even that both tosses of the coin give the same result is $\{(H,H),(T,T)\}$. Events are denoted by capital letters. A can denote this event, then A $= \{(H,H),(T,T)\}$. An event can be just one outcome, or a list of outcomes or even no outcomes at all.

To find the probability of an event, look at the sample space and add the probabilities of the outcomes which make up the event. For example, to toss a coin twice, the sample space will be $\{(H,H),(T,T),(H,T),(T,H)\}$ and the probability of each outcome will be $\frac{1}{4}$. The event A consists of two outcomes, so the probability of A, $\mathrm{P(A)} = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$.

If A is an event, the event "not A" is the event consisting of those outcomes in the sample which are not in A. Since the sum of the probabilities assigned to outcomes in the sample space is 1,

$$\mathrm{P(A)} + \mathrm{P(not\ A)} = 1.$$

The event "not A" is called the **complement** of the event A. The symbol A$'$ is used to denoted the complement of A. Therefore,

$$\mathrm{P(A)} + \mathrm{P(A')} = 1.$$

Example 3.2

The numbers $1, 2, \ldots, 9$ are written on separate cards. The cards are shuffled and the top is turned over. Calculate the probability that the number on this card is a prime number.

The sample space is $\{1,2,3,4,5,6,7,8,9\}$. Each outcome is equally likely and has a probability of $\frac{1}{9}$. Let B be the event that the card turned over is prime. Then B = {2,3,5,7}. The probability of B is the sum of the probabilities of the outcomes in B.

$$\mathrm{P(B)} = \frac{1}{9} + \frac{1}{9} + \frac{1}{9} + \frac{1}{9} = \frac{4}{9}.$$

and the probability of the card not being a prime number is

$$\mathrm{P(B')} = 1 - \frac{4}{9} = \frac{5}{9}.$$

Exercise 3.1

1. A fair 20-sided dice has eight faces coloured red, ten coloured blue and two coloured green. The dice is rolled:

   (a) Find the probability that the bottom face is red.

   (b) Let A be the event that the bottom face is not red. Find the probability of A.

2. A dice with 6 faces has been made from brass and aluminium and is not fair. The probability of a 6 is $\frac{1}{4}$, the probabilities of 2,3,4 and 5 are each $\frac{1}{6}$, and the probability of 1 is $\frac{1}{12}$. The dice is rolled.

   (a) Find the probability of rolling 1 or 6.

   (b) Find the probability of rolling an even number.

3.3 Addition of probabilities

When two events, A and B have no outcomes in common they are said to be **mutually exclusive**. The probability P(A or B) = P(A) + P(B) is known as the addition law of mutually exclusive events.

In Example 3.2 there are two events, event B consisting of all the prime numbers between 1 to 9. Let be A be the event of all the numbers consisting of non prime numbers.

$$\mathrm{A} = \{1,4,6,8,9\} \text{ and}$$
$$\mathrm{B} = \{2,3,5,7\}$$

are mutually exclusive between they do not have one or more outcomes that are the same. This can be seen in what is called a **Venn diagram** in Figure 3.1.

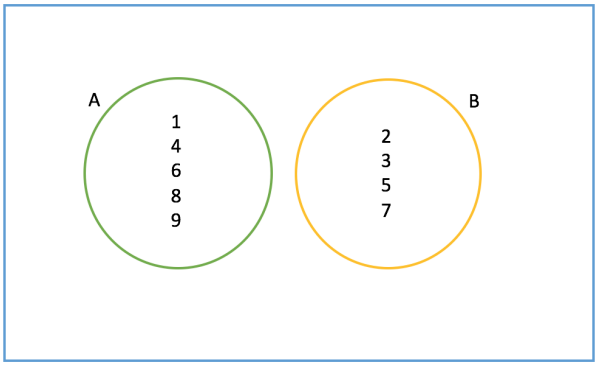


Figure 3.1: Mutually exclusive events A and B

When two events, A and B have outcomes in the sample space that are the same, they are not mutually exclusive. Using the same Example 3.2, the event C are all the even numbers in the sample space, C = {2,4,6,8}. The events A and C are not mutually exclusive, as there are outcomes that are the same, namely {4,6,8}. In this case the addition rule is not valid, P(A or C) $\neq$ P(A) + P(C).
The addition rule is modified as follows: P(A or C) = P(A) + P(C) - P(A and C).

The P(A and C) is called the **intersection** of sets of A and C. It can be denoted as: P(A $\cap$ C) and is illustrated by the region in the Venn diagram that has the outcomes {4,6,8} in Figure 3.2. The P(A or C) will therefore be the probability of outcomes in A plus the probability of all outcomes in C minus the probability of all outcomes in the intersection of A and C.

$$\begin{aligned} \text{P(A or C)} &= \text{P(A)} + \text{P(C)} - \text{P(A} \cap \text{C)} \\ &= \frac{5}{9} + \frac{4}{9} - \frac{3}{9} \\ &= \frac{6}{9}. \end{aligned}$$

Complete a Venn diagram with events A, B and C.

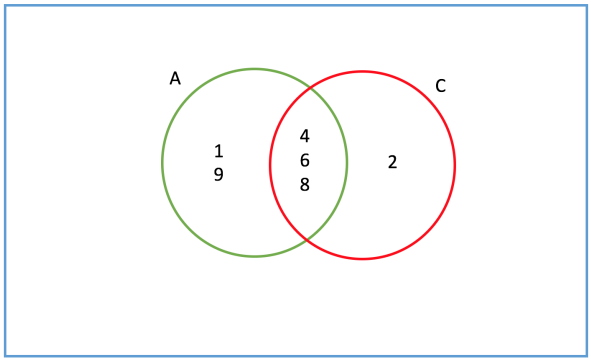


Figure 3.2: Intersection space of events A and C

Additional notation

The **union of sets** of events A and B is the set of all outcomes which belong to event A and event B. This is given by P(A or B) as done in the previous example, but it is denoted by $A \cup B$.

A set that contains no outcomes is called an **empty set** and is denoted by $\emptyset$ or $\{\}$.

An event A is said to be a **subset** of event B, if and only if each outcome of A is an outcome of B. It is denoted by $A \subseteq B$.

3.4 Conditional probability

Conditional probability is the probability of some event A given that some other event B has occurred. It is written as P(A|B) and is read as "the probability of A given B has occurred".

Consider a class of 20 students, of whom 12 are girls and 8 are boys. Suppose further that 7 of the girls and 2 of the boys are left handed. If a student is picked randomly from the class, then the chance that he or she is left handed is $\frac{7+2}{20} = \frac{9}{20}$.

However, the probability of selecting a student from the group of girls that is left handed is $\frac{7}{12}$ and the probability of selecting a student from the group of boys that is left handed is $\frac{2}{8} = \frac{1}{4}$. These probabilities have been calculated on the basis of an extra condition, which is selecting the student from a certain group. This is an example of **conditional probability**.

This can be written as follows:

$$\text{P(left handed} \mid \text{girl)} = \frac{7}{12} = \frac{\frac{7}{20}}{\frac{12}{20}}$$

which is essentially:

$$\text{P(left handed} \mid \text{girl)} = \frac{\text{P(left handed and girl)}}{\text{P(girl)}}$$

Complete the probability of selecting a left handed student given that the student is a boy.

The equation can be generalised as follows:
If A and B are two events and $\text{P(A)} > 0$, then the conditional probability of B given A is

$$\text{P(B|A)} = \frac{\text{P(A and B)}}{\text{P(A)}}. \tag{3.1}$$

Rewriting equation 3.1 gives

$$\text{P(A and B)} = \text{P(A) x P(B|A)}, \tag{3.2}$$

which is known as the **multiplication law of probability**.

Example 3.3

Weather records indicate that the probability that a particular day is dry is $\frac{3}{10}$. The South African football team Bafana Bafana show a record of success is better on dry days than on wet days. The probability that the team wins on a dry day is $\frac{3}{8}$, whereas the probability that they win on a wet day is $\frac{3}{11}$. The team is due to play their next match in a few days.

1. What is the probability that the team will win?

2. Three Saturdays ago, the team won their match, what is the probability that it was a dry day?

The sequence involves first the type of weather and then the result of the football match. In cases of conditional probability like in this example, one can make use of a **tree diagram**, as in Figure 3.3.

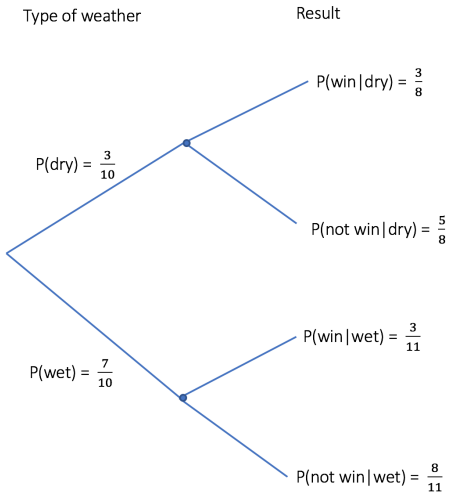


Figure 3.3: Tree diagram of conditional probabilities

Notice the probabilities on the first layer of branches is the probability of the type of weather, wet or dry. The probabilities on the second layer of branches are the conditional probabilities. You can use the tree diagram to calculate any of the four possibilities:

- Probability of winning given the weather is dry.
- Probability of not winning given the weather is dry.
- Probability of winning given the weather is wet.
- Probability of not winning given the weather is wet.

1. The probability of winning can happen in two different ways: P(win) = P(dry & win) or P(wet & win).

$$\begin{aligned} \text{P(win)} &= \text{P(dry \& win)} + \text{P(wet \& win)} \\ &= \text{P(dry) x P(win|dry)} + \text{P(wet) x P(win|wet)} \\ &= \frac{3}{10}\text{x}\frac{3}{8} + \frac{7}{10}\text{x}\frac{3}{11} \\ &= \frac{9}{80} + \frac{21}{110} \\ &= \frac{267}{880} = 0.303. \end{aligned}$$

2. In this case, you have been asked to calculate a conditional probability. However, the sequence of events has been reversed and you want to find out P(dry|win).

$$\text{P(dry|win)} = \frac{\text{P(dry \& win)}}{\text{P(win)}} = \frac{\frac{9}{80}}{\frac{267}{880}} = \frac{99}{267}.$$

Think of P(dry|win) as being the proportion of times that the weather is dry out of all the times that the team wins.

Exercise 3.2

1. The Prosecutor's fallacy. An accused prisoner is on trial. The defence lawyer asserts that in the absence of further evidence, the probability that the prisoner is guilty is 1 in a million. The prosecuting lawyer produces further piece of evidence and asserts that if the prisoner were guilty, the probability that this evidence would be obtained is 999 in 1000, and if he were not guilty would be only 1 in 1000. Assuming that the court order the legality of the evidence, and that both lawyers' figures are correct, what is the probability that the prisoner is guilty?

Bayes' Theorem

Bayes' Theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event.
Bayes' theorem is stated mathematically as the following equation:

$$\text{P(B|A)} = \frac{\text{P(A and B)}}{\text{P(A)}}. \tag{3.3}$$

where A and B are events and $P(B) \neq 0$.

- $P(A|B)$ is a conditional probability: the likelihood of event A occurring given that B is true.
- $P(B|A)$ is also a conditional probability: the likelihood of event B occurring given that A is true.
- $P(A)$ and $P(B)$ are the probabilities of observing A and B independently of each other; this is known as the marginal probability.

The examples done in this section of conditional probability uses Bayes' Theorem.

3.5 Independent events

Independent events are events which have no effect on one another. For two independent events, A and B,

$$\text{P(A and B)} = \text{P(A) x P(B)} \tag{3.4}$$

This result is called the **multiplication law for independent events**.

Example 3.4

In a carnival game, a contestant has to first toss a fair coin and then roll a fair cubical dice whose faces are numbered 1 to 6. The contestant wins a prize if the coin shows heads and the dice score is below 3. Find the probability the contestant wins the prize.

The two events of tossing a coin and rolling a dice are independent. The outcome of rolling the dice does not depend on the outcome of tossing the coin. Therefore the probability of winning is :

$$\begin{aligned}\text{P(prize won)} &= \text{P(coin shows heads) and P(dice score is lower than 3)}\\ &= \text{P(coin shows heads) x P(dice score is lower than 3)}\\ &= \frac{1}{2}\text{x}\frac{2}{6} = \frac{1}{6}.\end{aligned}$$

3.6 Relative frequency approach to probability

The relative frequency approach to probability is based on the number of times an event has occurred over all possible number of occurrences it could have occurred.

$$\text{P(event)} = \frac{\text{Number of times an event has occurred}}{\text{Total number of events}} \tag{3.5}$$

Example 3.5

Consider data on the mode of transport of 92 students to campus everyday:

1. The probability of randomly selecting a student that is a male

$$\begin{aligned}\text{P(male)} &= \frac{\text{Number of males}}{\text{Total number of students}}\\ &= \frac{51}{92} = 55.4\%\end{aligned}$$

Mode of transport / Gender	Male	Female	Total
Car	20	15	35
Taxi	24	18	42
Bicycle	5	3	8
Walked	2	5	7
Total	51	41	92

2. The probability of randomly selecting a student that is a female and travels to campus using a car

$$\begin{aligned} \text{P(female travelling by car)} &= \frac{\text{Number of females travelling by car}}{\text{Total number of students}} \\ &= \frac{15}{92} = 16.3\% \end{aligned}$$

3. Calculate the probability of selecting a male that rides a bicycle to campus.

4. Calculate the probability of selecting a female that walks to campus.

3.7 Counting methods

This section is about the number of arrangements of different objects, and the number of ways you can choose different objects.

3.7.1 Permutations

In the previous section, you could count the number of outcomes in a sample space. When the number of outcomes is fairly small, it is quite straightforward, but in certain instances counting the possible number of outcomes can be cumbersome. Think of listing the 5 different cards from a pack of 52 playing cards.

Suppose you have 3 letters: A, B and C written on 3 separate cards. There are different ways of arranging these cards.

ABC ACB BCA BAC CAB CBA

There are 3 choices for the first position: either A, B or C.

There are 2 choices for the second position:
B or C if A has been used
C or A if B has been used

A or B if C has been used

There is 1 choice for the third position:
only C if A and B have been used
only B if C and A have been used
only A if B and C have been used

Therefore altogether there are 3x2x1 = 6 possible ways of arranging the three cards.

Similarly the number of ways to arrange the letters A, B, C and D will be 4x3x2x1 = 24. The different arrangements of objects (they need not be letters) are called **permutations**. The number of permutations of n distinct objects is $n!$, where $n! = n$ x $(n-1)$ x $(n-2)$ x ... x 2 x 1. The expression $n!$ is called n **factorial**.

The formula for permutations is further extended to give the number of different permutations of r objects which can be made from n distinct objects:

$$^nP_r = \frac{n!}{(n-r)!} \tag{3.6}$$

Equation 3.5 is used for example in the case of arranging 4 letters out of 7 letters A, B, C, D, E, F and G. That is having the arrangements:

ABCD ABCE ABCF ABCG ...and so on.

The number of arrangements will be:

$$^7P_4 = \frac{7!}{(7-4)!} = \frac{7!}{3!} = 840.$$

3.7.2 Permutations when objects are not distinct

In the previous section, objects are distinctly different from each other. Suppose now that in a set of n objects there are k subgroups with n_1 in the first group and n_2 in the second group, up to n_k in the kth group such that $n_1 + n_2 + \ldots + n_k = n$. Then the number of distinguishable permutations is:

$$\frac{n!}{n_1!n_2! \cdots n_k!} \tag{3.7}$$

Example 3.6

Find the number of distinct permutations of the letters of the word MISSISSIPPI.

There are 11 letters, of which 4 S's, 4 I's, 2 P's and 1 M. The number of distinct permutations of the letters is therefore

$$\frac{11!}{4!\text{x}4!\text{x}2!\text{x}1!} = 34650.$$

Note in permutations, the order of objects is significant when counting the number of arrangements

3.7.3 Combinations

Combinations is the case when the order of objects does not matter in counting the number of different arrangements. For example, if you were dealt a hand of 13 cards from a pack of 52 cards, you would not be interested in the order in which you received the cards.

The number of different combinations of r objects selected from n distinct objects is:

$$^nC_r = \frac{n!}{(n-r)!\,\text{x}\,r!} \tag{3.8}$$

Example 3.7

How many ways can you arrange three letters from the word ATMOSPHERIC, where the order is not important?

$$^{11}C_3 = \frac{11!}{(11-3)!\,\text{x}\,3!} = 165.$$

Since the order is not important, selecting the arrangements ATM, AMT, TMA, TAM, MTA and MAT are all counted as just one selection. Try computing the permutation of this example.

Example 3.8

A team of 5 people, which must contain 3 men and 2 women, is chosen from 8 men and 7 women. How many different teams can be selected.

The number of different teams of 3 men which can be selected from 8 is $^8C_3$, and the number of different teams of 2 women which can be selected from 7 is $^7C_2$. Any

of ${}^8C_3$ teams can join up with ${}^7C_2$ to make a team of 5. The number of possible teams is

$$ {}^8C_3 \text{ x } {}^7C_2 = 1176. $$

Exercise 3.3

1. How many different arrangements can be made of the letters in the word STATISTICS?

2. (a) Calculate the number of arrangements of the letters in the word NUMBER?

   (b) How many arrangements in (a) begin and end with a vowel?

3. A bag contains 20 chocolates, 15 toffees and 12 peppermints. If three sweets are chosen at random, what is the probability that they are:

   (a) all different

   (b) all chocolates (excludes 1 or 2 chocolates)

   (c) all the same

   (d) all not chocolates

Acknowledgements

This chapter has been adapted from Advanced Level Mathematics Statistics 1, written by Steve Dobbs and Jane Miller in 2002.

Chapter 4

Random Variables and Distributions

4.1 Introduction

We have been looking at variables, describing them through samples and populations. Now, we will proceed to look at the distributions of data. There are a set/class of distributions where most data fall into. We will discuss these distributions, for **continuous** and **discrete** data. This discussion should help us understand the structure of populations better.

4.2 Discrete Random Variables

These are variables that take on distinct values, e.g. the number of games won by a team 0, 1, 2, e.t.c., up to the total number of games played.

Probability distribution

Let X be a random variable and let x_i, $i = 1, 2, \cdots, k$ denote the k distinct values that X may assume. Since each X corresponds to a basic outcome of a random trial, a **probability distribution** for the sample space will associate a probability value with each x_i. The probability that random variable X will assume value x_i will be denoted by $P(X = x_i)$ for example, $P(X = 0)$ is the probability that $X = 0$.

Example 4.1

In the game problem mentioned above earlier, suppose a team plays 3 games, and has the following probabilities of winning the games.

Prob(Winning 0 games)=0.65,
Prob(winning 1 game)=0.15,
Prob(Winning 2 games)=0.10,
Prob(winning 3 games)=0.10.

X	0	1	2	3
P(X=X)	0.65	0.15	0.10	0.10

This represents a probability distribution, i.e., the probability associated with each possible outcome.

Properties

Any probability distribution for a discrete random variable X has the properties.

1. $0 \leq P(X = x_i) \leq 1$ for $i = 1, 2, \cdots, k$,
2. $\sum_{i=1}^{k} P(X = x_i) = 1$.

The distribution function can also be referred to as a **probability function** or **probability mass function**.

Cumulative probability distribution

The **cumulative probability distribution** is the probability that a random variable X takes a value less than or equal to some specified value x, denoted by $P(X \leq x)$.

$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$ or
$P(X < 2) = P(X = 0) + P(X = 1)$

In Example 4.1, the cumulative distribution is given by;

X	$P(X = x)$	$P(X \leq x)$
0	0.65	0.65
1	0.15	0.80
2	0.10	0.90
3	0.10	1.00

Expected value

The mean value of a random variable in many trials is known as the **expected value**. For a discrete random variable X, the expected value is denoted by;

$$E[X] = \sum_{i=1}^{k} x_i P(x_i) \text{ where } P(x_i) = P(X = x_i)$$

Variance

The outcomes of a discrete random variable will vary, and as such it is useful to have a measure of their variability. As we discussed earlier a key measure of variation is the **variance**, defined;

$$\begin{aligned} \sigma_x^2 &= Var[X] \\ &= \textstyle\sum_{i=1}^{k}(x_i - E[X])^2 P(x_i) \\ &\textit{or} \\ &= E[X - E[X]]^2 \\ &= E[X^2] - [E[X]]^2 \end{aligned}$$

The standard deviation of a random variable X is given by

$$\sigma_x = \sqrt{Var[X]}.$$

Example 4.2

Find E[X] and Var[X] in the above example.

Solution

$$\begin{aligned} E[X] &= 0 \times 0.65 + 1 \times 0.15 + 2 \times 0.10 + 3 \times 0.10 \\ &= 0.15 + 0.20 + 0.30 \\ &= \underline{0.65} \end{aligned}$$

$$\begin{aligned} Var[X] &= (0 - 0.65)^2 \times 0.65 + (1 - 0.65)^2 \times 0.15 + (2 - 0.65)^2 \times 0.10 + (3 - 0.65)^2 \times 0.10 \\ &= \underline{1.03} \end{aligned}$$

Exercise 4.1

Consider a man who tosses a coin once, the probability of getting a head is $\frac{1}{2}$, the probability of getting a tail is 1-$\frac{1}{2}$=$\frac{1}{2}$. Let $X = 1$, if he gets a head and $X = 0$ if he gets a tail. What is the distribution function of X? Find E[X].

Moment Generating Functions

The expected values $E[X], E[X^2], \ldots, E[X^k]$ are called **moments**. As you've seen previously, $\mu = E[X]$ and $\sigma^2 = Var[X] = E[X^2] - \mu^2$, which are functions of moments.
Special functions called moment generating functions can sometimes make finding the mean and variance of a random variable simpler.

Definition 1 mgf

The moment generating function (mgf) of a discrete random variable X is defined to be;

$$M_X(t) = E(e^{tX}) = \sum_{x \in \mathscr{X}} e^{tx} P(x)$$

where $\mathscr{X}$ is the set of possible X values. The mgf$(\cdot)$ exists if $M_X(t)$ is defined for the $t \in [-h, h]$ and $h > 0$.

Note:

Some of the merits of using the mgf include the following properties.

1. If the mgf exists then, when $t = 0$, $M_X(0) = 1$ for any random variable.

2. If the mgf exists and is the same for two distributions then the two distributions are the same. This means mgf uniquely identify probability distributions.

3. If $mgf(\cdot)$ exist then;

$$E(X^r) = \frac{d^r}{dt^r} M_X(t)|_{t=0} = M_X^{(r)}(0)$$

So we can say the mean of X can be found by evaluating the first derivative of the mgf at $t = 0$. That is, $\mu = E[X] = M'(0)$.

The variance of X can be found by evaluating the first and second derivatives of the mgf at $t = 0$. That is, $\sigma^2 - E[X^2] - (E[X])^2 = M''(0) - (M'(0))^2$.

4. Let X have mgf $M_X(t)$ and let $Y = aX + b$. Then $M_Y(t) = e^{bt} M_X(at)$

4.3 Continuous Random Variables

A **continuous random variable** on the other hand is a random variable that can take on a continuum (any value in an interval). An example of this is the height of students. Remember the height can have an infinite number of possible values.

Here we talk of values on an interval as opposed to particular values. Thus the probability density function of a continuous random variable X is a mathematical function for which the area under the curve corresponding to any interval is equal to the probability that X will take a value in that interval. The **probability density function** is denoted by $f(x)$. The value $f(x)$ is called the probability density at x. For example

$$f(x) = \begin{cases} 12x(1-x)^2 & 0 \leq x \leq 1, \\ 0 & \text{elsewhere} \end{cases}$$

For instance, the probability density at $x = 0.5$ is

$$f(0.5) = 12(0.5)(1-0.5)^2 = 1.5.$$

Example 4.3

Consider the continuous random variable X, which represents the yield of a crop in tons per acre. Suppose the yield can take any value between 0 and 1 ton, and that X has the density above.

Then

$$\begin{aligned} P(0.5 \leq X \leq 0.7) &= \int_{0.5}^{0.7} 12x(1-x)^2 dx \\ &= \int_{0.5}^{0.7} 12x - 24x^2 + 12x^3 dx \\ &= [6x^2 - 8x^3 + 3x^4]_{0.5}^{0.7} \\ &= 6 \times 0.7^2 - 8 \times 0.7^3 + 3 \times 0.7^4 - 6 \times 0.5^2 - 8 \times 0.5^3 + 3 \times 0.5^4 \\ &= \underline{0.2288} \end{aligned}$$

See Figure 4.1. This means the probability that the crop yield will be between 0.5 and 0.7 is 0.2288.

Properties

As in the discrete case above;

1. $f(x) \geq 0$ for x ϵ $[a, b]$ interval of definition.
2. $\int_a^b f(x) = 1$.

Note that, in this particular case,
$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b)$.

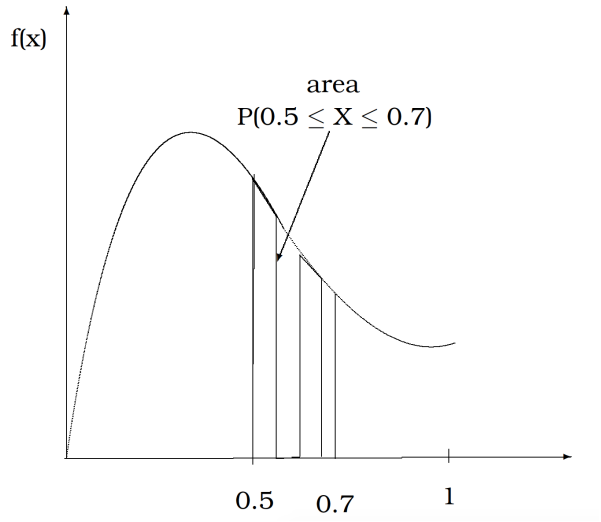


Figure 4.1: The area representing the probability

Cumulative Probability Function

The **cumulative probability function** of a continuous random variable is denoted by $F(x)$ and defined by; $F(x)$=P$(X < x)$ for x $\epsilon(-\infty, \infty)$.

The probability that $X < x_0$ = P$(X < x_0)$=F(x_0) is the area under the curve $f(x)$ to the left of x_0. i.e., using calculus $F(x_0)$=$\int_{-\infty}^{x_0} f(x)dx$. This is the cumulative distribiton function.

Expected value

The **expected value of a continous variable** is defined as

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

Example 4.4

In Example 4.3, we calculate the expected value from,

$$\begin{aligned} E[X] = & \int_0^1 xf(x)dx \\ = & \int_0^1 x \times 12x(1-x)^2dx \\ = & \int_0^1 12x^2 - 24x^3 + 12x^4dx \\ = & [4x^3 - 6x^4 + \tfrac{12}{5}x^5]_0^1 \\ = & 4 - 6 + \tfrac{12}{5} \\ = & \underline{0.4} \end{aligned}$$

Variance

The variance, as in the discrete case is calculated with the summation replaced by the integral, as follows;

$$\begin{aligned} Var[X] = & \int_{-\infty}^{\infty}(x - E[X])^2 f(x)dx \\ = & \int_{-\infty}^{\infty} x^2 f(x)dx - E[X]^2 \\ = & E[X^2] - [E[X]]^2 \end{aligned}$$

Example 4.5

$$\begin{aligned} Var[X] = & \int_0^1 (x - E[X])^2 f(x)dx \\ = & E[X - E[X]]^2 \\ = & E[X^2] - [E[X]]^2 \\ = & \int_0^1 x^2 \times 12x(1-x)^2 dx - 0.4^2 \\ = & \int_0^1 12x^3 - 24x^4 + 12x^5 dx - 0.4^2 \\ = & [3x^4 - \tfrac{24}{5}x^5 + 2x^6]_0^1 - 0.4^2 \\ = & \underline{0.04} \end{aligned}$$

Exercise 4.2

In Example 4.3, find the cumulative distribution function of X.

Moment generating function

Definition 2 mgf

The moment generating function (mgf) of a continuous random variable X is defined to be;

$$M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f(x)$$

The mgf() exists if $M_X(t)$ is defined for the $t \in [-h, h]$ and $h > 0$.

4.4 Empirical distributions

There are a wide variety of probability distributions. However, only a limited number of types, or families, of probability distributions are used in practical applications. These fall under the two categories discrete and continuous.

There are many other distributions not discussed in this chapter. However, the few distributions in the next two sections fall under a class of distributions we call empirical distributions.

4.5 Discrete distributions

Bernoulli trials

A random variable X which has two possible outcomes say 0 and 1 is called a Bernoulli random variable.

The probability distribution of X is,

$$\begin{aligned} P(X=1) &= p \\ P(X=0) &= 1-p \\ i.e., \quad P(X=0) &= 1-P(X=1) \\ &= \underline{1-p} \end{aligned}$$

Example 4.6

Tossing a fair coin, you get a head or a tail each with probability $p = \frac{1}{2}$. Thus if a head is labelled 1 and a tail 0, the random variable X representing the outcome takes values 0 or 1 if the probability that $X = 1$ is p, then we have that

$$\begin{aligned} p = P(X=1) &= \tfrac{1}{2} \\ P(X=0) &= 1-p = 1-\tfrac{1}{2} \\ &= \tfrac{1}{2} \end{aligned}$$

Since events $X = 1$ and $X = 0$ are mutually exclusive.

Binomial Distribution

Suppose in an experiment there are two possible outcomes (failure and success) and that the probability of success is p. Suppose also that the experiment is repeated n times, the probability of x successes follows a Binomial distribution.

Let $X = X_1 + X_2 + \cdots + X_n$ where X_i are independent and identically distributed Bernoulli random variables, then X is called a binomial random variable. Thus;

$$\begin{aligned} P(x) &= P(X=x) \\ &= \tbinom{n}{x} p^x (1-p)^{n-x} \end{aligned}$$

i.e., for $x = 0, 1, \cdots, n$ and $0 < p < 1$, $\binom{n}{x} = \frac{n!}{x!(n-x)!}$

The quantities n and p are called **parameters** and they specify the distribution. Let us look at one application of the Binomial with parameters n and p i.e Bin(n, p).

Example 4.7

Bin(n, p) here means Binomial distribution with parameters n and p.

Find
(i) probability of getting 4 heads in 6 tosses of a fair coin.
(ii) E[X] and Var[X].

Solution

$$\begin{aligned} P(X=4) &= \binom{6}{4}\tfrac{1}{2}^4(1-\tfrac{1}{2})^{6-4} \\ &= 15 \times \tfrac{1}{16} \times \tfrac{1}{4} \\ &= \underline{\tfrac{15}{64}}. \end{aligned}$$

$$\begin{aligned} E[X] &= \sum x\binom{n}{x}p^x(1-p)^{n-x} \\ &= \underline{np} \\ Var[X] &= \sum x^2\binom{n}{x}p^x(1-p)^{n-x} - (np)^2 \\ &= \underline{np(1-p)} \end{aligned}$$

Poisson Distribution

A Poisson random variable is a discrete random variable that can take integer values from 0 up to ∞. The parameter for this distribution is λ i.e., $P_0(\lambda)$.

An example of the application of the Poisson distribution follows, The number of individuals arriving at a bank teller per quarter hour X is a poisson random variable.

The Poisson probability function is

$P(x) = \frac{\lambda^x e_{-\lambda}}{x!}$, where $\quad P(x) = P(X = x)$, $x = 0,\ 1,\ \cdots, \infty$, and $0 < \lambda < \infty$

Example 4.8

The number of students arriving at a take away every 15minutes is a poisson random variable with parameter $\lambda = 0.2$. Find the probability that zero, one and two students arrive at the take away.

$$\begin{aligned} P(X=0) &= \tfrac{(0.2)^0 e^{-0.2}}{0!} \\ &= 0.8187 \text{ (no students arrive)} \end{aligned}$$

$$\begin{aligned} P(X=1) &= \tfrac{(0.2)^1 e^{-0.2}}{1!} \\ &= 0.1637 \end{aligned}$$

$$\begin{aligned} P(X=2) &= \tfrac{(0.2)^2 e^{-0.2}}{2!} \\ &= 0.0164 \end{aligned}$$

Properties

$$\begin{aligned} E[X] &= \sum_{x=0}^{\infty} \frac{x\lambda^x e^{-\lambda}}{x!} \\ &= \sum \frac{\lambda^x e^{-\lambda}}{(x-1)!} \\ &= \lambda \sum \frac{\lambda^{x-1} e^{-\lambda}}{(x-1)!} \\ &= \lambda \times 1 \\ &= \lambda. \end{aligned}$$

$$\begin{aligned} \text{Similarly } Var[X] &= \sum \frac{x^2\lambda^x e^{-\lambda}}{x!} - \lambda^2 \\ &= \lambda^2 + \lambda - \lambda^2 \\ &= \lambda. \end{aligned}$$

Hint: Let $y = x - 1$.

Poisson Distribution linked with the Binomial Distribution

The Poisson distribution can be viewed as a limiting case of a Binomial distribution, when the number of trials n gets large and the the probability p the probability of success gets small.

In the Binomial distribution $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$ we denote, $E(X) = np = \lambda$ so that

$$p = \frac{\lambda}{n}$$

which means

$$(1 - p) = 1 - \frac{\lambda}{n}$$

Thus, $P(X = x) = \binom{n}{x} \left(\frac{\lambda}{n}\right)^x (1 - \frac{\lambda}{n})^{n-x} \to \frac{e^{-\lambda}\lambda^x}{x!}$ as $n \to \infty$

Exercise 4.3

Show that the limiting distribution of the Binomial distribution is the Poisson distribution.

$$\lim_{n \longrightarrow \infty} P(x) = \frac{e^{-\lambda}\lambda^x}{x!}.$$

The Hypergeometric Distribution

We will at this stage mention a distribution related to the Binomial distribution. Suppose again we are looking at a population consisting of N items. Suppose there are two possible outcomes Success (S) and Failure (F) and that k of out comes are successes. If a random sample of n items is drawn and x of the items are successes, then x follows a **Hypergeometric distribution** given by;

$$P(X = x) = \frac{\binom{S}{x}\binom{N-S}{n-x}}{\binom{N}{n}}$$

Discrete Uniform distribution

The discrete Uniform random variable is a variable that can take on integer values within a given interval with equal probabilities.

For example let x be a discrete random variable which can assume s values.

Thus, the discrete Uniform Probability function is defined

$$P(X = x) = P(x) = \frac{1}{s},$$

$x = a,\ a + 1,\ ...,\ a + (s - 1)$ where $s > 0$ is the number of terms and a the first term.

Properties

E[X]= a+ $\frac{s-1}{2}$.
Var[X]= $\frac{s^2-1}{12}$.

Example 4.9

Let x be a discrete random variable which can assume 5 values. Then

X = 0 1 2 3 4,

in this case, $s = 5$, therefore,

X	=	0	1	2	3	4
P(X)	=	0.2	0.2	0.2	0.2	0.2

That is, the discrete Uniform Probability function. One way is to use the formula,

$$P(X = x) = P(x) = \frac{1}{5}$$

$x = 0,\ 0 + 1,\ 0 + 2,\ 0 + 3,\ 0 + (5 - 1) = 4$, thus $s = 5$.

Exercise 4.4

Show that for the discrete Uniform distribution

E[X]=$a + \frac{s-1}{2}$, and Var[X]=$\frac{s^2-1}{12}$

Hint: $Var(X) = \sum(X - E[X])^2 P(X = x)$.

4.6 Continuous distributions

Continuous Uniform Distribution

This distribution is also known as the **rectangular distribution**. This is the continuous equivalent to the discrete Uniform distribution, discussed above. A continuous Uniform variable has a Uniform probability density over an interval. This distribution is of the form,

$$f(x) = \begin{cases} \frac{1}{b-a} & a < x < b, \\ 0 & \text{elsewhere.} \end{cases}$$

Example 4.10

The marks from a certain exam are uniformly distributed over to 50 to 75. The density function for the marks is given by

$$f(x) = \begin{cases} \frac{1}{25} & 50 < x < 75, \\ 0 & \text{elsewhere.} \end{cases}$$

Properties

E[X]=$\frac{1}{2}(b+a)$

Var[X]=$\frac{(b-a)^2}{12}$

For the example above E[X]=62.5.

Exercise 4.5

For the Continuous discrete Uniform distribution, show that

E[X]=$\frac{1}{2}(a+b)$,
and Var[X]=$\frac{(b-a)^2}{12}$.

The Normal distribution

This is one of the most important probability distributions in ststistics. The normal probability density function is

$$f(x) = \begin{cases} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]} & \text{where } -\infty < x < \infty, \ -\infty < \mu < \infty, \ \sigma > 0 \\ 0 & \text{elsewhere} \end{cases}$$

We shall use the normal distribution to a great extent in inference. This is a two parameter distribution, usually denoted by $N(\mu, \sigma^2)$. A random variable z is said to be standard normal if $Z = \frac{X-\mu}{\sigma}$ has mean $\mu = 0$ and variance $\sigma^2 = 1$, for X a normal random variable with mean μ and variance σ^2. Tables which give normal cumulative probabilities are widely available.

Properties

Let X be a normally distributed random variable with mean μ and variance σ^2, i.e., $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma}$ is normally distributed with mean 0 and variance 1. We say Z has a standard normal distribution and $\frac{X-\mu}{\sigma}$ is called standardisation.

The Exponential probability distribution

An exponential random variable is a continuous random variable that can take on any positive value and is usually used to describe the time between events.

Example 4.11

Let X be the length of a long distance telephone call. Then X has an exponential distribution.

The density function is

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } 0 \leq x < \infty, \ \& \ \lambda > 0 \\ 0 & \text{elsewhere.} \end{cases}$$

Properties

$$E[X] = \frac{1}{\lambda}, \ and.$$

$$Var[X] = \frac{1}{\lambda^2}.$$

The parameter here is λ.

Memoryless Property

The memoryless property of a given probability distribution is mostly associated with the distribution of times. Consider tossing a coin, if your outcome was a head (H) and you toss the coin again does it mean at the second toss it is going to be a tail (T)? i.e., $P(T|H) = P(T)$ in other words ;

$$P(X_{t+1}|X_i, X_2, \ldots, X_x) = P(X_{t+1})$$

since these events are independent. The **memoryless** property means the future is independent of the past. There are certain probability distributions which have this property.

We will mention two in this course, in the discrete case a product of independent Bernoulli trials see coin tossing example.

The only memoryless continuous probability distribution is the exponential distribution. Suppose $X \in [0, \infty)$ is a continuous random variable, the probability distribution of X is said to be memoryless if for any real numbers t, $a \in [0, \infty]$

$$P(X > t + a | X > t) = P(X > a)$$

Exercise 4.6

1. Let X be Exponentially distributed. Show that $E(X) = \frac{1}{\lambda}$.

2. Prove that the Exponential distribution is memoryless.

4.7 Conclusion

We have briefly looked at some types of distributions. Some kinds of data tend to follow certain distributions. These distributions are not so many, rarely do we meet data with an unknown distribution, especially when the sample size is large.

The number of people arriving at a certain bus stop in specified periods of time will generally follow a Poisson distribution. The time taken by radio active material to decay generally follows the exponential distribution. Such natural phenomenon make it necessary for us to study Empirical distributions.

Exercise 4.7

1. The heights (in metres) of children aged between 10 and 14 are recorded below

   $$1.4, \quad 1.5, \quad 1.6, \quad 1.2, \quad 1.63$$

   .

   (a) Find $\overline{x}$ and s^2, and use them to estimate μ and σ^2 respectively.

   (b) If the height is normally distributed, find $P(X < 1.5)$.

2. Find the cumulative probability distributions of the Exponential, Poisson and the continuous uniform distribution.

3. Suppose that X, the number years of schooling a student completes beyond the age of 14, is distributed normally with a mean of 4 and a standard deviation of 2.

   (a) What is the probability that a student completes more than 8 years of schooling beyond the age 14?

   (b) What is the probability that a student completes 2 to 8 years of schooling beyond the age 14?

4. The probability that an individual gets a loan from a bank is 0.25. If 12 people applied for loans what is the probability that

   (a) at least 10 get loans,

   (b) more than 2 but less than 9 get loans,

   (c) exactly 5 get loans?

5. A new typist makes on average 1 error per page on her typing. What is the probability that she will make

   (a) no errors on a page?

   (b) at least 4 errors on at page?

   (c) makes no errors on a 5 page document? [hint: her error rate in now 5 per document]

Chapter 5

Estimation and Hypothesis Testing

5.1 Introduction

The objective of statistics is to make inferences about a population based on information contained in a sample. Statistical inference is mainly concerned with making inferences about population parameters. Methods of making inferences about parameters fall into two categories, making decisions concerning the value of a parameter or estimating/predicting the value of the parameter. The relevant information in a sample can be used to estimate the likely values of their associated population parameters.

We will often need to test the truth of some claims made about a population this will be covered under hypothesis testing.

5.2 Estimation

When a single statistic is used to estimate a population parameter we call it a **point estimator**. A good estimator of a population parameter should at least be an unbiased estimator. The value that the estimator takes, calculated from a sample is called a **point estimate**.

A statistic is called an **unbiased estimator** of a parameter if the expected value of the statistic, for all samples, equals the parameter. This technical definition simply means, if we were to take a representative sample from a population and compute a statistic, and continue taking samples and calculating the corresponding statistics, the average of these statistics should be equal to the population parameter. Estimation can be applied to many population parameters, but in this chapter we

will concentrate on estimation of the population mean and the population standard deviation.

Point Estimator of the Population Mean

A point estimator of the population mean μ is given by the statistic $\overline{x}$. Suppose we have a random sample of observations of size n, $x_1, x_2, \ldots, x_n$. The point estimator of μ is given by $\hat{\mu}$ where

$$\hat{\mu} = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{5.1}$$

One of the most important properties of the sample mean $\overline{x}$, is that it is an **unbiased** estimator of the population mean.

Point Estimator of the Population Standard Deviation

The point estimator of the population variance is given by

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left[\sum_{i=1}^{n} x_i\right]^2}{n}\right] \tag{5.2}$$

This means if you have a representative sample you can estimate the variance of the population. Similarly, the point estimator of population standard deviation is given by,

$$\hat{\sigma} = s = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left[\sum_{i=1}^{n} x_i\right]^2}{n}\right]} \tag{5.3}$$

Point Estimator of the Population Proportion

Suppose we wish to estimate a proportion of the population that have a cellphone in a large city. We may not have enough time to go to every individual to gather the information on cellphone ownership in order to calculate the population proportion π.

If a representative sample of size n is taken, and the number x of individuals with cellphones in the sample is recorded, we can estimate the population proportion using

$$\hat{\pi} = p = \frac{x}{n} \tag{5.4}$$

Example 5.1

A company which manufactures and bottles chemicals, collected a bottle from each batch of 20, which they dispatched to their quality control lab for quality control check-ups. They measured the volumes of chemicals and found the following results:

350, 351, 348, 352, 350, 356, 348, 347, 348, 352, 354

1. Estimate the mean volume of the whole consignment.

2. Estimate the standard deviation of the volume of chemicals.

3. Estimate the proportion of bottles which have less than 350ml in the whole consignment.

Solution

1. The point estimate of the consignment mean is

$$\begin{aligned}\hat{\mu} &= \overline{x} \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ &= \frac{1}{11} \times 3856 \\ &= 350.545\end{aligned}$$

2. The estimated standard deviation is

$$\begin{aligned}\hat{\sigma} &= s \\ &= \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)} \\ &= \sqrt{\frac{1}{10} \times \left(1351782 - \frac{3856^2}{11}\right)} \\ &= 2.8058\end{aligned}$$

3. The estimated proportion is

$$\begin{aligned}\hat{\pi} &= p \\ &= \frac{x}{n} \\ &= \frac{4}{11} \\ &= 0.3636\end{aligned}$$

5.3 Confidence Intervals (Interval Estimation)

Since we used a sample to calculate a point estimate, we can not be sure or state with any confidence on the degree of accuracy of our estimate. The sample mean is one point in the sample space and may not accurately estimate the true mean of the population. For this reason, the interval estimation technique is useful and appropriate for providing some level of confidence on the estimates of population parameter values.

When computing confidence intervals for our population parameters, we need to consider the following:

1. the sample size, n, we are dealing with, and

2. whether we know the population variance σ^2 (standard deviation σ) or not.

We can define a **confidence interval** or an **interval estimate** as a range of values in which the probability that the population parameter will lie in it is known. In confidence interval estimation we will be using the value $z_{\frac{\alpha}{2}}$. This value is found using normal distribution tables. $\frac{\alpha}{2}$ is the value (probability) corresponding to z such that $\mathrm{P}(Z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$.

You will get the value of α from the question. For instance, if you are asked to construct a 95% confidence interval, then

$$\begin{aligned}95\% &= 0.95 \times 100\% \\ &= (1 - 0.05) \times 100\%\end{aligned}$$

so that $\alpha = 0.05$. This then means that $\frac{0.05}{2} = 0.025$. We now do the opposite of what we did in the last chapter, that is, we want to find the value of k that gives $P(Z > k) = 0.025$. Looking down column $\phi(z)$ (depending on the tables), we

get the probability closest to 0.025. In this case the value is 1.96. So in this case $k = z_{\frac{\alpha}{2}} = z_{\frac{0.05}{2}} = z_{0.025} = 1.96$.

z	0.00	$\cdots$	0.06	$\cdots$	0.09
-3.4	0.0003	$\cdots$	0.0003	$\cdots$	0.0002
$\vdots$	$\vdots$		$\vdots$		$\vdots$
-1.9	0.0287	$\cdots$	0.025	$\cdots$	0.0233
$\vdots$	$\vdots$		$\vdots$		$\vdots$
3.4	0.9997	$\cdots$	0.9997	$\cdots$	0.9998

Some normal quantiles,

$Z_{0.10}$	1.28
$Z_{0.05}$	1.645
$Z_{0.025}$	1.96
$Z_{0.01}$	2.33
$Z_{0.005}$	2.58

Case 1: Confidence Interval for μ, when σ^2 is known

A $(1-\alpha)100\%$ confidence interval for μ, when σ^2 known, is given by

$$\overline{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \tag{5.5}$$

where $\overline{x}$ is the mean of a sample of size n from a population with known variance σ^2 , and $z_{\frac{\alpha}{2}}$ is a value in the standard normal distribution that leaves an area of $\frac{\alpha}{2}$ to the right of the normal curve.

Case 2: Confidence Interval for μ, when sample size is large, (i.e. $n > 30$) and σ^2 is unknown

A $(1-\alpha)100\%$ confidence interval for μ, when σ^2 unknown and the sample large is given by

$$\overline{x} - z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \leq \mu \leq \overline{x} + z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \tag{5.6}$$

where $\overline{x}$ is the mean of a sample of size n from a population with an estimated sample variance s^2 and $z_{\frac{\alpha}{2}}$ is the standard normal distribution leaving an area of $\frac{\alpha}{2}$ to the right.

Case 3: Confidence Interval for μ, when the sample size is small, (i.e. $n < 30$) and σ^2 is unknown

If the sample size is small and σ^2 is unknown, we use a Student's t-distribution.

A $(1-\alpha)100\%$ confidence interval for μ, when σ^2 unknown and the sample is small is given by

$$\overline{x} - t_{(n-1,\frac{\alpha}{2})}\frac{s}{\sqrt{n}} \leq \mu \leq \overline{x} + t_{(n-1,\frac{\alpha}{2})}\frac{s}{\sqrt{n}}, \tag{5.7}$$

where $\overline{x}$ and s are the mean and standard deviation, respectively, of a sample of size $n < 30$ from an approximately normal population, and $t_{(n-1,\frac{\alpha}{2})}$ is the value of the t-distribution, with $n-1$ degrees of freedom, leaving an area of $\frac{\alpha}{2}$ to the right.

5.3.1 Using Student's t Tables

If X is normally distributed with mean μ and variance σ^2 then $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ will also be normally distributed with mean 0 and standard deviation 1.

If σ^2 is unknown then we can not standardise X to get Z, instead, we compute a new statistic called the Student's t, where we substitute s for σ, such that

$$T = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}. \tag{5.8}$$

This statistic follows what is called the **Student's** t distribution with $n-1$ degrees of freedom. This is the statistic whose value being used in Equation 5.7. We look up this value from our Student's t tables which look like Table 5.1:

Suppose you want to look up $t_{(n-1,\frac{\alpha}{2})}$ where n=7 and $\alpha = 0.05$. We have:

$$\begin{aligned} t_{(n-1,\frac{\alpha}{2})} &= t_{(7-1,\frac{0.05}{2})} \\ &= t_{6,0.025} \\ &= 2.45 \qquad \text{See arrows} \end{aligned}$$

As the sample size n increases the t-distribution tends to the standard normal distribution. So, for a large sample size, the values of $t_{(n-1,\frac{\alpha}{2})} \approx Z_{\frac{\alpha}{2}}$. For example $z_{\frac{\alpha}{2}} = z_{\frac{0.05}{2}} = z_{0.025} = 1.96$, (see last row column 5).
A few examples are given, each of which should help you appreciate when to apply which formula.

Table 5.1: An example of the Student's t distribution table

	Amount of α in one tail					
				$\downarrow$		
df	0.25	0.10	0.05	0.025	0.01	0.005
1	1.000	3.08	6.31	12.7	31.8	63.7
2	0.816	1.98	2.92	4.30	6.97	9.92
3	0.765	1.64	2.35	3.18	4.54	5.84
4	0.741	1.53	2.13	2.78	3.75	4.60
5	0.727	1.48	2.02	2.57	3.37	4.03
$\rightarrow$6	0.718	1.44	1.94	2.45	3.14	3.71
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
29	0.583	1.31	1.70	2.05	2.46	2.76
For n> 30 z	0.0674	1.28	1.65	1.96	2.33	2.58

Example 5.2

A sample of 36 people at a rave night club revealed an average age of $\overline{x} = 19.38$ years, sample standard deviation $s = 4.760$ years. Determine and interpret a 95% confidence interval for the true mean age of individuals at the rave night club.

Solution
n is large ($n > 30$)
σ is unknown, so we use case 2.

A 95% confidence interval for μ is given by

$$\overline{x} - z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \leq \mu \leq \overline{x} + z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

$$19.38 - (1.96)(\tfrac{4.760}{\sqrt{36}}) \leq \mu \leq 19.38 + (1.96)(\tfrac{4.760}{\sqrt{36}})$$

$$17.83 \leq \mu \ \leq 20.93$$

We are 95% confident that the true mean age lies is at least 17.83 years, and at most 20.93 years.

Example 5.3

The scores below were recorded after Statistics examination scripts for 50 students were marked. The standard deviation of the marks for this examination in the past is known to be 7.93%

73	52	67	53	51	61	49	66	41	48
52	47	65	46	71	67	48	66	47	44
63	65	44	46	61	52	55	54	51	56
49	62	57	56	47	45	56	59	59	47
48	57	48	52	53	52	51	63	68	53

Determine and interpret the 98.44 confidence interval for the true mean of the marks.

Solution

$n = 50,\ \overline{x} = 54.86\%,\ s = 7.95\%,\ \sigma = 7.93\%$

But σ is known, and n is large, so we use case 1.

Then a 98.44% confidence interval for the mean mark is

$$\overline{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

$$54.86 - (2.42)(\tfrac{7.93}{\sqrt{50}}) \leq \mu \leq 54.86 + (2.42)(\tfrac{7.93}{\sqrt{50}})$$

$$52.15 \leq \mu \ \leq 57.57$$

We are 98.44% confident that the true mean mark is at least 52.14% , and at most 57.57%.

Example 5.4

The weights of seven similar containers of a chemical are recorded below.

277.83 289.17 294.84 277.83 283.5 289.17 272.16

Find a 95% confidence interval for the true mean weight of such containers, assuming an approximate normal distribution for the population of weights.

Solution
$\overline{x} = 283.5, n = 7, s = 8.0186$

σ is unknown, n is small, so we use case 3.
$t_{6,0.025} = 2.45$

Then a 95% confidence interval for μ is

$$\begin{aligned} \left(\overline{x} - t_{6,0.025}\frac{s}{\sqrt{n}} \quad ; \quad \overline{x} + t_{6,0.025}\frac{s}{\sqrt{n}}\right) \\ \left(283.5 - (2.45)(\tfrac{8.0186}{\sqrt{7}}) \quad ; \quad 283.5 + (2.45)(\tfrac{8.0186}{\sqrt{7}})\right) \\ (283.5 - 7.4253 \quad ; \quad 283.5 + 7.4253) \\ (276.0747 \quad ; \quad 290.9253) \end{aligned}$$

There is 95% confidence that the true mean weight for such containers is at least 276.0747, and at most 290.9253.

Confidence interval estimation can be extended to cover other parameters like proportions, the standard deviation as well as differences between two means and a ratio of standard deviations.

The ideas used in the construction of confidence intervals can be extended to help determine the size of a sample that will lead us to estimate the mean to any desired degree of accuracy.

5.4 Sampling to a Desired Precision

Let $x_1, x_2, \ldots, x_n$ be a random sample from a population following $N(\mu, \sigma^2)$. Let $\overline{x}$ be sample mean and σ be population standard deviation, respectively. Then the minimum sample size required for $\overline{X}$ to be within ϵ of the true mean μ with a $100(1-\alpha)\%$ probability is

$$n = \left[\frac{Z_{\frac{\alpha}{2}} \times \sigma}{\epsilon}\right]^2$$

Example 5.5

The manager of a department wants the margin of error of the mean number of calculation errors in reports in his department to be within $\pm$ 3 points for the year 2016 trainees. The extent to which this error is likely to occur is 0.95. What sample size should he take of reports if the standard deviation is known to be 9.2?

Solution

$$\begin{aligned} n &= \left[\frac{z_{\frac{\alpha}{2}} \times \sigma}{\epsilon}\right]^2 \\ &= \left[\frac{1.96 \times 9.2}{3}\right]^2 \\ &= 36.128 \end{aligned}$$

Therefore, 36 reports is the minimum needed sample size.

Exercise 5.1

1. Show that $\overline{x}$ is an unbiased estimator of μ.
2. Show that $var(\overline{x}) = \frac{\sigma^2}{n}$.
3. Show that $\hat{s}^2$ is an unbiased estimator of σ^2.
4. Suppose

   22.4 31.6 -3.8 34.0 -35.2 -21.5 23.0 10.8 23.2

   is a random sample from $N(\mu, \sigma^2)$. Construct a 90% confidence interval for the mean μ in each of the following cases.

   (a) $\sigma^2 = 7.24$ is known.
   (b) σ^2 is unknown.

5. A sample of 44 retired gentlemen was drawn from a group of pensioners and their weekly cigarette expenses recorded in order to estimate the population mean cigarette expenses. A mean of R325.00 and a standard deviation of R125.00 was found from the sample.

   (a) Estimate μ, the mean amount spent on cigarettes during the past year.
   (b) Find a 99% confidence interval of the true mean expenditure.
   (c) How large a sample is needed so that one is
      i. 95% confident that the sample mean will be within 0.1 of the true mean?
      ii. 98% confident that the sample mean will be within 0.5 of the true mean?

We can use confidence intervals to test the validity of a claim. Tests of claims are best addressed through hypothesis testing.

5.5 Hypothesis Testing

Often before you engage in any study or eventually make any recommendations, you will need supporting evidence. It has become generally acceptable that recommendations and findings be accompanied by statistical evidence. This is often in the form of statistical tests.

The testing of a statistical hypothesis is perhaps the most important part of decision making.

What is a Statistical Hypothesis?

A statistical hypothesis is an assumption or statement, which may or may not be true, concerning one or more populations.

The truth or falsity of a statistical hypothesis is never known with certainty unless we examine the entire population. A random sample is taken from the population of interest and the information contained in this sample is used to decide whether the hypothesis is likely to be true or false. Evidence that is inconsistent with the stated hypothesis leads to a rejection of a claim whereas evidence supporting the hypothesis leads to failure of its rejection/its acceptance. For example, one might want to test the the hypothesis that the pass rate for an exam is always 40%, or test the claim that women live longer than men.

There are two types of hypothesis: a **null** hypothesis and an **alternative** hypothesis.

Null Hypothesis, H_0

The null hypothesis, usually denoted H_0, is a claim that is held as true about a population parameter or characteristic.

Alternative Hypothesis, H_1

The alternative hypothesis, usually denoted by H_1, expresses the way in which the value of a population parameter or characteristic may deviate from that specified in the null hypothesis.

We fail to reject the null hypothesis if we do NOT have sufficient evidence to reject it. There is however a chance that reaching this decision maybe erroneous. This then leads us to a discussion of two possible outcomes of this nature; we can reject the null hypothesis when it is in fact correct, or we can fail to reject it when it is in fact wrong. A wrong decision could lead to (a) wrong conclusion(s).

There are two main types of errors:

1. **Type I error**: H_0 is rejected when it is in fact true.
2. **Type II error**: H_0 is not rejected when it is in fact false.

The probability of committing a Type I error is known as the **the level of significance** of the test, denoted by α. This is the value we were using in confidence interval construction. The value of α can vary at the researcher's discretion. However a value of $\alpha = 0.05$ is a common benchmark for reasonable doubt.

The probability of committing a Type II error is denoted by β. This probability helps us determine the power of our test. The **power of a test** can be defined as the probability of a test to make a correct decision. So we can say the power of a test = 1 - P(making a Type II error) = 1 - β

Formulating a Hypothesis Test

When testing a hypothesis, the following steps are followed;

1. **State the hypotheses**
   Null Hypothesis (H_0): There is no difference between population parameter and given parameter.
   Alternative Hypothesis (H_1): There is a difference between population parameter and given parameter.

   Note: The alternative hypothesis will indicate whether a *1-tailed* or a *2-tailed* test is utilized to reject the null hypothesis.

2. **Set the rejection criteria**
   This determines how different the parameters and/or statistics must be before the null hypothesis can be rejected. This "region of rejection" is based on α the significance level. The point of rejection is known as the **critical value**.

3. **Compute the test statistic**
   The collected data are converted into standardized scores for comparison with the critical value.

4. **Decide results on null hypothesis**
   If the test statistic equals or exceeds the region of rejection bracketed by the critical value(s), the null hypothesis is rejected.

5. **Conclude**
   Make a conclusion based on the results obtained.

Terms used in Hypothesis Testing

A **critical value** is a value that separates the acceptance region from the rejection region.

An **acceptance region** is a range of values of the sample statistics centered about the null hypothesised population parameter that would lead to the failure of H_0 rejection/acceptance of the null hypothesis for values of the sample statistic which fall within its limits.

A **rejection region** is the range of values of sample statistic value that would lead to the rejection of the null hypothesis for values of the sample statistic which fall within its limits.

A **two-tailed test** has an area of rejection both below and above the hypothesised value.

A **one-sided and right-tailed test** has an area of rejection which lies above the null hypothesised value of the population parameter

A **one-sided and left-tailed test** has the area of rejection which lies below the null hypothesised value of the population parameter.

A **test statistic** is a value calculated from the sample data which is used to decide whether or not H_0 should be rejected.

Hypothesis Testing for a Single Population Mean

In this chapter we will confine our discussion of hypothesis testing to tests concerning the population mean. The hypotheses we will discuss are:

1. Two-sided test: $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$
2. One-sided test (right-tailed): $H_0 : \mu = \mu_0$ versus $H_1 : \mu > \mu_0$.
3. One-sided test (left-tailed): $H_o : \mu = \mu_0$ versus $H_1 : \mu < \mu_0$.

Let us now proceed to test a hypothesis. When calculating the test statistic of a single mean, one has to ascertain whether the variance is known or not, and also have an appreciation of the sample size.

The following test statistics are used depending on whether σ^2 is known.

Case 1: When σ^2 is known: Use Z, where

$$z = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{5.9}$$

Case 2: When σ^2 is unknown and $n < 30$: Use T where

$$t = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}} \tag{5.10}$$

Case 3: When σ^2 is unknown and n large Use Z where;

$$z = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}} \tag{5.11}$$

5.5.1 Two-sided tests

Testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ when σ^2 is known

The test statistic is

$$Z_{calc} = \frac{\overline{X} - \mu_o}{\frac{\sigma}{\sqrt{n}}} \tag{5.12}$$

and $|Z_{calc}|$ tends to be large if H_0 is false and small otherwise. An α-size test is to reject H_0 if $|Z_{calc}| > z_{\frac{\alpha}{2}}$

Testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ when σ^2 is unknown $n < 30$

The test statistic is

$$T_{calc} = \frac{\overline{X} - \mu_o}{\frac{s}{\sqrt{n}}} \tag{5.13}$$

and $T \sim t_{n-1}$. An α-size test is to reject H_0 if $|T_{calc}| > t_{n-1,\frac{\alpha}{2}}$

Testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ when σ^2 is unknown, n Large

The test statistic is

$$Z_{calc} = \frac{\overline{X} - \mu_o}{\frac{s}{\sqrt{n}}} \tag{5.14}$$

An α-size test is to reject H_0 if $|Z_{calc}| > z_{\frac{\alpha}{2}}$

Example 5.6

The ages of a sample of employees who come to work very early were observed in one financial institute. The following information from the sample was obtained: $\overline{x} = 26.3$ and $n = 36$. Also, $\sigma^2 = 9$. Test the hypothesis that the mean age is 25. Assume $\alpha = 0.005$.

Solution
Since σ^2 is known, we are using Case 1, the Z distribution. The test is a two tailed test.

$H_0 : \mu = 25$
$H_1 : \mu \neq 25$
$\alpha = 0.005 \Rightarrow Z_{0.0025} = 2.81$

The critical region is $Z < -2.81$ and $Z > 2.81$. Thus, reject H_0 if $|Z_{calc}| > 2.81$.

The test statistic is

$$\begin{aligned} Z_{calc} &= \frac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \\ &= \frac{26.3 - 25}{\frac{3}{\sqrt{36}}} \\ &= 2.6 \end{aligned}$$

Since $|Z_{calc}| < 2.81$, that is, $2.6 < 2.81$, we fail to reject H_0.

We state that there is insufficient evidence at 0.5% level of significance to conclude that the average age of early comers is significantly different from 25.

Example 5.7

It is desired that a statistician test the hypotheses given below and make correct recommendations based on the information gathered from a sample of measurements of 40mm pipes which were supplied by a new supplier. Assume $\alpha = 0.01$
$H_0 : \mu = 40$, $H_1 : \mu \neq 40$, $n = 18$, $\overline{x} = 34$, and $s^2 = 64$.

Solution
Since n is small and σ unknown, we use the t-distribution. The test is a two tailed test.

$H_0 : \mu = 40$
$H_1 : \mu \neq 40$
$\alpha = 0.01 \Rightarrow t_{n-1,\frac{\alpha}{2}} = t_{17,0.005} = 2.90$

The critical region is $T < -2.8982$ and $T > 2.8982$. Thus reject H_0 if $|T_{calc}| > 2.8982$.

The test statistic is

$$\begin{aligned} T_{calc} &= \frac{\overline{X} - \mu_o}{\frac{s}{\sqrt{n}}} \\ &= \frac{34 - 40}{\frac{8}{\sqrt{18}}} \\ &= -3.182 \end{aligned}$$

Since $|T_{calc}| > 2.90$, that is, $3.182 > 2.90$, we reject H_0.
We state that there is sufficient evidence at a 1% level of significance to conclude that mean is significantly different from 40. The pipes are not 40mm.

Exercise 5.2

1. A geologist is testing the hypothesis that the melting point of an unusual carbon substance is 1946^{o}C. He makes 7 determinations and obtained the values of 1944, 1947, 1945, 1947, 1949, 1946 and 1944^{o}C. What conclusions can you draw at a significance level of 0.05?

2. An insurance broker claims that teachers spend an average of R65.00 per month life insurance. To test this hypothesis a random sample of 124 teachers was taken and it was observed that on average teachers spend about R52.24 with a standard deviation of R11.64. Using a 0.01 significance level, does this support the broker's claim.

3. A baker stated that on average the number of loaves bread sold daily is 3 000 with a standard deviation of 300. An employer want to test the accuracy of this statement. A random sample of 36 days showed the average daily sales were 3 150. Test at the 1% level of significance if the bakery's statement can be accepted.

5.5.2 One-sided tests (right-tailed)

The ideas discussed in Section 5.5, cases 1, 2 and 3 can be extended to testing this one sided test. The test statistic to be used in this test will be determined by

whether we know σ^2 or not, and by the sample size n (whether its a small sample or a large sample), see Section 5.5.

We wish to test the hypotheses
$H_0 : \mu = \mu_0$ versus
$H_1 : \mu > \mu_0$

Depending on the sample size and knowledge of σ^2, we reject H_0 if *test statistic* > *tabulated value*$_{(\alpha)}$

1. If σ^2 is known, then use $z = \frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$, and reject H_0 if $z > z_\alpha$
2. If σ^2 is unknown and $n < 30$, then use $t = \frac{\overline{x}-\mu}{\frac{s}{\sqrt{n}}}$, and reject H_0 if $t > t_\alpha(n-1)$
3. If σ^2 is unknown and $n > 30$, then use $z = \frac{\overline{x}-\mu}{\frac{s}{\sqrt{n}}}$, and reject H_0 if $Z > z_\alpha$

Example 5.8

The average distance traveled by a small engined vehicle on 10 litres of petrol is 162.5 kms with a standard deviation of 6.9 kms. Is there reason to believe that adding a new additive to petrol increases the distance travelled on 10 litres if a random sample of 50 small cars has an average of 165.2 kms per 10 litres, at the 5% level of significance?

Solution
n = 50, $\overline{x} = 165.2$ and $\sigma = 6.9$. Since σ is known we use the Z distribution. The test is a one-tailed test.

$$H_o : \mu = 162.5$$

$$H_1 : \mu > 162.5$$

$$\alpha = 0.05 \Rightarrow z_\alpha = z_{0.05} = 1.645$$

The critical region is $Z > 1.645$. Thus reject H_0 if $Z_{calc} > 1.645$.

The test statistic is

$$\begin{aligned} Z_{calc} &= \frac{\overline{X} - \mu_o}{\frac{\sigma}{\sqrt{n}}} \\ &= \frac{165.2 - 162.5}{\frac{6.9}{\sqrt{50}}} \\ &= 2.7669 \end{aligned}$$

Since $z_{calc} > z_{0.05}$, that is, $2.7669 > 1.645$, we reject H_0.

We state that there is sufficient evidence at a 5% level of significance to conclude that the the additive significantly increases the distance travelled on 10 litres of petrol.

Exercise 5.3

1. A supermarket claims that customers to its stores spend on average 25 minutes carrying out their purchases. A consumer body wants to verify this claim. They observed then entry and departure times from supermarkets in the chain of 24 random selected customers. The sample average time was half an hour with a standard deviation of 14.1 minutes. Test the validity of the supermarket's belief at the 2.5% level of significance.

2. It is claimed that an automobile is driven on average less than 12 000kms per year. To test this claim a random sample of 100 automobile owners are asked to keep a record of the kilometres they travel. Would you agree with this claim if the random sample showed an average of 14 500 kilometres and a standard deviation of 2 400kilometres? Use a 0.01 level of significance.

5.5.3 One-sided tests (left-tailed)

The test statistic to be used in this test will be determined by whether we know σ^2 or not, and by the sample size n (whether its a small sample or a large sample), as we did in Section 5.5.2.

We wish to test the hypotheses
$H_0 : \mu = \mu_0$ versus
$H_1 : \mu < \mu_0$

Depending on the sample size and knowledge of σ^2, we reject H_0 if *test statistic* $<$ *tabulated value*$_{(\alpha)}$

The critical values, for particular tests, sample sizes and significance levels, are available available in tables. Remember, sometimes, we have to extrapolate the critical values from tables.

1. If σ^2 is known, then use $z = \frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$, and reject H_0 if $z < -z_\alpha$
2. If σ^2 is unknown and $n < 30$, then use $t = \frac{\overline{x}-\mu}{\frac{s}{\sqrt{n}}}$, and reject H_0 if $t < -t_{n-1,\alpha}$
3. If σ^2 is unknown and $n > 30$, then use $z = \frac{\overline{x}-\mu}{\frac{s}{\sqrt{n}}}$, and reject H_0 if $z < -z_\alpha$

Example 5.9

The average length of time spent in a bank queue has been 50 minutes. A new banking system is testing a new software. If a random sample of 12 clients had an average banking time of 42 minutes with a standard deviation of 11.9 minutes under the new system, test the hypothesis that the population mean is now less than 50 using a level of significance of

(i) 0.05, and

(ii) 0.01.

Assume the population of times to be normal.

Solution.
Let $H_0 : \mu = 50\ minutes$ versus $H_1 : \mu < 50\ minutes$.

(i) Using t-distribution at $\alpha = 0.05 \Rightarrow t_{n-1,0.05} = t_{11,0.05} = 1.80$. Reject H_0 if $T < -1.80$, that is, critical region is $[-\infty, -1.80]$. The test statistic is

$$\begin{aligned} T &= \frac{\overline{X} - \mu_o}{\frac{s}{\sqrt{n}}} \\ &= \frac{42 - 50}{\frac{11.9}{\sqrt{12}}} \\ &= -2.3288 \end{aligned}$$

Since $-2.3288 < -1.80$, that is, the test statistics is in the critical region, we reject the null hypothesis at 5% level of significance.

(ii) Using t-distribution at $\alpha = 0.01 \Rightarrow t_{11,0.01} = t_{11,0.01} = 2.72$ Reject H_o if $T < -2.72$, that is critical region is $[-\infty, -2.72]$. Since $-2.33 > -2.72$, we do not reject H_0 at 1% level of significance.

Exercise 5.4

1. A cement manufacturer packs 50kg bags. To see if the manufacturer puts enough cement in the bags, the contents of random sample of 200 such bags were weighed. The average contents turned out to be 48kgs with a standard deviation of 0.12. State the null and research hypothesis for this problem. Hence test at the 0.05 significance level, whether or not the manufacturer satisfies the requirement?

2. A scientist claimed that mice with an average life span of 28 months will live to be about 43 months old when 45% of the calories in their food are replaced by vitamins and proteins. Is there any reason to believe that the mean age is less than 43 months if 54 mice that are placed on this diet have an average life of 38 months with a standard deviation of 5.8 months? Use a 0.025 level of significance.

3. Redo number 2, and use a sample of size 20 mice instead of 54 mice that are placed on this diet.

Note

In inference we use samples to make inferences about a population. You will never have to calculate the population standard deviation. You are either given its value or you do not know it.

The p-value

What is the p-value? You will find this value under a variety of names, some of these are; **critical level**, **the probability value** and the **associated probability**.

Suppose, for a given null hypothesis H_0, we calculate a test statistic, say k_{test}, then

$$\text{p-value} = P(k \geq k_{test}|H_0)$$

where k is a real number. Generally we reject H_0 at the α level, if the p-value$< \alpha$.

You can can also say the p-value is the smallest value of α for which test results are statistically significant.

The p-value is commonly used in statistical computer packages. Once calculated you do not need statistical tables.

Exercise 5.5

1. Suppose it is known that a variable X is a normally distributed with a mean of 340 minutes. If a random sample of 20 observations has an average of 332 with a standard deviation of 43 minutes. Test the hypothesis at the 0.025 level of significance at that $\mu = 340$ minutes against the alternative $\mu < 340$ minutes.

2. Suppose the observations

   1.4 -2.6 1.3 2.1 -2.2 -3.6 3.2 1.8 2.4

   are independent and identically distributed as $N(\mu, \sigma^2)$. Test the hypothesis at 5% significance level that the mean of the population is 2.2.

3. A random sample of size 14 from a normal distribution has a mean $\overline{X} = 33.2$ and a standard deviation of $s = 5.41$. Does this suggest, at the 0.05 level of significance, that the population mean is greater than 32?

4. A manufacturer claims that the average life of batteries produced by his firm is at least 30 months. You disagree, contending that the average life of the batteries is less than 30 months. A random sample of 12 batteries has a mean of 38.7 months and a standard deviation of 18 months. Perform the appropriate hypothesis test. Use a significance level of 0.05.

5. The manufacturer of an over the counter pain reliever claims that its product brings pain relief to headache sufferers in less than 3.5 minutes on average. To be able to make this claim in its television advertisements the manufacturer was required by a particular television network to present statistical evidence in support of the claim. The manufacturer reported that for a random sample 50 headache sufferers, the mean time to relief was 3.3 minutes and the standard deviation was 66 seconds. Does this data support the manufacturer's claim. Test using $\alpha = 0.05$.

6. What is the advantage of using the p-value over the critical value?

5.6 Summary

In this chapter we discussed two concepts: estimation and hypothesis testing. These two concepts fall under the topic known as Statistical Inference. In statistical inference, we use sample data to make inferences about the population from where the sample is drawn.

You should now be able to calculate the point and interval estimates of a population. You should also be able to test a variety of hypotheses, depending on the question.

Chapter 6

Correlation and Regression Analysis

6.1 Introduction

Correlation analysis is the study of the strength of the relationship between **any pair** of variables. Correlation measures how strongly pairs of variables are related. **Regression** analysis on the other hand describes a collection of statistical techniques that quantifies how change in one variable depends on a particular another (or several other variables). Regression analysis is now perhaps the most widely used method of modelling relationships.

6.2 Scatter Plots

The easiest way to visualise the nature of the relationship between variables is to use a scatter plot. A **scatter plot** is a plot of one variable against another. You may use three variables to get a three-dimensional plot. By looking at the scatter plot, you can get an idea of the relationship between variables, that is, whether the variables are **linearly** related or related in some other way.

Note:
A scatter plot is made up of the axes and a plot of the points where values meet. The points should not be connected together using a line. An example of a scatter plot is shown in Figure 6.1.

Example 6.1

The incomes of a sample of individuals and amounts they spent on entertainment at a bar was recorded, and a scatter plot constructed.

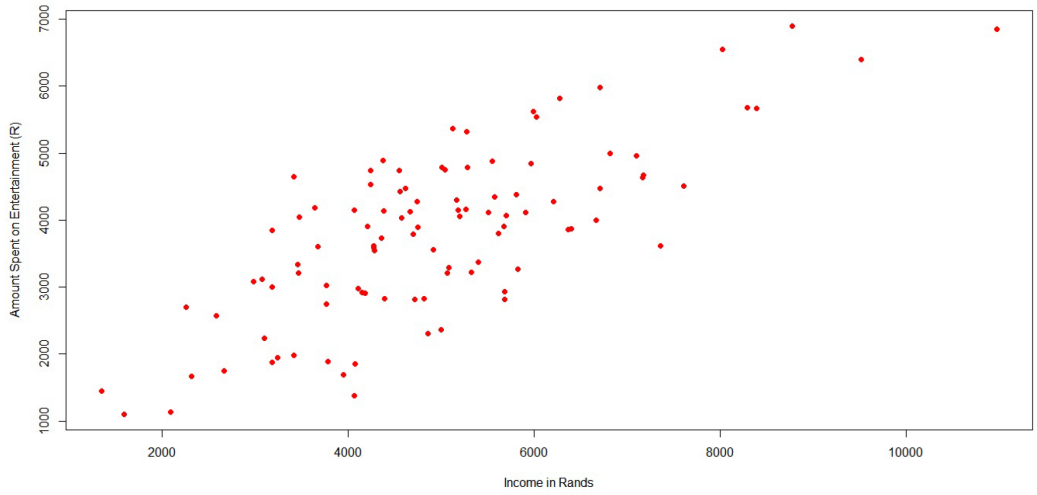


Figure 6.1: Scatter plot of incomes versus entertainment

The scatter plot shows that as income increases the amount spent on entertainment also tends to increase.
It is important to construct a scatter plot in order to get an understanding of what you will expect the relationship to be.

6.3 Correlation

Correlation is the **intensity** or **strength** of the relationship between two variables. It is a measure of the extent to which variables are related or associated. If the correlation between two variables is zero, then the two variables are not related. On the other hand, a correlation of 1, means that there is a perfect relationship between the two variables, be it linear or otherwise. Here, "perfect" means an exact relationship.

In this course, we concentrate on the linear relationship.

6.3.1 Correlation Coefficient

Symbolically, the population correlation between two variables X and Y is given by the correlation coefficient ρ_{XY}. This is defined by:

$$\rho_{XY} = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}}.$$

We rarely deal with population parameters as they stand. We often estimate the population parameters on the basis of samples. In this particular case, ρ_{xy} is esti-

mated from a sample, say (x_1, y_1), (x_2, y_2), ..., (x_n, y_n), to give $\hat{\rho}_{xy}$.

This quantity is calculated using the formula given in Equation 6.1:

$$\hat{\rho}_{xy} = \frac{\sum_{i=1}^{n}[x_i - \bar{x}][y_i - \bar{y}]}{\sqrt{\sum_{i=1}^{n}[x_i - \bar{x}]^2 \times \sum_{i=1}^{n}[y_i - \bar{y}]^2}} \tag{6.1}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2\right)\left(\sum_{i=1}^{n} y_i^2 - \frac{1}{n}(\sum_{i=1}^{n} y_i)^2\right)}} \tag{6.2}$$

There are other measures of association, but for now we shall work with the one above. ρ_{xy} is confined to the interval $[-1 \leq \rho_{xy} \leq 1]$. If $\rho_{xy}=$ -1, then there is an exact inverse relationship between x and y. If x increases, then y decreases and vice-versa.

Example 6.2

Consider the number of expensive goods n sold by a company (which got the goods at a very cheap cost price) and the profit p. $\rho_{np} \approx 1$ in this case. This means that the more the sales, the higher the profit. ("$\approx$" here means approximately) If $\rho_{np} \approx 0$, then we would conclude that, there is no relationship between the profit and number of sales.

On the other hand, $\rho_{np} \approx -1$ suggests that the more the sales the lower the profit. The product is probably being sold at less than the cost price.

6.3.2 Correlation Matrix

You will notice that the correlation coefficient is defined for two variables. Now let us assume we have many variables. Take, for instance, p variables, then how do we handle the correlation problem? Let these variables be $X_1,\ X_2,\ \cdots,\ X_p$. Then we can express the different combinations of correlation coefficients in a matrix, known as a **correlation matrix**, as follows:

$$\rho = \begin{pmatrix} 1 & \rho_{x_1x_2} & \cdots & \rho_{x_1x_p} \\ \rho_{x_2x_1} & 1 & \cdots & \rho_{x_2x_p} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ \rho_{x_px_1} & \rho_{x_px_2} & \cdots & 1 \end{pmatrix}$$

$\rho_{x_i x_j}$ for $i,\ j = 1,\ 2,\ \cdots,\ p$, represents the correlation between variables X_i and X_j. Therefore by looking at the correlation matrix, we can tell which variables are correlated. Notice that the correlation between X_k and itself is always 1. For example, let us look at the sample correlation coefficient for any k, $\hat{\rho}_{x_k x_k}$. Then:

$$\begin{aligned}
\hat{\rho}_{x_k x_k} &= \frac{\sum_{i=1}^{n}[X_{ki}-\overline{X_k}][X_{ki}-\overline{X_k}]}{\sqrt{\sum_{i=1}^{n}[X_{ki}-\overline{X_k}]^2 \times \sum_{i=1}^{n}[X_{ki}-\overline{X_k}]^2}} \\
&= \frac{\sum_{i=1}^{n}[X_{ki}-\overline{X_k}]^2}{\sqrt{[\sum_{i=1}^{n}[X_{ki}-\overline{X_k}]^2]^2}} \\
&= \frac{\sum_{i=1}^{n}[X_{ki}-\overline{X_k}]^2}{\sum_{i=1}^{n}[X_{ki}-\overline{X_k}]^2} \\
&= 1
\end{aligned}$$

This gives us the matrix:

$$\hat{\rho} = \begin{pmatrix} 1 & \hat{\rho}_{x_1 x_2} & \cdots & \hat{\rho}_{x_1 x_p} \\ \hat{\rho}_{x_2 x_1} & 1 & \cdots & \hat{\rho}_{x_2 x_p} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ \hat{\rho}_{x_p x_1} & \hat{\rho}_{x_p x_2} & \cdots & 1 \end{pmatrix}$$

Example 6.3

Suppose you are given three variables z_1, z_2 and z_3 representing the coded tensile strength, melting point and amount of Titanium in a new alloy respectively. Use the following data to calculate the correlation matrix.

z_1	z_2	z_3
12	3	0.2
23	7	0.8
9	2	0.1
30	10	1.0

$$\begin{aligned}
\hat{\rho}_{z_1 z_2} &= \frac{\sum_{i=1}^{n}[z_{1i}-\overline{z_1}][z_{2i}-\overline{z_2}]}{\sqrt{\sum_{i=1}^{n}[z_{1i}-\overline{z_1}]^2 \times \sum_{i=1}^{n}[z_{2i}-\overline{z_2}]^2}} \\
&= 0.999 \\
\hat{\rho}_{z_1 z_3} &= \frac{\sum_{i=1}^{n}[z_{1i}-\overline{z_1}][z_{3i}-\overline{z_3}]}{\sqrt{\sum_{i=1}^{n}[z_{1i}-\overline{z_1}]^2 \times \sum_{i=1}^{n}[z_{3i}-\overline{z_3}]^2}} \\
&= 0.993 \\
\hat{\rho}_{z_2 z_3} &= \frac{\sum_{i=1}^{n}[z_{2i}-\overline{z_2}][z_{3i}-\overline{z_3}]}{\sqrt{\sum_{i=1}^{n}[z_{2i}-\overline{z_2}]^2 \times \sum_{i=1}^{n}[z_{3i}-\overline{z_3}]^2}} \\
&= 0.988
\end{aligned}$$

$$\hat{\rho} = \begin{pmatrix} 1 & 0.999 & 0.993 \\ 0.999 & 1 & 0.988 \\ 0.993 & 0.988 & 1 \end{pmatrix}$$

The correlation between any two of the variables is very high meaning that the variables are highly correlated. They either increase or decrease together. Addition of Titanium increases the tensile strength and melting point of the alloy. The process seems to produce an alloy whose strength can be increased or decreased, by changing the amount of Titanium.

Exercise 6.1

Consider the following data collected on two variables which are suspected to be related. The variable x represents the grade scores of students in a class and y the number of students with the same grade.

Grade, x	2	5	9	3	6	7
Frequency, y	34	45	59	40	50	48

The data above shows the tabulation relating the two variables x and y. From this illustration, it is clear that there is a linear relationship between the two variables x and y. As one variable increases, the other variable also increases. This means that as the grade increases the number of students with high grades also increases.

Calculate the correlation coefficient of the grade and the frequency. Comment on your result.

When we calculated the correlation coefficient we quantified the strength of the relationship between two variables. In regression analysis, we study how one variable (called the **dependent** variable) depends on other variables (called the **independent** variable). In this course we will introduce the case of one independent variable.

In real life situations we are usually interested in the relationship between variables. For example, it is difficult to study the impact or effect of salary increments, food price increases, etc. on inflation by using descriptive techniques. Thus, regression analysis equips you with a way of studying such situations.

Note: Descriptive techniques generally look at **one variable** at a time, while regression analysis looks at the dependency **between variables**.

6.4 Simple Linear Regression Analysis

Simple regression analysis is seldom used in applied research because the workings of most socio-economic systems cannot be adequately represented by such a simple formulation. However, knowledge of simple regression analysis is a good foundation for understanding multiple regression analysis [which is usually used in applied research].

The term **simple** implies that a single independent variable x is involved in determining the change in y, and the term **linear** implies linearity in the parameters.

Consider the following equation which relates Y and X:

$$Y = f(X) \tag{6.3}$$

where f is a function showing how Y is related to X. In literature, the response variable, i.e., the variable of interest Y, and the explanatory or independent variable X, which is used to explain Y, have several names. We give these names in the table below;

Table 6.1: Names given to response and explanatory variables in Regression Analysis

	Y	X
(a)	Predictand	Predictor
(b)	Regressand	Regressor
(c)	Dependent variable	Independent variable
(d)	Effect variable	Causal variable
(e)	Endogenous variable	Exogenous variable
(f)	Target variable	Control variable

Each pair of the above terms is appropriate for a particular use of regression analysis. For example, the terminology in (a) is often used if the purpose of the regression is prediction; pairs (b), (c) and (d) are used by different applied researchers in their discussion of **regression models**; (e) is usually used in studies of causation or causality; while pair (f) is more appropriate in control problems.

6.4.1 Types of Relationship

There are two main types of relationship between or among variables, namely **exact** and **statistical** relationships.

Exact Relationships

An exact relationship is a relationship of the form:

$$y_i = \beta_0 + \beta_1 x_i \tag{6.4}$$

where the subscript i (for $i = 1, 2, \cdots, n$) refers to the i^{th} observation. What does this mean? Well, for any value of x_i the y_i value will be equal to some constant value β_0 added to the $\beta_1 \times x_i$. β_0 and β_1 are constants. y_i is determined by x_i, i.e., a unit change in X causes a change equal to β_1 in Y.

Note
The variables X and Y can be random or deterministic i.e., non-random.

Generally this equation can be expressed as follows:

$$y = \beta_0 + \beta_1 x \tag{6.5}$$

where x and y are possible values of X and Y respectively. An example of an exact relationship follows.

Exercise 6.2

The relationship between the area of a square A and the length of one side L is given by:

$$Area = \beta_0 + \beta_1 \times length^2$$

where $\beta_0 = 0$ and $\beta_1 = 1$. This is an exact relationship. Figure 6.2 gives a plot of the relationship between Area $[A]$ and the square of the length $[L^2]$.
If we had plotted A against L, we would have obtained the relationship shown in Figure 6.3

These are both exact relationships. In Figure 6.2, we plotted A versus L^2 which gives a linear relationship of the form $y = \beta_0 + \beta_1 x$. On the other hand, in Figure 6.3 we plotted A versus L, we notice that, this time, there is a quadratic relationship.

Statistical Relationships

A statistical relationship, unlike an exact relationship, is not a perfect one, that is, it does not give unique values of Y for a given value of X, but can be described exactly in probabilistic terms. For instance, consider the following **regression model** showing a statistical relationship between Y and X which is no longer exact because

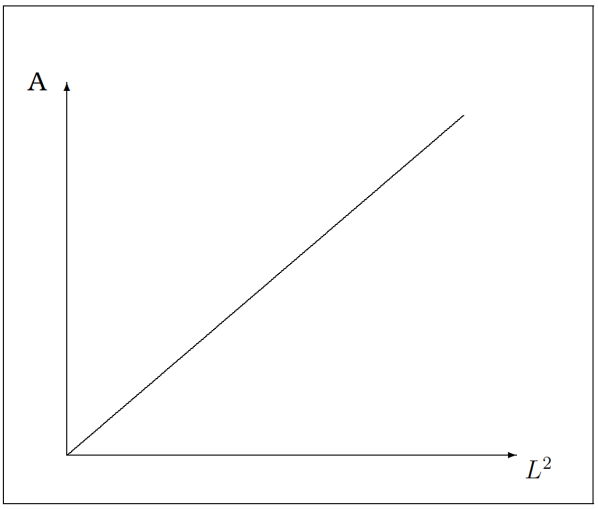


Figure 6.2: The relationship between A and L^2

of the error term ϵ_i:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i. \tag{6.6}$$

The variable ϵ_i is a value added to the equation to make the two sides of the equation equal. The term ϵ_i is called the **error term**. The error term is usually assumed to have a normal distribution with mean 0 and variance σ^2. The relationship between Y and X in Equation 6.6 is called a **stochastic** or **statistical** relationship because of the presence of the random error term needed to make the equation exact.

Definition 3 *A regression model is a statistical relationship between a dependent variable, say Y, and some explanatory variable(s), say X. The model is said to be deterministic if the explanatory variable X is non-random.*

Note: The following term from Equation 6.6:

$$\beta_0 + \beta_1 X_i \tag{6.7}$$

is the **deterministic component** of Y_i, and β_0 and β_1, are called **regression coefficients** or **regression parameters**. ϵ_i is the **stochastic** or **random disturbance term**, it takes care of all the variation of Y which is not explained by the deterministic component $\beta_0 + \beta_1 X_i$. We will discuss how to estimate the regression coefficients from a given data later.

Let us look at a practical example of a statistical relationship in order for us to appreciate the differences between an exact relationship and a statistical relationship.

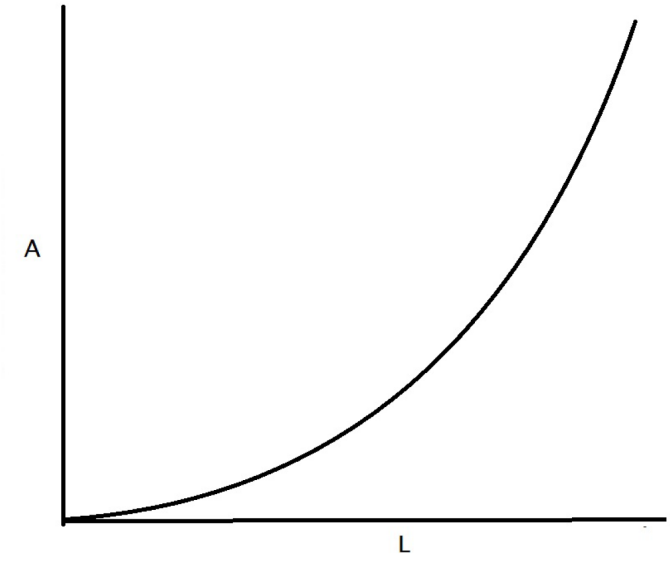


Figure 6.3: The relationship between A and L

Exercise 6.3

[Statistical relationship] A group of students are interested in evaluating the advantages and disadvantages of different study patterns and their effect on their performance. Consider Y, the mark a student gets after an examination, and X_1, the number of hours the student puts into reading for the examination.

The variable X_1 was chosen by the students because it seemed [appeared] to contribute a lot to the examination mark. A possible equation to represent the relationship between Y and X_1 is given as:

$$y = \beta_0 + \beta_1 x_1 + \epsilon \tag{6.8}$$

β_0 and β_1 are unknown constants or regression parameters. If a student puts 0 hours into studying for the examination, then we expect him/her to get β_0 marks. On the other hand, if a student increases his/her study time by one hour, the model suggests that the mark should change by β_1. Please note that, as in Equations 6.8, we will index y and x to y_i and x_i respectively, when we have the actual observations $x_1, x_2, \cdots, x_n$ and $y_1, y_2, \cdots, y_n$ at hand.
Notice how the points in Figure 6.4 are not always on the line. This is because the relationship between examination mark and the time spent on studying is not exact.

The students could have added x_2, the number of books the student consulted as a variable, since this appears to have an impact on the final examination mark. This

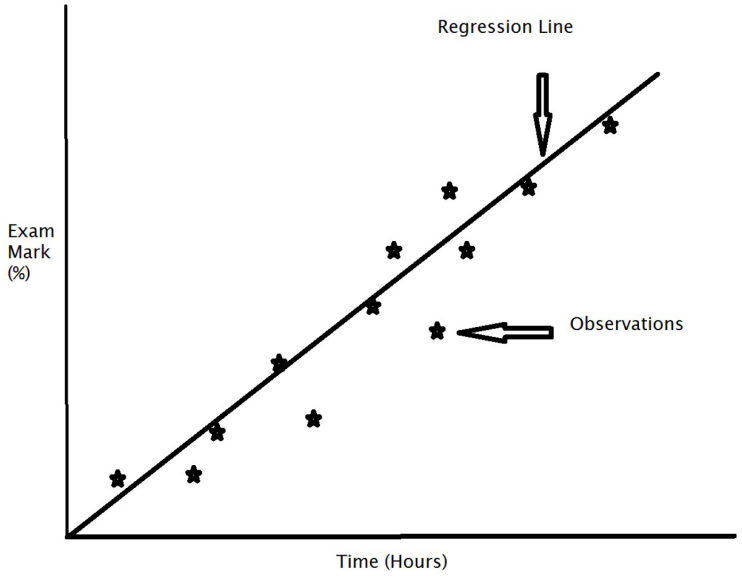


Figure 6.4: Examination mark versus time spent studying

would give Equation 6.9.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \tag{6.9}$$

Procedures embraced by regression analysis concern themselves with drawing conclusions about these coefficients. An example of the implications of these coefficients follows from the fact that a positive coefficient means that the more the hours spent studying, the higher the examination mark etc. The term ϵ in the equation is added to account for the fact that the equation is not exact. If there are p explanatory variables, a regression equation can be expressed more generally as in Equation 6.10.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \tag{6.10}$$

This is referred to as **multiple linear regression**.

When to apply Regression Analysis

There are conditions which must be satisfied before we can apply regression analysis. The first condition to be met is that the variables of concern should be related to each other, otherwise the idea of regression collapses. The second condition is that one variable should change in response to the other, i.e there should be a dependence relationship. How do we check on these requirements? This is often done by:

1. constructing a **scatter plot**.
2. calculating the correlation of the variables.

Exercise 6.4

Figure 6.5, shows examples of a scatter plots showing the relationship between an independent variable and a dependent variable. The variable y represents the dependent variable, while x represents the independent variable. In the first plot we have a linear relationship. This could be the relationship between Intelligence Quotient (IQ), x, and the mark obtained in an achievement test, y. In the second plot, we have a quadratic relationship. This could be the relationship between time x and the vertical distance position, y, (from the source) of an object thrown up into the air.

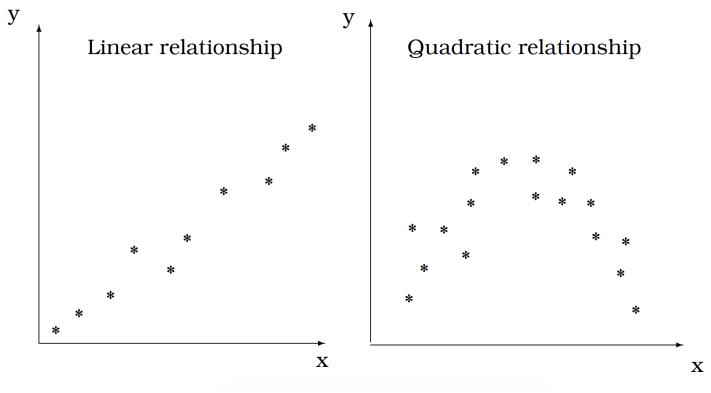


Figure 6.5: Scatter plots showing two possible relationships between x and y

So, by constructing a scatter plot, we are guided in our decision or choice of the equation to use.
All procedures and conclusions drawn in regression analysis depend, at least indirectly, on the assumptions of the **regression model**. A model is what the data analysts perceive as the mechanism that generates the data on which the regression analysis is conducted.

The term **fitting the model** to a set of data involves estimation of the regression coefficients and formulation of a **fitted regression model** [i.e the model with the estimated coefficients]:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{6.11}$$

Some Uses of Regression Analysis

Uses of regression analysis fall into the following categories:

1. Prediction: using the data availabe in conjunction with the model to predict/estimate future outcomes.
2. Variable screening: removing unnecessary variables from the a multiple linear regression model. In this way we can find variables that are important to a dependent variable.
3. Explaining: system explanation, which variables contribute the most and how they contribute to the dependent variable.
4. Inference: estimation of the parameters in the model.
5. Planning and control: if we have an appropriate model we can explain the physical system and thus plan ahead and control the system.

Regression techniques are also widely used in econometrics.

Regression Assumptions

We shall use the **Least Squares** (LS) procedure to estimate the parameters in the model given by Equation 6.6. Although there are other procedures available for estimating the parameters, we shall, however, only use the LS procedure. We will also make the following assumptions (these are necessary for inference):

- There is a linear relationship between the independent variable x_i and the dependent variable y_i.
- The x_i's are non-random and are observed with negligible error.
- The ϵ_i's are random variables with mean zero and constant variance. This is called the **homogeneous** variance assumption. Mathematically, this assumption is:

$$E(\epsilon_i) = 0 \text{ and } Var(\epsilon_i) = E(\epsilon_i^2) = \sigma^2.$$

- The ϵ_i's are uncorrelated, i.e.

$$E(\epsilon_i, \epsilon_j) = \begin{cases} 0 \text{ if } i \neq j \\ \sigma^2 \quad if \quad i = j \end{cases}$$

fo $i, j = 1, 2, \cdots, n$

- The normal theory assumption is imposed on the ϵ_i's. This is the assumption that the ϵ_i's are normally distributed with mean zero and variance σ^2. Mathematically, this is stated as

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

From above, it follows that:

$$E(y_i) = \beta_0 + \beta_1 x_i \qquad i = 1, 2, \ldots, n, \tag{6.12}$$

since $E(\epsilon_i) = 0$. Thus, we can use the notation:

$$E(y|x) = \beta_0 + \beta_1 x. \tag{6.13}$$

This is the expected value of y for a given value of X, i.e.. given $X = x$.

6.5 The Least Squares Technique

The method of least squares, attributed to K. Gauss, a German scientist, is perhaps the most extensively used technique for estimating the parameters β_0 and β_1, to give the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively. These parameters have to be estimated because we do not have the actual values available. When we use these estimates, we get the fitted model given by:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i. \tag{6.14}$$

We call this the fitted model because the model now has estimated parameters. Generally in Statistics, the 'hat' notation is used to indicate an estimate. Notice that we don't have the error term in Model 6.14. The relationship between x and $E(Y|x)$ is now an exact one.

Definition 4 *(***Residual***)*
Let $r_i = y_i - \hat{y}_i$, This difference is called a **residual**.

The distinction between the residual r_i and the error term ϵ_i is important. The former measures the deviation of y_i from $\hat{y}_i$. Since ϵ is usually unknown, it is estimated by r_i. The residuals are needed not only for estimating the magnitude of the random variation in the y_i's, but also for assessing the appropriateness of the regression model employed. We shall discuss this later.

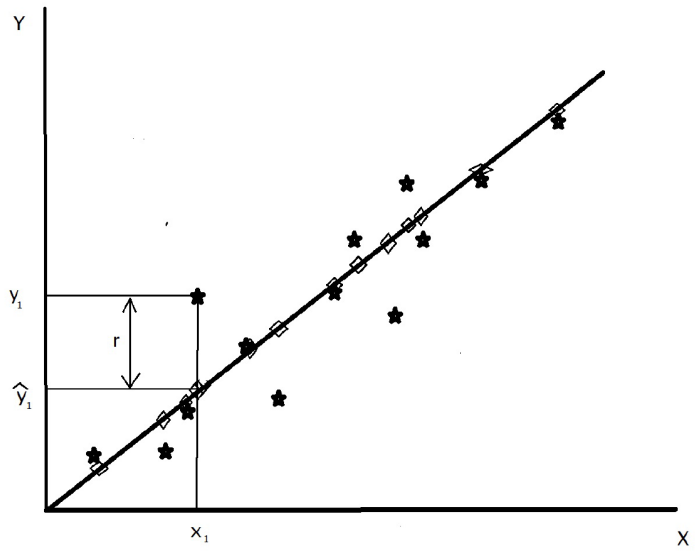


Figure 6.6: The observed values y_i (marked '★'), residuals r_i and fitted values $\hat{y}_i$ (marked '◊')

Least Squares Estimation

We want the fitted values $\hat{y}_i$ to be as close as possible to y_i. To achieve this, let us consider $\sum_{i=1}^{n} r_i^2$, the **Residual Sum of Squares** (RSS). It makes good sense to minimise RSS, since a good fit should produce the smallest possible sum of squares. This is the basis of the least squares technique.

Figure 6.6 illustrates what is really happening, the points marked by the '★'s represent the observed values while the fitted values lie on the line indicated by '◊'s. The residual r_i is shown clearly as the difference between the observed value y_i and the fitted value $\hat{y}_i$.
To minimise the RSS, $\hat{\beta}_0$ and $\hat{\beta}_1$ must satisfy the conditions:

$$\begin{aligned} \frac{\partial}{\partial \hat{\beta}_0}\left(\sum_{i=1}^{n} r_i^2\right) &= 0, \text{ and} \\ \frac{\partial}{\partial \hat{\beta}_1}\left(\sum_{i=1}^{n} r_i^2\right) &= 0. \end{aligned}$$

Thus:

$$\begin{aligned} \frac{\partial}{\partial \hat{\beta}_0}\left(\sum_{i=1}^{n} r_i^2\right) &= \frac{\partial}{\partial \hat{\beta}_0}\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 \\ &= -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) \\ &= -2\sum_{i=1}^{n} y_i + 2n\hat{\beta}_0 + 2\hat{\beta}_1 \sum_{i=1}^{n} x_i \\ &= 0. \end{aligned}$$

So that:

$$\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i \quad = 0.$$

By dividing by n leads us to:

$$\overline{y} \quad = \hat{\beta}_0 + \hat{\beta}_1 \overline{x} \tag{6.15}$$

For β_1 we have:

$$\begin{aligned} \frac{\partial}{\partial \hat{\beta}_1} \left(\sum_{i=1}^{n} r_i^2\right) &= \frac{\partial}{\partial \hat{\beta}_1} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 \\ &= -2 \sum_{i=1}^{n} x_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) \\ &= -2 \sum_{i=1}^{n} y_i x_i + 2\hat{\beta}_0 \sum_{i=1}^{n} x_i + 2\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \\ &= 0. \end{aligned}$$

This simplifies to,

$$\sum_{i=1}^{n} x_i y_i \quad = \quad \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \tag{6.16}$$

Equations 6.15 and 6.16 are called normal equations. Solving for $\hat{\beta}_0$ in Equation 6.16 gives us the following estimate for β_0 in terms of β_1. $\sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i$ can be solved to give:

$$\begin{aligned} \hat{\beta}_0 &= \frac{1}{n} \sum_{i=1}^{n} y_i - \hat{\beta}_1 \frac{1}{n} \sum_{i=1}^{n} x_i \\ &= \overline{y} - \hat{\beta}_1 \overline{x} \end{aligned} \tag{6.17}$$

Substituting for $\hat{\beta}_0$ in equation 2.8 gives us

$$\begin{aligned} 0 &= \sum_{i=1}^{n} x_i y_i - \left[\frac{1}{n} \sum_{i=1}^{n} y_i - \hat{\beta}_1 \frac{1}{n} \sum_{i=1}^{n} x_i\right] \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \\ &= \hat{\beta}_1 \left[\frac{1}{n} \left(\sum_{i=1}^{n} x_i\right)^2 - \sum_{i=1}^{n} x_i^2\right] + \sum_{i=1}^{n} x_i y_i - \frac{1}{n} \left(\sum_{i=1}^{n} y_i\right) \left(\sum_{i=1}^{n} x_i\right) \end{aligned}$$

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n} (x_i y_i) - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} \\ &= \frac{\sum_{i=1}^{n} (x_i y_i) - n\overline{x}\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \\ &= \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \end{aligned}$$

We often state the above as:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \tag{6.18}$$

where $S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$ and $S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2$. Note that we first find $\hat{\beta}_1$ and then find $\hat{\beta}_0$, using $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x}$.

The fitted line is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. This fitted line is referred to by many different names in statistics. Some of the names are: the least squares line, fitted regression line, estimated regression line or just the fitted model.

Example 6.4

The starting salary S per year of people of different educational background (Ed) has always been of interest to people going to university. They have always tried to find out the relationship between these. We expect the starting salary to be directly related to the educational level, i.e., as the educational level increases, so does the salary.

As this may not be the case, we shall investigate this suspicion using regression analysis. Suppose that an individual's educational level is given a score, then an appropriate model is given by:

$$S = \beta_0 + \beta_1 Ed + \epsilon$$

The data on S and Ed were collected and recorded as shown below. Find the estimates for β_0 and β_1 and discuss your results.

Number	Annual salary ($)	Educational level score
1	20 000	2.8
2	24 500	3.4
3	23 000	3.2
4	25 000	3.8
5	20 000	3.2
6	22 500	3.4

Solution:

Calculations give

Number	S_i	Ed_i	S_iEd_i	Ed_i^2
1	20 000	2.8	56 000	7.84
2	24 500	3.4	83 300	11.56
3	23 000	3.2	73 600	10.24
4	25 000	3.8	95 000	14.44
5	20 000	3.2	64 000	10.24
6	22 500	3.4	76 500	11.56
Total	135 000	19.8	448 400	65.88

$$\begin{aligned}
\hat{\beta}_1 &= \frac{S_{Ed,S}}{S_{S,S}} \\
&= \frac{448400 - \frac{(19.8)(135000)}{6}}{65.88 - \frac{19.8^2}{6}} \\
&= \frac{2900}{0.54} \\
&= \$5370.
\end{aligned}$$

$$\begin{aligned}
\text{Therefore,} \quad \hat{\beta}_0 &= \frac{135000}{6} - 5370\left(\frac{19.8}{6}\right) \\
&= \$4779
\end{aligned}$$

Thus, we have the fitted model $\hat{S}_i = 4779 + 5370Ed_i$. How do we interpret this model? The starting salary is predicted to be \$4 779, when the Educational level score is zero. This may not say much since an educational score of zero does not apply to this group of people by virtue of their being at University.

Perhaps, of primary interest is the slope (coefficient) which indicates that for a one-unit increase in educational score, the predicted salary increases by \$5370. For example, for an educational level score of 2.8, the predicted salary is $\hat{P} = 4779 + 5370 \times 2.8 = \$19815$.

Exercise 6.5

1. In Example 6.4, remove the last (sixth) number and estimate β_0 and β_1.
2. Predict the salary for someone with an educational score of 2.8.

6.6 Properties of Estimates

We shall briefly mention, without proving, some of the properties of the estimates we derived earlier. We shall then use these properties in assessing how **good** our model is.

The Least Squares estimates of β_0 and β_1 are unbiased so that we have the following properties.

The Sampling Distribution of β_0 is given by:

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right]\right), \tag{6.19}$$

where σ^2 is the variance of the error term.

The sampling distribution of $\hat{\beta}_1$ is given by;

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right). \tag{6.20}$$

Now, using the properties of the sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$, inferences about β_0 and β_1 can be made. First, however, an estimate of one other unknown parameter in the regression model is needed. This is an estimate of σ^2. This estimate is given by:

$$\begin{aligned}\hat{\sigma}^2 \approx s^2 &= \frac{1}{n-2}\sum_{i=1}^{n}(r_i - \overline{r})^2 \\ &= \frac{1}{n-2}\sum_{i=1}^{n}(r_i)^2 \\ &= \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.\end{aligned} \tag{6.21}$$

Since $\sum_{i=1}^{n} r_i = 0$ and $\overline{r} = 0$.

Note that s^2=MSE, the Mean Square Error and SSE=$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ the Error Sum of Squares.

We will use s^2 to estimate σ^2 in our expressions of $Var(\hat{\beta}_0)$ and $Var(\hat{\beta}_1)$ to get:

$$\widehat{Var(\hat{\beta}_0)} = s^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \right) \tag{6.22}$$

and

$$\widehat{Var(\hat{\beta}_1)} = \frac{s^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \tag{6.23}$$

respectively.

Lets discuss how we can make inferences about the regression parameters before proceeding to investigate how well the fitted line fits the data.

Exercise 6.6

1. Deduce the variance of y_i.
2. Assuming $\hat{\beta}_1$ and $\hat{\beta}_2$ are independent, find the variance of $\hat{y}_i$.

6.7 Making Inferences about β_0 and β_1

The following procedure can be used to test hypotheses about the slope and the intercept.

1. Establish the null and alternative hypotheses.
   (a) The null hypothesis is that: There is no linear relationship between Y and X, that is, $H_0 : \beta_1 = 0$.
   (b) The alternative hypothesis is that: There is a linear relationship between Y and X, i.e. $H_1 : \beta_1 \neq 0$.
2. Determine the tolerance α for a type I error probability.
3. Identify an appropriate test-statistic. Using $H_0 : \beta_1 = 0$, the test statistic is given by:

$$t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{s^2 / \sum_{i=1}^{n}(x_i - \overline{x})^2}} \tag{6.24}$$

   Note that t has a student's t-distribution with $n - 2$ degrees of freedom.

4. State the assumptions under which this test statistic has a t-distribution. The assumptions about the error must be valid for the conclusions or inferences to be valid.

5. Determine whether these assumptions are satisfied.

6. Find the values for the test statistic that allow a rejection of the null hypothesis. We find the critical values $t_{n-2,\frac{\alpha}{2}}$, such that we reject H_0 if $t < -t_{n-2,\frac{\alpha}{2}}$ or if $t > t_{n-2,\frac{\alpha}{2}}$.

7. Compute the t-value based on sample data.

8. Interpret the result from a statistical viewpoint.

Example 6.6

We assume that a two-tailed test is appropriate and we use the data in Example 6.4 to test the hypothesis
$H_0 : \beta_1 = 0$ versus
$H_1 : \beta_1 \neq 0$
at $\alpha = 0.05$.

First we calculate s as follows:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(S_i - \hat{S}_i)^2}{n-2}}$$

Number	S_i	$\hat{S}_i$	$S_i - \hat{S}_i$	$(S_i - \hat{S}_i)_i^2$
1	20 000	19 815	185	34.225
2	24 500	23 037	1,463	2 140 369
3	23 000	21 963	1,037	1 075 369
4	25 000	25 185	-185	34 225
5	20 000	21 963	-1,963	3 853 369
6	22 500	23 037	-537	288 369
Total			0	7 425 926

$$s = \sqrt{\frac{7425926}{4}} \approx 1363$$

Calculation of the variance estimate:

Number	Ed_i	$Ed_i - \overline{Ed}$	$(Ed_i - \overline{Ed})^2$
1	2.8	-0.5	0.25
2	3.4	0.1	0.01
3	3.2	-0.1	0.01
4	3.8	0.5	0.25
5	3.2	-0.1	0.01
6	3.4	0.1	0.01
Total			0.54

We calculate $\sqrt{\widehat{Var(\hat{\beta}_1)}}$ as follows:

$$\sqrt{\widehat{Var(\hat{\beta}_1)}} = \frac{1363}{\sqrt{0.54}} = 1855$$

Our test statistic t is found to be:

$$t = \frac{5370 - 0}{1855} = 2.895$$

The critical value is $t_{n-2,\frac{\alpha}{2}} = t_{4,0.025} = 2.78$. Therefore, we reject H_0 since t exceeds the critical value $t_{n-2,\frac{\alpha}{2}}$. We therefore conclude that there is a significant statistical relationship between the starting salary and the educational level score. See Figure 6.7 for the decision-making diagram.

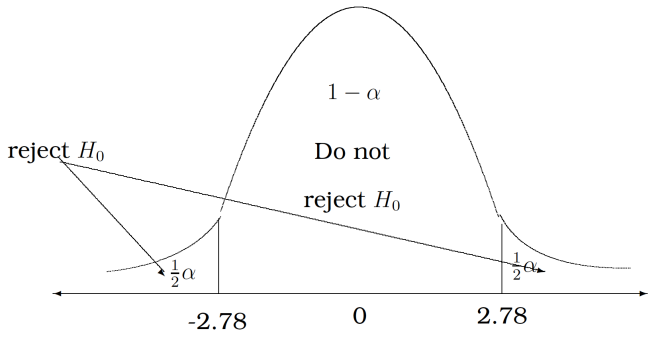


Figure 6.7: t-Distribution: when to reject, or fail to reject H_0

A similar approach can be used to test the significance of the intercept.

Exercise 6.7

1. If $\hat{\beta}_1$ had been -0.32, what conclusions would you have drawn?
2. Test the null hypothesis that the intercept is zero.

6.8 Analysis of Variance of Simple Linear Regression

Analysis of variance (ANOVA) is a highly useful and flexible mode of analysis for regression models. We will use ANOVA to compute σ^2 and to measure the degree of linear association between X and Y in the sample data.

Partitioning the Total Sum of Squares

The uncertainty associated with a prediction is related to the variability of the Y observations around their mean, as measured by the following deviations:

$$Y_i - \overline{Y}$$

The greater the variability in the data, the larger will be the deviations, $Y_i - \overline{Y}$, and the greater is the uncertainty associated with a prediction Y_i, without utilising knowledge of X_i.

Conventionally, the measure of variability of the observations is expressed in terms of the sum of squares of the observations $Y_i - \overline{Y}$ and is denoted by:

$$SST = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 \tag{6.25}$$

where SST stands for Total Sum of Squares. If there is a lot of variability in the Y_i, then SST is large.

Error Sum of Squares

The uncertainty associated with a prediction is related to the variability of the Y_i around the fitted regression line as measured by the following deviations:

$$r_i = Y_i - \hat{Y}_i$$

If all the Y_i values fall on the regression line, all the deviations r_i, will be zero. The larger the deviations r_i the greater the uncertainty associated with a prediction utilising knowledge of the independent variables X_i.

The conventional measure of variability around the fitted regression is the Error Sum of Squares (SSE) which is calculated as follows:

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} r_i^2 \tag{6.26}$$

If all the Y_i values fall on the regression line, SSE will be zero.

Regression Sum of Squares

The reduction in the variability associated with the utilisation of the knowledge of the independent variable X_i is another sum of squares known as Regression Sum of Squares (SSR). Figure 6.8 illustrates how each of these components arises.

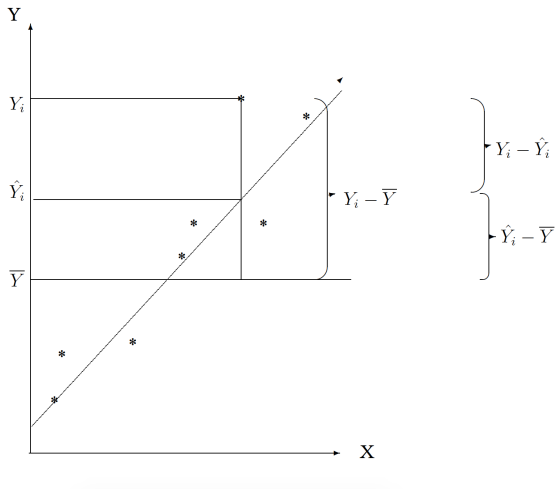


Figure 6.8: Variability on the regression line

$$SSR = SST - SSE \tag{6.27}$$

We can show that SSR is the sum of squares involving the deviations:

$$\hat{Y}_i - \overline{Y},$$

which represent the fitted value and the mean of the fitted value.

$$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 \tag{6.28}$$

SSR can be viewed as a measure of the effect of the regression relation in reducing the variability of Y_i. If SSR $= 0$, the regression calculation will not reduce variability at all. SSR can be interpreted as the proportion of variation in Y explained by the regression.

A more mathematical approach is to say we are partitioning the variability of the Y_i's. Thus, for Simple Linear Regression, the decomposition of the SST into two components is achieved as follows:

$$SST = SSR + SSE \tag{6.29}$$

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{6.30}$$

The computational formulas are put as follows:

$$SST = \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2 \tag{6.31}$$

$$SSR = \frac{(\sum_{i=1}^{n} Y_i X_i - n\overline{YX})^2}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2} \tag{6.32}$$

and

$$SSE = SST - SSR \tag{6.33}$$

Partitioning of Degrees of Freedom

(Let: $\sum = \sum_{1=1}^{n}$)

A sum of squares has an associated number of degrees of freedom. Recall that the variance estimate s^2 in Equation 3.3 has a denominator $n - 2$. These are the degrees of freedom associated with the numerator sum of squares in s^2.

Corresponding to the partitioning of the sum of squares is a partitioning of the degrees of freedom. SST has $n - 1$ degrees of freedom (df) associated with it. Why? SST has n deviations, namely $Y_i - \overline{Y}$. However, there is one constraint on these deviations, namely $\sum(Y_i - \overline{Y}) = 0$, so we lose one degree of freedom, to remain with $n - 1$ degrees of freedom in the n deviations.

SSE has $n - 2$ degrees of freedom, since we imposed constraints on the r_i's during the estimation of β_0 and β_1. As SSR has 1 df, there are two parameters in

the regression function, but the deviations $\hat{Y}_i - \overline{Y}$ are subject to the constraint $\sum(\hat{Y}_i - \overline{Y}) = 0$. Thus, the degrees of freedom are additive and given by $n - 1 = 1 + n - 2$.

Mean Squares

A sum of squares divided by the degrees of freedom is called a mean square. For example, $s^2 = MSE$. The two important mean squares are the **regression mean square** denoted by MSR and the **error mean square** denoted by MSE.

Thus:

$$MSR = \frac{SSR}{1} \tag{6.34}$$

and

$$MSE = \frac{SSE}{n-2} = s^2 \tag{6.35}$$

Some Properties of Mean Squares

It can be shown that the expectations of the mean squares are given by:

$\mathrm{E}[MSE] = \sigma^2$

It can also be shown that: $\mathrm{E}[MSR] = \sigma^2 + \beta_1^2 \sum(X_i - \overline{X})^2$

Thus, when $\beta_1 = 0$, $\mathrm{E}[MSR] = \sigma^2$, both MSE and MSR have the same expected value under this condition. On the other hand, when $\beta_1 \neq 0$, the term $\sigma^2 + \beta_1^2 \sum[X_i - \overline{X}]^2$ will be positive and $\mathrm{E}[MSR] > \mathrm{nE}[MSE]$. Hence, if $\beta_1 \neq 0$, MSR will tend to be larger than MSE.

Exercise 6.8

1. What does it mean if the fitted model gives you SSE = 0.

2. If SSR is zero, what does it tell you about the model?

3. In the Simple Linear Regression model, suppose that SST has 14 degrees of freedom. Deduce the SSE and SSR degrees of freedom.

6.9 The Basic ANOVA Table

It is useful to collect the sum of squares, degrees of freedom and mean squares in an ANOVA table for regression analysis. Table 6.2 gives the structure and the appearance of the basic ANOVA table.

Source of variation	SS	df	MS	F
Regression	$SSR = \sum(\hat{Y}_i - \overline{Y})^2$	1	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$
Error	$SSE = \sum(Y_i - \hat{Y}_i)^2$	$n-2$	$MSE = \frac{SSE}{n-2}$	
Total	$SST = \sum(Y_i - \overline{Y})^2$	$n-1$		

Table 6.2: The basic ANOVA table for simple linear regression

From the ANOVA table, we can get the variance s^2 and test the hypothesis that there is a regression relationship. How do we do this? The ratio F in the ANOVA table has what we call the Fisher's distribution with 1 and $n-2$ degrees of freedom if the assumptions of the model hold.

If F is near 1, then MSR and MSE are approximately equal. $F > 1$, suggests that $\beta_1 \neq 0$. Thus, an upper-tail test is appropriate.

The hypotheses we are testing here are as follows:
$H_0 : \beta_1 = 0$ versus
$H_1 : \beta_1 \neq 0$
at level α. Our decision rule here is as shown in Figure 6.9.

The decision rule is given by:
Fail to reject H_0 if $F \leq F_{1,n-2;1-\alpha}$
Reject H_0 if $F > F_{1,n-2;1-\alpha}$.

6.10 The Coefficient of (simple) Determination

The **coefficient of determination** R^2 is a measure of the effect an independent variable X has in the regression model in explaining the total variability in Y.

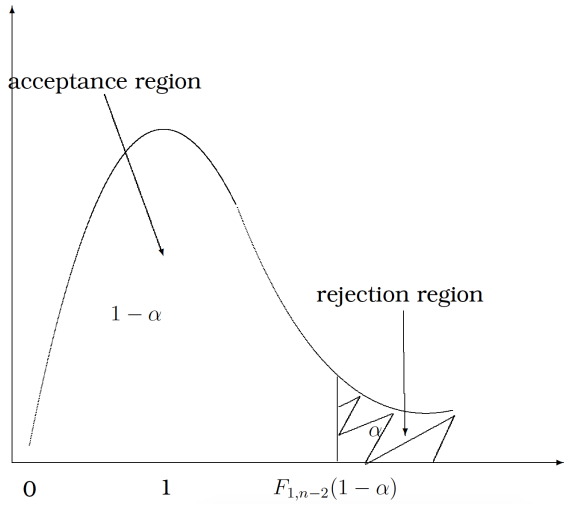


Figure 6.9: The general form of the statistical decision rule for an F-test

The coefficient of determination, denoted by R^2, is defined as follows:

$$R^2 = \frac{SSR}{SST} \tag{6.36}$$

$$= 1 - \frac{SSE}{SST} \tag{6.37}$$

Thus, R^2 measures the proportionate reduction in SST associated with the use of an independent variable.

In the Simple Linear Regression case, we usually refer to the coefficient of Determination as the **Coefficient of Simple Determination** (R^2). Note that R is the simple correlation coefficient of the independent and dependent variables.

R^2 takes values between 0 and 1. We obtain $R^2 = 0$ when $\beta_1 = 0$, and $R^2 = 100\%$ when the Y_i's fall directly on the regression line. A value of $R^2 > 80\%$ or sometimes 70%, suggests that the model has a good fit.

Adjusted R^2

One phenomenon found on adding terms to a regression model is that the R^2 increases. Although this may be an indication that the extra terms improve the regression equation, it is may also be a reflection of the fact that one is using more variables to predict the same number of data points. This problem may be taken

into account by examining not only the actual value of R^2, but also the value of the adjusted R^2. This statistic takes into account the number of data points and variables in the regression equation, by replacing the SSE and $\sum_{i=1}^{n}(y_i - \overline{y})^2$ by the corresponding MSE's, giving

$$\overline{R}^2 = 1 - \frac{SSE/(n-2)}{\sum_{i=1}^{n}(y_i - \overline{y})^2/(n-1)} \tag{6.38}$$

which can also be written as $\overline{R}^2 = 1 - \frac{(n-1)}{(n-2)}(1 - R^2)$.

It is so scaled that, if a second added variable results in a non-significant improvement in the regression fit, the adjusted R^2 will decrease.

Example 6.7

The price of a kilogram of flour (Y) at a market place in a particularly busy township seems to vary according to what the vendor thinks is your salary X (in thousands of Rands). Use the data supplied to investigate if this suspicion is true.

induvidual	X	Y
1	2	8.74
2	2	10.53
3	2	10.99
4	2	11.97
5	3	12.83
6	3	14.69
7	3	14.69
8	3	15.30
9	4	16.11
10	4	16.31
11	4	16.46
12	4	17.69
13	5	19.65
14	5	18.86
15	5	19.93
16	5	20.51

Solution

The model is given by: $y = \beta_0 + \beta_1 x + \epsilon$ ($price = \beta_0 + \beta_1 salary + \epsilon$)

Table 6.3: The ANOVA Table

SOURCE	df	SS	MS	F
Regression	1	177.668	177.668	183.921
Error	14	13.526	0.966	
Total	15	191.194		

Fitting the model gives us;

$$\begin{aligned}\widehat{price} &= \hat{\beta}_0 + \hat{\beta}_1 salary \\ &= 4.8970 + 2.9805 salary\end{aligned}$$

Let us now construct the ANOVA table. First we calculate SSR, SST, and SSE.

$$\begin{aligned}SST = &\sum y_i^2 - n\overline{y}^2 \\ = &\ 197.195,\end{aligned}$$

$$\begin{aligned}SSR = &\sum_i^{16} (\hat{y})_i - n\overline{y}^2 \\ = &\ 177.668,\end{aligned}$$

$$\begin{aligned}\text{and } SSE = &\ SST - SSR \\ = &\ 191.194 - 177.668. \\ = &\ 3.527\end{aligned}$$

From the above, the ANOVA table can be constructed as follows.

We can see from this ANOVA table that F is quite large. Infact it leads us to a rejection of the hypothesis of no regression relationship (verify).

We can compute the coefficient of Multiple determination from Table 6.3 above. Thus:

$$\begin{aligned}R^2 &= \tfrac{SSR}{SST} \\ &= \tfrac{177.668}{191.195} \\ &= 0.929.\end{aligned}$$

Thus, about 92.9% of the variation in prices (Y) is explained by the regression model. So, the salary estimates do seem do determine the price of flour. The adjusted $R^2 = 92.4\%$.

The correlation coefficient (r) measures the strength of the linear relationship between the dependent variable and all the independent variables. It is computed from the formula:

$$r = +\sqrt{R^2} \tag{6.39}$$

$$= +\sqrt{\frac{SSR}{SST}}, \tag{6.40}$$

where SSR is the Regression Sum of Squares and SST the Total Sum of Squares. R close to 1 means that there is a good linear relationship between the dependent and independent variables.

Exercise 6.9

A student recorded the 6 test marks she obtained after devoting a particular number of hours of study. The marks are:

Marks (%)	60	50	40	100	10
Time (hrs)	2	1.5	0.5	3	0

1. Estimate β_0 and β_1 in a simple linear regression model.

2. Construct the ANOVA table.

3. Show that F = 42.89, and test the null hypothesis that $\beta_1 = 0$ at $\alpha = 0.05$.

6.11 Conclusion

In this chapter, we focused on Simple Linear Regression. We discussed the estimation of the parameters using the Least Squares technique. There are other methods availabe for estimating these parameters. We shall meet these in future modules.

We went on to discuss how to check if any of the assumptions which enable us to use the least squares approach have been violated. This is often ignored by "pseudo-statisticians". Some blame this abuse of Regression Analysis on computers which allow you to use statistical computer packages without looking at the underlying theory behind the techniques.

Exercise 6.10

1. Find the relationship between the correlation coefficient r and β_1.

2. A farmer is carrying out an experiment on the effect of soil acidity levels on yield. The farmer records the yields of 10 fields of equal size but different acidity levels. [A negative acidity level here indicates that the acidity is less than the standard level pegged at 0.]

Field:	1	2	3	4	5	6	7	8	9	10
Level:	4.5	17.7	-16.6	-14	18.6	-10.6	5.8	-8.1	-5.2	7.8
Yield:	75	112	38	120	105	52	116	118	105	110

   (a) Examine graphically the relationship between Level and Yield.

   (b) Is this a statistical relationship? Explain.

   (c) Which is the dependent variable, and which is the independent variable? Explain.

   (d) Fit a simple linear regression model to this data.

   (e) Investigate if any assumptions were violated.

   (f) What do you expect the yield to be for a field with acidity level 0?

   (g) Do you think the field would make a better independent variable?

3. A recent review of salaries at a company in Harare made the recommendation that, depending on the number of years of service X, the minimum salary Y of an employee should be as follows:

x:	5	10	15	20	25	30
y:	39.5	49.0	58.5	68.0	77.5	87.0

   (a) Examine graphically the relationship between number of years of service and minimum salary.

   (b) Do these recommendations represent an exact or statistical relation? Explain using statistics like the sample correlation coefficient.

4. For each of the following pairs of variables, explain whether an exact or statistical relation would most likely hold:

   (a) X=number of beds in a hotel; Y=hotel's annual operating cost.

   (b) X = Volume of a gas; Y = Pressure on the gas.

   (c) X = A departmental store's promotional and advertising expenditure; Y = the company profits.

5. Find the correlation matrix ρ for the data in Question 1 above.

6. The following information was recorded over 10 years, the amount of rainfall, maize production and maize price.

   (a) Which is the dependent variable? Explain.

   (b) What is the regression model linking these variables?

   (c) Which two variables are likely to have a high correlation?