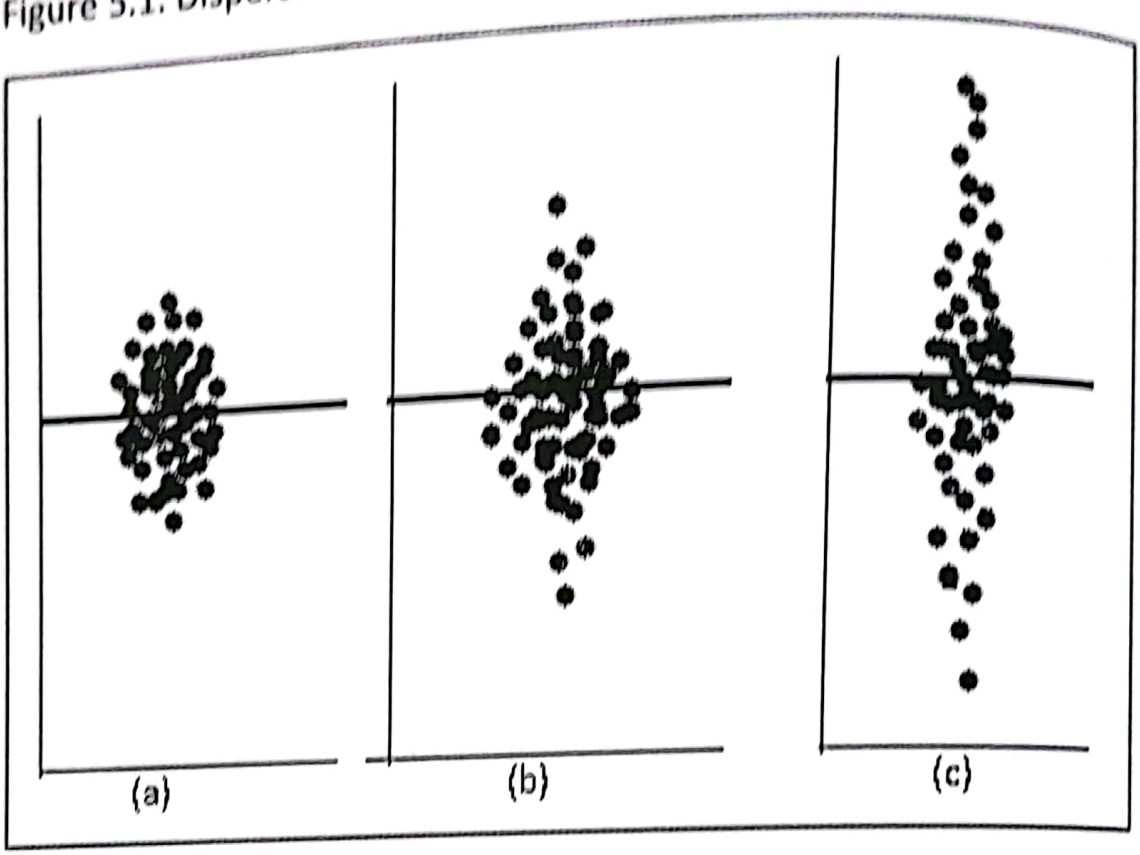
CHAPTER 5

5 MEASURES OF VARIATIONS IN ACHIEVEMENTS

In the preceding chapter, we looked at the measure of achievements. We looked for the level or average value or position of a group. The position or level was determined by a representative number arrived at in different ways. These were the arithmetic mean, mode and the median. All these produced a number that, when interpreted, tended to put all the observations together. This is necessary but not sufficient information to understand the scores in a test or the economic performance of a group of countries.

The missing bit is the description of the spread or clusteredness of the observations. This is the measure of variation in achievement or more precisely, the measure of dispersion. The aim is to get a picture of the position of the observation often relative to the measured level of achievement. This measure will give an interpretation of whether the observations are clinging to the average or are more spread out. See Figure 5.1 below.

Figure 5.1. Dispersion in data



In part (a) of the figure, the dots look so close to the central value. They are clustered around the mean. In part (c), there is a wide spread from the central value compared to both (a) and (b). We say there is high dispersion in part (c) compared to the other two. Alternatively, dispersion or spread around the central value is lowest in part (a) of the above figure.

In economics, the measure of average achievement will be indicative of efficiency in the level of economic performance whereas the measure of dispersion will be indicative of equity in the distribution of that performance. For example, in Zambia reasonable rates of sustainable economic growth have resulted in higher average incomes. But the distribution of these incomes has been widening, resulting in higher levels of inequality. Thus, while average economic has been improving over time, serious equity issues still remain since the benefits of income growth have not been uniformly spread over the population.

A higher spread entails a lot of disparities in values because there is a higher degree of deviation from the central measure. Some observations are way above the central level and some way below. A lower spread means most

observation are close to the central value, meaning the disparities are not very pronounced.

Like in the measure of achievement, there are also several measures or methods for measuring dispersion in the data. Some are very simple but not telling much or not robust enough. On the other hand, there are more complex and robust measures with rich information.

5.1 Range

The range, the simplest of all the measures of dispersion, is the difference between the maximum value and the lowest value. Often, when very little is known about a given data set, we are only able to say, "the age of UNZA students ranges from 16 to 35" or "the incomes of the people present at the entrepreneurship conference ranged from K800 to K3000". Whenever the range is used, it has connotations of the lowest and the highest observations. Other values are expected to fall anywhere between the two values.

The range gives an idea of the dispersion in the data because it gives the distance between the lowest and the highest. Its major weakness is that it depends only on two observations and is blind to changes in other observations. When more observations are added to the sample, the range will not be affected as long as the additional observations are neither lowest nor highest.

Example 5.1

Consider the following illustration. A company has two branches; one at Makeni shopping mall and another at Manda Hill. The Marketing Manager has asked for information on the volume of sales (number of customers buying something) for each month in the second half of 2013. The information is presented in the table below.

Branch	July	Aug	Sep	Oct	Nov	Dec
Makeni	45	65	53	56	48	61
Manda Hill	51	87	47	55	60	54

Sales at the Makeni mall branch range from 45 (the lowest) to 65 (the highest). The range in sales denoted by R is therefore $R = 65 - 45 = 20$. For the Manda Hill branch, the sales don't look very different from the others except for one extreme case in August. This affects the range $R = 87 - 47 = 40$. The range for Manda Hill is twice that of Makeni branch.

Suppose now the high sales in the month of August shown in the above table are as a result of SADC games that were hosted by the University of Zambia (UNZA). With this information, does the calculated range give a fair representation of sales at the Manda Hill branch? Well, maybe not. Sales under normal circumstance actually range from 47 to 60. When the outlier is ignored, because it is a one off occurrence, the range changes to $R = 60 - 47 = 13$. This is a one off event and giving a range based on this 'exceptional' value may not give a true picture. This amplifies the major weakness in the range as a measure of dispersion namely, it is affected by outliers. It will suffice to mention here that both the arithmetic mean (under measures of achievement) and the range (under measures of variations) are affected by outliers or extreme values but for different reasons. For the arithmetic mean, it is because it gives equal weight to each observation, including outliers. It leaves out no observation, even when it may be appealing to do so.

For the range, the reason is that is depends only on the two end-values, the lowest and the highest. If there is an outlier in the sample, it is either the lowest or the highest or both in some cases. This makes the range quite vulnerable.

Absolute differences do not always give a good picture of dispersion in data. For instance, two farmers have the following assets.

Farmer	Number of cattle	Cash at Bank
A	5	K 9,581
B	7	K 9,579

From the table, farmer B has more cattle than farmer A. the former exceeds the latter by two cattle. On the finance side, farmer A now exceeds the other by K2. The difference is two in both cases. This may, however, have misleading interpretation. We say it may be misleading because the difference in cattle is more significant than in cash owing to the values at play. The difference of K2 when talking of K9000 is insignificant. The differences will therefore add more value when analysed in relative terms.

Related to the range is the *Coefficient of Range*. This is a relative measure involving the two extreme values of a sample. It is given by

$$Coef of Range = \frac{Max - Min}{Max + Min}$$

A lower value of the coefficient indicates low variations in the data. A coefficient closer to unity is associated with high variability.

For the sales data given, the coefficient of range for the Makeni mall branches is

$$CoR_{Makeni} = \frac{65 - 45}{65 + 45} = \frac{20}{110} = 0.18$$

$$CoR_{Mandahill} = \frac{87 - 47}{87 + 47} = \frac{40}{134} = 0.30$$

The calculated coefficients of range show that there is more variation in sales at the Manda Hill branch. Of course, the CoR is higher because of the outlier value at the Manda Hill branch.

5.2 Interquartile range

Since the range suffers from the problem of outliers, the *Interquartile range* is devised to obtain a simple measure of dispersion that is free of outliers. Data is arranged in order of magnitude and divided into four quarters. The divisions makes use of five values which form boundaries for the four quarters. The first is the minimum which is at the lower end of the first

quarter. On the upper end of the first quarter is the value known as the *first quartile*. It separates the first and second quarters and is denoted by Q_1.

Between the second and the third quarter is the second quartile denoted by Q_2. Below it will be the first and second quarter while the third and fourth quarters will lie above. The second quartile therefore divides the data into two equal halves, two quarters below and two quarters above. This is the definition that was assigned to the median in the preceding chapter. Therefore, the second quartile is the same as the median and can be used interchangeably. For convenience sake, median is used when dealing with measures of achievement and second quartile when in dispersion.

The *third quartile* Q_3 is the value separating the third quarter and the fourth quarter. It has the third quarter on the lower side and the fourth on the upper side. Like in the calculation of the median in the preceding chapter, the quartile values may be part of the sample or may be mid-points of two boundary values. This will depend on the number of observations. If the number is one less than a multiple of four such as 11, 19, 31, all the quartiles will be observations from within the group. When the number is even but not divisible by 4 such as 14, 18, 34 etc, the first and third quartiles will be elements within the sample. The maximum will form the boundary for the fourth quarter. The interquartile range is then defined as the difference between the third and first quartiles.

$$QR = Q_3 - Q_1$$

Fifty percent or half of the observations fall within the interquartile range. Because it bounds fifty percent of data, it is sometimes known as the *midspread* or *middle fifty*.

A related measure is known as the *Quartile Deviation* denoted by QD. The quartile deviation is the interquartile range divided by two. For this reason, it is also known as the semi-interquartile range.

$$QD = \frac{Q_3 - Q_1}{2}$$

Roughly, a quarter of the observations will be within QD above the median and another quarter will lie roughly within QD below the median. Suppose the number of tourists visiting the South Luangwa National Park in the month of March of a given year is as shown in the table below.

Table 5.1. Number of tourist per day

Day	No.
1	134
2	127
3	136
4	127
5	119
6	116
7	126
8	138
9	131
10	136

Day	No.
11	128
12	117
13	126
14	119
15	141
16	115
17	88
18	137
19	92
20	135

Day	No.
21	141
22	124
23	128
24	124
25	137
26	105
27	114
28	129
29	125
30	92
31	104

As can be seen, the values are not in any order. They keep fluctuating from one day to another. To get the quartile values, the data must be arranged in an order of magnitude. We arrange it in ascending order below.

Figure 5.2. Number of tourists per day (ordered)

Day	No.
17	88
19	92
30	92
31	104
26	105
27	114
16	115

14	119
22	124
24	124
29	125
7	126
13	126
2	127
4	127

9	131
1	134
20	135
3	136
10	136
18	137
25	137
8	138

6	116
12	117
5	119

11	128
23	128
28	129

15	141
21	141

After ordering, we are able to tell that the lowest number of tourists (88 tourists) was observed on the 17^{th} day of the month while the highest numbers (141 tourists) occurred on the 15^{th} and 21^{st} of the month. This information is enough to get the range of the numbers.

$$R = 141 - 88 = 53$$

For the quartile range or deviation, we need to identify the positions for the quartile values. Since the number of observations is one less than 32, a multiple of 4, all the quartile values will be amongst the observations. With 31 observations, the 16^{th} observation will be the median. It will have 15 observations below and 15 above.

With 15 observations below the median, the 8^{th} will separate the lower part of the data into two equal half or quarters of the original number. This is the first quartile. Thus $Q_1 = 116$. On the upper side, we should look for the number that will divide the 15 observations above the median into two halves. This is the 8^{th} observation above the median or 24^{th} from the bottom. Therefore $Q_3 = 135$. The two values needed for the measure of dispersion have been identified.

$$\begin{aligned} QR &= Q_3 - Q_1 \\ &= 135 - 116 \\ &= 19 \end{aligned}$$

For the quartile deviation,

$$\begin{aligned} QD &= \frac{Q_3 - Q_1}{2} \\ &= \frac{135 - 116}{2} \\ &= \frac{19}{2} \\ &= 9.5 \end{aligned}$$

Since the relative values of the two measures are equal, only one measure is used. This is the *Coefficient of Quartile Deviation*. It is based on the Quartile Deviation.

$$CoefofQD = \frac{\frac{Q_3-Q_1}{2}}{\frac{Q_3+Q_1}{2}} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

For the data on number of tourist presented above, the Coefficient of Quartile Deviation is

$$\begin{aligned} CoefofQD &= \frac{Q_3 - Q_1}{Q_3 + Q_1} \\ &= \frac{135 - 116}{135 + 116} \\ &= \frac{19}{251} \\ &= 0.08 \end{aligned}$$

The first two measures of dispersion, namely the range and quartile deviation, are measures based on two values only. They ignore the positions of all the other observations. They fail to take into account all the values in the sample.

5.3 Mean Absolute Deviation

The rational way to look at the spread of observations from the central value is to consider how each deviates from the central value. For each observation, calculate its deviation from the central value, the arithmetic mean. This is called the *deviation,* denoted by d.

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

5.3.1 Mean Absolute Deviation for Ungrouped data

For each observation, subtract the mean to get its deviation from the mean. There will be as many deviations as are observations. This is the measure of the observation's position relative to the mean. Since some observations will fall below the mean and others above, some deviations will be negative

and some positive. Theoretically, the sum of all the deviations is always zero. That is

$$\sum_{i=1}^{n} d_i = 0$$

In rare circumstances, the deviations may be taken from the median as opposed to the arithmetic mean.

Given the n observations, the task is to get their average. This should give the average deviation of each observation from the mean. This measure is more inclusive than the earlier measures because it takes into account all the observations. The problem, however, is that the average requires the sum of all the deviations. Since this would be zero, such a measure becomes meaningless.

To obviate this limitation, a measure based on the absolute value of the deviation is used. Instead of dealing with the deviation which are sign sensitive, we could just look at the absolute value, ignoring the sign. This is a candid measure because of the interest to know how far observations are from the mean. The direction will really not matter, whether below or above the mean will be immaterial. This measure is known as *Mean Absolute Deviation* (MAD).

$$MAD = \frac{\sum|d_i|}{n} = \frac{\sum|x_i - \bar{x}|}{n}$$

The modulus lines inside the summation indicate that only the absolute value is considered.

Example 5.2

Consider an example from agriculture. An Agriculture Extension Officer is supervising nine small-scale farmers. At the end of the farming season, each farmer indicates his/her harvest in number of 50-kg bags of maize. The data is provided in the table below.

Farmer	1	2	3	4	5	6	7	8	9
Yield	36	85	93	38	64	34	52	49	71

a. What is the mean number of bags per farmer?

b. Calculate the Mean Absolute Deviation of the data.

We present the data in a table so that for each observation, the deviation from the mean is calculated

Farmer	Yield (x_i)	Deviation d_i	Abs Dev $\lvert d_i \rvert$
1	36	-22	22
2	85	27	27
3	93	35	35
4	38	-20	20
5	64	6	6
6	34	-24	24
7	52	-6	6
8	49	-9	9
9	71	13	13
Total	**522**	**0**	**162**

Mean:

$$\bar{x} = \frac{\sum x_i}{n}$$
$$= \frac{522}{9}$$
$$= 58$$

Mean Absolute Deviation: the values of the deviation and absolute deviation are shown in column 3 and 4 of the above table.

$$MAD = \frac{\sum |d_i|}{n}$$
$$= \frac{162}{9}$$
$$= 18$$

The relative measure of the Mean Absolute Deviation, called *Coefficient of Mean Deviation* is obtained by dividing the Mean Absolute Deviation by the corresponding average. If the arithmetic mean was used in calculating the MAD, then the Coefficient of Mean Deviation will involve dividing by the arithmetic mean. It will divide by the median if the MAD is based on the median.

$$Coeff\ of\ MD = \frac{MAD\ from\ Mean}{Mean}$$

$$= \frac{18}{58} = 0.31$$

Or

$$Coeff\ of\ MD = \frac{MAD\ from\ Median}{Median}$$

5.3.2 Mean Absolute Deviation for Grouped Data

When data is grouped such that only group boundaries are known, the actual deviations will also be unknown. Nonetheless, like it is done for the arithmetic mean in grouped data, the deviations can also be calculated on the basis of class midpoints. The class midpoints represent all the unknown observations in a group. Instead of considering the deviation of each element from the mean, we look at how each class deviates from the mean.

With grouped data, the mean absolute deviation is given by

$$MAD = \frac{\sum f_i \times |x_i - \bar{x}|}{\sum f_i}$$

In the formula, the x_i no longer represent an individual observation. Instead, it is the class mark or midpoint of the i^{th} class. It represents a class average. The f_i is the frequency of the i^{th} class.

Example 5.3

The following data shows the performance in a class of 84 students.

Mark	Number
20-29	3
30-39	8
40-49	13
50-59	25
60-69	27
70-79	7
80-89	1

a. Find the mean.

b. Using the arithmetic mean, calculate the Mean Absolute Deviation.

c. Calculate the Coefficient of Mean Deviation from the Mean.

The measure of achievement should be pretty easy by now. The formulas are provided in the preceding chapter. The first step is to complete the table by finding the true class boundaries, the class marks or midpoints and the product of frequency and midpoints.

Mark	Frequency f	Class Boundary	Midpoint x	fx
20-29	3	19.5-29.5	24.5	73.5
30-39	8	29.5-39.5	34.5	276
40-49	13	39.5-49.5	44.5	578.5
50-59	25	49.5-59.5	54.5	1362.5
60-69	27	59.5-69.5	64.5	1741.5
70-79	7	69.5-79.5	74.5	521.5
80-89	1	79.5-89.5	84.5	84.5
Total	**84**			**4638**

$$Mean\ \bar{x} = \frac{\sum f x_i}{\sum f}$$

$$= \frac{4638}{84}$$

$$= 55.2$$

For the Mean Absolute Deviation, the table must be expanded to include columns on deviation and absolute deviation. We replicate the table below.

Mark	Freq f	Class Boundary	Mid-point x	fx	Abs dev $\|d\|$	$f\|d\|$
20-29	3	19.5-29.5	24.5	73.5	30.7	92.1
30-39	8	29.5-39.5	34.5	276	20.7	165.7
40-49	13	39.5-49.5	44.5	578.5	10.7	139.3
50-59	25	49.5-59.5	54.5	1362.5	0.7	17.9
60-69	27	59.5-69.5	64.5	1741.5	9.3	250.7
70-79	7	69.5-79.5	74.5	521.5	19.3	135.0
80-89	1	79.5-89.5	84.5	84.5	29.3	29.3
Total	**84**			**4638**		**830**

With this information in the table, the Mean Absolute Deviation

$$MAD = \frac{\Sigma f|d|}{\Sigma f}$$
$$= \frac{830}{84}$$
$$= 9.88$$

This means on average, each observation deviates from the mean by about 9.9 percentage points. This is either above or below the mean.

The last part is the coefficient of mean deviation. Since the mean was used when finding the *MAD*, we must divide the *MAD* by the mean.

$$Coef\ of\ MD = \frac{MAD}{\bar{x}}$$
$$= \frac{9.88}{55.2}$$
$$= 0.18$$

The Mean Absolute Deviation is superior to the Range and Interquartile range as a measure of dispersion. Unlike the other two, it is based on all the values or observations. It is also easy to calculate and has a fairly good interpretation. Its major drawback is that its calculation involves ignoring algebraic signs.

5.4 Variance and Standard Deviation

In the MAD discussed in the preceding section, the major problem encountered was that the deviations in their natural values sum to zero. The summation of all the deviations from the mean always collapses to zero. As such, the mean deviation fails to provide a measure of dispersion. The MAD thus ignored the algebraic sign and only considered the absolute distance or deviation from the central value, traditionally the mean. There are two ways of ensuring that the negative values in the deviation do not offset the positive values. In the MAD, this was achieved by taking the absolute values only.

The alternative way of addressing the problem created by negative values is to square the numbers. If the deviations are squared, summed and then obtain a square root, the sum will no longer collapse to zero. We also avoid ignoring any detail and the negative values are well handled by squaring. This will give us two measures of dispersion: the *variance* which will be the arithmetic mean of the squares of the deviations measured from the mean and the standard deviation which is the positive square root of the variance. This is tantamount to calculating the positive square root of the variance.

5.4.1 Variance and Standard deviation in ungrouped data

The formula for the variance is given by.

$$Var(x) = s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

The standard deviation is then obtained as .

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

It can be shown that the above formula for the standard deviation can be simplified to.

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

However, when the number of observations is 'small', there is a bias in the standard deviation (as well as the variance) formula given above as it tends to understate the true standard deviation. The problem to live with though is that there is no agreed benchmark for the sample to be regarded as large enough. In most social sciences including Economics, it is generally acceptable to consider a sample size above 30 to be large. Thus for sample sizes above 30, there is no need to worry since the bias becomes negligible owing to a large denominator.

When the sample size falls below 30, the bias could be significant and hence some adjustment in the formula may be required. One commonly used adjustment is known as *Bessel's correction*, named after a German astronomer and mathematician Friedrich Wilhelm Bessel (1784-1846). This adjusted formula uses $n-1$ instead of n in the denominator. The correction has the effect of scaling up the estimated standard deviation in smaller sample. The standard deviation is scaled up by $\sqrt{n/n-1}$ which approaches a unity for large values of n.

For small samples, the standard deviation and variance are given by:

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

$$Var(x) = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Example 5.4

Ten small scale farmers are asked to state the number of draught cattle each has. The information is provided in the table below.

Farmer	1	2	3	4	5	6	7	8	9	10
No. of cattle	14	7	18	25	10	21	13	8	11	13

a. What is the mean number of animals per farmer?

b. What is the standard deviation of the number of cattle per farmer?

Since the sample is of size 10 in this example, it must be treated as a small sample. The required calculations are shown in the table below.

No of cattle x	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
14	0	0
7	-7	49
18	4	16
25	11	121
10	-4	16
21	7	49
13	-1	1
8	-6	36
11	-3	9
13	-1	1
140	**0**	**298**

For the mean,

$$\bar{x} = \frac{\sum x_i}{n} = \frac{140}{10} = 14$$

For the standard deviation,

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{298}{9}} = \sqrt{33.11} = 5.75$$

5.4.2 Variance and Standard deviation in grouped data

When data is grouped, we may not be able to see, let alone sum, individual values from a sample or population. Instead, a single value or class mark will be known, representing all the elements in a class. The summation therefore in the formula must take into account the fact that some known values must be added repetitively as they represent many unknown observations. They must be added as many times as are the represented values. In the formula, this will show the class frequencies multiplying with the class marks deviations.

$$Var(x) = s^2 = \frac{1}{n-1}\sum_{i=1}^{n} f_i(x_i - \bar{x})^2$$

Where x_i is the class mark (midpoint) of the i^{th} class and f_i is the corresponding class frequency. The n remains the total sample size, which in this case will be the summation of frequencies for all the classes. Consequently, the standard deviation for grouped data will be

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} f_i(x_i - \bar{x})^2}$$

Example 5.5

A reproductive health specialist is studying 50 couples. For each couple, data on the difference in age between the husband and wife is given in the table below. Particularly, the data is on how older the husband is over the wife.

Age diff	Number f
0-2	8
3-5	12
6-8	15
9-11	10
12-14	3
15-17	2
Sum	**50**

Calculate:

a. The mean age difference between husband and wife.

b. The standard deviation.

The sample size $n = 50$ is greater than the set benchmark of 30. Therefore, in the calculation of the standard deviation, it is permissible to treat it as a large sample.

Age diff	Number f	Mid-point x	$f_i x_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$f_i(x_i - \bar{x})^2$
0-2	8	1	7	-6.1	36.7	257.1
3-5	12	4	44	-3.1	9.4	103.0
6-8	15	7	105	-0.1	0.0	0.1
9-11	10	10	100	2.9	8.6	86.4
12-14	3	13	65	5.9	35.3	176.4
15-17	2	16	32	8.9	79.9	159.8
Sum	**50**		**353**			**782.8**

The formulas for grouped data are

$$\bar{x} = \frac{\sum f_i x_i}{n} = \frac{353}{50} \cong 7.1$$

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n} f_i(x_i - \bar{x})^2} = \sqrt{\frac{1}{50}(782.8)} = \sqrt{15.66} = 3.96$$

5.4.3 Properties of Variance (Standard deviation)

I. The variance of a constant is zero, $Var(\alpha) = 0$. This emanates from the fact that a constant does not vary. Since the variance measure the degree of variation in observations, it will be zero if the variable does not vary.

II. If a constant is added to each observation, the variance remains unaffected. It is not affected by change of location or level.

III. $Var\ (X \pm \alpha) = Var(X)$
Adding a constant to each observation (or subtracting from each observation) merely changes the location of all observation without affecting their spread.

IV. If all observations are scaled by a constant, the variance scales by the square of that constant. That, if all values of X are scales by a factor α,

V. $Var(\alpha X) = \alpha^2 Var(X)$
Since the standard deviation is a square root, it will scale by the same factor.

VI. The variance of a combination of two random variables, X and Y, is given by:

$$Var(\alpha X + \beta Y) = \alpha^2 Var(X) - 2\alpha\beta Cov(X,Y) + \beta^2 Var(Y)$$
$$Var(\alpha X - \beta Y) = \alpha^2 Var(X) + 2\alpha\beta Cov(X,Y) + \beta^2 Var(Y)$$

Where $C(X,Y)$ is the covariance of X and Y. Covariance is expected to be zero if the two variables are independent. In general,

$$Var\left(\sum X_i\right) = \sum Var(X_i)$$

This is known as Bienayme formula, named after its formulator, a nineteenth century French statistician Irenee-Jules Bienayme

5.4.4 Coefficient of Variation

The relative measure for the standard deviation is known as *coefficient of variation* (cv). It is the standard deviation expressed as a percentage of the corresponding mean.

$$cv = \frac{s}{\bar{x}}$$

In Example 5.5 above, we can calculate the coefficient of variation since both the mean and standard deviation are known

$$cv = \frac{s}{\bar{x}} = \frac{3.96}{7.06} = 0.56$$

Example 5.6

A company is considering selecting one of the two mutually exclusive projects; A and B. The expected net present value (NPV) for the two projects are $K360{,}000$ and $K500{,}000$ resepectively. The standard deviation for eth two projects are $K278{,}000$ and $K320{,}000$ resepctively. Help the company select the less risky project.

Though the standard deviation is indicative of risk, it is in absolute terms and fails to account for the different 'levels' of the different variable. This selection must therefore be made based on the coefficient of variation. The coefficient of variations for the two projects are:

$$cv_A = \frac{278{,}000}{360{,}000} = 0.772$$

$$cv_B = \frac{320{,}000}{500{,}000} = 0.640$$

Project B has a lower coefficient of variation despite having a higher standard deviation compared to project A. therefore, Project B is less risky and must be chosen.

The coefficient of variation is an important measure of relative variation that is often used in economics. It is the standard measure of volatility in variables, which signals instability in the variable. Consider a case study which illustrates the use of the coefficient of variation as a measure of volatility in economic growth rates presented in the box below.

Box 5-1. Case Study on Sustained Economic growth in Africa

1. Title of the case study

Sustained Economic Growth in Africa

2. Objective of the case study

Through analysing sustained economic growth in Africa, this case study purports to make an economic policy analyst to appreciate two important statistical measures that are widely used in economic analysis. These are the arithmetic mean and the coefficient of variation.

3. Summary description of the case study theme

Sustained economic growth can be analysed for a single country over a period of time (through time series data) or a comparative analysis can be done between two or more countries over a period of time (using panel data).

Sustained economic growth has two defining dimensions: the average growth rate and the volatility of the growth rate. On the basis of these dimensions, countries can fall, for instance, into four possible categories:

1. Those with low growth rates and high volatility;
2. Those with low growth rates and low volatility;
3. Those with high growth rates and high volatility;
4. Those with high growth rates and low volatility.

Of course, one can have more than two categories for each dimension. For instance, growth rates can be high, moderate or

low. Likewise, volatility can be low, moderate or high. In this instance, there will be nine categories (3 x 3).

The best-performing countries would be those with high growth rates with low volatility while the worst-performing countries would be those with low growth rates and high volatility.

There are no theoretical rules for determining the empirical categorization of countries on the basis of the average growth rate or volatility. For instance, what constitutes low growth? Is it 2% or below or is it 4% or below? Again, what constitutes high growth? Is it more than 2% or is it more than 5%? And so on.

The empirical categorization will have to be determined by the researcher on the basis of the data that are available. For instance, if there are no countries in the chosen sample having average growth rate over a certain period that exceeds 3%, then there is no point in setting 5% or more as constituting high average growth rate since comparative performances are always relative. So a certain amount of subjective but circumspect discretion will have to be exercised by the researcher in forming the empirical categories of the variables. This will be seen in Section 5.

Policy analysts and policy decision-makers would have to first understand as to what the situation is in respect of sustained economic growth in their respective countries. Is there a problem of low average economic growth or is there a problem of high volatility of the growth rate or is there a problem of both? Depending on the perceived nature of the problem, one has then to come up with measures to either raise the level of economic growth or reduce its volatility or both. And to do this, one has to identify the factors that are either constraining growth or producing high fluctuations in growth over time.

NB: Even if a country is seen to do well on both counts (high growth, low volatility), one can still explore the possibility of

doing even better in the future. For instance, suppose average growth rate in a country has been 6%, is there scope to raise it further to say 8% or higher without adversely impacting on volatility?

Also, some additional caveats can be included to obtain more stringent or more relaxed definitions of sustainability. This will again be illustrated in Section 5.

4. Summary descriptions of the statistical measures

From classroom lectures, one would be familiar with the following measures;

- The arithmetic mean (AM) which is used to measure average growth rate over a period of time. The AM is calculated by dividing the sum of all the values of the variable by the number of values.
- The coefficient of variation (CV) which is used to measure the volatility of the growth rate over a period of time. The CV is calculated as the ratio of the standard deviation to the mean multiplied by 100 to express it as a percentage. It is a measure of dispersion (or fluctuations or volatility in a given variable).

5. The Case Study

This case study is drawn from the *Economic Report on Africa 2011* published by the United Nations Economic Commission for Africa (UNECA). The relevant portions have been extracted from Chapter 4 of the Report.

Over the period 1960-2007, 16 African countries (accounting for about 18 percent of Africa's population) had average annual real per capita GDP growth rates in excess of 2 percent; 11 countries (accounting for 15 percent of the continent's population) recorded negative growth rates; and 26 countries recorded

positive growth rates of less than 2 percent, and 12 of them less than 1 percent.

Among the major features of the African growth process, especially those of sub-Sahara Africa, is their relatively high volatility. Measuring volatility by the coefficient of variation, and using a value of one or less as a benchmark for very low volatility (as in the case of Malaysia), it is found that none of the growth processes of the African countries was characterized by very low volatility over the entire period 1960-2007. Low volatility, which is defined as a coefficient of variation of greater than one but less than three, was recorded for 12 countries. The lowest volatility was recorded for Botswana, with a coefficient of variation of 1.1 (table 1).

Moderate volatility, defined as a coefficient of variation of three but less than six, was recorded for 16 countries; high volatility, defined as a coefficient of variation of six but less than ten, was recorded for 13 countries; and very high volatility, defined as a coefficient of variation of 10 and greater, was recorded for the remaining 12 countries, with the highest volatility recorded for Zambia with a coefficient of variation of about 70 (resulting from an average growth rate of real per capita GDP of 0.15 percent a year and a standard deviation of 10.46).

Table 1: Growth and volatility in Africa, 1960-2007

Volatility (coefficient of variation)	Average annual real per capita GDP growth rates (%)			
	Less than 0	0-1	1-2	2+
Low (1-3)			Tanzania, United Rep. of (2.8; 1.5) South Africa (1.5; 1.5)	Botswana (1.1; 5.5) Cape Verde (2.0; 3.2) Egypt (1.6; 3.2) Equatorial Guinea (2.8; 8.4) Lesotho (2.5; 2.9) Mauritius (2.1; 3.2) Morocco (2.1; 2.8) Seychelles (2.1; 4.0) Swaziland (2.8; 3.5) Tunisia (1.2; 3.4)
Moderate (3-6)	Central African Rep. (4.4; -1.0) Congo, Dem. Rep. of (3.4; -2.6) Somalia (4.7; -1.6)		Benin (3.7; 1.2) Burkina Faso (5.1; 1.2) Mali (4.9; 1.3) Mozambique (3.7; 1.7) Namibia (4.0; 1.1) Nigeria (4.9; 1.8) Sudan (4.3; 1.9)	Angola (5.3; 2.1) Congo (3.9; 2.8) Gabon (4.0; 2.2) Ghana (5.4; 2.9) Malawi (4.4; 2.0) Mauritania (4.2; 2.6)
High (6-10)	Djibouti (6.5; -1.5) Niger (7.8; -0.7) Senegal (9.7; -0.4)	Cameroon (6.6; 0.8) Comoros (6.5; 0.7) Côte d'Ivoire (7.5; 0.7) Kenya (9.7; 0.4) Uganda (8.2; 0.6)	Algeria (7; 1.2) Chad (8.0; 1.2) Eritrea (6.3; 1.3) Ethiopia (7.1; 1.0) Guinea-Bissau (7.9; 1.6)	
Very High (10+)	Liberia (13.8; -1.6) Libyan Arab Jamahiriya (10.7; -1.1) Madagascar (57.6; -0.2) Sao Tome and Principe (27.0; -0.3) Zimbabawe (20.6; -0.5)	Burundi (20.7; 0.3) Gambia (34.1; 0.2) Guinea (17.2; 0.2) Rwanda (26.0; 0.5) Sierra Leone (19.3; 0.4) Togo (24.1; 0.2) Zambia (69.7; 0.2)		

Source : Calculations by UNECA based on World Bank, World Development Indicators (2010)

Note: The first entry in parentheses is the coefficient of variation (the ratio of the standard deviation to the absolute value of the average annual growth rate); the second entry is the average annual per capita GDP growth rate as a percentage.

Based on table 1, a sustained growth process may be defined as one that requires an average annual real per capita GDP growth of 2 percent or more over the period 1960-2007, maintained for each of the three sub-periods (1960-1972, 1973-2000 and 2000-2007), with low volatility for the entire period, where low volatility may be defined by a coefficient of variation for the growth rates of one to less than three. Using this definition of sustainability, only six African countries recorded sustained growth over the period in question: Botswana (with an average annual real per capita GDP growth rate of about 5.5 percent and a standard deviation of about 6.2 percentage points); Cape Verde (3.2 percent and 6.4 percentage points); Egypt (3.2 per cent and

5.2 percentage points); Equatorial Guinea (8.4 percent and 23.6 percentage points); Lesotho (2.9 percent and 7.4 percentage points); and Tunisia (3.4 percent and 4.3 percentage points).

Combining the sustainability and volatility of the African growth process, a sustained, low volatility growth country will be classified as having achieved a *classical structural transformation* of its economy during 1970-2007 if the respective GDP shares of the three sectors of agriculture, industry and services, and of the manufacturing subsector, obey the stylized paths of structural transformation as real per capita GDP increases. According to the available information, only one African country, out of the six countries that achieved sustained growth over the period since 1960, was able to satisfy the requirements of a classical structural transformation during 1970-2007-Tunisia.

A more relaxed definition of a sustained African growth process would require maintaining an average annual real per capita GDP growth rate of 2 percent or more for the entire period, as well as for two sub-periods, and a positive annual rate of growth for the third sub-period; together with low volatility for the entire period. Such a relaxed definition adds four countries: Mauritius (with an average annual real per capita GDP growth rate of 0.46 percent in 1960-1972); Morocco (1.54 percent in 1972-2000); Seychelles (0.23 percent in 2000-2007); and Swaziland (1.36 percent in 2000-2007).

Case Study by: Seshamani, Venkatesh and Frank C. Chansa

5.5 Further Note on Achievement and Dispersion

The measures of dispersion, MAD and standard deviation, have been based on the central value. This does not however imply that the value of dispersion is dependent on the central value. It is possible to have two samples with the same mean yet different values of dispersion. The level would be the same but one having a higher spread than another. In the same

way, two samples would have the same standard deviation yet differ in the mean. Consider the following scenario.

A quarrying company wants to measure the amount of quarry produced in a day. A tipper truck is used to haul quarry from the plant. In order to know how much quarry is produced, the truck is made to go over a weigh bridge and the measurement is recorded. After ten (10) trips, the measurement are as follows.

Trip	1	2	3	4	5	6	7	8	9	10
Weight	43	39	36	40	45	38	43	41	37	36

What is the mean weight and standard deviation?

In order to make the calculations, the data is replicated in the table below so that more columns can be added.

Trip	Weight	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	43	3.2	10.24
2	39	-0.8	0.64
3	36	-3.8	14.44
4	40	0.2	0.04
5	45	5.2	27.04
6	38	-1.8	3.24
7	43	3.2	10.24
8	41	1.2	1.44
9	37	-2.8	7.84
10	36	-3.8	14.44
Total	**398**		**89.6**

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{398}{10} = 39.8$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{9}(89.6)} = \sqrt{9.95} = 3.155$$

Suppose now after presenting these results, we are told that the weight measurement erroneously include the vehicle's net weight of 15 tonnes. What is the true mean and standard deviation per trip?

The implication of this information is that each measurement must be reduced by 15, to account for the truck's own weight. Subtract a constant from each measurement. With this at the back of the mind, theory says that the mean will reduce by the same constant. For the standard deviation, it must remain unchanged. The reasoning is simple. Since all measurements have shifted by the same distance as the mean, their relative position with respect to the mean will be unchanged. Each maintains its distance from the mean. Since the standard deviation is a measure of deviations from the mean, producing a different value will be parodying.

CHAPTER 6

6 MEASURES OF SYMMETRY AND PEAKEDNESS IN ACHIEVEMENT

In addition to the measures of central value and dispersion discussed in preceding chapters, there may also be interest in getting an idea on the positions of observations. If we say the variation in data is high, we do not in any away make comment on whether the far lying observations are one sided or fall on both sides of the central value. This chapter discusses the concepts of symmetry and peakedness in achievement and how they can be measured.

6.1 Symmetry

Every distribution of data has the lower and upper end, it has the left and right side with measures of achievements defining the middle area. There are observations that will fall above the average and others below. In the case of symmetry, the interest is to understand how observations are shared between those above and those below average. It is the symmetry of the frequency graph.

When observations are equally distributed on both sides of the central value, the corresponding frequency curve will be *symmetrical* or *bell-shaped*. This was noted earlier in Chapter 3 and occurs when observations equidistant from the central value have the same frequency. By equidistant we mean observations that are of the same distance from the average - mean, median or mode – on the right and left side have the same frequency. That is, observations that are one unit on the right of the mean have the same frequency as observations that are one unit on the left side of the mean. When the distribution is symmetrical, the three measures of achievement – the mean, median and mode – are equal.

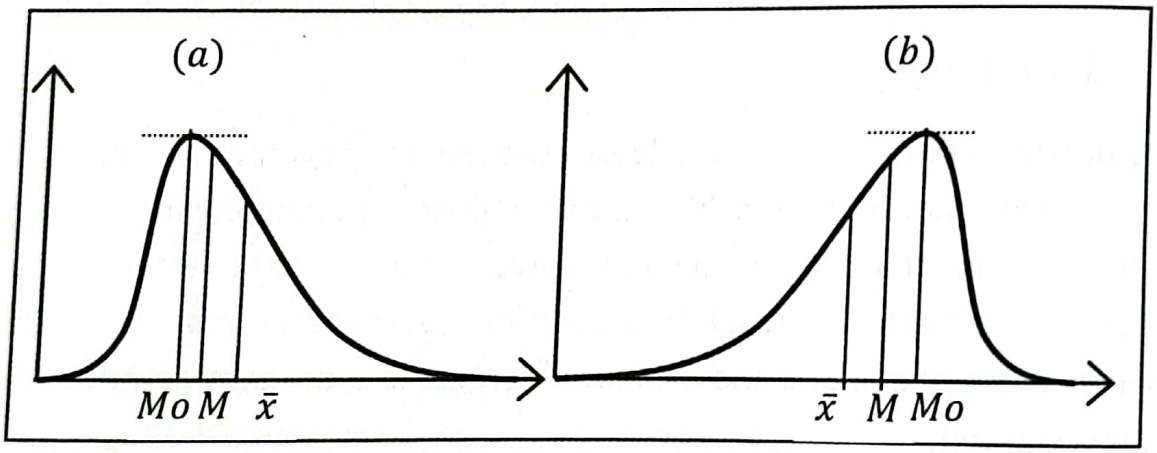
When the distribution is not symmetrical, it is said to be *asymmetrical*. It is also said to be *skewed*, to the right or left as the case may be. It is negatively skewed or skewed to the left when the longer tail is on the left and positively skewed or skewed to the right when the longer tail is on the right side. When the distribution is skewed, the three measures of achievement do not coincide.

Two measures of skewness or symmetry in distributions are used. The first in based on the comparison of the three measures of achievement. The second is based on the definition of the quartiles.

6.1.1 Average-Based Measure of Skewness

Consider the following frequency curves for two asymmetrical distributions.

Figure 6.1. Skewness - average based



Since the mean takes into consideration all the observations, it is affected by extreme values. When the distribution is skewed, having extreme values on one side, the mean will be pulled in that direction. In a right skewed distribution, the mean will lie on the right of the midpoint (median) and the mode. We can then think of the measure of skewness being based on the difference between the mean and the mode (or the median). When the distribution is symmetrical, the difference is zero and therefore no amount of skewness. When the difference is positive ($\bar{x} > Mo$)

Based on the above reasoning, skewness is measured by the following formula.

$$sk = \bar{x} - Mo$$

The smaller the *absolute value*, the less skewed the distribution is. It is an indication of a more symmetrical distribution. When the absolute value is very high (negative or positive), the distribution is highly skewed. There is a very big disparity between the right and left side of the middle. One end tail is longer than the other.

A negative value of skewness mean the distribution is negatively skewed. There are more outliers on the left than are on the right side of the central value. When positive, the distribution is positively skewed or skewed to the right. It has more outliers on the right or the tail is much longer on the right than on the left.

In some cases, the mode may be ill-defined such that it does not give a true picture of the distribution. This may necessitate the use of the median in place of the mode. This should not be interpreted to mean replacing the mode with the median, the two are not always equal. Instead, it means using an alternative equation which avoids the mode. Recall in the measure of central value that there is a relationship among the three measures – mean, median and mode. That is

$$\bar{x} - Mo = 3(\bar{x} - M)$$

Based on this relationship, when the mode is undesirable in the measure of skewness, we could still measure it using

$$sk = 3(\bar{x} - M)$$

The above measures of skewness are absolute, they will be misleadingly higher if larger numbers are used and lower when smaller numbers are used. To control this anomaly, the relative measure, known as the Coefficient of Skewness is used. This divides the absolute measure by the distribution's standard deviationσ. This is also known as Pearsonian coefficient of skewness, named after an English mathematician and biometrician Karl Pearson (1857 – 1936).

$$Coef\ of\ Sk = \frac{\bar{x} - Mo}{\sigma}$$

When the mode is undesirable, the measure is

$$Coef\ of\ Sk = \frac{3(\bar{x} - M)}{\sigma}$$

6.1.2 Quartile-based Measure of Skewness

When the distribution is symmetrical, the median must roughly be the midpoint of the lower and upper quartiles also known as the first and third quartiles. The two quartiles must roughly be equidistant from the median. That is

$$M - Q_1 = Q_3 - M$$

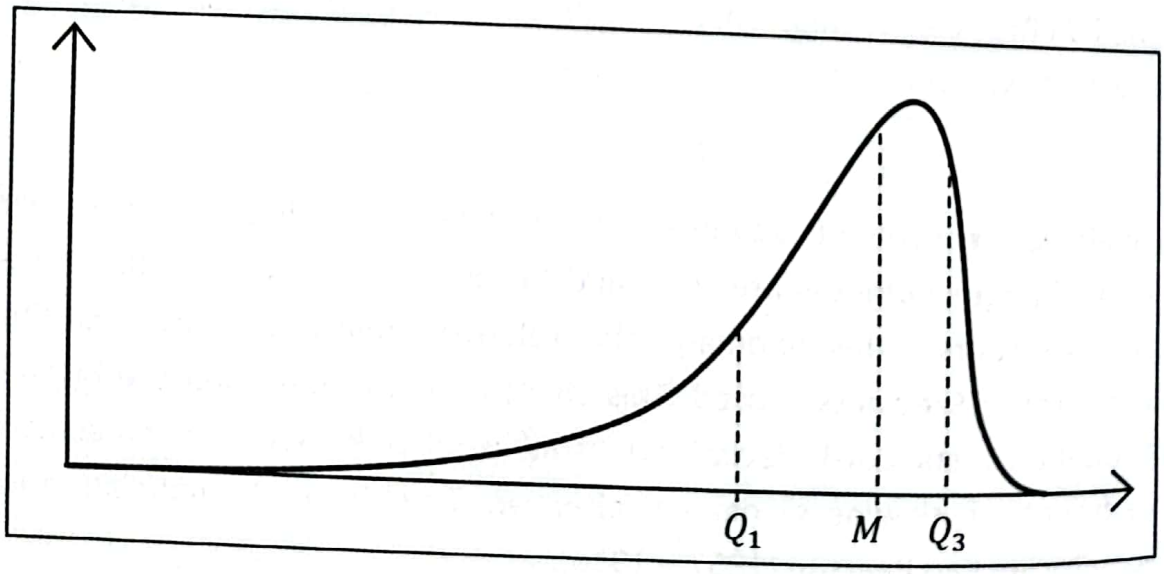
Where Q_1 and Q_3 denote the first and third quartiles respectively while M is the median. The implication of the above equation is that

$$Q_1 + Q_3 = 2M$$

$$Q_1 + Q_3 - 2M = 0$$

With an asymmetric distribution however, one quartile will be closer to the median than the other. Consider the frequency curve given in the figure below.

Figure 6.2. Skewness - Quartile based



In the above figure, Q_3 is much closer to the median than Q_1. The average of the two or the midpoint cannot equal the median. In this particular case, it will be less than the median. The average will be less than the median in a

left skewed distribution because the higher spread on the left causes the lower quartile to be far from the median compared to the upper quartile. When the average of the two quartiles is computed, it is less than the median. In the same way, the average of the lower and upper quartiles will be greater than the median in a positively skewed distributions.

Therefore,

$$Q_1 + Q_3 - 2M$$

provides a useful measure of skewness in distributions. The relative measure of skewness based on quartile is achieved by dividing the above absolute measure by the interquartile range. This is known as Bowley's coefficient of skewness.

$$Coef\ of\ sk = \frac{Q_1 + Q_3 - 2M}{Q_3 - Q_1}$$

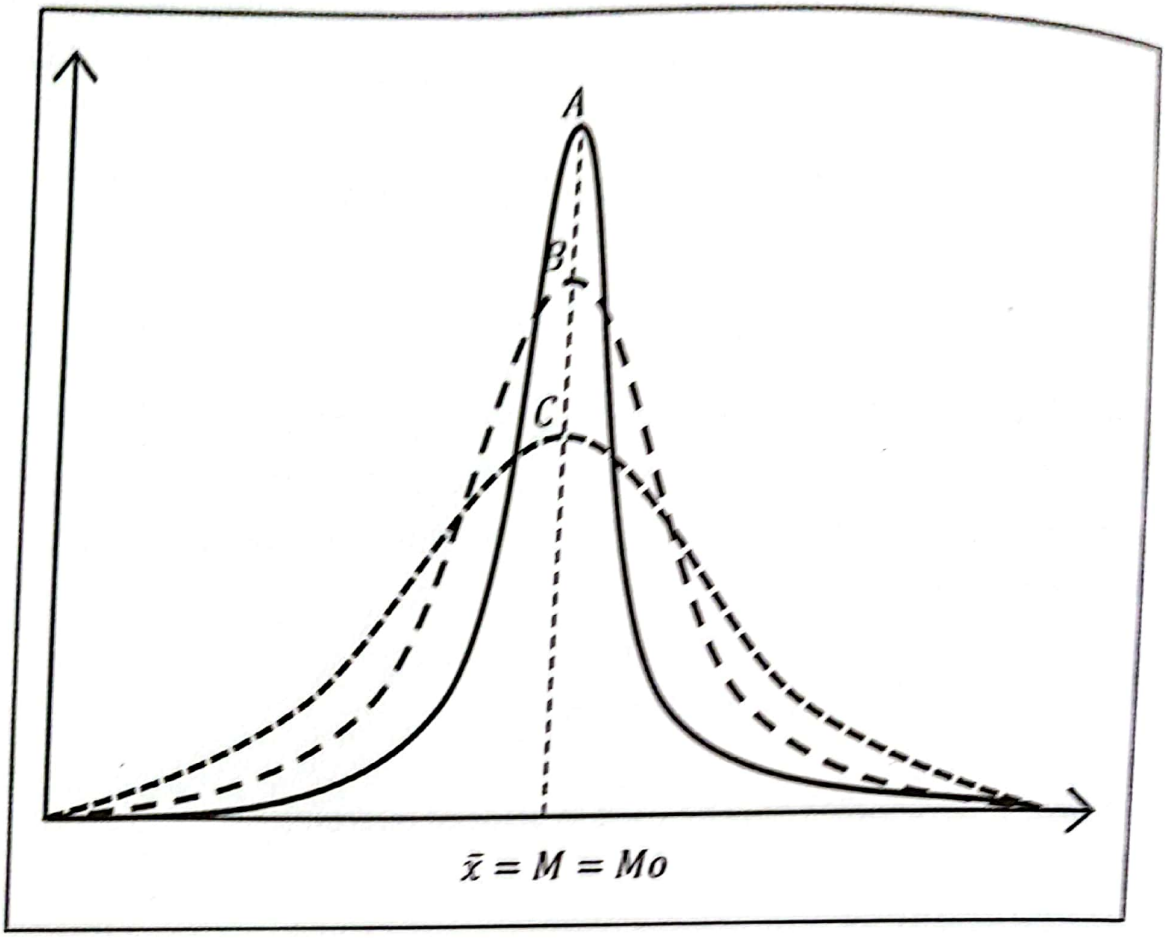
Alternatively, percentiles may be used in place of quartiles.

6.2 Peakedness

Besides the skewness, we are also interested in the peakedness of the frequency curve. Since the frequency curve takes the shape of a hill, the peakedness refers to how steep the hill climbs. Are the sides steep or are they somewhat flat and climb slowly? This is also known as *Kurtosis*, derived from a Greek word for the degree of peakedness of a frequency distribution.

Consider the following figure in which we compare three distributions each represented by a frequency curve.

Figure 6.3. Kurtosis



The three distributions are represented by the three frequency curves shown in the table. The three distributions are symmetrical, all their means are theoretically equal. However, the peakedness representing the concentration of observations around the central value is visibly higher in A than it is in the other two. It is lowest in C. Distribution B is said to be higher when compared to C and lower when compared to A. it is a case of 'who you are standing with'.

This is a relative measure. To standardise the comparison, the measure of peakedness is done in comparison to a normal frequency curve which is said to be *mesokurtic*. This is a Greek formulation for middle kurtosis. The distribution has a high peak if it is more peaked than a normal curve. This is known as *leptokurtic*. The prefix 'lepto' in Greek meaning 'slender, thin or narrow'. It is used here to refer to a frequency distribution with a high peakedness, making it slender or narrow. A leptokurtic distribution is therefore one with a higher kurtosis.

When the distribution is less peaked compared to the normal distribution, it is said to be *platykurtic*. Again, the prefix 'platy' is derived from Greek for 'broad or flat'. It is used here to refer to frequency curves that are flat compared to the 'marker', the normal distribution.

In order to measure kurtosis, we must first define moments. In general, the r^{th} moment about the mean is

$$\mu_r = \frac{\sum(x_i - \bar{x})^r}{n}$$

The first four moments are:

$$\mu_1 = \frac{\sum(x_i - \bar{x})}{n} = 0$$

$$\mu_2 = \frac{\sum(x_i - \bar{x})^2}{n} = \sigma^2$$

$$\mu_3 = \frac{\sum(x_i - \bar{x})^3}{n}$$

$$\mu_4 = \frac{\sum(x_i - \bar{x})^4}{n}$$

Kurtosis is then measured by the coefficient

$$\beta_2 = \frac{\mu_4}{(\mu_2)^2} = \frac{\mu_4}{\sigma^4}$$

This has a middle value of $\beta_2 = 3$, the mesokurtic distribution. When $\beta_2 < 3$, the distribution is platykurtic or less peaked. It is leptokurtic or highly peaked and narrow when $\beta_2 > 3$. An alternative coefficient $\gamma_2 = \beta_2 - 3$ allows the comparison to be done with respect to zero. The distribution will be platykurtic, mesokurtic and leptokurtic if γ_2 is positive, zero or negative respectively.

Example 6.1

The following table shows the number of times that each of the 100 Members of Parliament had lost a parliamentary election before first getting elected.

x	f	fx
0	8	0
1	17	17
2	36	72
3	22	66
4	11	44
5	5	25
6	1	6

The mode and median number of losses is 2 while the mean number of losses is 2.3. calculate

a. Pearson's coefficient of skewness

b. The measure of kurtosis

c. Comment on the symmetric and peakedness of the distribution.

In order to answer the above questions, we extend the table. The Pearson's coefficient of skewness requires that we calculate the absolute measure of skewness $(\bar{x} - Mo)$ and the standard deviation $(x_i - \bar{x})$. Kurtosis on the other hand requires the calculation of the second the forth moments about the mean.

$$sk = \bar{x} - Mo = 2.3 - 2 = .3$$

$$\sigma = \sqrt{\frac{1}{n}\sum f(x_i - \bar{x})^2} = \sqrt{\frac{1}{100}167} = \sqrt{1.67} = 1.29$$

$$Coef\ of\ sk = \frac{\bar{x} - Mo}{\sigma} = \frac{0.3}{1.29} = 0.233$$

Kurtosis on the other hand requires the calculation of the second and forth moments about the mean. The moments are defined as follows

$$\mu_2 = \sigma^2 = 1.67$$

$$\mu_4 = \frac{1}{n}\sum f(x_i - \bar{x})^4 = \frac{1}{100}823.01 = 8.23$$

The measure of kurtosis is then

$$\beta_2 = \frac{\mu_4}{(\mu_2)^2} = \frac{8.23}{(1.67)^2} = 2.95$$

From the above calculations, we see that the measure of skewness is quite low, closer to zero. We say there is little positive skewness in the distribution. For kurtosis, the result must be compared to $\beta_2 = 3$ for the normal distribution. Clearly, the calculated value is just slightly below this middle value. It is safe to conclude that the distribution is mesokurtic.

Moments can also be used to measure skewness. In particular, the moment coefficient of skewness is

$$\gamma_1 = \frac{\mu_3}{\sigma^3}$$

CHAPTER 7

7 MEASURES OF STRENGTHS OF RELATIONSHIPS – CORRELATION ANALYSIS

In life, we often link many variables on the basis of behaving in a coordinated manner. In the farming communities, higher temperatures are considered precursors to good rains – and as a consequence, good harvest. A school teacher will expect that pupils who are absent from school frequently will perform poorly in assessment. Good governance will help attract foreign direct investment (FDI). In all these illustrations, the two variables move in the same direction.

Again, increasing investment in the country is expected to lead to reduction in the level of unemployment. Higher price of a commodity will generally be associated with a lower quantity demanded of it. In economics, according to a theory represented by a well-known Philips curve, higher inflation is generally associated with lower levels of unemployment. In these illustrations, although there is association between the two variables, the movements are in opposite direction.

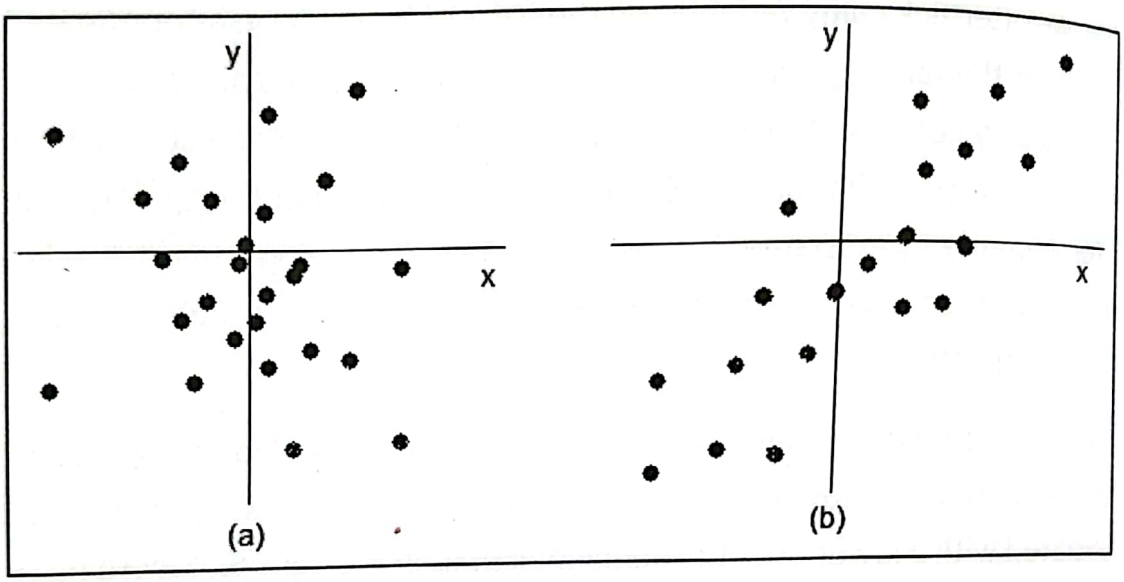
It is therefore paramount that in each relationship, there is an objective way of measuring the strength of the relationship. The measures used for this purpose are known as *measures of correlation*.

Correlation may be defined as the degree and type of relationship between any two variables which vary together. A positive correlation exist where the high values of one variable are associated with the high values of the other variable. A negative correlation means association of high values of one variable with the low values of the other. It should be noted here that while correlation suggests possible connections or associations between variables, it does not prove or disprove any cause-and-effect relationships

between them. Such causal relationships are analysed through regression, which is dealt with in the next chapter.

Consider Figure 7.1 below which shows two scatter plots, (a) and (b).

Figure 7.1. Scatter Plot to show Covariation



In part (a) of the plot, no pattern in the dots is visible. In (b) however, it is clear that higher values of y are associated with higher values of x. We see a pattern in which values of y that are above the average (in whatever way calculated) are associated with values of x that fall above the average. Thus, when x is higher, so is y and when x is lower, so will y be. We can conclude that there is some covariation in part (b) of the figure which is not the case with figure (a).

7.1 Explaining Causation

When the price of maize is high, the producer or farmer has the income increase. That is, holding all other factors constant, the increase in maize prices causes incomes to go up. These are relationships that have been covered in the introductory economics courses. We can also say the increase in prices causes the quantity demanded of a good to drop. An increase in general wages causes demand in an economy to go up. In all these cases, the relationships are very clear, one variate causing changes in another.

In statistics, this is known as *causation*. It is a strong case and quite rarely established. Causation occurs when changes in one variable cause changes in another. An increase in temperature causes the volume of air to increase. This is the principal behind the combustion engine. Since it is a case of causality, there is guarantee that once heated, the volume of air will increase and turn the engine.

Thus in a causation kind of relationship, we see the two variables changing in a synchronised manner. When one changes, the other also follows, either simultaneously or with a lag. An increase in supply will not cause the quantity supplied increase instantly but will show its effect over time. The demand on the other side may be quicker to respond to changes in the price.

The above type of analyses is done under regression.

7.2 Genuine and Spurious Correlation

As already stated, simultaneous or related changes in variables may not always be indications of causation between them. Two variables may move together by pure coincidence. This is known as *spurious correlation*. By accident, the increase in one variable coincides with increase in another and so on. At times, it may be due to common links with a third variable.

Suppose when temperatures are rising in Britain, the number of tourists visiting the Victoria Falls is also seen to increase and when the temperatures in Britain are dropping, the number of tourist also shows signs of falling. In this case, it may not be prudent to conclude that temperatures in Britain are related to the number of tourists that visit the famous Victoria Falls. This is an example of a spurious relationship.

A plausible explanation for this seeming association is that changing temperatures and water levels at the falls occur almost at the same time of the year. Temperatures in the northern hemisphere are lowest in December. This is the month when rains feeding the falls start and the water levels are at the lowest. The falls is not very attractive and will record the lowest visits. As the temperature is increasing, so is the water level on the falls. The temperatures are highest around June and this is the time the falls

is probably most attractive, hence the link between the two variables. In other words, the seeming association between the two variables is because of the common relationship to a third variable, namely, temporal synchronisation of weather patterns in the two countries.

The link between the two variables is not explained by any theory or empirical evidences. There is no justification to suspect that one of the two causes the other. Nonetheless, the relationship is indisputable. The two variables vary in a related way. They are said to *covary*. Thus, there is *covariation* between the two variables.

7.2.1 Measuring Covariance

Assume two variables, $X$and Y, the former measured in the horizontal axis and the latter in the vertical axis. Each variable has values which may be positive or negative wholly or in part. This means the points of the scatter plot of the two variables will fall in any of the four quadrants. For instance, if we are looking at a person's height (x) and body weight (y) both variables are strictly positive. All the points of the scatter plot will fall in the first quadrant.

However, when the deviation from the mean is used in place of the absolute values, each variable will have a fair share of negative and positive values.

A measure of covariance is defined as the average of the product $(x - \bar{x})(y - \bar{y})$. Thus,

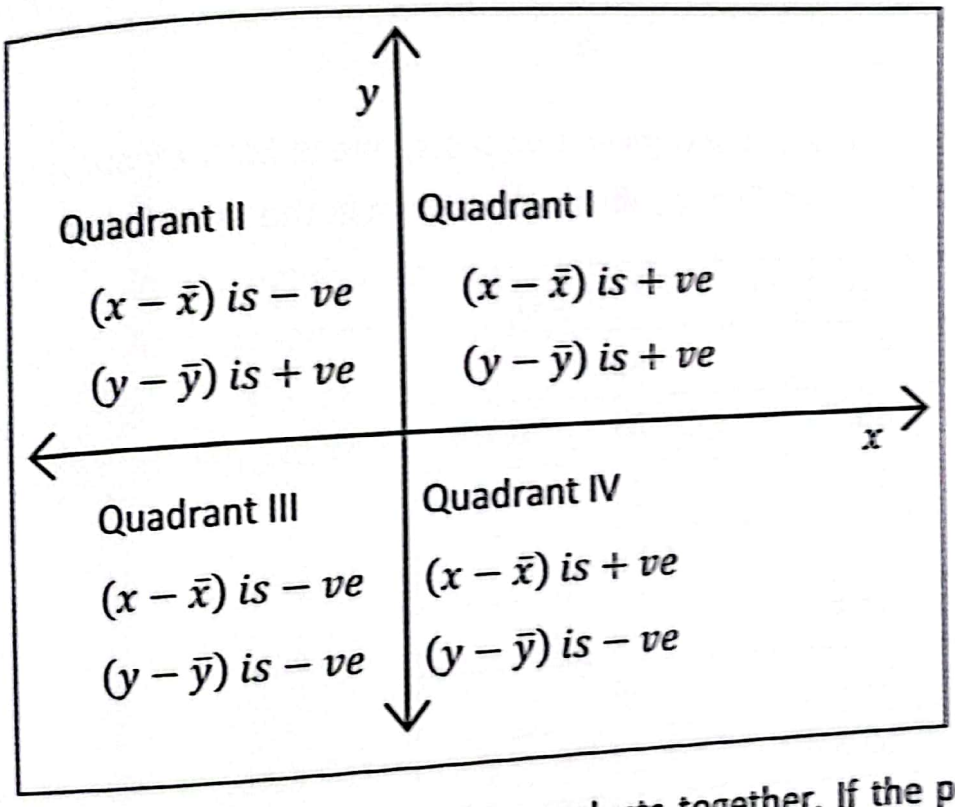
$$Cov(x, y) = \frac{\sum(x - \bar{x})(y - \bar{y})}{n}$$

where n is the number of pairs. When both $(x - \bar{x})$ and $(y - \bar{y})$ in a pair are of the same sign, falling in the first or third quadrant, the product $(x - \bar{x})(y - \bar{y})$ is positive. When the signs differ, one is positive and another negative, the product will be negative and the point will fall either in the second or fourth quadrant depending on which variable is positive and negative.

When the sample size is small, ideally below 30 pairs, we have to make use of Bessel's correction mentioned in the discussion of the variance. With a small sample size, the above formula tends to understate the true covariance. This is corrected by reducing the denominator by unity, known as Bessel's correction.

$$Cov(x,y) = \frac{\sum(x-\bar{x})(y-\bar{y})}{n-1}$$

Figure 7.2. The four quadrants.



The summation brings all the products together. If the positive products outweigh the negative, the sum will be positive, indicating a positive covariance. This occurs when most scatter points fall in the first and third quadrants (where $(x-\bar{x})$ and $(y-\bar{y})$ are of the same sign). When a *best fit* line is fitted, it will tend to be upward sloped, with a positive gradient. We say there is a positive covariance between the two variables.

When most points fall in the second and fourth quadrants, the negative products will outweigh the few positive products falling in the first and third quadrants. The summation will be negative and so will the measure of covariance. A best fit line will be negatively sloped. The covariance between the two variables is said to be negative.

When the scatter points are just dotted around the origin, often falling equally in all the four quadrants, the negative and positive products often tend to be equal. In this case, the summation will be zero or near zero. We say the covariance is zero or almost zero depending on how it compares to zero.

However, the actual value of the covariance will depend upon the number and magnitude of the deviations that are distributed among the four quadrants.

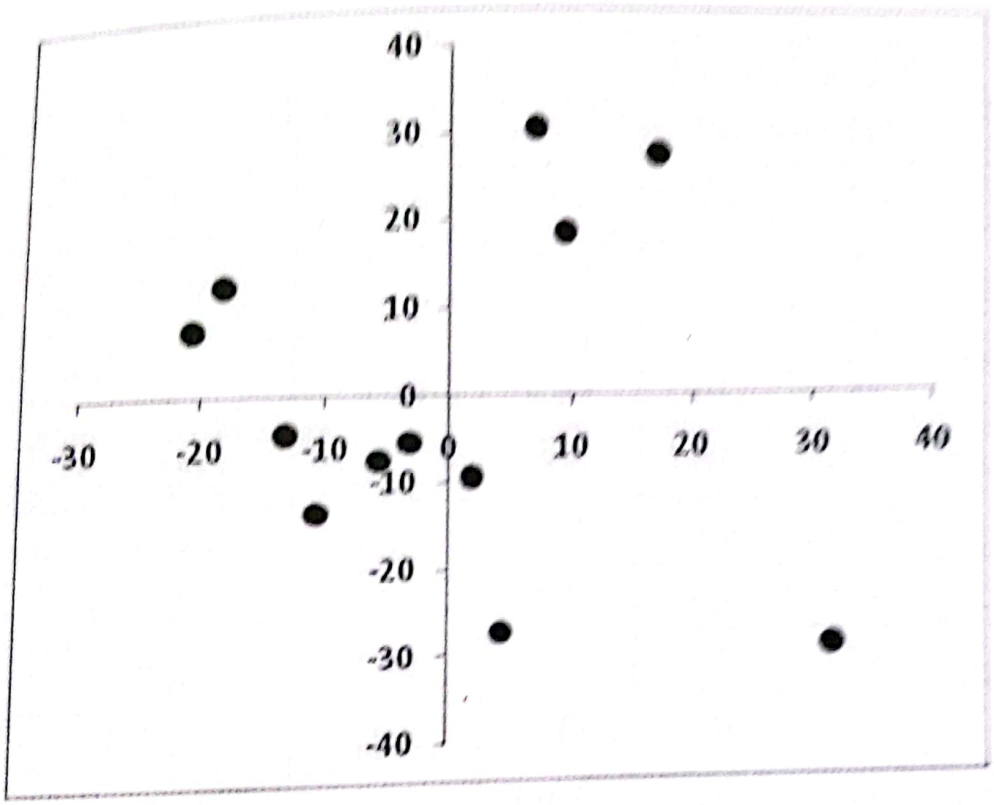
Example 7.1

A class of twelve pupils was given two tests, one in Mathematics and another in Accounts. The results are as shown in the table below.

S/N	Test 1 (x)	Test 2 (y)
1	60	41
2	65	81
3	90	21
4	75	78
5	37.5	58
6	67.5	69
7	62.5	23
8	40	63
9	55	45
10	45	46
11	47.5	37
12	52.5	43

Calculate the covariance between the two tests.

Before going into calculation, it is always helpful to start with a scatter plot. We present this below.



In the above scatter plot, the deviations are scattered among all the four quadrants. However, the value of the covariance could be positive or negative, large or small depending upon the number and magnitude of the individual deviation.

S/N	Test 1 (x)	Test 2 (y)	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$
1	60	41	1.9	-9.4	-17.7
2	65	81	6.9	30.6	210.3
3	90	21	31.9	-29.4	-937.7
4	75	78	16.9	27.6	465.5
5	37.5	58	-20.6	7.6	-156.4
6	67.5	69	9.4	18.6	174.2
7	62.5	23	4.4	-27.4	-119.9
8	40	63	-18.1	12.6	-228.1
9	55	45	-3.1	-5.4	16.9
10	45	46	-13.1	-4.4	58.0
11	47.5	37	-10.6	-13.4	142.6
12	52.5	43	-5.6	-7.4	41.7
Total	697.5	605.0	0.0	0.0	-350.6

Since the sample size is small, we use the corrected formula for the covariance.

$$Cov\,(x,y) = \frac{\sum(x-\bar{x})(y-\bar{y})}{n-1}$$

$$= \frac{-350.6}{11} = -31.87$$

We get a negative value. For now, we are able to comment on the sign (direction of the relationship) but we can provide no further interpretation of the number itself (strength of relationship). There are no normative benchmarks against which it can be evaluated. Hence, we have a more refined measure which allows such an interpretation. This is the *coefficient of correlation*.

7.3 Measuring Correlation

There are several methods to calculate the coefficient of correlation. Here we discuss only two of them, namely, Pearson's coefficient of correlation r and Spearman's coefficient of rank correlation r_{rank}.

7.3.1 Pearson's method

The Pearson's coefficient of correlation, named after an English mathematician and biometrician Karl Pearson (1857 – 1936), uses the product moment to calculate a relative measure of covariance.

This relative measure is obtained by using relative deviations in place of the absolute deviations. To get the relative deviations, each deviation is divided by the respective standard deviation. That is, instead of using $x - \bar{x}$ as in the calculation of covariance, we use $\frac{x-\bar{x}}{\sigma}$ where σ is the standard deviation of the variable x. The same applies for y. The result is the Pearson's coefficient of correlation or the *product moment coefficient of correlation* because of its use of the product moment or covariance. It is denoted by the lower case letter r.

$$Coef\ of\ Corr = r = \frac{\sum\left(\frac{x-\bar{x}}{\sigma_x}\right)\left(\frac{y-\bar{y}}{\sigma_y}\right)}{n}$$

Since the two standard deviations are constant, they can come behind the summation symbol, effectively changing the presentation of the formula.

$$r = \frac{\frac{1}{\sigma_x\sigma_y}\sum(x-\bar{x})(y-\bar{y})}{n}$$

$$= \frac{\sum(x-\bar{x})(y-\bar{y})}{n\sigma_x\sigma_y}$$

$$= \frac{\frac{1}{n}\sum(x-\bar{x})(y-\bar{y})}{\sigma_x\sigma_y}$$

$$= \frac{Cov\,(x,y)}{\sigma_x\sigma_y}$$

The Product moment coefficient of correlation is the covariance divided by the product of the square root of the two variances. It is a pure number, independent of the unit of measurement. It is negative if the variables are negatively correlated and positive if they are positively correlated.

The product moment coefficient of correlation (as other measures of correlation) has a normative range of -1 to 1. The sign (positive or negative) is indicative of direction of association. It provides information on whether the variates are moving together or moving in opposite directions. The strength or closeness of association is based on the absolute value. It is absent when $r = 0$ and is perfect when $|r| = 1$. The table below describes the degree of correlation for given levels of the absolute value of r.

Table 7.1. Interpreting the absolute coefficient of correlation$r$.

Level of $|r|$	**Degree of correlation**
$\|r\| = 1$	Perfect
$0.75 < \|r\| < 1$	High
$0.25 \leq \|r\| \leq 0.75$	Moderate
$0 < \|r\| \leq 0.25$	Low
$\|r\| = 0$	Absent

However, two things have to be born in mind. While the above a rules of thumb to measure the strength of correlation, whether the value of r is deemed to be low or high depends on the kind of data that is used. For example, in the case of time series data, a value of $r < 0.8$ may not be regarded as very satisfactory; while in the case of cross section data, a value of $r = 0.4$ may be regarded as good.

Second, r may be high or low but if it is calculated from a sample and not from a population, one will have to go by the statistical significance of the r value. A relatively high value of r in a calculated from a sample could turn out to be insignificant in some circumstance while a lower value of r calculated from a sample could turn out to be significant in other circumstance.

Example 7.2

Using the data provided in Example 7.1, calculate the Pearson's coefficient of correlation and comment on the results.

We know, based on the solution to the example that $Cov\,(x,y) = -31.87$. All that remains now is the calculation of the two variances, for x and y.

S/N	Test 1 (X)	Test 2 (Y)	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
1	60	41	1.9	-9.4	3.5	88.7
2	65	81	6.9	30.6	47.3	935.3
3	90	21	31.9	-29.4	1016.0	865.3
4	75	78	16.9	27.6	284.8	760.8
5	37.5	58	-20.6	7.6	425.4	57.5
6	67.5	69	9.4	18.6	87.9	345.3
7	62.5	23	4.4	-27.4	19.1	751.7
8	40	63	-18.1	12.6	328.5	158.3
9	55	45	-3.1	-5.4	9.8	29.3
10	45	46	-13.1	-4.4	172.3	19.5
11	47.5	37	-10.6	-13.4	112.9	180.0
12	52.5	43	-5.6	-7.4	31.6	55.0
Total					2539.1	4246.9

$$Var\,(x) = {\sigma_x}^2 = \frac{2539.1}{11} = 230.8$$

$$\sigma_x = 15.2$$

$$Var\,(y) = \sigma_y{}^2 = \frac{4246.9}{11} = 386.1$$

$$\sigma_y = 19.6$$

Given all this information,

$$r_{xy} = \frac{Cov\,(x,y)}{\sigma_x \sigma_y}$$

$$= \frac{-31.87}{15.2 \times 19.6}$$

$$= -0.11$$

It can be seen that the covariance calculated from the data seemed high at -31.87. The value of r shows that it is very low.

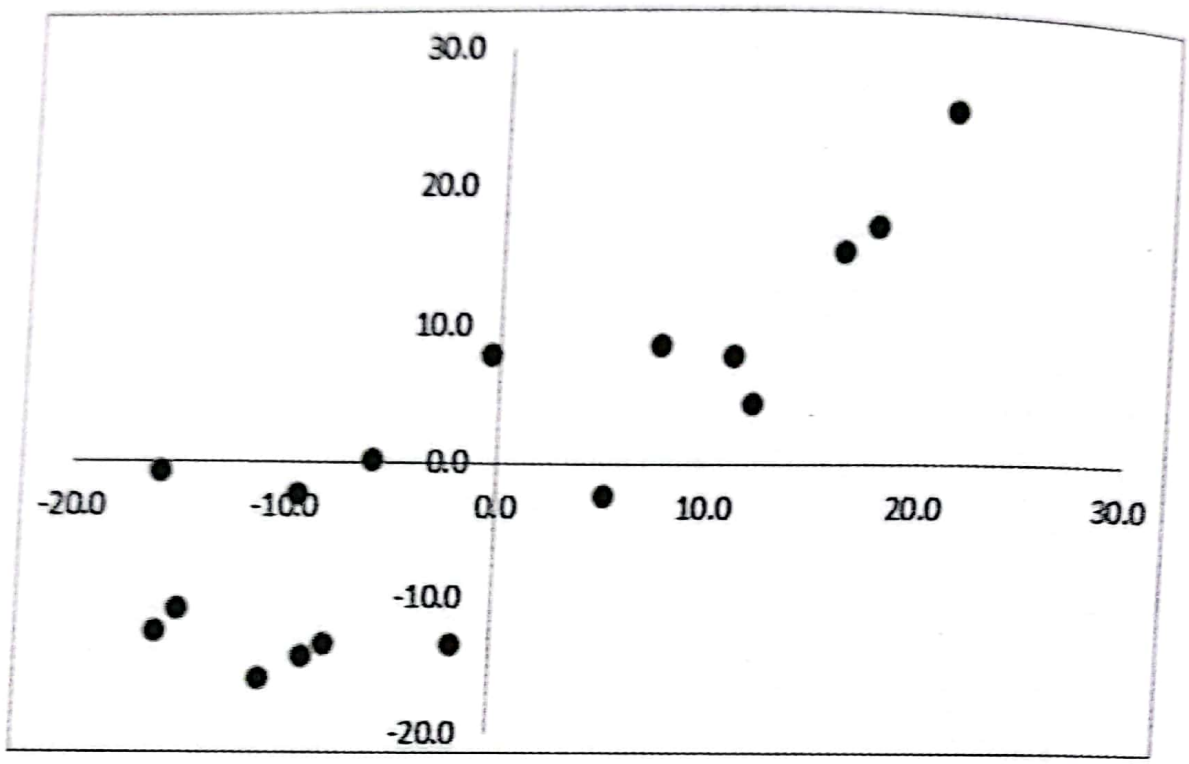
Example 7.3

An Agriculture Extension Officer is interested in knowing whether farmers who grew more maize also tended to grow more or less cotton. A total of 18 peasant farmers were surveyed to get data on the amount of maize and cotton grown. The data is provided in the table below in tonnes.

	Maize	Cotton
1	6	8.5
2	20	7.65
3	6	20.4
4	7	10.2
5	11	5.1
6	14	7.65
7	13	6.8
8	16	21.25
10	12.5	18.7
11	27	18.7
12	29.5	29.75
13	34	25.5
14	33	28.9
15	21.5	28.9
16	38	36.55
17	39.5	38.25
18	43	46.75

Use the coefficient of correlation to determine whether there is a relationship between the quantities of maize and cotton.

Before going into the calculations, it is always helpful to know the shape of the data by having a scatter plot. This is given in the figure below.



In the above scatter plot, there are more dots in the first and third quadrants (positive products) than are in the second and forth quadrants (negative products). We expect, on this basis, the coefficient of correlation to be positive.

Based on the formula

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}}$$

we generate the table below.

s/n	x	y	$x - x$	$y - \bar{y}$	$(x - \bar{x})^2$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$
1	6	8.5	-15.8	-12.7	250.4	160.0	200.2
2	20	7.65	-1.8	-13.5	3.3	182.3	24.6
3	6	20.4	-15.8	-0.8	250.4	0.6	11.9
4	7	10.2	-14.8	-11.0	219.7	119.9	162.3
5	11	5.1	-10.8	-16.1	117.1	257.6	173.7
6	14	7.65	-7.8	-13.5	61.2	182.3	105.6
7	13	6.8	-8.8	-14.4	77.9	205.9	126.6
8	16	21.25	-5.8	0.1	33.9	0.0	-0.6
10	12.5	18.7	-9.3	-2.5	86.9	6.0	22.8
11	27	18.7	5.2	-2.5	26.8	6.0	-12.7
12	29.5	29.75	7.7	8.6	58.9	74.0	66.0
13	34	25.5	12.2	4.4	148.3	18.9	53.0
14	33	28.9	11.2	7.8	124.9	60.1	86.6
15	21.5	28.9	-0.3	7.8	0.1	60.1	-2.5
16	38	36.55	16.2	15.4	261.7	237.2	249.1
17	39.5	38.25	17.7	17.1	312.5	292.4	302.3
18	43	46.75	21.2	25.6	448.4	655.4	542.1
Total	371.0	359.6	0.0	0.0	2482.5	2518.5	2111.1

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}}$$

$$= \frac{2111.1}{\sqrt{2482.5 \times 2518.5}}$$

$$= \frac{2111.1}{2500.4}$$

$$= 0.84$$

As predicted on the basis of the scatter plot, the coefficient of correlation is positive. The absolute value is also high. There is a strong correlation between the quantities of maize and cotton

produced. Farmers that grow a lot of maize also tend to grow a lot of cotton. Though theory may suggest otherwise on the basis that maize and cotton are substitutes in production, it is not strange to find a positive correlation between the two outputs. The alternative view is that output depends on the overall scale of each farmer. Big farmers will produce more of both crops as smaller farmers also produce less of everything.

The Pearson's coefficient of correlation is based on all the values in the data. If one value changes, it also changes. It is appealing because it is based on all the data. The major weakness is that it is blind to nonlinear correlations. That is, a zero correlation does not necessary mean the variables are not correlated. It simply means there is no linear correlation. If a best fit line is fitted, it would be horizontal, indicating that on average, the vertical variable does not change with changes in the horizontal variable.

In addition, some relationships tend to be lagged. For instance, the effect of a product price on its output in the agriculture sector will have a one or two year lag. Output values for this year are not associated with this year's price. Instead, it is the previous price that matters. This will pose a serious problem when the lag lengths are not known. When the lengths are known, appropriate adjustments in pairing the values eliminates this problem.

7.3.2 Spearman's Rank method

The measure of correlation as suggested by Pearson is no doubt complex. It involves squaring of variables in deviation form and summations. The alternative and much simpler way to get the measure of correlation is to use the ranking of the magnitudes of the variables. This is based on the fact that positive correlation implies that larger values in one variable are associated with larger values in the other. In the same way, smaller values in one variable are associated or paired with smaller values in the other variable. Negative correlation on the other hand will imply that larger values from one variable are associated with smaller values and vice versa.

By dealing with ranks only, we are able to derive a measure of correlation. This is known as Spearman's rank order correlation. It is named after its pioneer, an English Psychologist Charles Edward Spearman (1863 – 1945).

Without disturbing the pairings, each value is replaced by a numerical value representing the rank in the variable. The largest value in the X is assigned one, the second largest assigned two and so on until the smallest which will be in the n^{th} position. The same applies for variable Y. This produces pairs of ranks as opposed to the original values. In cases where duplicate values exist, there will be ties in position. The rank assigned is equal to the average of the positions they would occupy. For instance, if the second position in ascending order has two values, we know that the next value will be fourth in rank. This is because the two equal values occupy the second and third positions. A common rank of $\frac{2+3}{2} = 2.5$ is assigned to the two equal values. If they are three, the average of the three positions they would occupy is assigned. Ranks can also be assigned from the lowest value to the highest.

Obviously, the first fascination for this method is that we deal with smaller numbers. The numbers deal with the position in the sample, with the largest value n, the sample size. In addition, the values are always positive. This makes the calculation much simpler even for variables involving large values.

For each pair, the difference d, is calculated. We know that in perfectly correlated variables, the highest value in one variable will be paired with the highest in the other. The ranks will be the same in both variables. In this case, all the differences will be zero. Similarly, when there is perfect negative correlation, the lowest in one variable are paired with the highest in the other. Though not all the differences will be zero, their sum will be equal to zero. The negative differences simply cancel the positive ones. This is very simple to understand. Let x_i be the rank of the i^{th} X and y_i be the rank of the i^{th} Y. The difference is defined as $d_i = x_i - y_i$. Then

$$\sum d_i = \sum (x_i - y_i)$$
$$= \sum x_i - \sum y_i$$

$$= 0$$

This is zero because X and Y have the same values, the counting numbers $1, 2, \ldots n$. To circumvent this limitation, Spearman proposes using the squares of the difference d_i^2. When the squared values are added, the sum will be zero in perfect positive correlation because the paired or subtracting ranks will be the same. This will give a list of zero differences and hence the zero sum.

The sum of square difference is highest in perfect negative correlation. There is an inverse link between the summation of the squares of the differences and the correlation. With some scaling down through division, summation of the squares subtracted from one provides a descent measure of correlation, the Spearman's rank order correlation.

The Spearman's rank correlation coefficient is given by

$$r_{rank} = 1 - \left[\frac{6\sum d_i^2}{(n-1)n(n+1)}\right]$$

This is reduced by recognising the expanded difference of two squares.

$$r_{rank} = 1 - \left[\frac{6\sum d_i^2}{n(n^2-1)}\right]$$

Spearman's rank coefficient of correlation has the same range as the Pearson's product moment coefficient of correlation in the preceding subsection. The rank correlation can also be interpreted on the basis of Table 7.1.

When some ranks repeat because of duplicate values, corresponding to every such repeated rank which repeats m_j times, add $\left(\frac{m_j^3 - m_i}{12}\right)$ to $\sum d_i^2$ in the formula

$$r_{rank} = 1 - \left[\frac{6\left[\sum d_i^2 + \frac{1}{12}\sum(m_j^3 - m_j)\right]}{n(n^2-1)}\right]$$

Where m_j is the number of times the jth tied rank repeats.

Example 7.4

The Zambia Wildlife Authority (ZAWA) operates national parks in the country. Information is collected from two of the many national parks in order to check if the number of visits is in anyway correlated.

Park A	Park B
65	87
46	80
82	49
54	72
59	45
72	43
74	23
46	38
56	76
55	48

The easiest way to workout Spearman's rank order correlation is to create additional columns in which the observations are ordered. The original columns however need to be retained because the ranks will have to be assigned to the original pairings. We can call these ordered columns.

Park A	Park B	A rank	B rank	d	d^2
65	87	7	10	-3	9
46	80	1.5	9	-7.5	56.25
82	49	10	6	4	16
54	72	3	7	-4	16
59	45	6	4	2	4
72	43	8	3	5	25
74	23	9	1	8	64
46	38	1.5	2	-0.5	0.25
56	76	5	8	-3	9
55	48	4	5	-1	1
Sum					200.5

In the data, one position or rank is tied by two values ($m = 2$), therefore the adjusted formula is used.

$$r_{rank} = 1 - \left[\frac{6\left[\sum d_i^2 + \frac{1}{12}\sum(m_j^3 - m_j)\right]}{n(n^2-1)}\right]$$

$$= 1 - \left[\frac{6\left[200.5 + \frac{(2^3-2)}{12}\right]}{10(100-1)}\right]$$

$$= 1 - \left[\frac{6[200.5 + 0.5]}{10(100-1)}\right]$$

$$= 1 - 1.218$$

$$-0.218$$

In spearman's rank correlation, it is the rank and not the actual values of observations. As a consequence, even when observations change, the correlation value remains the same as long as the relative standing of observations in each variable remain unaffected.

Its major drawback lies in its inability to detect small changes in the values. Even when it records a perfect correlation, it is perfect correlation in ranks and may only be near perfect correlation in the actual values.

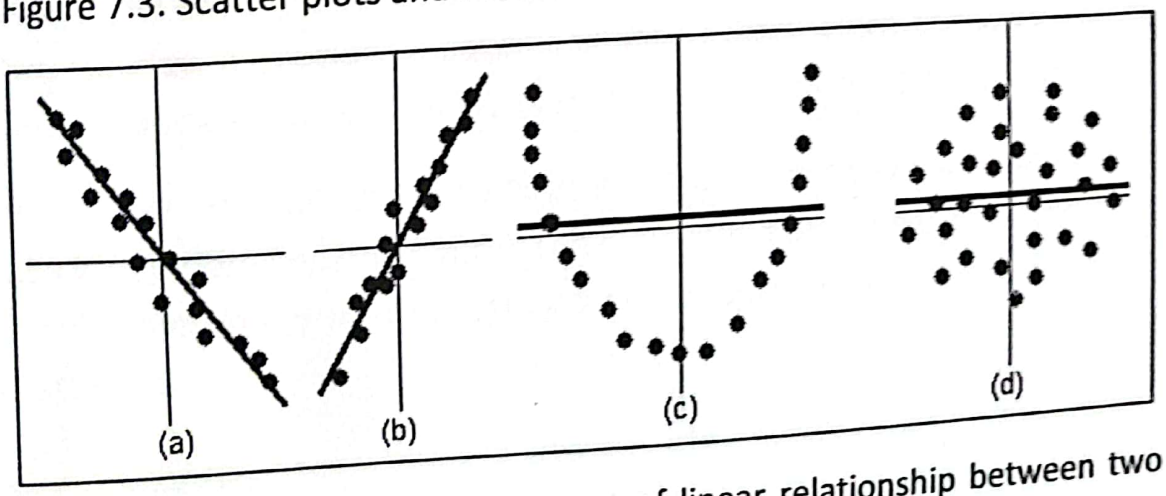
7.4 Some Weaknesses of Correlation Measures

There are two major weaknesses or limitations to the use of measure of correlations in general. The first is based on the fact that the measures of correlations are only indicative of direction and not magnitude. Correlations measures the extent to which observations fit in a line. To what extent do observations form a single line? The steepness or flatness of this line is however not an issue. This weakness is dealt with in the next Chapter on Regression.

The second weakness emanates from its assumption of linear correlation. The measures of correlation consider all relationships to be linear. It ignores others forms of relationship between two variable. For instance, correlations will show a zero correlation even when it is clear from say the scatter plot that a relationship exists between the two variables.

We explain the two weakness using the figure below. The figure has four scatter plots with a best fit line.

Figure 7.3. Scatter plots and Measures of Correlation.



Subfigures (a) and (b) are clear cases of linear relationship between two variables. These are captured correctly by the measure of correlations. The reason is simply that the relationship is linear. In particular, (a) would have

a negative correlation coefficient while (b) is positive. The best fit line slopes downward in the former and upwards in the latter.

In figure (c), we see a clear case of a curvature scatter plot. All dots seem to fit quite well on a best fit 'curve'. This actually looks like a quadratic function. However, because the measure of correlation forces a linear function, we end with a horizontal best fit line. In this case, the coefficient of correlation will be zero. The zero means that the two variables are not correlated. But clearly, this at variance with the picture which shows good fit. This is a serious anomaly with measures of correlation. They are blind to nonlinear relationships. Part (d) is not problematic. The dots are sprinkled around the origin without any sign of relationship. A correlation coefficient of zero as indicated by the best fit line is acceptable and normal.

CHAPTER 8

8 MEASURES OF DIRECTION OF RELATIONSHIPS – REGRESSION ANALYSIS

In the preceding chapter, we looked at the measures of correlation, measuring how variables are changing together or how a change in one variable is related to another. We are able to tell on this basis, for instance, that when one variable goes up, the other should be going up as well (positive correlation) or going down (negative correlation). However, the chapter was not able to predict the magnitude of change in one variable (y) as another (x) changes, say by a unity. As such, correlation cannot be used in forecasting or predicting one variable when the other is known. For instance, even if age and weight are variables that are known to be correlated, this knowledge is not sufficient to predict the weight at a given age.

Regression is a method by which the value of one variate can be predicted on the basis of the known value of the other. Regression allows a clear and concise statement of the relationship in the form of an equation linking the two variates. It provides a specific equation from which the value of one variate can be estimated on the basis of the value of the other variate.

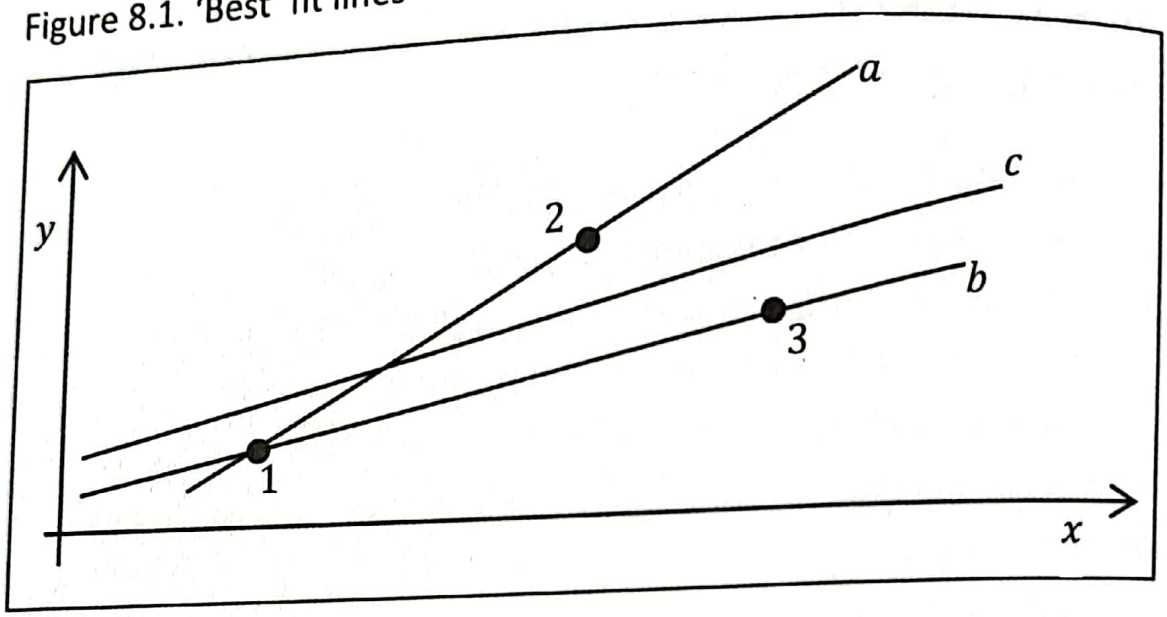
When the points on a scatter plot form a line or closely fit a line, we say the two variates are linearly related. The corresponding regression is known as *linear regression*. It involves fitting a line to describe all the points on the scatter plot.

8.1 Best fit

Consider a scatter plot with three points only. If these fall on a straight line, well, we would have no problem fitting a line describing all the three points. A single straight line would touch the three points. When the three points

do not fall in a line, then fitting a single line becomes impossible. They can only be joined by multiple lines. But we are looking for a single line describing the three points. Consider the figure below.

Figure 8.1. 'Best' fit lines



Since a single line cannot touch the three dots, how then can it be used to describe the three points? What is required in such as case is a line that *best* fits the points. Such a line need not touch all the points. In principle, many such lines can be drawn to fit the three points. Three such lines are shown in the preceding figure.

Line (a) touches points 1 and 2, line (b) touches points 1 and 3 and line (c) does not does any of the points. At first sight, line c may appear to be the worst fitting line of the three points and yet, when well placed, this line may in fact be closer to all the points than the other two lines that touch two of the points. Since we seek a line that best fits all the points, there could be more appeal for a line that is closest to all the points than one touching a few and completely ignoring some others. For this line, the sum of distances from each point is minimised, it will be closest to all the points.

Consider the Cartesian coordinates of the three points. These are (x_i, y_i), $i = 1,2,3$. The y_i is the actual y value corresponding with $x = x_i$. In the fitted line however, the y value corresponding to $x = x_i$ is $\hat{y}_i$. It is an estimated, as opposed to the true value of y. It could be equal to the true value but it need not be the case. On the basis of this information, we are

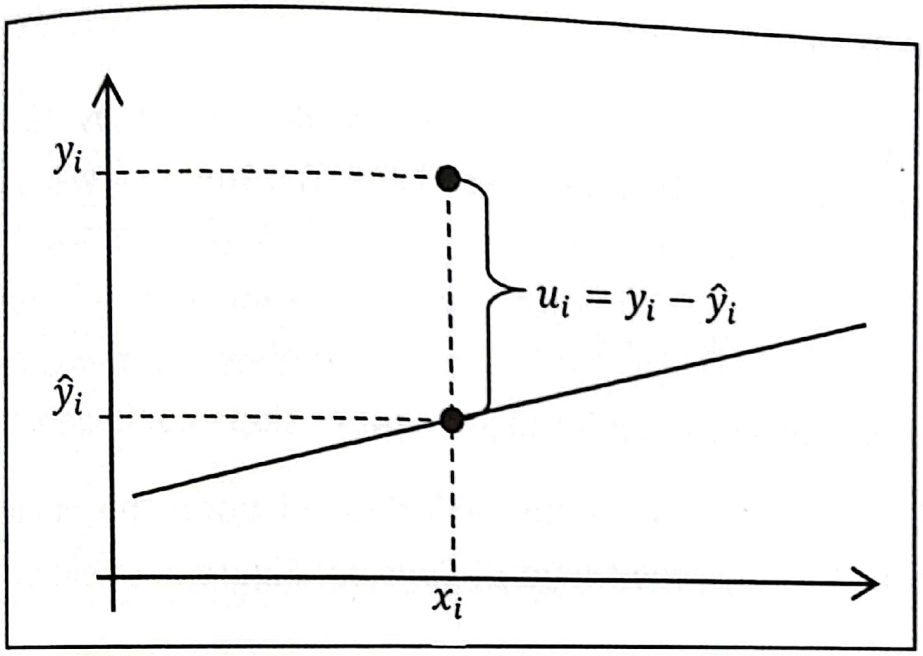
able to establish the deviation of each point from the best fit line. It is the true value less the estimated.

Let

$$u_i \doteq y_i - \hat{y}_i$$

Where u_i is the *residual of estimation.*

Consider the figure below for the illustration.



At the value of $x = x_i$, the actual value of y is y_i while the value estimated on the basis of the regression line is $\hat{y}_i$. The difference between the two values is the residual, u_i. The logic of the best fitting line is to identify one that minimises e_i for every point in the data, namely, $\sum u_i = \sum(y_i - \hat{y}_i)$

However, it will be noted that for any such line, some actual points will fall above the line and some others below the line. For points that fall above the line, the residual will be positive and for those that fall below the line, it will be negative. It is therefore possible that the sum of the residuals could be zero or close to it since the positive and negative residuals may offset one another giving a misleading impression that there is little or no residual.

There are two way of addressing this issue. One is to use the sum of the absolute values of the differences ($\sum|y_i - \hat{y}_i|$) and the other is to sum the squares of the differences ($\sum(y_i - \hat{y}_i)^2$). The line that yields the lowest sum

is the best fitting line. In other words, the best fitting line is the one that meets any one of the following equations:

$$\sum |u_i| = \sum |y_i - \hat{y}_i|$$

$$\sum u_i^2 = \sum (y_i - \hat{y}_i)^2$$

8.2 Regression line

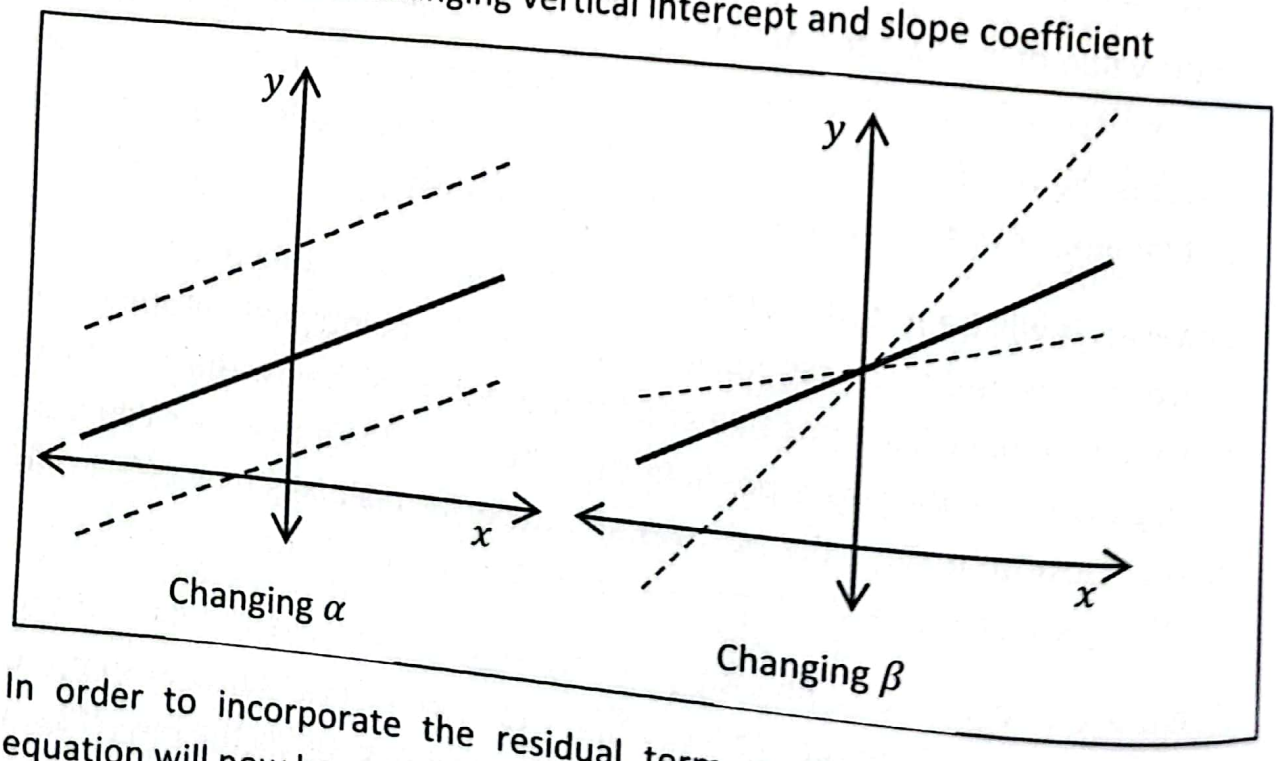
As is generally known, a linear function or a line is defined by two parameters: the slope (or gradient) and the intercept. The function will be of the form

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

Where α is the vertical intercept and β is the gradient or slope coefficient.

To remind oneself, the position of the line will depend upon the value vertical intercept α and the slope coefficient β. Consider Figure 8.2 below.

Figure 8.2. Effects of changing vertical intercept and slope coefficient



In order to incorporate the residual term u, the above mathematical equation will now be written as

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

where e is the residual term as defined above. Such a line is called the line of regression of y on x. For such a line, the residual now can be restated as:

$$u = y - \alpha - \beta x$$

We shall look at the method of obtaining the best fit line by minimising the sum of the squared residuals. This is tantamount to finding the values of $\alpha$and β that will lead to

$$\min_{\alpha,\beta} \sum u^2$$

This is a problem in optimisation. This method is known as *Ordinary Least Squares* (OLS) estimation because it is based on minimising the sum of squared residuals. It can be shown that the solution for this optimisation problem will yield the following formulae for $\alpha$and β

$$\hat{\beta} = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}$$

It can be shown that it is equivalent to

$$\hat{\beta} = \frac{n\sum xy - n^2\bar{x}\bar{y}}{n\sum x^2 - n^2\bar{x}^2} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{\sum (y - \bar{y})(x - \bar{x})}{\sum (x - \bar{x})^2} = \frac{Cov\,(x,y)}{Var\,(x)}$$

$\hat{\alpha}$ will be obtained as:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

An alternative method, known as the *maximum likelihood* estimators are based on maximising the likelihood of obtaining the sample at hand. Under the foregoing assumptions, the ML method produces the same estimators.

Example 8.1

An Economics student gathered the following data in order to estimate the demand function for a particular commodity.

Price	50	76	87	86	98	91	49	93	85	75	66	61	54	59	85

Quantity	78	66	30	49	44	26	100	41	52	64	71	87	70	88	45

The demand function is an expression of quantity demanded as a function of the price. Using q for quantity and P for price, we express the demand function as

$$q = \alpha + \beta P + e, \qquad \beta < 0$$

We now seek to find the parameters $\alpha$and β. We first find the latter using

$$\hat{\beta} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

s/n	P	q	P^2	Pq
1	50	78	2,500	3,900
2	76	66	5,776	5,016
3	87	30	7,569	2,610
4	86	49	7,396	4,214
5	98	44	9,604	4,312
6	91	26	8,281	2,366
7	49	100	2,401	4,900
8	93	41	8,649	3,813
9	85	52	7,225	4,420
10	75	64	5,625	4,800
11	66	71	4,356	4,686
12	61	87	3,721	5,307
13	54	70	2,916	3,780
14	59	88	3,481	5,192
15	85	45	7,225	3,825
Sum	**1,115**	**911**	**86,725**	**63,141**

$$\hat{\beta} = \frac{(15)(63141) - (1115)(911)}{15(86725) - (1115)^2}$$

$$= \frac{947115 - 1015765}{1300875 - 1243225}$$

$$= \frac{-68650}{57650}$$

$$\hat{\beta} = -1.19$$

For every unit increase in price, the quantity demanded drops by about 1.19 units. The quantity drops because the coefficient β is negative. This is the normal demand functions that was expected. In order to get the constant α, we make use of the formula

$$\alpha = \frac{\sum y}{n} - \beta \frac{\sum x}{n}$$

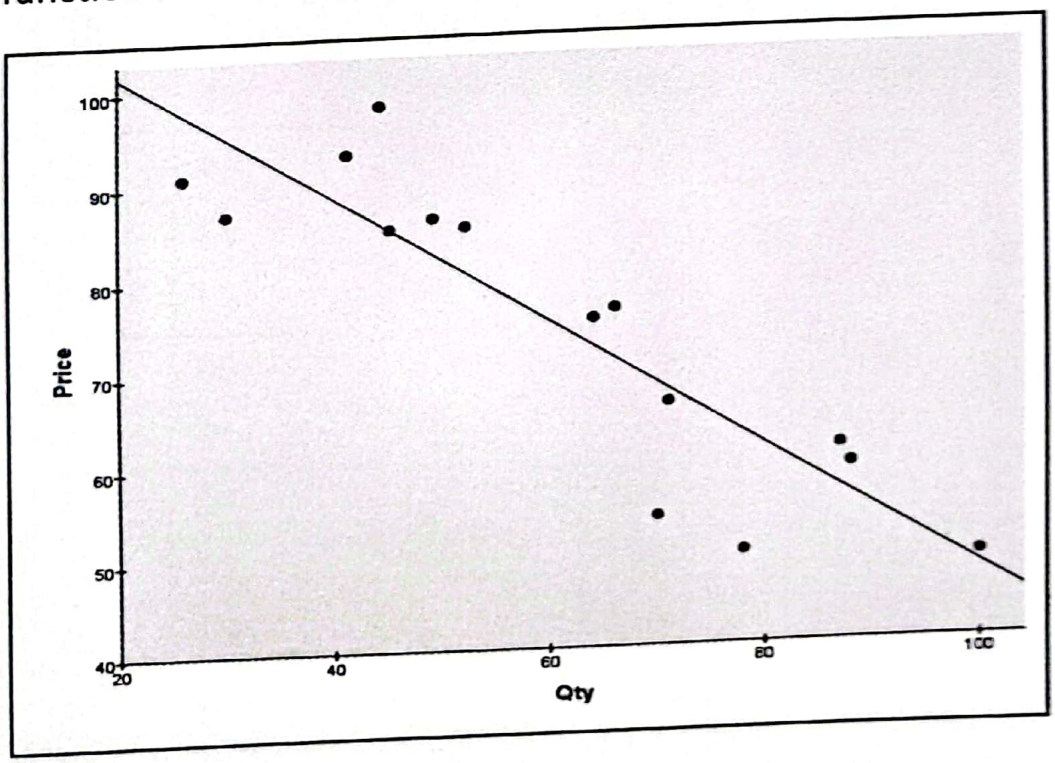
$$\alpha = \frac{911}{15} - (-1.19)\frac{1115}{15}$$

$$\alpha = 149.19$$

With the two parameters established, the demand function is

$$q = 149.19 - 1.19P$$

On the basis of this function, the scatter plot with the demand function fitted is given in the figure below.



Example 8.2

Mubita, afourth year Economics students gathers data on the daily price of tomato and the quantity (number of crates) supplied each day for the whole month of March. The aim is to estimate the supply function of the commodity. The students intends to use the formula

$$\beta = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

The table is generated below.

Day	Price$P$	Qty$Q$	P^2	$P \times Q$
1	75	166	5625	12450
2	70	174	4900	12180
3	65	176	4225	11440
4	70	152	4900	10640
5	55	151	3025	8305
6	85	130	7225	11050
7	50	182	2500	9100
8	85	85	7225	7225
9	50	191	2500	9550
10	75	93	5625	6975
11	85	184	7225	15640
12	65	204	4225	13260
13	90	107	8100	9630
14	60	216	3600	12960
15	55	123	3025	6765
16	75	128	5625	9600
17	110	176	12100	19360
18	85	244	7225	20740
19	65	162	4225	10530
20	55	108	3025	5940
21	80	136	6400	10880
22	55	162	3025	8910
23	80	125	6400	10000
24	50	170	2500	8500
25	60	88	3600	5280
26	105	120	11025	12600
27	110	235	12100	25850
28	105	264	11025	27720
29	55	241	3025	13255
30	105	124	11025	13020

31	55	237	3025	13035
sums	**2,285**	**5,054**	**179,275**	**372,390**

On the basis of this output from Excel, the student proceeded to calculate the regression coefficient $\hat{\beta}$ as.

$$\hat{\beta} = \frac{31(372390) - (2285)(5054)}{31(179275) - (2285)^2}$$
$$= \frac{11544090 - 11548390}{5557525 - 5221225}$$
$$= \frac{-4300}{336300}$$
$$= -0.013$$

The results are stunning for the students. First the coefficient is negative when the student is convinced it is supposed to be positive. A normal supply function is upward sloped. Second, the absolute value of the coefficient looks too small. It is almost zero. He comes to a conclusion that there is no relationship between what is supplied and the price. Quantity supplied is independent of price.

Mweemba, another student insists that the relationship exist, only that it is lagged. "You don't think the farmers bringing tomato today are responding to today's price, do you?" asks Mweemba. "Ah, I see. It is the previous day's price that matters," agrees Mubita. The student re-pairs the observations by pairing quantity to previous day's price and proceeds to recalculate the coefficient. In doing this, one observation is lost since he has to start with the second day's quantity which is paired with the first day's price. There is no price to be paired with the first day's quantity since no information is available for the preceding day.

Price x		Qty y		price squa	price qty
Date	P	Date	Q	P^2	$P \times Q$
1	75	2	174	5625	13050
2	70	3	176	4900	12320
3	65	4	152	4225	9880
4	70	5	151	4900	10570
5	55	6	130	3025	7150
6	85	7	182	7225	15470
7	50	8	85	2500	4250
8	85	9	191	7225	16235
9	50	10	93	2500	4650
10	75	11	184	5625	13800
11	85	12	204	7225	17340
12	65	13	107	4225	6955
13	90	14	216	8100	19440
14	60	15	123	3600	7380
15	55	16	128	3025	7040
16	75	17	176	5625	13200
17	110	18	244	12100	26840
18	85	19	162	7225	13770
19	65	20	108	4225	7020
20	55	21	136	3025	7480
21	80	22	162	6400	12960
22	55	23	125	3025	6875
23	80	24	170	6400	13600
24	50	25	88	2500	4400
25	60	26	120	3600	7200
26	105	27	235	11025	24675
27	110	28	264	12100	29040
28	105	29	241	11025	25305

29	55	30	124	3025	6820
30	105	31	237	11025	24885
sums	**2230**	**495**	**4888**	**176250**	**389600**

$$\beta = \frac{30(389600) - (2230)(4888)}{30(176250) - (2230)^2}$$

$$= \frac{11688000 - 10900240}{5287500 - 4972900}$$

$$= \frac{787760}{314600}$$

$$= 2.5$$

Mubita is now happy with the result and can now proceed to calculate the constant α.

$$\alpha = \frac{\sum y}{n} - \beta\frac{\sum x}{n}$$

$$\alpha = \frac{4888}{30} - (2.5)\frac{2230}{30} = -22.9$$

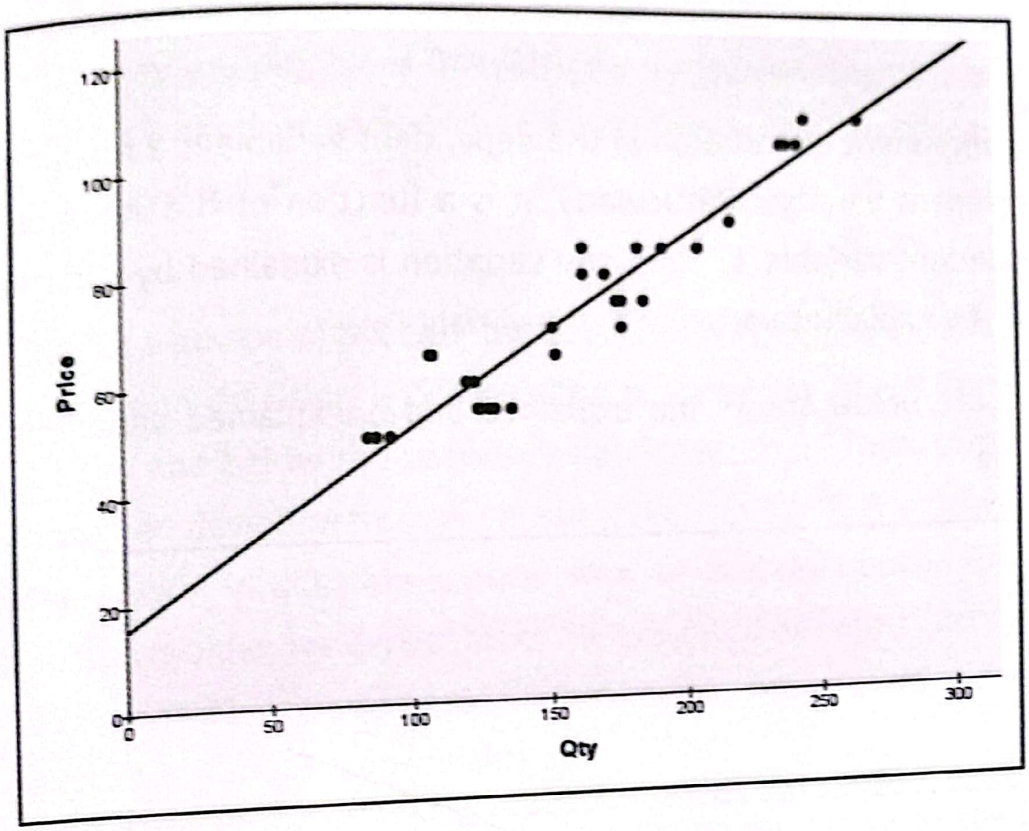
The supply function is

$$Q_t^S = 2.5P_{t-1} - 22.9$$

Because the quantity supplied is influenced by a price level in another period, it is important to show this by indicating the time periods of the two variable. In the above case, the quantity supplied in the current period, time t, is a function of the price in the immediate past period, time $t - 1$. When both the quantity and the price are of the same time period, as is often the case with the demand function, the emphasis on time can still be included but is not necessary.

The message that comes out from this example is that one has to be very careful with the specification of the regression equation. A wrong specification can lead to very misleading results.

The scatter plot together with a best fit line, an estimation of the supply function is given in the figure below.



8.3 Explained and Unexplained variation

With a regression equation, the deviation of the individual observations from the central value is no longer treated as a random variation. Some of it is explained by the explanatory variable x. The total deviation of each observation from the central value is broken down as follows.

$$y_i - \bar{y} = (y_i - \hat{y}) + (\hat{y} - \bar{y})$$

The first part shows the deviation of each observation from the estimated level. The second is the deviation of the estimated level from the average. We express the two below.

$$y_i - \hat{y} = \alpha + \beta x_i + e_i - (\alpha + \beta x_i)$$

$$= u_i$$

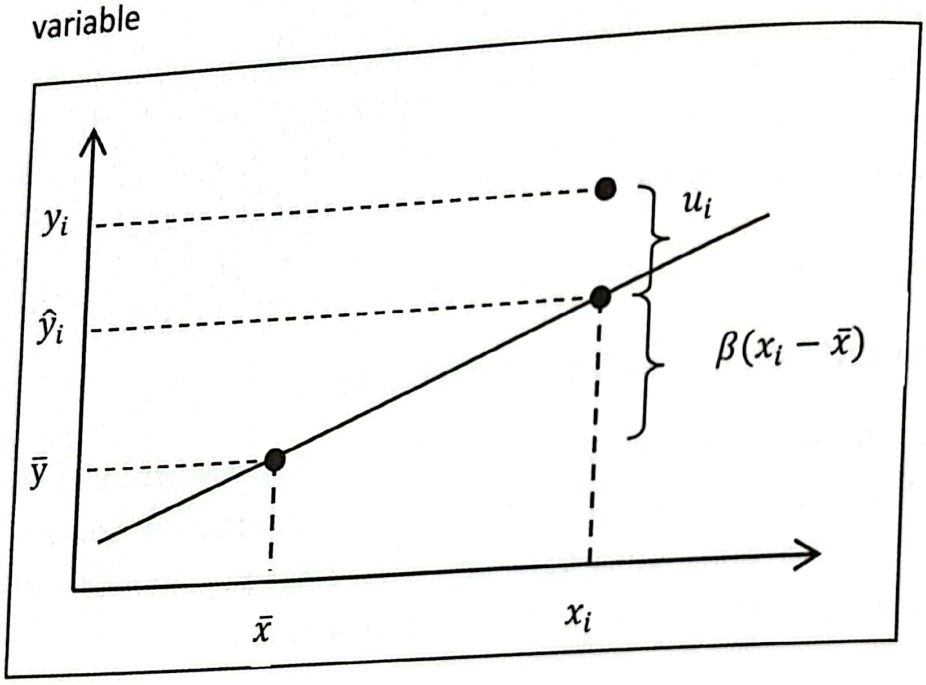
The expression shows that the deviations of observations from the estimated level is a random term, the residual. This is the unexplained variation in the dependent variable.

The latter is

$$\hat{y} - \bar{y} = \alpha + \beta x_i - (\alpha + \beta \bar{x})$$
$$= \beta(x_i - \bar{x})$$

This component of variation in the dependent variable is a function of the independent variable. Particularly, it is a function of the variation in the independent variable x. Thus, the variation is explained by variation in x. This is the explained variation in the variable y.

The figure below shows the explained and unexplained variations in the variable



The total deviation is

$$y_i - \bar{y} = \beta(x_i - \bar{x}) + u_i$$

Example 8.3

Take for instance a regression of body weight (y) and age (x) with averages $\bar{y} = 39$ and $\bar{x} = 16$ respectively. Assume that after a regression, the following equation is estimated.

$$\hat{y}_i = 3.8 + 2.2x_i$$

This is an estimated line which shows the estimated weights for each age. For a person of average age, we expect that their weight should

also be around the average. But for individual older than the average, it is only natural to expect them to be heavier than an average person. Their weighing above the average is explained by the fact that they are old, older than an average person. Suppose that one person is aged 21 and weighs $54kgs$. This weight deviates from the group average by

$$54 - 39 = 15$$

But for a person of this age, the expected weight based on the above regression equation is $50kgs$. This is above the group average by $11kgs$ and below the observed weight by $4kgs$. Thus, of the 15kg observed deviation from the mean, 11kg is attributed to the fact that the person in question is older than an average person. Since the person is older, we expect his or her weight to be higher, precisely by $11kgs$ at 21 years of age. This is what has been referred to as the explained variation (in weight).

The remaining 4kgs is beyond what can be attributed to age alone. Obviously, we know there are many factors that go into determining body weight. These include genetics, standard of living or diet and many others. But these are not in the model, they are not part of the regression. This part of the deviation is referred to as the residual. It is how much the expected misses the actual value. It is the deviation that cannot be explained by the regressand. Thus the residual accounts for all deviations that are not explained by explanatory variables in the model.

Taking squares and summing on both sides, we have[7]

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \beta^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n} {u_i}^2$$

Which is equivalent to

[7] The summation of the product of the regressand and the error term will be zero given the independence of the two, $\sum \beta(x_i - \bar{x})u_i = \beta \sum x_i u_i = 0$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y} - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y})^2$$

On the left is the sum of squared total deviations of the dependant variable from the overall average and is referred to as the *total sum of squares* (TSS). It represents the overall variation in the dependent variable. The right hand side has two sums. The first is the sum of squared deviation of the estimated values from the average. Since estimated values are based on the explanatory variable, this variation is said to be explained by the explanatory variable X and is referred to as the *explained sum of squares* (ESS).

The last term is the sum of squared residuals. This is the deviation of actual values of the dependent variable from the estimated values. The sum of the squared residuals is referred to as the *residual sum of squares* (RSS). This implies that;

$$TSS = ESS + RSS$$

The disaggregated measures of variation are important in the determination of how good the estimated results are, the coefficient of determination.

8.4 Coefficient of Determination

In the preceding subsection, we disaggregated the total deviations in the dependent variable into the explained and residual and obtained the equation

$$TSS = ESS + RSS$$

Dividing on both sides by TSS, the equation transforms to

$$1 = \frac{ESS}{TSS} + \frac{RSS}{TSS}$$

Since both the ESS and RSS are not greater than TSS, the ratios on the right will individually lie between zero and one. They are proportions. More precisely, the $\frac{ESS}{TSS}$ is the proportion of total variation in the dependent variable that is explained by the explanatory variable. This is known as the coefficient of determination and denoted by r^2 in a single explanatory

variable and R^2 for multiple explanatory variable models. Conceptually, the two are the same and the latter is often used even for single explanatory variable models.

$$R^2 = \frac{ESS}{TSS}$$
$$= \beta^2 \frac{\sum(x_i - \bar{x})^2}{\sum(y_i - \bar{y})^2}$$

Alternatively, the R^2 can be expressed as

$$R^2 = 1 - \frac{RSS}{TSS}$$
$$= 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$
$$= 1 - \frac{\sum u^2}{\sum(y_i - \bar{y})^2}$$

A high R^2 (closer to one) indicates a strong model in which most of the variation in y and explained by x. A lower coefficient means the model is not explaining much of the variation or the chosen explanatory variables are only explaining a small proportion of variation in the dependent variable. Though the coefficient of determination is important in assessing the explanatory power of the model, it need not be the basis for rejecting or accepting models. The researcher must also examine the indications of other statistics such as the coefficient t-statistics and the F-statistics.

The major weakness of the coefficient of determination is that it is a monotonic function of the number of regressors. As more regressors are added, the R^2 also increases towards one. This has the potential to favour models with many regressors (abet with minimal marginal contribution) over models with few but important regressors. This misnomer is corrected by looking at the Adjusted Coefficient of Determination denoted by $\bar{R}^2$. This is the coefficient of determination that is adjusted for the number of explanatory variables.

In the adjusted R^2, the RSS and TSS are both divided by their respected degrees of freedom, that is, $(n - k)$ and $(n - 1)$ respectively. Thus,

$$\bar{R}^2 = 1 - \frac{\sum u^2/_{n-k}}{\sum (y_i - \bar{y})^2/_{n-1}}$$

$$= 1 - \frac{(n-1)}{(n-k)} \frac{\sum u^2}{\sum (y_i - \bar{y})^2}$$

In the fraction on the right hand side of the equation, the numerator is the unbiased estimator of the variance of the residuals. The denominator is also the sample variance of the dependent variable Y. Therefore,

$$\bar{R}^2 = 1 - \frac{\hat{\sigma}^2}{s_y^2}$$

At this stage, it is possible to show the relationship between the adjusted and unadjusted coefficients of determination. It was demonstrated earlier that

$$R^2 = 1 - \frac{\sum u^2}{\sum (y_i - \bar{y})^2} \Leftrightarrow 1 - R^2 = \frac{\sum u^2}{\sum (y_i - \bar{y})^2}$$

Substituting this formulation into the equation for adjusted R^2, we show the relationship between the adjusted and unadjusted R^2.

$$\bar{R}^2 = 1 - \frac{(n-1)}{(n-k)} \left[\frac{\sum u^2}{\sum (y_i - \bar{y})^2} \right]$$

$$= 1 - \frac{(n-1)}{(n-k)} (1 - R^2)$$

If $k = 1$, the adjusted and unadjusted coefficients of determination are equal. For all other values of k, that is $k = 2,3,4, \ldots$, the adjusted coefficient of determinationis strictly less than the unadjusted coefficient. That is,

$$\bar{R}^2 < R^2, \forall\, k > 1$$

Second, while the unadjusted coefficient is non-negative, the adjusted coefficient of determination can actually take on negative values which are devoid of meaning. For this reason, the adjusted coefficient is truncated at zero.

8.5 Regression and extrapolation

Regression establishes a relationship between two variables, expressed in an equation form. The equation shows the explained variable (y) as a function of the explanatory variable (x). We have explained already that the independent variable can also be time, in which case it becomes a time series. This is an equation showing how a variable behaves with time. This equation is formed based on the regression of available data.

Extrapolation then involving estimating beyond the observation range using an established relationship with another variable, in this case time. For instance, GDP for Zambia is known for the past periods, say past 14 years. Using t to represent time period in year and y for GDP, an equation linking the two can be estimated using data for the past period. Assume $t = 1$ for 2001, $t = 2$ for 2002 up to $t = 14$ for 2014. The equation is as follows.

$$\hat{y} = \alpha + \beta t$$

The equation can be used to estimate the level of GDP for various time periods t. If for whatever reason we wish to estimate the level of GDP for any year between 2001 and 2014 inclusive, the equation becomes handy. This is known as *interpolation*. It is estimation within the observed period of 2001 to 2014.

In most cases, however, these equations are used to make predictions into the future. The level of GDP is probably known for all the years in question and leaves no justification for interpolation. Instead, having completed 2014, planners or policy makers are interested in knowing beforehand how the next year will fair. What will be GDP for the coming year? This is an estimation beyond or outside the observed period. It is referred to as *extrapolation*. Using the above equation, the level of GDP for 2015, 2016 and so on can be estimated by using $t = 15$, $t = 16$ and so on.

8.6 Correlation analysis and regression analysis

The correlation analysis discussed in the preceding chapter shares a lot of commonality with regression. Four point are worth remembering when dealing with the two:

I. The same type of relationship holds for both regression and correlation analysis. The correlation coefficient r takes the same sign as the regression coefficient β.

II. The correlation coefficient r is a measure of the closeness of fit of the regression line. The closer the absolute value of r is to unity, the more useful the regression equation as a prediction devise.

III. A given value of r is consistent with an infinite number of regression lines. This is because regression is a directional method but correlation is not. Therefore, regression should be used when one variable is clearly dependent of the other; correlation is still useful even when neither of the two variables can be considered as a consequence of the other. That is:

   a. Regression implies causation while correlation implies association

   b. Causation implies association but association need not imply causation because

      i. It may be a result of pure chance

      ii. Two variables may be influenced by a third common factor

IV. Sometimes one cannot make out which is the cause and which is the effect. For instance, there may be a high correlation between grades and class attendance. This could mean that greater class attendance increases the amount learned and thus cause high grades. Alternatively, it could mean good grades motivate students who obtain them to attend classes more frequently.

8.7 A note on Regression

A regression coefficient is a derivative (or partial derivative) in the regression equation. So it can provide information only about how small changes in the

explanatory variable relate to changes in the dependent variable. For example, suppose we have a regression equation of income (y) on education (E). Now consider the time when the University of Zambia was the only university in Zambia. In this situation, an additional university graduate would significantly affect income. But today when many universities have sprung in the country with a big increase in the supply of university graduates, an additional graduate is unlikely to earn a great deal more. The value of the education regression coefficient may even turn from highly significant to insignificant.

CHAPTER 9

9 PROBABILITY

In the many Presidential elections that Zambia has held, opinion polls have been conducted predicting that a particular candidate would win the elections. It is not the interest of this chapter to look at the credibility of these predictions, rather how they compare with the actual outcome after the elections. We know that some pointed to the correct winners, and some cases, the predicted candidates were not the winners. The Zambia national soccer team went to 2012 edition of the African Cup of Nations without so much hope in fans. Most soccer commentators looked up to other countries scooping the tournament. In the end, it was Zambia. There are always predictions for future events.

In economics, most countries experience economic turbulences in run-up to elections. The currency depreciates along reduced inflow of investments. Investors become hesitant to bring in money. The reason is that they are not certain of events that will unfold during or after elections. Though they get assurances that elections will be peaceful, they always know of a possibility of adverse events that may affect their investments. One of the core functions of the Food Reserve Agency (FRA) in Zambia is to hold strategic foods reserves. The whole idea of reserves come in because of the notion of an 'unknown tomorrow'. Even though the harvests are consistently good, government will not turn a blind-eye to a possibility of a poor harvest in the coming season.

There are several other areas of application. For example, when the bank is giving out a loan, they expect an interest from the borrower, in addition to the principal or the amount loaned. However, there is a chance that the borrower may not commit to the agreement either due to negligence or inability. This is referred to as a default, the borrower is said to have

defaulted on the loan. In response, some banks now require that loans are insured.

Common to all these illustrations is the presence of uncertainty of events. We often know that some event might occur but we are seldom sure. It is a game of chances. Sometimes goes our way and at other times, lucky is not on our side. The study of these chances is known as probability and is the preoccupation of this chapter.

9.1 Definitions

Defining probability starts with a look at a simple coin. It has two sides, one called the Head and the other Tail. When tossed, we cannot tell with certainty which side will be up. But we be sure if the coin is fair, it will show one of the two sides. In other words, we rule out the possibility standing on its edge. Each of the possible outcomes, Head (H) or Tail (T) is known as a *sample point*. Each sample point has a certain chance or possibility of occurrence. In a simple coin case, we assume that each side will occur half of the times. When all the sample points are put together, they form the *sample space*.

When a coin is tossed once, there are two sample points: H and T. The sample space is therefore H, T. These are the possible outcomes from tossing a coin once. When tossed twice, a different picture emerges. There will be four possible outcomes: it may show Heads in both tosses (HH) and tails in both tosses (TT). Alternatively, we may get a Head in the first toss and a Tail in the second (HT) or Tail first and then Head (TH). These are the four sample points. The sample space will be HH, HT, TH, TT.

When a sample space has countable elements or sample points, it is known as a discrete sample space. The case of tossing a coin is one example of a discrete sample space because the possible outcomes are discrete and countable. In some variables however, the sample points may not be countable. This is the case with continuous variable which can assume all real numbers within certain limits. Recall that the sample space is a list of all possible outcomes and not necessarily the actual outcomes. This means the

number of sample points will not be limited to the sample size. When the number of sample points is finite, it is referred to as *finite sample space*, otherwise it is an *infinite sample space*.

Probability is the measure or quantification of the likelihood or the chance that a particular event will occur. It ranges between 0 for an impossible event and 1 for a certain event. The lower the probability, the lower the chances of the event occurring and the higher the probability, the higher the chances of the event occurring. For an event A within the samples space S, $A \subset S$.

$$0 \leq P(A) \leq 1$$

In an opinion poll, we can measure the likelihood that the predicted candidate will win an election. The bank is interested in the probability that a given customer will default on a loan. Note that this is a number representing the level of likelihood. The higher it is, the more likely that the event will actually occur. It does not mean the event will certainly occur. If this was the case, the bank would know who will default and deny them the facilities.

Formally, given that an event would occur a number of times and would not occur b equally likely number of times, the probability that it will occur is

$$P(a) = \frac{a}{a+b}$$

The numerator is the number of times that the event occurs. The denominator is the sum of both the occurrence and non-occurrence of the event. In other words, it is the number of all possible outcomes that constitute an exhaustive listing. Note that the probability of the sample space itself $P(S) = 1$.

For any two events A and B, the probability of either event occurring is the union of the two events and is given by

$$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Where $P(A \cap B)$ is the probability that both events occur at the same time.

However, if the two events are *mutually exclusive* (the two events cannot occur at the same time), $P(A \cap B) = 0$. The probability of the union is then given as

$$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B)$$

For instance, the probability of getting a 1 or 2 from tossing a single die. The two events are mutually exclusive since they cannot occur together. If the two events A and B are independent, then the probability of the intersection also known as the *joint probability* is the product of the individual probabilities.

$$P(A \text{ and } B) = P(A \cap B) = P(A) \times P(B)$$

For example, when two coins are tossed, the probability of getting Head from both tosses is the product of the probabilities of getting head from the toss of each coin.

Finally, when two or more events are *collectively exhaustive*, their union constitutes the sample space. The probability of the union is one. For example, the following probabilities indicate exhaustiveness of the respective sets.

$$P(A \text{ or } B) = P(A \cup B) = 1$$
$$P(A \text{ or } B \text{ or } C) = P(A \cup B \cup C) = 1$$

Example 9.1

Consider the tossing of two fair coins.

a. What is the sample space?

b. What is the probability that:

   i. Both coins have Head;

   ii. There is only one Tail;

   iii. There is at least one Head?

The sample space will be as shown in the table below.

		Second die	
		H	T
First die	H	HH	HT
	T	TH	TT

The sample space as shown in the above table has four sample points. These are HH, HT, TH and TT. The outcomes HT and TH may be considered the same in cases where the order of appearance is immaterial. Nonetheless, we need to show them so that the counting of events is not complicated. In calculating the probabilities, the denominator will be 4, the number of all possible outcomes.

The probability that both coins have Head. This is only occurring once out of four possible outcomes.

$$P(HH) = \frac{1}{4}$$

The probability that there is only one tail makes no mention of which coin produces the only tail. It can come from the first or second coin. This gives two events with only one tail in the sample space. One where it comes from the first coin and anther where it comes from the second coin. There is a restriction on the number of Tail which eliminates the both tails outcomes.

$$P(one\ T) = \frac{2}{4} = \frac{1}{2}$$

The probability of at least on Head puts no upper restriction on the Heads. This includes events with only one Head as well as where both coins produce Heads. From the sample space, there are two with only one Head and one with two Heads. This gives a total of three out of four possible outcomes.

$$P(at\ least\ one\ H) = \frac{3}{4}$$

Example 9.2

A box contains 24 balls, each with a unique serial number from 1 to 24. If a ball is drawn at random, find the probability that

a. It is an odd numbered ball.

b. It is a single digit number.

The probability is calculated as a fraction of the subpopulation to the total population. In this particular case, there are 24 balls in the box. This constitutes the total population. The subpopulations will then depend on what we are looking for.

Between 1 and 24 inclusive, there are 12 odd numbers. This means there are 12 odd numbered balls out of 24 balls in the box. The probability of selecting such as ball is then

$$P(odd\ numbered) = \frac{12}{24} = \frac{1}{2}$$

The number of balls with single digit numbers will be 9. These are balls from 1 to 9 inclusive out of a total of 24 balls. The probability of selecting such as ball is

$$P(single) = \frac{9}{24} = \frac{3}{8} = 0.375$$

9.2 Addition law of probabilities

Since the probability is the measure of likelihood, it is smaller the more unlikely that a certain event will occur or that a particular element will be selected. The probability is always proportionate to the number of specified items in the box. For instance, if there are more red balls compared to blue balls in a box, the chances of picking a red ball must be higher than the chances of picking a blue ball.

Since probability is proportionate to the number of items, it is additive just as are number of items. Consider a basket of balls with three primary colours: red, blue and green. The basket only has these colours and each colour has an equal number of balls. The probabilities or likelihoods that each of the three colours is selected in a single pick is one third. That is,

$$P(R) = P(G) = P(B) = \frac{1}{3}$$

Each pick can result in any of the three colours. Let us look at the red balls for instance. The probability of picking a red ball is

$$P(R) = \frac{1}{3}$$

Suppose we want to know the probability of selecting either a green or blue ball. That is, we want to know the value of $P(G \text{ or } B)$. This would be equal to

$$P(G \text{ or } B) = P(G) + P(B)$$

$$= \frac{1}{3} + \frac{1}{3} = \frac{2}{3}$$

Alternatively, the same probability can be calculated as the probability of not picking a red ball. Since the sample space consists of red, green and blue balls, the probability of not picking a red ball would be

$$P(Not\ R) = P(S) - P(R)$$

$$= 1 - \frac{1}{3} = \frac{2}{3}$$

Theorem: Given k mutually exclusive sets $A_1, A_2, \ldots, A_k$, the probability of the union of all the sets is the summation of the probabilities of the individual sets. That is:

$$P(A_1 \cup A_2 \cup \cdots \cup A_k) = P(A_1) + P(A_2) + \cdots + P(A_k)$$

The above theorem is useful when dealing with partitions. These are subdivision of a bigger set so that each element falls in one and only one

partition. Think of the division of a geographical area of a country into provinces. Each point is identified with a one and only one province. In Zambia for instance with ten provinces, each point will be associated with one province. The provinces do not intersect, they do not hold any common elements. As such, they are partitions.

In the military, each officer or soldier holds one rank at any given time. Therefore when a division is made on the basis of one's rank, no one will say, 'but I belong to two ranks, so which one do I join?' The ranks are mutually exclusive.

Theorem: Given the sample space S and two subsets A and B, the probability of the union set is the sum of the probabilities of the two sets less the probability of the intersection set. Algebraically, this is written as:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

We know that the union of two sets is synonymous to summing or putting together the two sets but not entirely so. This applies only when the two sets have now common elements. When some elements belong to both sets, an addition would double count these; they are added from the two sets. To correct this, the intersection has to be subtracted from the sum of the two sets.

Example 9.3

The distribution of the number of pupils on the basis of grade at a given secondary school is as shown below.

Grade	8	9	10	11	12	Total
Number of students	140	135	70	68	67	480

What is the probability that a pupil selected at random will be:

a. In Grade 10;

b. Will be in Junior secondary (Grade 8 & 9)?

The total number of pupils in this example is 480. Thus the sample space consists of 480 pupils. The number of pupils in Grade 10 is 70. Therefore

$$P(G10) = \frac{n(G10)}{Total}$$

$$= \frac{70}{480} = 0.146$$

The probability of selecting a junior secondary school puts two classes or sets together. It is the union of the two Grades. Since grades are mutually exclusive, the theorem states that

$$P(Junior) = P(G8) + P(G9)$$

$$= \frac{n(G8)}{Total} + \frac{n(G9)}{Total}$$

$$= \frac{140}{480} + \frac{135}{480}$$

$$= 0.292 + 0.281$$

$$= 0.573$$

9.3 Conditional probability

Suppose we consider residents in Zambia as citizens or foreigners and working or not working. Two variables are defined here: citizenship with two possible values, Zambian (Z) and non-Zambian (Z'). The second is the variable on whether one is engaged in work (W) or not (W'). This produces four classes: Zambians who are working; Zambian not working; Non Zambians working and Non Zambians who are not working. Each of these classes has some probability depending on the number in each class. This is summarized in the table below.

	Nationality	
	Z	Z'
Working?		
W	$P(W \cap Z)$	$P(W \cap Z')$
W'	$P(W' \cap Z)$	$P(W' \cap Z')$

The probabilities provided in the table are bivariate. The classes are defined by two variables jointly. As such, the probabilities are known as *joint probabilities*. They provide probabilities for the possible outcomes of two variables jointly.

This is not to say the information in the table should only be looked at from a two variable angle. It is possible to ignore one variable and just consider one. For instance, there may be no reason at times to know whether someone is working or not. What may be paramount is their nationality. What is the probability that a randomly chosen person is Zambian? This is irrespective of whether one is working or not. In the table, Zambian are divided between two mutually exclusive groups; those working and not working. The set Zambian is therefore a union of these two sets.

$$P(Z) = P(W \cap Z) + P(W' \cap Z)$$

The resulting probability is referred to as a *marginal probability*.

Some events are not mutually exclusive, the occurrence of one does not preclude the occurrence of another. Consider again two events A and B. When the two are independent, the occurrence of one says nothing about the probability of how the other will turn out. In medicine, there is a common belief that the probability of having a girl child or boy child is independent of the genders of the older children. That is, even if we know the gender of the first child for instance, it provides no clue as to whether the second child will be male or female. The two are independent.

Now suppose the two events are the mock examination and the final examination. In each one, there is a probability that a student will pass or fail. The probability of passing the mock examination need not be equal to the probability of passing the final examination. Often one is tougher. Suppose the failure rate in the final examination is 20 percent. A randomly

selected pupil has a probability of failure $P(F) = 0.2$. Though admission into college is on the basis of the final examination only, many colleges now offer provisional admission on the basis of the mock examination results. Do these colleges have a 20 percent risk of enrolling someone that may not make it in the final? Well, the risk is there but certainly not 20 percent.

Colleges are able to make provisional admission because they are aware that performances in the two examinations are not independent. A pupil that passes in the mock examination has higher chances of passing the final examination also. And for a pupil that fails in the mock, they also stand higher chances of getting similar results in the final. Even if the probability of passing for a randomly selected pupil is only $P(Pass) = 0.8$, the probability for someone who has passed the mock examination is much higher. Therefore, the mock examination is indicative of performance in the final examination. This is known as *conditional probability*. It is a probability with some condition(s) attached. In our example, the condition is that the pupil passed in the mock examination. This is denoted as

$$P(B/A)$$

This is the probability of B occurring given that A has occurred. In the case of the mock and final examination illustration, the knowledge that a randomly selected student has actually passed the mock examination helps in making a more informed judgement on the probability that the student will pass the final examination.

The conditional probability of A occurring given that B has occurred is

$$P(B/A) = \frac{P(A \cap B)}{P(A)}$$

For illustration, consider the following example on the mock and final examination.

Example 9.4

A class of 100 students sat for the mock and final examination. For each student, assume the results only show that the student passes or fails the examination. The results are shown in the table below.

	Final (B)		
Mock (A)	Pass	Fail	Total
Pass	76	6	82
Fail	4	14	18
Total	80	20	100

In the final examination 80 pupils passed and 20 failed. This makes the probability that a randomly selected pupil passed the final examination $P(Pass\ Fin) = 0.8$. Suppose now it is known that the randomly selected pupil passed the mock examination, what is the probability that the pupil passed the final examination? The prior knowledge, that the pupil passed the mock examination, alters the sample space. With this knowledge, the search will not be from all the 100 pupils but only those that passed the mock examination. This is the information at hand. We are certain the pupil is not coming from the 18 that failed the mock examination. Instead, the sample space is now restricted to only 82 pupils that passed the examination. The selected pupil is coming only from the 82 that passed the mock.

Let M and F be the event that a pupil passes the mock and final examinations respectively. If only those that passed the mock are in the sample space, the probability that a randomly selected pupil is the ratio of those that passed the final examination (meaning they passed both the mock and final in order to be in the sample space) to the number of pupils that passed the mock examination.

$$P(F/M) = \frac{n(F \cap M)}{n(M)}$$

In the numerator is the number of pupils that passed in both examinations, the intersection of the two sets. If both the numerator

and denominator are divided by the total number of hundred, we get the probabilities of the two sets.

$$P(F/M) = \frac{\frac{n(F \cap M)}{Total}}{\frac{n(M)}{Total}}$$

$$P(F/M) = \frac{P(F \cap M)}{P(M)}$$

$$= \frac{0.76}{0.80}$$

$$= 0.95$$

If the selected pupil has passed the mock examination, the probability that they also passed the final examination is actually 95 percent.

The above example illustrates a case where the two events' occurrence is not independent. As such, the probability of one event occurring changes when information about the other event is revealed. If however the probability of one event occurring is irrespective of what is known about another event, the two are said to be *independent events*. Suppose two expecting mothers go to the hospital to deliver. Each mother has a 50 percent chance of delivering a baby boy and a 50 percent chance of delivering a baby girl.

If the first mother delivers and it is now known that she delivers a baby boy, what is the probability of the second mother now delivering a girl or boy? This information is not helpful at all to the 'still expecting' mother. Even for the husband, there is nothing to take from the fact that the other mother has delivered a boy. The two events are independent. The probability of the second mother delivering a boy remains unchanged even after knowing about the first mother.

$$P(B/A) = P(B)$$

In independent events A and B, the knowledge of the occurrence of A does not affect what we should think of B. The independence of events does not mean they cannot occur jointly. That is, it is possible for the two mothers to

both have girls or both have boys. The independence simply means the occurrence of one event does not affect the likelihood of another occurring.

9.4 Multiplicative law of probabilities

Sometimes the interest in probability goes beyond the probability of one event occurring. We may be interested in knowing the probability that two events will occur. For instance, for Brian to marry Precious, it takes two decisions and therefore two events. The first is Brian proposing to marry Maria and second, Maria accepting to marry him. If Brian makes the proposal but Precious declines, no marriage will take place. Similarly, if Brian does not propose, even if Precious would accept, there would be no marriage still. In this case, the interest may be to find the probability that Brian will marry Precious. This is a probability of two event occurring; Brian proposing and Maria accepting.

Let B be the event that Brian decides and therefore proposes to marry Maria. Let M be the event that Maria decides and therefore accepts the proposal for marriage. But is Maria's decision independent of Brian's steps? Well, less likely. We cannot think that Maria would respond to an absent proposal. She may not have Brian in mind. Only when a proposal has been made can we think of assessing Maria's probability of saying 'Yes'. Therefore, while we look at Brian's probability of making the proposal, $P(B)$, Maria's probability is looked at on condition that Brian has made the proposal $P(M/B)$. The probabilities of the two occurring jointly is the product of the two.

$$P(B \cap M) = P(B) \cdot P(M/B)$$

In general, given two events A and B, the probability that the two will occur simultaneously is

$$P(A \cap B) = P(A) \cdot P(B/A)$$

The reader will recall that this similar to the expression given for the conditional probability$B$ given A. Only the subject of formula has changed. Since an intersection set is commutative, the order of the two sets can be swapped without affecting the validity of the expression

$$P(A \cap B) = P(B \cap A)$$

then

$$P(A \cap B) = P(A) \cdot P(B/A) = P(B) \cdot P(A/B)$$

When the two events are independent, then $P(A/B) = P(A)$ and $P(B/A) = P(B)$. The formula for the probability of joint occurrence changes accordingly.

Example 9.5

A lot contains 80 good and 20 defective articles. Two articles are picked, one after the other without replacement. What is the probability that out of the two articles chosen;

a. Both are good.

b. One is good and another defective.

Let A be the event that the first article is good and B be the event that the second article is good. The complements of the two events, A' and B' are the events that the first and the second respectively are defective (not good). While the probability of the first being good or defective is fixed, the probability of the second depends on the outcome of the first. The outcome of the first determines the remaining number of good and defective articles prior to drawing the second one.

$$P(A) = \frac{80}{100}, \quad P(B/A) = \frac{79}{99}, \quad P(B'/A) = \frac{20}{99}$$

$$P(A') = \frac{20}{100}, \quad P(B/A') = \frac{80}{99}, \quad P(B'/A') = \frac{19}{99}$$

The probability that both are good

$$P(A \cap B) = P(A) \cdot P(B/A)$$

$$= \frac{80}{100} \times \frac{79}{99}$$

$$= \frac{316}{495}$$

The probability that only one is good and the other is defective occurs in two ways: either a good comes out first followed by a defective or the defective comes out first followed by a good article. When either outcome is counted, the probability is the sum of the two alternatives.

$$P(only\ one\ good) = P(A \cap B') + P(A' \cap B)$$
$$= P(A) \cdot P(B'/A) + P(A') \cdot P(B/A')$$
$$= \frac{80}{100} \times \frac{20}{99} + \frac{20}{100} \times \frac{80}{99}$$
$$= \frac{16}{99} + \frac{16}{99}$$
$$= \frac{32}{99}$$

The multiplicative law of probability can be extended to more than two events. Consider three events A, B and C each with a non-zero probability of occurring.

$$P(A \cap B \cap C) = P(A) \cdot P(B/A) \cdot P(C/A \cap B)$$

It is the product of the probability that A occurs, the conditional probability that B occurs given that A has occurred and the conditional probability that C occurs given that both A and B have occurred.

Example 9.6

There are two urns in the house. The first urn contains 3 white and 2 black balls. The second contains 4 white and 5 black balls. A ball is taken at random from the first urn and placed into the second urn. After mixing the balls in the second urn, a ball is taken out from thence. What is the probability that this ball is white?

The ball is being drawn from the second urn whose initial content is known. The problem however is that the ball drawn from the first into the second is unknown, it could be white or black. If it is white, then the second urn would contain 5 of each colour. If the ball is

black, the second urn would then contain 4 white and 6 black. A white ball would be picked if either of the two scenarios occur. We sum the probabilities of the two. Let W_1 be the event that the ball from the first urn is white and B_1 if otherwise. Then W_2 is the event that the ball drawn from the second urn is white and B_2 if otherwise. Our problem is to find $P(W_2)$, the probability that the ball drawn from the second urn is white.

$$P(W_2) = P(W_1) \cdot P(W_2/W_1) + P(B_1) \cdot P(W_2/B_1)$$

$$= \frac{3}{5} \times \frac{5}{10} + \frac{2}{5} \times \frac{4}{10}$$

$$= \frac{3}{10} + \frac{4}{25}$$

$$= \frac{23}{50}$$

The multiplicative law of probability is useful in dealing with joint occurrence of events. It applies for dependent as well as independent events.

9.5 Bayes theorem

Consider a retailer receiving goods from k suppliers, each supplying $A_j, j = 1,2, \ldots, k$. The sum is the total amount supplied.

$$A = \sum_{j=1}^{k} A_j$$

From each supplier, there are defective and non-defective units. The retailer actually knows the probability of having a defective unit from each supplier. Using D to denote defective items, $P(D/A_j)$ is the probability of getting a defective item given that the batch is coming from the j^{th} supplier. For a randomly selected item, the knowledge of the source sheds some light on the probability of being defective. If a supplier has a high degree of accuracy, the probability of a defective commodity from that supplier will be less and higher for supplier with many defectives.

The presentation of the formula changes to

$$P(A_j/D) = \frac{D_j}{D} = \frac{P(D/A_j) \cdot A_j}{\sum_{i=1}^{k} P(D/A_i) \cdot A_i}$$

Now divide both in the numerator and denominator by the total number of units, the overall sample space.

$$P(A_j/D) = \frac{P(D/A_j) \cdot {}^{A_j}/_{A}}{\sum_{i=1}^{k} P(D/A_i) \cdot {}^{A_i}/_{A}}$$

The ratio of goods from the j^{th} supplier to the total received is the probability that a randomly selected item will be from the j^{th} supplier.

$$P(A_j/D) = \frac{P(D/A_j) \cdot P(A_j)}{\sum_{i=1}^{k} P(D/A_i) \cdot P(A_i)}$$

This is Bayes theorem. The denominator is the probability that a randomly selected items is defective. As such the theorem can also be written as

$$P(A_j/D) = \frac{P(D/A_j) \cdot P(A_j)}{P(D)}$$

Where

$$P(D) = \sum_{i=1}^{k} P(D/A_i) \cdot P(A_i)$$

Example 9.7

A bank divides its customers into four categories based on the perceived level of riskiness. Each customers is assessed and placed in one of the four categories. Category A customers are low risk customers with an average default rate of 2 percent. Category B customers have a default rate of 6 percent. Categories C and D are considered to be risk and have the probability of default set at 15 and 30 percent respectively. The bank has given out loans to 100 customers categorised as follows:

Category	A	B	C	D
Number	14	25	37	24

a. What is the probability that a randomly selected customer will default on the loan?

b. Suppose it is known that a selected customer has defaulted, what is the probability that they are?

i. Category B customer

ii. Category D customer.

The first step with this sort of lengthy questions is to start by summarising or collating the given data.

$$P(A) = \frac{14}{100}, \quad P(B) = \frac{25}{100}, \quad P(C) = \frac{37}{100}, \quad P(D) = \frac{24}{100}$$

Let F be the event that a customer defaults (Fails to pay back the loan). The conditional probabilities of default given the categories are

$$P(F/A) = 0.02, \quad P(F/B) = 0.06, \quad P(F/C) = 0.15, \quad P(F/D) = 0.30$$

These probabilities are provided in the statement above as the default rates for the respective categories of customers. The probability of default for all the customers is the weighted average of default probabilities for the respective classes. The weights are represented by the class probabilities, that is, the probability that a randomly selected customers is of a particular category. This is the ratio of the number of customers in that category to the total number of customers. Therefore,

$$P(F) = P(F/A).P(A) + P(F/B).P(B) + P(F/C).P(C) + P(F/D).P(D)$$

$$= 0.02\frac{14}{100} + 0.06\frac{25}{100} + 0.15\frac{37}{100} + 0.3\frac{24}{100}$$

$$= \frac{0.28}{100} + \frac{1.5}{100} + \frac{5.55}{100} + \frac{7.2}{100}$$

$$= \frac{14.53}{100}$$

$$P(F) = 0.1453$$

There is a 14 percent chance that a randomly selected customer will default.

For a defaulting customer, the probability that they are from a particular category is provided by the Bayes theorem.

For category B

$$P(B/F) = \frac{P(F/B) \times P(B)}{P(F)}$$

$$= \frac{0.06\frac{25}{100}}{\frac{14.53}{100}}$$

$$= \frac{1.5}{14.53}$$

$$P(B/F) = 0.1032$$

For category D

$$P(D/F) = \frac{P(F/D) \times P(D)}{P(F)}$$

$$= \frac{0.3\frac{24}{100}}{\frac{14.53}{100}}$$

$$= \frac{7.2}{14.53}$$

$$P(D/F) = 0.4955$$

The probability that the customer who has defaulted is from category B is about 10 percent while the probability that the customer is from D is 50 percent. This is despite the almost equal representation of the two categories, that is, the two categories have almost the same number. The reason for the huge difference lies in the respective

categories' chances of defaulting. The default rate for B is only 6 percent compared to D at 30 percent. This means customers in category B are less likely to be amongst defaulters. Category D with a high default rate is likely to dominate the group of defaulting customers.

Example 9.8

A textile company receives the following quantities of thread from three ginneries:

Ginnery	Mumbwa	Choma	Kapiri	Total
Quantity	300	480	220	1000

Out of the sourced quantities, some units get rejected because they do not meet the quality standard for the machine. Suppose on average 5 percent of supplies from Mumbwa is rejected, 2 percent of supply from Choma is rejected and 12 percent of supply from Kapiri is rejected.

a. What is the percentage of supply that is rejected?

b. Of the rejected units, give an estimated breakdown of source.

The percentage of supplies that are rejected is the probability that a randomly selected unit is rejected. Let D be the event that a unit is rejected (because it is defective) and using M for Mumbwa, C for Choma and K for Kapiri Mphoshi. The probabilities that a randomly selected unit is from one of the three ginneries is the proportion of that ginnery's supply to the total supplied.

$$P(M) = \frac{300}{1000} = \frac{3}{10}, \qquad P(C) = \frac{480}{1000} = \frac{12}{25},$$

$$P(K) = \frac{220}{1000} = \frac{11}{50}$$

The probability that a randomly selected unit is rejected is:

$$P(D) = P(D/M)P(M) + P(D/C)P(C) + P(D/K)P(K)$$

[illegible]

$$= [illegible]$$

$$= [illegible]$$

$$= [illegible]$$

$$P(B) = [illegible]$$

[illegible]

[illegible]

[illegible]

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

$$= \frac{[illegible]}{[illegible]}$$

$$= \frac{[illegible]}{[illegible]}$$

$$= [illegible]$$

[illegible]

[illegible]

$$P(C|B) = \frac{P(B|C) \cdot P(C)}{P(B)}$$

$$= \frac{[illegible]}{[illegible]}$$

$$= \frac{[illegible]}{[illegible]}$$

$$= [illegible]$$

The probability that a randomly selected unit from the rejected batch is from the Kapiri Mposhi ginnery is:

$$P(K/D) = \frac{P(D/K) \cdot P(K)}{P(D)}$$

$$= \frac{.12\frac{11}{50}}{0.051}$$

$$= \frac{.0264}{0.051}$$

$$= .518$$

Therefore, the estimated breakdown of the rejected units in terms of sources is as follows. Roughly 30 percent will be from Mumbwa, just under 20 percent from Choma and slightly over half will be from Kapiri. Again, this is in contrast to the fact that the textile company receives the least supply from Kapiri.

The Bayes theorem is very useful in logistical planning. The information from the above example for instance can be useful to the purchasing or warehouse manager for managing logistics of returning the rejected units from the respective suppliers. In this particular example, the manager is able to estimate that roughly half of the rejected items should be going to Kapiri even before information about actual source is confirmed.

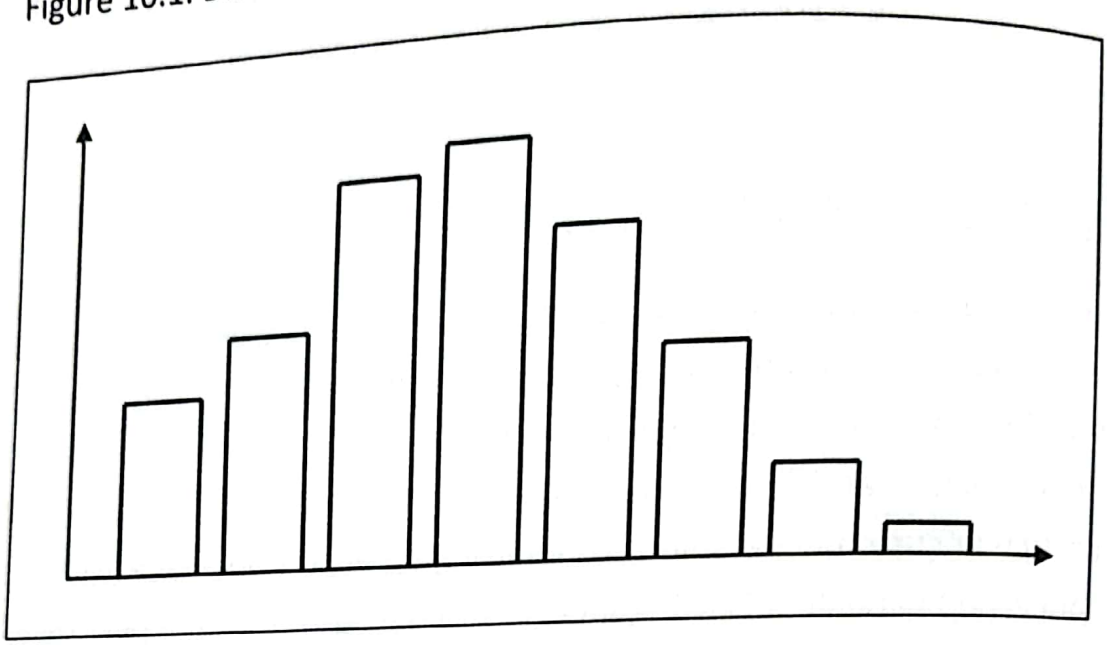
CHAPTER 10

10 DISCRETE PROBABILITY DISTRIBUTIONS

Assume an event with only two possible outcomes; failure or success, head or tail in a coin toss. There are many such examples that one can think of from various disciplines. In medicine, surgeons will have an idea of the success or failure rate of a particular surgery, a bank will deal with chances of a particular customer defaulting or servicing a loan, etc. In all these cases, the concerned variable X is categorical or discrete. The variable only takes on countable and disjoint values. A discrete variable may have many possible outcomes. A dice for instance has six possible outcomes in each toss. Alternatively, a variable may have only two possible outcomes as in the tossing of a coin. The state of many variable often have two possible values $x = 0$ (failure, tail, default etc) or $x = 1$ (success, head, non-defective etc). These are known as *binary variables*. These will be at the centre of this chapter and are discussed in more detail in the succeeding section.

Since the event itself is done a multiple times, many loans issued out, there must be a way of estimating the number of those events that will fall on either size. How many will be on the fail side and how many will be on the success side. The probabilities of each possible number of specific outcome are shown using bar graphs.

Figure 10.1. Discrete probability density function



Various methods are used in estimating the probabilities of each outcomes after a number of trials. Since we are dealing with a discrete variable X, discrete theoretical distributions are used. There are many such theoretical distributions, each dealing with a particular situations.

10.1 Binomial Distribution

The binomial distribution is defined for a binary outcome. This is an outcome in each the possible outcomes are defined into two exhaustive and mutually exclusive classes. In many cases, the outcomes themselves may be binary, that is, only two outcomes are possible. For instance, the state of being present or not-present at a meeting. There are only two outcomes and each member is defined by one only. One is either present or they are not present. The reason for not being present are not necessary. Another example include the tossing of a coin. If we rule out the possibility of a coin standing, then only two mutually exclusive outcomes remain, one called the Head and the other Tail.

Suppose now an event with a binary outcome is tried n times. In each of the trials, there is a fixed probability of either of the two possible outcomes occurring. If a fair coin is tossed n times and we define a random variable x as the number of Heads in the n tosses. This variable has many possible

numbers. It is possible, though not highly probable, that all the outcomes are tails or they are all heads. The number of Heads will therefore range between 0 and n inclusive. It is also discrete since it is a counting number.

Take a simple example of throwing a fair or balanced coin n times. In each of the tosses, the coin has a constant probability of falling either head or tail up. We then define a variable X as the number of times the coin shows a head. As mentioned earlier, the variable is discrete and will range from 0 when it is all tails to n, when it is all heads. The probability that a particular number of heads show is estimated using a binomial distribution.

Let us start with a simple illustration of tossing a coin. A coin has two sides and we designate one of the sides as a success and the other a failure. Let the occurrence of a head be success with a probability of occurring in any toss p. The occurrence of a tail will be failure with a probability q. Since the occurrence of a head and tail are mutually exclusive and exhaustive events, the respective probabilities sum to one.

$$p + q = 1 \Leftrightarrow q = 1 - p$$

Suppose there are four tosses of a coin, what is the probability that there is exactly one head (exactly three tails)? Since there are four tosses, there are many ways in which a single head would occur. The following are the possibilities.

$$HTTT,\ THTT,\ TTHT\ \ or\ \ TTTH$$

Now

$$\begin{aligned} P(HTTT) &= P(H) \times P(T) \times P(T) \times P(T) \\ &= p \times q \times q \times q \\ &= p^1 q^3 \end{aligned}$$

This is the probability of the first scenario where the head comes from the first toss. The reader will realise that the probability will be the same for the other three scenarios. It is always a product of one p and three q, only that the position of p in the product will vary depending on the position of the Head in the sequence of tosses. The probability of one head $P(X = 1)$

without restriction on the position will be the above probability multiplied by 4 since there are four possible ways it can occur.

$$P(X = 1) = 4p^1q^3$$

What if we are now interested in the probability of two heads occurring? The possibilities now will be

$$HHTT,\ HTHT,\ THHT,\ THTH,\ TTHH,\ or\ HTTH$$

The reader now knows that the probability of a particular order of occurrence of the two heads (and so two tails) such as head, head, tail and tail is.

$$P(HHTT) = p^2q^2$$

Next is to identify the number of possible combinations that give two heads out of four tosses. We figured out the case of one head without much trouble but for two heads, knowing that they are actually six (6) as shown above is not always easy. The reader should realise that this is a *combination*. There are four tosses identified by position in the sequence. Of the four positions (tosses) what is the possible number of ways that two heads can be distributed? This is similar to the famous question of number of two-member committees from four available people. In general, this is given by

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

Where n is the number of tosses and x is the number of heads. In the particular case of four tosses two heads,

$$Number = {}^4C_2 = \binom{4}{2} = \frac{4!}{2!\,(4-2)!} = 6$$

And the probability of having two heads is

$$P(X = 2) = 6p^2q^2$$

In general, the probability of a particular number of heads or successes out of n tosses or trials is

$$P(X = x) = \binom{n}{x} p^x q^{n-x}$$

Where p is the probability of 'head' occurring in a single toss and q is the probability that a head does not occur and $\binom{n}{x}$, from a family of positive integers called binomial coefficients, is the possible number of ways that x heads would occur in n tosses. This is a binomial distribution with parameters n and p. Consider the following example for example.

A variable following a binomial distribution $X \sim B(n, p)$ has a mean of

$$\bar{X} = np$$

The variance is.

$$Var\ (X) = npq$$

Example 10.1

A bank gives out loans to five (5) people with similar characteristics and so with equal chance of default. The bank estimates that each customer has a 20 percent chance of defaulting on the loan.

a. What are the possible numbers of defaulters?

b. What is the probability that:

   i. None defaults;

   ii. Exactly two default?

Each customer has an equal chance of defaulting on the loan and this is independent of the outcome with other customers. This means it is possible that all the customers default or only some default or none defaults. In this example, $n = 5$, $p = 0.2$ and $q = 0.8$. Assuming x is the number of defaulters, then the possible values are $x = 0, 1, 2, 3, 4, or\ 5$.

The probability that none of the customers defaults in the probability that the number of defaulting customers is nil, $P(x = 0)$. Using

$$P(X = x) = \binom{n}{x} p^x q^{n-x}$$

Then

$$P(X = 0) = \binom{5}{0} 0.2^0 0.8^5$$

$$= 0.32768$$

In the second part, the question asks for the probability that only two customers default. This places no restrictions on which of the five customers default. The probability that two customers default is

$$P(X = 2) = \binom{5}{2} 0.2^2 0.8^3$$

$$= 0.2048$$

It should be noted that the use of the words 'success' and 'failure' does not adhere to the intrinsic meaning of the two words. Success here means the occurrence of an event of interest such as head in a coin toss, default in loans or defective in output. It is simple the outcomes we are counting. Failure on the other hand is used to denote the occurrence of the other possible outcome. This may be tail in a coin, non-defective is a product and so on. The binomial distribution is a useful tool for estimating probabilities of events occurring in a particular way.

Example 10.2

A military fortification has four possible entrances, each manned by a sentry who has to screen all entrants to prevent all threats to the fortification. As long as all the sentries do not fail (failing to detect and therefore prevent a threat), the fortification is considered secure and safe. However, if any of the sentries fails the whole fortification is doomed. A research has shown that each sentry has a 4 percent chance of failure which authorities reluctantly accept since it is below a benchmark of 10 percent. You are required to interpret these findings to the Commander and what they mean on the vulnerability of the fortification.

According to the information provided, the fortification has four entry points whose probability of failure is both independent and constant. So, $n = 4$ and $p = 0.04$. The fortification is said to be at risk if any of the four fortification fails, as long as one allows the threat in. Therefore, the probability of having a threat is the probability that at least one of the sentries fails $P(x \geq 1)$. This will be a sum of the probability of one, two, three and four sentries failing. Since this involves many numbers, and given the knowledge that the complement of this is the probability that no sentry fails, it is easier to calculate the latter and subtract from one. That is

$$\begin{aligned} P(x \geq 1) &= 1 - P(x = 0) \\ &= 1 - \binom{4}{0}(0.04)^0(0.96)^4 \\ &= 1 - 0.8493 \\ &= 0.1507 \end{aligned}$$

This is a vulnerability risk of 15 percent on the fortification. Clearly, the fortification is highly vulnerable since its probability of security failure is above the set benchmark.

Example 10.3

A farmer who owns a piggery has a newly born set of 15 piglets. The farmer is concerned that some of them may not survive beyond a year and asks a veterinary for advice. All the veterinary is able to ascertain is that each piglet has a 95 percent chance of surviving past its first birthday. The farmer asks the services of a statistician to tabulate, on the basis of the above information, the probabilities that none of the piglets die and the probability that no more than two die.

In this example, there are 15 piglets each with a uniform and presumably independent chance of surviving beyond a year. Thus, $n = 15$ and $p = 0.95$. In the first question, the probability that none dies is the probability that all of them survive.

$$P(X = 15) = \binom{15}{15}(0.95)^{15}(0.05)^0$$

$$= 0.4633$$

With a probability less than half, it seems less likely that all the fifteen piglets would survive beyond one year.

The probability that no more than two die is the sum of the probability that 15 survive, 14 survive or 13 survive.

$$P(x \geq 13) = P(x = 13) + P(x = 14) + P(x = 15)$$

$$= \binom{15}{13}(0.95)^{13}(0.05)^{2} + \binom{15}{14}(0.95)^{14}(0.05)^{1} + 0.4633$$

$$= 105(0.0013) + 15(0.0244) + 0.4633$$

$$= 0.9658$$

There is a 96 percent chance that more than 12 piglets will survive beyond one year. This is more encouraging for a farmer.

The binomial distribution is defined by two parameters; the number of trials (n) and the probability of success (p). It is restricted to experiments or trials that are done a finite number of times.

10.2 Poisson Distribution.

Some variables in life have a time dimension. For instance, while there is a non-zero probability that it may for instance rain on a particular time of the day, the probability is affected by the time length. The probability of raining in a week is greater than the probability of raining on a particular day, say Thursday, of the week. The shorter the time interval, the less likely that the event will occur. We become more certain with a longer time horizon. For instance, a shop keeper may be interested in the probability that a customer will walk-in within a given time interval of say a minute. This probability may be close or even be as low as zero. If the time interval is increased, so is the probability. The longer the time interval, the more certain that a customer will walk in. The probability that a customer will walk in between the time the shop opens and closes is arguably 100 percent or very close to it.

In the same way, when the interval is shortened, the probability also reduces. With shorter time intervals, observations have to be made more often. Taking each infinitesimal time interval as a trial, there will be a million trials in say a day each with a presumably constant probability that a particular event will occur. The probability of the event occurring is very small but the trials are many. This is estimated using the Poisson distribution. The Poisson distribution is also referred to as the law of rare events because it deals with trials with very low probability of the occurrence of the event.

The Poisson distribution is defined by the parameter $\lambda \equiv np$ in a binomial distribution and is denoted by

$$X \sim Poi(\lambda)$$

The probability distribution is given by

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

Since p is assumed to be very small and its converse q approaching one, the mean of the binomial distribution $\bar{X} = np$ will generally equal the variance $Var(X) = npq$. Therefore, in the Poisson distribution, the mean and variance are equal to the parameter.

$$\bar{X} = Var(X) = \lambda$$

The specific probabilities for particular values of $X = x$ are given for various of values of the parameter in the Poisson table.

Example 10.4

A shop owner has estimated that on average, 8 customers will buy something from the shop every day. What is the probability that on a particular day:

a. Nothing is sold?

b. Exactly 8 customers buy from the shop?

In the statement $\lambda = 8$, and the first question asks for the probability that the number of customers served is zero.

$$P(X = 0) = e^{-8}\frac{8^0}{0!}$$

$$= e^{-8}$$

$$= 0.000335$$

In the second part, it is the probability that the number of customers buying equals the average of 8.

$$\mathrm{P(X = 0)} = \mathrm{e}^{-8}\frac{8^8}{8!}$$

$$= 0.000335 \times \frac{16{,}777{,}216}{40320}$$

$$= 0.1394$$

Sometimes there is a tendency for people to think that the number will always be the average. For instance, with the calculated average number of customers, the shop owner may have difficulties accepting that on a particular day, no customers bought anything. The two probabilities above has demonstrated that even with an average of 8 customers a day, there is only a probability of 0.1394 that exactly 8 customers will be served.

CHAPTER 11

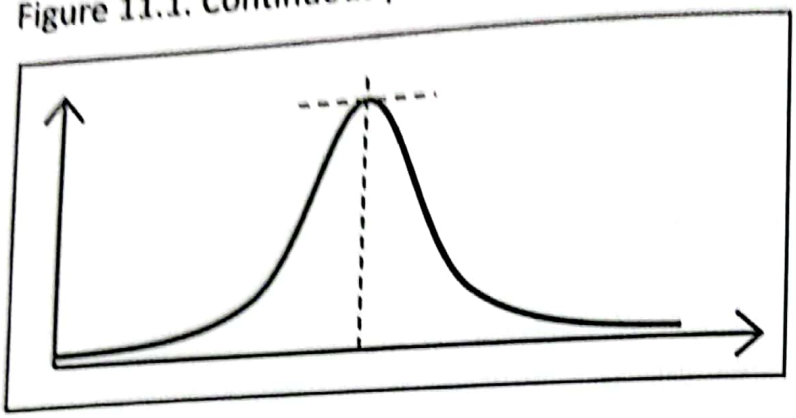
11 CONTINUOUS PROBABILITY DISTRIBUTIONS

It was stated in the preceding chapter that when the variable in the question takes only particular values or is discrete, then discrete probability distributions are used. Some variables are continuous and need to be estimated using continuous probability distributions. Income or salary for instance is a continuous variable. It has an infinite number of possible values and as such, it is not plausible to get the probability of a particular value of income. Theoretically, such probability is zero. Other continuous variables include a person's age, agricultural output, amount of rains etc. The probabilities of these variables can only be approximated using continuous probability distributions.

Even in cases where the variable only takes on integers but the range of values is great, discrete distributions become inappropriate. For instance, the number of pupils or students in a learning institution is discrete. But because of large number of students say in universities, calculating the probabilities that the number of students at a selected university using discrete distribution is inappropriate. Such variables are estimated using continuous probability distributions.

The probability distribution of continuous variable are shown using frequency curves. These will take on many shapes depending on the variable in question. Nonetheless, the frequency curve will generally take the following shape.

Figure 11.1. Continuous probability mass function



Like the discrete probability distributions, there are also several continuous probability distributions, each applying to particular context. This will depend on the assumptions made on the distribution of values between the minimum and the maximum or the two limits. Common distributions include the normal, uniform, chi-square, etc. Of these, the normal and its variant the standard normal are the most widely used distributions. In here, we concentrate on the normal and its standardised variant.

11.1 Normal distribution

The normal distribution is defined by two parameters: the mean (μ) and variance (σ^2). The mean gives the central measure or the position of the data as a group. The variance on the other hand gives the spread or variation in the data. It is an indication of how the data is deviating from the mean. Therefore, if

$$X \sim N(\mu, \sigma^2)$$

The probability mass function is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Where $e = 2.7183$ and $\pi \cong 3.1416$. The distribution assumes an equal distribution of values on both sides of the mean, giving a symmetrical bell-shaped probability curve. The area under the curve is unity and any sub area bound by two distinct points a and b is the probability that selected observations will fall in the interval (a, b). Roughly 68 percent of observations will fall within one standard deviation distance from either side

of the mean, 95 percent of observations will fall within two standard deviation distances and 99 percent will be within three standard deviation distances.

When a normal distribution has a zero mean and unity variance, it is called a *standard normal distribution* denoted by:

$$Z \sim N(0,1)$$

Consequently, the standard normal distribution has a probability function.

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$$

Every variable X following a normal distribution can be converted to a variable Z which follows a standard normal distribution. The transformation is explained in the following steps and principles.

Start with a normally distributed variable

$$X \sim N(\mu, \sigma^2)$$

The variable has a mean μ and variance σ^2. Recall under the properties of the arithmetic mean that if a constant is subtracted or added to each variable, the mean also subtracts or adds by the same constant. That is, since μ is the mean of X, then the mean of $Y = X - \mu$ is zero. The mean reduces by its own value. This has not affected the variance (standard deviation) because the relative standing of the observations remain the same. Each one has simply reduced by the value of the mean which maintains the dispersion or spread in the observations.

$$Y = X - \mu \sim N(0, \sigma^2)$$

Under the properties of the variance or the standard deviation, it was stated that if a given variable Y has a variance of σ^2, then for $Z = kY$ where k is any arbitrary constant has a variance of

$$Var(Z) = k^2\sigma^2$$

If this constant k is set to $k = \frac{1}{\sigma}$, then the variance of $Z = kY$ will be.

$$Var(Z) = \left(\frac{1}{\sigma}\right)^2 \sigma^2 = 1$$

Using this principle, $Y = X - \mu$ is transformed to

$$Z = kY$$

$$= \frac{1}{\sigma}(X - \mu)$$

$$Z = \frac{X - \mu}{\sigma}$$

It has been demonstrated that Z has zero mean and a unit variance. The mean of zero remains unchanged by multiplication. Therefore, the new variable Z follows a standard normal distribution, $Z \sim N(0,1)$. This will have roughly 68 percent of observations within the interval $(-1, 1)$, 95 percent of observations falling within $(-2, 2)$, and about 99 percent of observations within $(-1, 1)$.

The area bound between the curve and any two distinct points z_1 and z_2 is the probability that the observed Z value will fall within that interval. The Z can be transformed to corresponding X values and vice versa for the same probability using the linking equation.

$$z_i = \frac{x_i - \mu}{\sigma}$$

Though the evaluation of probability can be accomplished using the concepts of definite integrals, an easier way is available through use of statistical tables.

Example 11.1

A maize buyer weighs each bag of maize brought for sale. A bag is accepted and bought if its weight is within the interval $(49, 52)$. The buyer refuses to buy anything below $49kg$ and the seller will not sell any bag that exceeds $52kg$. Assuming the bags of maize have a true mean of $\mu = 51kg$ and standard deviation of $\sigma = 1.8$, find the probability that a randomly selected bag:

a. is bought (B).

b. is rejected (R) by the buyer.

For the bag to be bought, it must fall within the range acceptable to both the buyer and seller. Thus, the probability that a bag is bought is:

$$P(B) = P(49 \leq X \leq 52)$$
$$P(B) = P\left(\frac{49-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{52-\mu}{\sigma}\right)$$
$$P(B) = P\left(\frac{49-51}{1.8} \leq Z \leq \frac{52-51}{1.8}\right)$$
$$P(B) = P(-1.11 \leq Z \leq 0.56)$$

Since Z follows a standard normal distribution, this probability is the definite integral of the standard normal distribution. That is,

$$P(-1.11 \leq Z \leq 0.56) = \int_{-1.11}^{0.56} f(Z)\, dZ$$
$$= \int_{-1.11}^{0} f(Z)\, dZ + \int_{0}^{0.56} f(Z)\, dZ$$

Though the probability at this state can be evaluated using techniques of integration, they can also be found from standard distribution table with much ease. Many versions of the table are in use and the reader will need to familiarise oneself on how to read the probabilities from different versions of the table. In the above case,

$$\int_{-1.11}^{0} f(Z)\, dZ = 0.3665$$

And

$$\int_{0}^{.56} f(Z)\, dZ = 0.2123$$

Therefore,

$$P(B) = P(49 \leq X \leq 52) = 0.5788$$

For the second part of the question, it is worth noting that the rejection by either buyer or seller is mutually exclusive. The buyer rejects underweight bags while the seller refuses to sell overweight bags. Thus, the probability that the bag is rejected by the buyer is the probability that its weight falls below the minimum acceptable weight.

$$\mathrm{P(R)} = \mathrm{P(X < 49)}$$

$$\mathrm{P(R)} = \mathrm{P}\left(\frac{\mathrm{X} - \mu}{\sigma} < \frac{49 - \mu}{\sigma}\right)$$

$$\mathrm{P(R)} = \mathrm{P}\left(\mathrm{Z} < \frac{49 - 51}{1.8}\right)$$

$$\mathrm{P(R)} = \mathrm{P(Z < -1.11)}$$

$$\mathrm{P(R)} = \int_{-\infty}^{-1.11} f(Z)\, dZ$$

This is an improper integral, it has $-\infty$ as the lower limit. The result from the table is

$$\mathrm{P(R)} = \mathrm{P(Z < -1.11)} = 0.1335$$

Example 11.2

A company uses a machine to manufacture heavy duty ball-bearings. The machine produces bearing of diameter $12mm$ with a standard deviation of $sd = 0.4mm$. The equipment where the bearings are supplied allows a maximum tolerance in the diameter of the bearings from $11.5mm$ to $12.5mm$. Any bearing falling outside the tolerance region is rejected and considered defective. Assuming the diameters are normally distributed, find rejection rate of the machine's output.

The rejection rate in this context is the probability of a bearing being rejected $P(R)$ or the proportion of rejected bearings. To be rejected, the diameter of the bearing must be outside the tolerance region.

$$P(R) = P(X < 11.5) + P(X > 12.5)$$

The first part is the probability that a bearing is rejected because it is too small and the second, because the bearing is too big. These are mutually exclusive and can be summed without problem

$$P(R) = P\left(\frac{X-\mu}{\sigma} < \frac{11.5-\mu}{\sigma}\right) + P\left(\frac{X-\mu}{\sigma} > \frac{12.5-\mu}{\sigma}\right)$$

$$P(R) = P\left(Z < \frac{11.5-12}{0.4}\right) + P\left(Z > \frac{12.5-12}{0.4}\right)$$

$$P(R) = P(Z < -1.25) + P(Z > 1.25)$$

$$P(R) = 0.1056 + 0.1056 = 0.2112$$

This translates to roughly 21 percent rejection rate. About 21 percent of manufactured ball bearings are rejected because their diameters are beyond what the machine can tolerate.

Example 11.3

A juice manufacturing company uses machine to fill containers. The machine is set with an average of two litres (2 ltr) and has a standard deviation of 0.015 ltr. Suppose the bureau of standards, a standards regulatory authority, requires that the volume in each bottle or container be at least 99 percent of the declared volume. To avoid severe sanctions, the firm has devised a mechanism of detecting and refilling under-fill bottles. Find the proportion of bottles that will be rerouted (R) for a refill.

A juice container is only rerouted when the volume falls below 99 percent of 2 litres or 1.98 ltr, which is 1 percent below the declared volume. The proportion of container being rerouted is equivalent to the probability of a containers falling below the set limit.

$$\mathrm{P(R)} = \mathrm{P(X < 1.98)}$$

$$\mathrm{P(R)} = \mathrm{P}\left(\frac{\mathrm{X}-\mu}{\sigma} < \frac{1.98-\mu}{\sigma}\right)$$

$$P(R) = P\left(Z < \frac{1.98 - 2}{0.015}\right)$$
$$P(R) = P(Z < -1.33)$$
$$P(R) = \int_{-\infty}^{-1.33} f(Z)\, dZ$$
$$P(R) = 0.0918$$

About 9 percent of the containers are rerouted for refill because their volumes fall below what the bureau of standard would permit.

11.2 Normal approximation of the Binomial Distribution

When the number of trials n in a binomial distribution becomes large, it becomes nearly impossible or too tedious to evaluate the probability of each possible occurrence of a random variable. Consider a machine producing n items with each having a probability p of being defective. Binomial distribution allows the calculation of probability of a specific number of items being defective. However, when n is large, there would be too many possible numbers to consider. For a large number of trials, the binomial distribution is approximated by a normal distribution.

Thus, $X \sim Bin(n, p)$ can be approximated using $X \sim N(np, npq)$. Using the mean and variance of the binomial distribution, a normal approximation of the binomial distribution is defined. This approximation can be used when the sample size is large and that neither np nor nq is below 5. This ensures that a symmetrical normal distribution is not used to approximate highly skewed binomial distributions.

Example 11.4

A hatchery has 100 crocodile eggs in incubators. Research has shown that, at the prevailing temperatures, each egg has a 40 percent chance of being male. What is the probability that there will be less than 35 males?

This is a binomial distribution case with $n = 100$ and $p = 0.4$.

If this problem is to be evaluated using the discrete binomial distribution, it would require evaluating the probabilities of each possible outcome. That is, the probability of each number below 35 and summing the probabilities. Now, there are 35 numbers (including zero) whose probabilities have to be evaluated. Even if the formula is ease, the number of cases to deal with is too numerous. It would not be a pleasant route to follow.

Since both np nor nq are greater than 5, this can be approximated using the normal distribution.

$$np = 100 \times 0.4 = 40$$

$$npq = 100 \times 0.4 \times 0.6 = 24$$

Therefore, we take X to have the distribution

$$X \sim N(40, 24)$$

The probability that there will be less than 35 males is.

$$P(X < 35) = P\left(\frac{X - \mu}{\sigma} < \frac{35 - \mu}{\sigma}\right)$$

$$P(X < 35) = P\left(Z < \frac{35 - 40}{\sqrt{24}}\right)$$

$$P(X < 35) = P(Z < -1.021)$$

$$P(X < 35) = \int_{-\infty}^{-1.021} f(Z)\, dZ$$

$$P(X < 35) = 0.1539$$

11.3 Normal approximation of the Poisson

If a random variable follows a Poisson distribution, $X \sim Poi(\lambda)$, both its mean and variance equal the parameter, $\bar{X} = Var(X) = \lambda$. For a sufficiently large λ, the distribution can be approximated by a normal distribution.

$$X \sim N(\lambda, \lambda)$$

Example 11.5

A manufacturer of light bulbs finds that 1.5 percent of light bulbs produced in a day are defective. Find the probability that from a day's production of 600 bulbs, the number of defective bulbs will be less than 10.

Though this is a discrete probability, the numbers involved are too numerous. Nonetheless, the probability can be approximated by the normal distribution with appropriate parameters. With 1.5 percent of 600 daily production defective, there are 9 defective bulbs in a day. Therefore, $\lambda = 9$ and

$$X \sim N(9, 9)$$

The probability of having less than 10 defectives in a day is

$$P(X < 10) = P\left(\frac{X - \mu}{\sigma} < \frac{10 - \mu}{\sigma}\right)$$

$$P(X < 10) = P\left(Z < \frac{10 - 9}{\sqrt{9}}\right)$$

$$P(X < 10) = P(Z < 0.33)$$

$$P(X < 10) = \int_{-\infty}^{0.33} f(Z)\, dZ$$

$$P(X < 10) = 0.6293$$

11.4 Other Distributions: The Chi-square

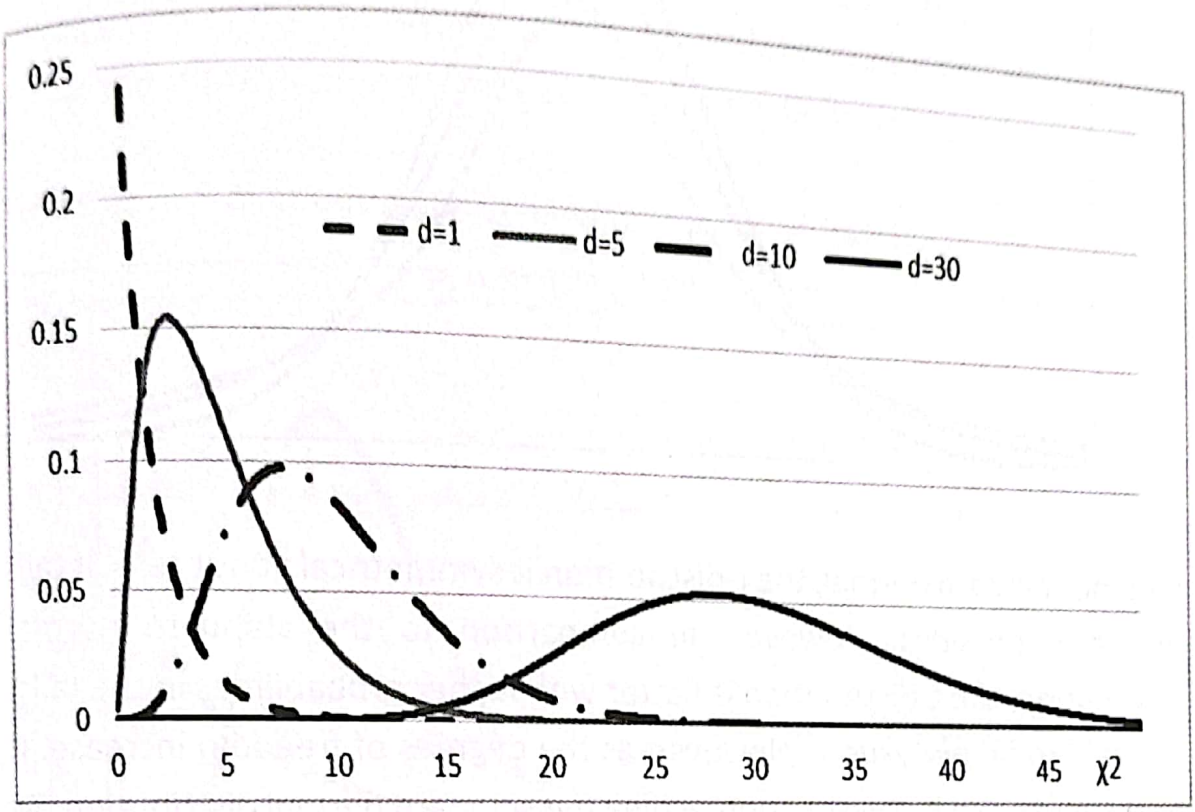
As such, we discuss the chi-square distribution before getting into the test.

If a random variable Z follows a standard normal distribution $Z \sim N(0,1)$, then the sum of k squared independently and identically distributed Zs will follow a chi-square distribution. That is

$$X = \sum_{i=1}^{k} Z^2 \sim \chi^2_{k-1}$$

Each degree of freedom defines a distribution. Therefore, a chi-square is often referred to as a family of distribution.

Figure 11.2. Chi-square distributions



As seen from the figure, the chi-square distribution with low degrees of freedom as skewed to the right. As the degree of freedom increases, the chi-square resembles the normal distribution. This conforms to the central limit theorem which was stated in earlier chapters. The chi-square distribution is mainly used in the test of *goodness of fit*. This is deal with under test of hypothesis.

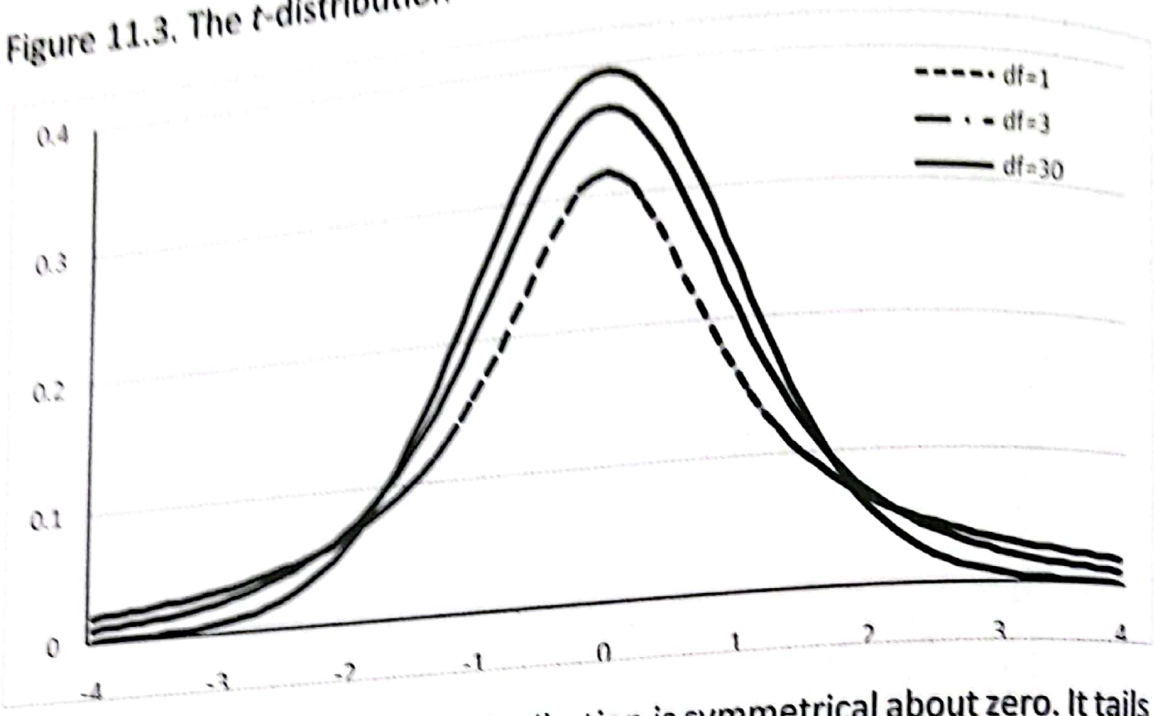
11.5 Other Distributions: The Student *t*-distribution

If Z and X are independent variables following a standard normal and chi-square distribution with k degrees of freedom respectively, then the statistic

$$T = \frac{Z}{\sqrt{X/k}}$$

follows a student t-distribution with k degrees of freedom. The degrees of freedom of the t-distribution will depend on the degrees of freedom of the respective chi-square distribution.

Figure 11.3. The *t*-distribution



Like the standard normal, the t-distribution is symmetrical about zero. It tails off on both sides. However, in comparison to the standard normal distribution, the t-distribution is flatter with higher probabilities in the tails. It is said to be *platykurtic*. However, as the degrees of freedom increase, it peakedness also increases approaching the standard normal distribution. As such, for higher degrees of freedom, the t-distribution can be approximated by the standard normal distribution. The t-distribution is used as an alternative to the normal distribution when certain requirements of the latter (the knowledge of the variance) are not fulfilled.

11.6 Other Distributions: The *F*-distribution

To define the F-distribution, consider two independent variables X_1 and X_2, each following a chi-square distribution with k_1 and k_2 degrees of freedom respectively. That is,

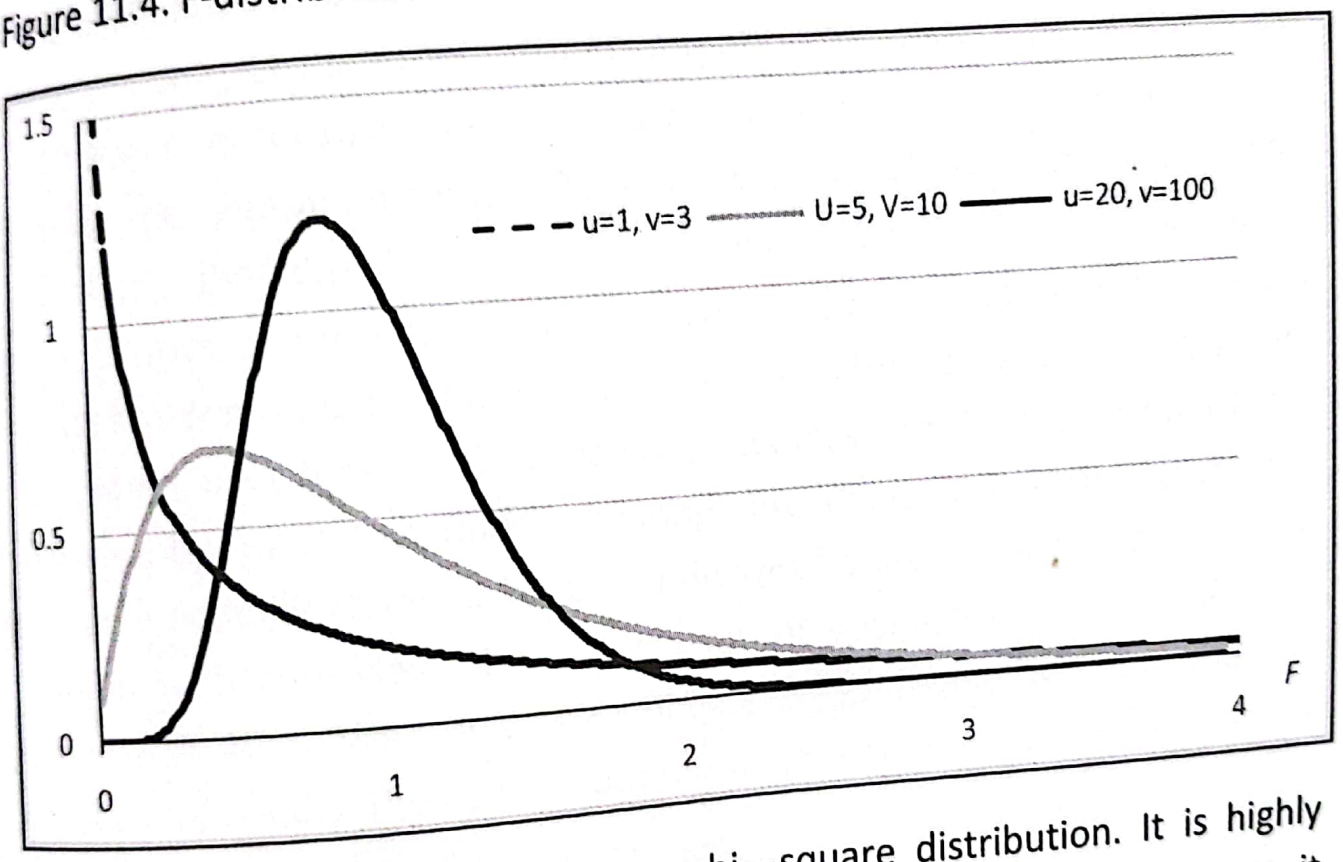
$$X_1 \sim \chi^2(k_1) \text{and} X_2 \sim \chi^2(k_2)$$

The ratio of the two chi-square variables, each divided by its degrees of freedom follows an F-distribution. That is, the variable

$$F = \frac{X_1/k_1}{X_2/k_2}$$

follows an F-distribution with k_1 and k_2 degrees on freedom. Because the first degree of freedom is associated with the numerator and the second with the denominator, the two cannot be swapped. They are often referred to as the numerator and dominator degrees of freedom respectively. Each combination of the two degrees of freedom defines an F-distribution. This makes the F-distribution a family of distribution. Consider Figure 11.4 below.

Figure 11.4. F-distribution with u and v degrees of freedom



The F-distribution is similar to the chi—square distribution. It is highly skewed at low degrees of freedom. As the degrees of freedom increase, it approaches the normal distribution. This distribution is used for hypothesis testing, particularly when testing the equality of many means. It will be used in later chapters.

CHAPTER 12

12 ESTIMATION

Economic theory is concerned with many variables that must be known. We often talk about the levels of income, the consumption levels, and demand for a commodity or the supply. The formulation of policy requires that these are known. In order to plan for the community, there is often emphasis on knowing the community itself. This knowledge includes the size of the population, the distribution of age, and many other demographic characteristics of the community. However, the measurements of these variables is complex. In addition, the variables themselves are not static, they change over time. If the process of establishing the true value of the parameter is undertaken, due to the long time it may involve, the parameter may have actually changed by the time results are released. This is very common with censuses. The last census of population for Zambia was conducted in 2010. The results were only ready some years later. At the time of release of results, the actual number along with many other parameters would have changed.

Given this complexity in determining the actual level of a parameter, methods have been devised to simply estimate them. This chapter therefore discusses the concept of estimation, the methods and their strengths and weaknesses.

12.1 What is estimation?

Consider the levels of consumption by a household or individuals. It will vary from one person to another. Since it may not be feasible to enumerate all the subjects in a population so that the true parameter is known, an estimation or approximation is made on the basis of a sample. Given a random variable X, its expected value also known as the average is

$$E(X) = \mu$$

Because only a sample is available, the true value of the parameter cannot be known. Nonetheless, it can be estimated. Multiple methods are used. For instance, the sample mean is an estimator of the population mean. The value of the sample mean will give an indication of the expected value of the variable.

In some cases, variation in a variable is driven by other variables. This is true with consumption which is a function of income. It is indisputable, under normal circumstances, that the consumption will be higher for persons or households with higher levels of income. In the same way, low income households will have low levels of consumption. Our interest then is to determine the level of consumption for a given level of income. The knowledge of a person's level of income will enable us determine the level of consumption. Assume the level of consumption is related to income and can be expressed as follows.

$$C = C(Y)$$

Where Y is the level of income. The relationship can take on various functional forms including *constant* if consumption is independent of income or linear, quadratic, logarithmic or exponential forms. The exact form will depend to a larger extent on variables in the question and the underlying assumptions.

If it is assumed that the *marginal propensity to consume* (MPC) is constant for all levels of income, then the consumption function will be linear. It will be defined by two parameters; the constant α representing the autonomous consumption and the coefficient β representing the MPC.

$$C = \alpha + \beta Y$$

The above relationship is true for all observations, it is based on the population. However, data is often very difficult or costly to gather from the entire population. Only sample data is available. Because we do not have all the data to determine the value of the parameters, we must approximate or estimate them on the basis of available sample data. Thus estimation is the process of finding an estimate or approximation of an unknown parameter.

The task of estimation is to estimate or approximate the true parameter based on the available information, the sample. There are two approaches to estimation: *point estimation* and *interval estimation*. The former provides a single value estimate of the parameter while the latter will provide a range or interval of probable values of the parameter. The two methods are discussed later in the chapter.

12.2 Definitions

Population is the entire collection of objects. Examples include the population of Zambia, all banks in the country, etc. and is often studied through census. In statistics, any characteristic of a population is known as a parameter and often denoted by greek letters. For instance, the population average is denoted by μ, the standard deviatiton is denoted by σ. We have already shown the two parameters of the linear consumption function, α and β. All these characteristics are based on the population or the entire set of objects we wish to study.

Because of the difficulty in gathering population-based data, statisticians resort to collecting samples. These are subsets or portions of the population from which unknown parameters can be estimated. Characteristics based on a sample are known as *statistics* and *statistic* in singular. The name is the same as the subject matter itself. The latter refers to the subject, statistics while in the former, it is a synonym of parameters but based on a sample.

An estimator is then a method that is used to estimate or approximate a parameter. For instance, the sample arithmetic mean is an estimator of the population mean. When we calculate the sample mean, we want to have an idea of what the true population mean is likely to be. Thus, an estimator is a method, a formula or way of approximating the parameter. For instance, $\bar{x} = \frac{\sum x_i}{n}$ is an estimator of the population mean μ. A specific value of an estimator, say $\bar{x} = 25$, is a *point estimate*. We estimate that the true mean is $\mu = 25$. Alternatively, the statement "the average between 22 and 26" is an *interval estimate* of the parameter.

12.3 Point Estimation

Point estimation is the most common method of estimation. We have encountered in many spheres of life when a single value is given as an estimate of the parameter. For instance, the presentation of GDP per capita is a point estimation. It gives a single number for each country or region. Other examples include the average age for the Zambian population which can be given as a single value. Thus, point estimation involves the calculation of a single value from a sample to serve as an estimate of unknown parameter.

Many methods have been devised, abet with different properties. Since the choice of which method to employ must be guided by the properties, it is imperative that we look at the desirable properties of estimator.

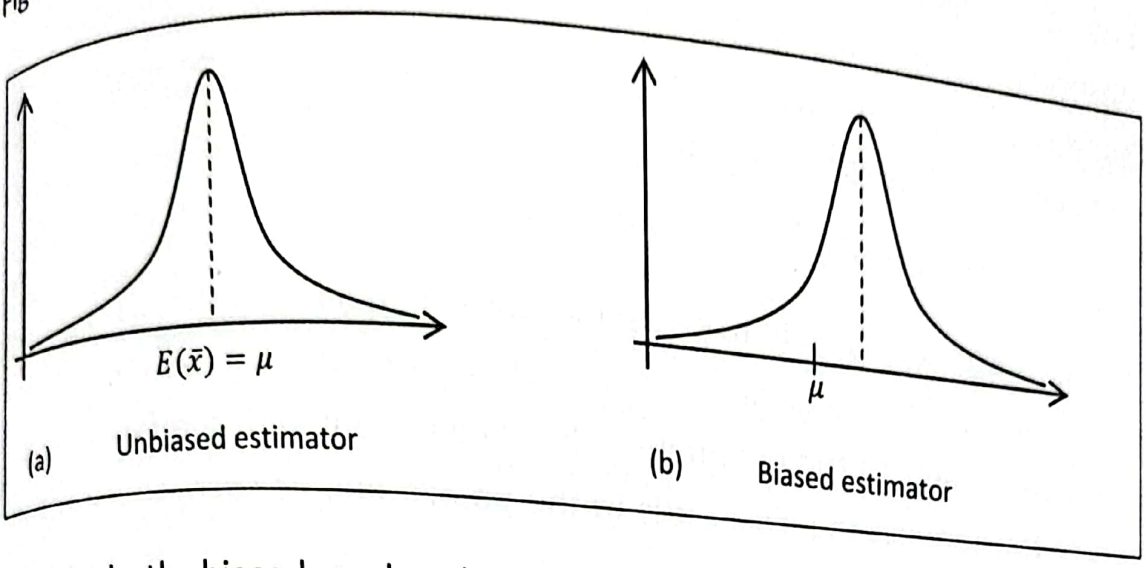
12.4 Desirable properties of "good" estimators

Each estimator has attributes or properties which define their advantages and disadvantages. We discuss four properties here:

12.4.1 *Lack of bias*

The primary rationale for estimation is to get insights on the position of the parameter. We want to know, based on sample data, what the value of the parameter is. Since the parameter is unknown, we are not able to compare the estimate to the true value of the parameter so as to see the amount of deviation or residual. If different samples are picked repeatedly, the average of the estimates must be equal to the parameter. In other words, an estimator is *unbiased* if the many estimates are dotted around the true mean such that their mean is the parameter. An unbiased estimator gives 'on the average' the true value of the parameter. The distribution of sample means must have the parameter as the mean.

Figure 12.1. Biased and unbiased estimators



Though both biased and unbiased estimators will normally miss the parameter, in several repetitions, the latter will have the parameter as its average. For a biased estimator, the mean of the distribution deviates from the parameter. For an unbiased estimator, the expected value of the mean would be equal to the parameter. That is,

$$E(\bar{x}) = \mu$$

If $E(\bar{x}) \neq \mu$, then the estimator is biased. Thus, the bias B is given as

$$B = \mu - E(\bar{x})$$

In theory, both the LSE and MLE are unbiased.

12.4.2 Efficiency

Efficiency refers to achieving a high quality results from the same number of samples. This is anchored on the need to minimal variations of samples. If in addition to being unbiased, an estimator also has the lowest variance, it is said to be an efficient estimator. Efficiency allows making stronger conclusions from the same sample. Few samples would be needed to get a reliable idea of the parameter than for an inefficient estimator.

12.4.3 Sufficiency

An estimator is said to be a sufficient estimator of the true population parameter if no other estimator can add any further information about the parameter. A sufficient estimator captures all the information in the data relevant for the parameter so that no other estimator would be able to add

any further information. For instance, the sample mean, capturing all the observations, is a sufficient estimator of the parameter μ. On the other hand, the mode and median are not sufficient estimators because they do not capture all the observations.

12.4.4 Asymptotically unbiased

A biased estimator is said to be asymptotically unbiased if the bias reduces as the sample size gets large. As the sample size approaches infinite, the estimator points to the parameter. This is consistent with the fact that the larger the sample, the closer it is to the population. As such, the statistics also approach the parameter.

Notationally, an estimator is asymptotically unbiased if

$$Plim\ \bar{x} = \mu$$

12.5 Interval estimation

Point estimates almost always miss the true parameter. They only get closer to the true parameter. Though with repeated samples, the estimate can be improved, we known this is not always feasible. Point estimates give a single 'point' with no clue on the direction of the true parameter. The alternative method is known as *interval estimation* and involves giving an interval in which the parameter must fall. It gives an interval within which the parameter lies with a certain level of confidence. It is the method of calculating the interval of probable values of the unknown parameter. This is known as the *confidence interval* for a parameter. Formally, an interval estimate for a real valued parameter θ is a pair of two functions of a sample, the lower limit a and the upper limit b. We can infer on the basis of the interval estimate that the true parameter lies between the two limits, albeit with a given level of confidence.

Suppose we wish to estimate the true mean of the population μ on the basis of the sample. The sample mean, though unbiased, will say little about where the true mean lies, whether above or below the estimate. Interval estimate instead provides an interval with a corresponding probability of the true parameter falling within the interval. The probability helps rank the

different estimates on the basis of their probability of capturing the parameter.

Central to interval estimation is the level of confidence associated with each estimate. This is proportionate to the probability that the given interval contains the true parameter. Interval estimation therefore involves getting the interval (a, b) so that there is a $(1-\alpha)$ probability that the true parameter lies within the interval where α is referred to as the level of significance. It is the probability that the interval misses the parameter. This can also be stated as $100(1-\alpha)$ percent chance that the parameter is captured by the interval.

Interval estimation rely on knowing the spread of estimates, assuming repeated samples. Though many methods are available for estimating the spread or dispersion in a random variable, the calculation of interval estimates is based on the standard deviation of the estimate. This is denoted by $sd(\hat{\theta})$, where $\hat{\theta}$ is the estimate of the parameter θ. Because the standard deviation is seldom known, it is estimated by the standard error of the estimate, denoted by $se(\hat{\theta})$. It should suffice to note that the calculation or estimation of the standard deviation will depend on the nature of estimate.

For instance, the estimator for the population mean μ is the sample mean $\bar{x}$. To get the standard error of the sample mean, we can rely on the Bienayme formula on properties of the variance.

$$Var\left(\sum x_i\right) = \sum Var(x_i)$$

Thus,

$$Var(\bar{x}) = Var\left(\frac{\sum x_i}{n}\right)$$

$$= \frac{\sum Var(x_i)}{n^2} = \frac{\sum \hat{\sigma}^2}{n^2} = \frac{n\hat{\sigma}^2}{n^2} = \frac{\hat{\sigma}^2}{n}$$

Where $\hat{\sigma}^2$ is the estimate of the variance of x and n is the sample size. Then the standard error, the square root of the variance is

$$se(\bar{x}) = \frac{\hat{\sigma}}{\sqrt{n}}$$

The estimated standard error will be a random variable, because it depends on a random sample. It will get smaller as the sample size gets larger.

When the population variance (standard deviation) is not known, it must be estimated based on the basis of a sample. But the sample standard deviation is not unbiased. It misses the true parameter. Nonetheless, the estimate is consistent. It approaches the true parameter as the sample size gets larger. The implication is that we are not to worry of the bias for as long as the sample size is large. However, for small samples, some adjustments have to be made.

12.6 Confidence Interval

Given the interval (a, b) the probability that it contains the parameter, also known as the *coverage probability* is

$$P(\theta \in (a, b)) = P(a < \theta < b)$$

In the above expression, it must be noted that it is the two limits, a and b that are random and not the parameter θ. As such, the expression must be modified so that the random variable is compared to a fixed point, the parameter. Thus,

$$P(\theta \in (a, b)) = P(a < \theta \ \& \ b > \theta)$$

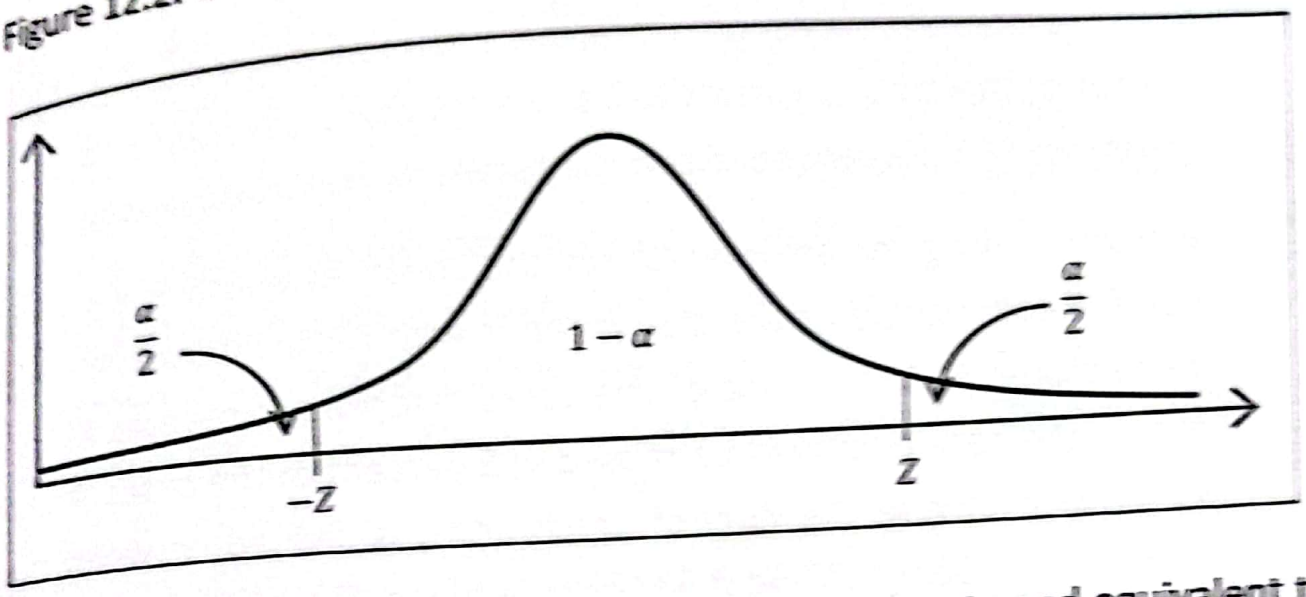
The reader will recall from previous chapters that this will become

$$P(\theta \in (a, b)) = P\left(\frac{a - \theta}{se(\hat{\theta})} < 0 \ \& \frac{b - \theta}{se(\hat{\theta})} > 0\right)$$

Where *se* is the standard error of the estimated limits.

We want to vary the two limits so that coverage probability is as predetermined. The two limits must yield a predetermined coverage probability. Consider the figure below

Figure 12.2. Coverage Probability



We know that the two limits a and b are proportionate and equivalent to the standardised values $-Z$ and Z respectively. The probability that a particular value is between the two standard values is $1-\alpha$. There is a $\alpha/2$ probability that the value will be below the lower limit or above the upper limit. With a symmetrical frequency curve, the absolute value of Z is the same for both limits. In order to emphasise that the Z value represents a probability of $\alpha/2$, we denote the respective values as $-Z_{\frac{\alpha}{2}}$ and $Z_{\frac{\alpha}{2}}$.

The interval estimate of the parameter is therefore

$$-Z_{\frac{\alpha}{2}} = \frac{a-\hat{\theta}}{se(\hat{\theta})}$$

$$a = \hat{\theta} - Z_{\frac{\alpha}{2}} \times se(\hat{\theta})$$

Corollary,

$$b = \hat{\theta} + Z_{\frac{\alpha}{2}} \times se(\hat{\theta})$$

Once the two limits are identified, there is a $100(1-\alpha)$ percent probability that the true parameter lies within the interval. There is a $100(1-\alpha)$ level of confidence in the estimate.

Example 12.1

A crop forecasting survey of 100 farmers, found a mean $\bar{x} = 720kgs.$ Output of the crop is known to have a standard deviation $\sigma = 260.$ Construct a 95 percent confidence interval for the true mean output.

We are looking for an interval for the mean in which we are 95 percent confident that the true mean is within the interval. It must be captured in 95 percent of the cases sampling is repeated. This means the parameter will be missed in 5 percent of the cases. The 5 percent, known as the level of significance, is divided between the 'below' and 'above'. If we miss the true mean, it is either above the interval or below it. Thus, there should be 2.5 percent that it falls below and 2.5 that it falls above the interval.

Using the sample mean as the estimate, the confidence interval(a, b), will be given by

$$\left(\bar{x} - Z_{\frac{\alpha}{2}} \times se(\bar{x}) , \bar{x} + Z_{\frac{\alpha}{2}} \times se(\bar{x})\right)$$

The standard normal value Z with 2.5 percent or 0.025 probability in the tail is $Z_{0.025} = 1.96.$

$$\left(720 - 1.96 \times \frac{260}{\sqrt{100}} , 720 + 1.96 \times \frac{260}{\sqrt{100}}\right)$$

$$(720 - 1.96 \times 26 , 720 + 1.96 \times 26)$$

$$(669.04 , 770.96)$$

When the population variance is estimated based on the sample, the confidence interval have to be calculated using the *t*-distribution. If Z and X are independent variables following a standard normal and chi-square distribution with k degrees of freedom respectively, then the statistic

$$T = \frac{Z}{\sqrt{X/k}}$$

follows a student t-distribution with k degrees of freedom. The degrees of freedom of the t-distribution will depend on the degrees of freedom of the respective chi-square distribution.

The t-distribution takes care of the inherent bias in the estimated standard error. The t-distribution has degrees of freedom $n-k$ where k is the number of parameters involved. With a single parameter, the degrees of freedom is $n-1$. The statistic is denoted by,

$$t_{n-1,\frac{\alpha}{2}}$$

It is nonetheless permitted to use the standard normal distribution since the sample, from which the variance is estimated, is large enough. But what constitutes a sufficiently large sample? Different disciplines, with different appetite for precision, have settled on different numbers. While natural science will generally require $n \geq 120$ to be considered large, in social sciences and other discipline, $n \geq 30$ is considered large. Economics falls in this category and will consider a sample of 30 to be large enough to allow the use of the standard normal distribution.

Example 12.2

A random sample of 25 villages in one province of Zambia gave a mean population per village of 628 with a standard deviation of 230. Construct a 99 percent confidence interval for the true mean population per village.

The reader should note the available standard deviation is based on a sample of size $n = 25$. This is less than the benchmark for the standard normal. Therefore, a t-distribution will be used. The 99 percent requires that there is 0.5 percent or 0.005 on either side of the interval. This gives a t-value of

$$t_{24,0.005} = 2.797$$

The confidence interval is given by

$$\left(\bar{x} - t_{24,0.005} \times se(\bar{x}), \bar{x} + t_{24,0.005} \times se(\bar{x})\right)$$

$$\left(628 - 2.797 \times \frac{230}{\sqrt{25}}, 628 - 2.797 \times \frac{230}{\sqrt{25}}\right)$$

$$(628 - 128.66, 628 - 128.66)$$

$$(499.34, 756.66)$$

We are 99 percent confident that the true mean population per village is anywhere between 499 and 757 people.

CHAPTER 13

13 THE DESIGN OF STATISTICAL TESTS

The preceding chapter dealt with estimation, both point and interval estimations. In some cases, however, we are encountered with situations where we have to compare sample outcomes with presumed population parameters. We often have to deal with officially held numbers and what is observed in samples. We have to judge whether observed deviations between observed statistics and presumed parameters is a sampling variation or the parameters are truly different from initially thought. To illustrate this, consider the following case.

The Ministry for Agriculture in Zambia announces the opening of the maize marketing season each year. This signals to, other buyers, the readiness of the commodity for purchase. Before flagging the season, the ministry has to satisfy itself that the moisture content (a measure of dryness in maize) is below a set benchmark. This is to minimise the chances of purchased maize rotting because of high levels of moisture.

Suppose now the ministry will not open the season until the moisture content mc is below 20 percent, $mc < 20\%$. If the maize is bought too early ($mc \geq 20\%$), it would be detrimental to the buyers and the economy as a whole as most of it would rot. If the marketing is delayed, it would delay the farmers' incomes. There would be pressure from farmer organisations to open the marketing season. Therefore, the ministry must open the season once the maize is dry enough but should be careful not to open the market before the maize is really dry.

In order to measure the moisture content, the ministry takes a sample of maize and measures the moisture content of the sampled maize. Suppose the sampled maize shows an average moisture content $mc = 19\%$. Should the ministry launch the marketing season or the one percentage point difference may be due to sampling variation? This type of question requires

a yes or no answer. In making the decision, the ministry must satisfy itself that the moisture content is indeed below 20%. We deal with this type of question under hypothesis testing. This involves putting the hypothesis that the moisture content is not below 20 percent to trial. Then determine if the available sample evidence is strong enough to 'convict' (reject) the hypothesis or 'acquit' (fail to reject) because the evidence is insufficient.

We define the *null hypothesis*, denoted by H_0, as a statement that the moisture content is not below 20 percent. We then put this hypothesis to the test on the basis of evidence available from a sample. We either fail to reject the null hypothesis if the evidence against it is not strong enough or reject it if there is strong evidence. If the null hypothesis is rejected, then the *alternative hypothesis*, denoted by H_a or H_1 is true. The alternative hypothesis simply states that the null is not true. It holds the contrary of the null hypothesis. The alternative in the above scenario states that the moisture content has changed, it is below 20 percent.

But if the ministry decides to launch the season on the basis of this sample evidence, it may be the correct decision or could be a mistake or error because the sampled maize understated the true parameter. This is known as Type I error, read as "type one error". Alternatively, if the ministry decides to wait further because they are not satisfied with the sample evidence, they could be making another mistake of not launching when the maize is actually ready. This is known as Type II error. We discuss the two errors in the next section.

13.1 Type 1 and type 2 errors

In the above scenario of moisture content in maize, there are two possible true states at a time. It is either the maize is not dry enough ($mc \geq 20\%$) in which no action should be taken or the maize is actually dry enough ($mc < 20\%$) and marketing should be launched. There is no other possible state. Since the true state may not always be known, a decision has to be made on the basis of a sample. For each true state, there are two possible decisions, of which one is correct and the other incorrect or an error.

Similarly in the first scenario the maize is actually not yet dry. The null hypothesis is true. If based on the sample, we also conclude that the maize is not dry enough, then a correct conclusion will have been made. Alternatively, the evidence may be deceiving so that we conclude otherwise. The error of concluding against the null hypothesis when it is in fact true is known as Type I error. This error would lead to the marketing of maize with moisture content higher than the benchmark.

The second scenario is where the null hypothesis is actually not true, the maize is dry enough. The correct conclusion in this case will be one that also rejects the null hypothesis. We may often make the correct decision by rejecting the null hypothesis. If however we fail to reject it, then a Type II error is committed. This is failing to reject the null hypothesis when it should actually be rejected. This conclusion would unnecessarily lead to a delay to commence the marketing season.

The table below illustrates when each of the two errors occurs.

Table 13.1. Type I and Type II Errors

		Conclusion/Decision	
		Accept Null	Reject Null
Actual	Null Hypothesis True	Correct decision	**Type I Error**
	Null Hypothesis False	**Type II Error**	Correct decision

Since the true state is seldom known, we are not able to tell whether we have made the correct conclusion or have committed either of the two errors. Nonetheless, we can calculate for each test the probability of committing either of the errors. Let us define α as the probability of committing Type I Error. That is,

$$\alpha = P(Reject\ H_0|H_0) = P(Type\ I\ Error)$$

This is also known as the *level of significance* of the test. A hypothesis test will have a pre-stated level of significance α. Most commonly used levels of significance are $\alpha = .05$ and $\alpha = .01$. The test provides a 5 percent and 1 percent chance of rejecting the null hypothesis when it is in fact true. At $\alpha =$

.05, the null hypothesis is rejected even without much evidence against it compared to $\alpha = .01$. The latter requires a high level of evidence against the null hypothesis.

Similarly, the probability of committing the Type II Error is denoted by β.

$$\beta = P(P(Accept\ H_0|H_1) = P(Type\ II\ Error)$$

The probability of Type II Error defines the *power of the test*. This is the ability of the test to detect the falsehood of the null hypothesis. That is, the power of a test for the parameter θ is.

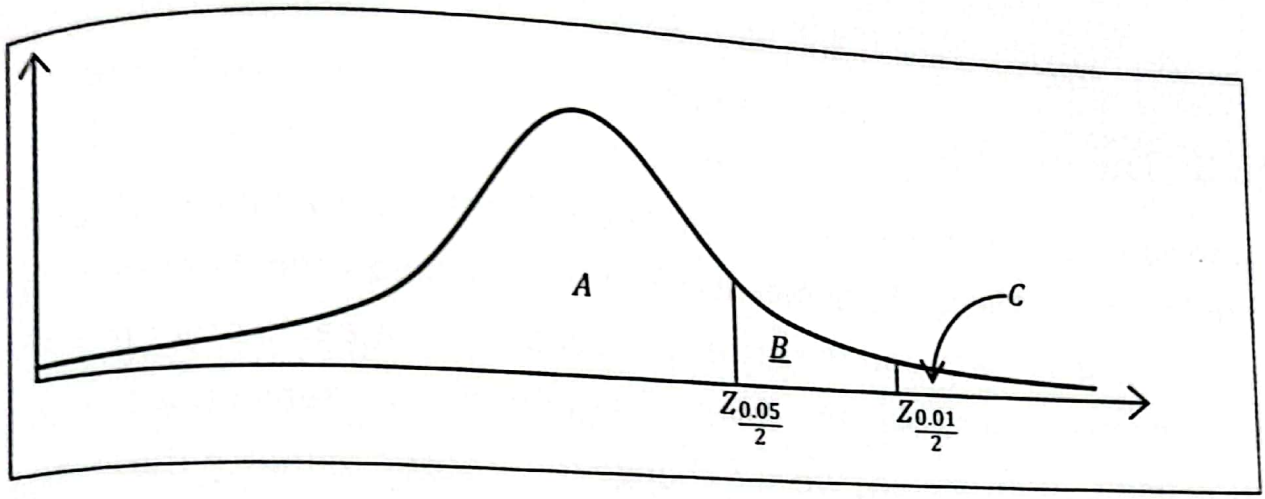
$$\pi(\theta) = 1 - \beta$$

Given the level of significance α, a test is chosen so as to maximise the power.

13.2 Critical Points

A critical point is the boundary point between the acceptance and rejection region. The former is the region that if the observed statistic falls in, we would fail to reject the null hypothesis. If the statistic falls in the rejection region, the null hypothesis is rejected. The critical point is calculated based on the desired level of significance and in some cases, the degrees of freedom. We mentioned in the preceding section that different levels of significance are commonly used. Since each level of significance is associated with a particular critical point, this gives a possibility of rejecting the null hypothesis at one level of significance and fail to reject it at another level of significance. In order to harmonise the two commonly used levels of significance, we deal with two critical points and hence three regions; the acceptance, the no decision and rejection regions. We know that the lower level of significance requires a high level of evidence against the null in order to reject. Therefore, insufficient evidence at say $\alpha = 0.05$ is also insufficient at $\alpha = 0.01$. However, sufficient evidence at a higher level of significance may not always be sufficient at a lower level of significance. Consider Figure 13.1 below.

Figure 13.1. Critical Points



Note that the levels of significance are divided by two in order to apportion the other half to the left tail. This is the case with two tailed tests. For one tailed test, the whole level of significance or probability of Type I error is on one tail.

Three regions are demarcated; A, B and C. Region A falls below the critical point at 5 percent level of significance, $-Z_{\frac{0.05}{2}}$. If the observed statistical fell in this region, we would accept the null hypothesis without hesitance. This is because its acceptance would not be affected by the change in the level of significance to 1 percent. The evidence against the null hypothesis would still be insufficient. Region C falls above the critical point at 1 percent level of significance $-Z_{\frac{0.01}{2}}$. In this region, the null hypothesis is rejected with a high degree of confidence. Again, a change of level of significance does not affect the conclusion.

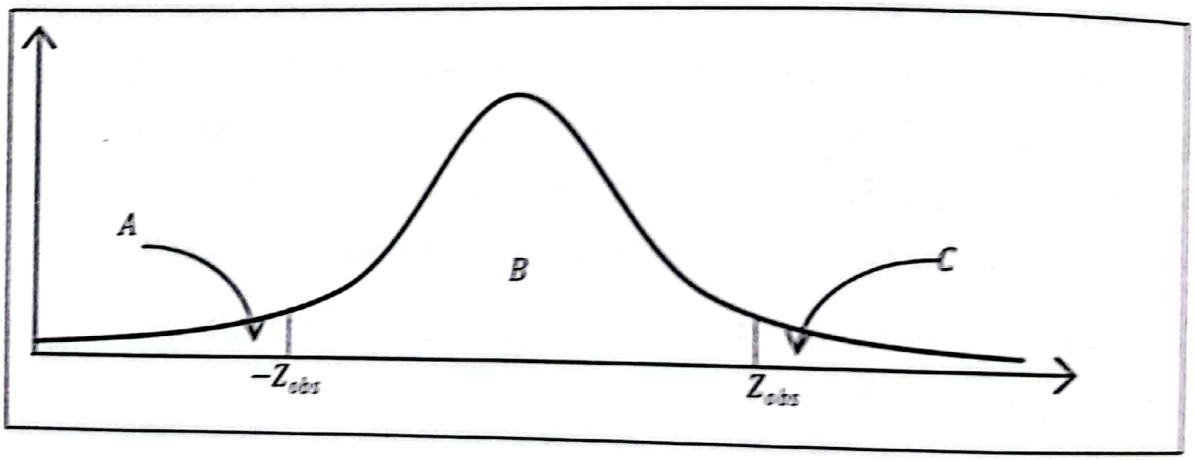
In region B however, the rejection or acceptance will depend on the level of significance used. When the observed statistic falls within region B, the null hypothesis is rejected if $\alpha = 0.05$ is used and is accepted if $\alpha = 0.01$ is used. This region introduces some subjectivity as an investigator is at liberty to alter the level of significance so as to obtain the desired results. In order to eliminate this subjectivity, we shall call this region a *neither reject nor accept* region. Alternatively, it is a *no-decision* region. If the observed statistic falls in this region, we simply recommend for further investigation. The available evidence does not point to either direction, to accept or reject the null

hypothesis. Emphasis must be made that decisions to accept or reject the null hypothesis must be made on the basis of high level of evidence.

13.3 The p-value

In the above method, we are comparing the Z-statistic from the sample and a predetermined level of significance. The two are measured differently, the former being a length and the latter a probability or area. In order to make the comparison possible, the level of significance is transformed into a critical point. Alternatively, we can change the statistic to match the level of significance. This means converting the observed statistic to a corresponding probability which is then compared to the level of significance for decision. The probability corresponding to each statistic is known as the *p-value*. This is the area in the tails. Consider Figure 13.2 below.

Figure 13.2. The p-values



In the above figures, the absolute values of the two observed statistics are the same. One is the negative of the other. We use two values because we want to demonstrate the two-tailed-test case. The two values have divided the area under the standard normal curve into three areas which sum to one. Based on the symmetrical nurture of the standard normal curves, the areas in the two tails are equal, $A = C$.

If

$$Z_{obs} = Z_{\frac{0.05}{2}} = 1.96$$

Then the sum of the two areas would equal the 5 percent, $A + C = 5$. A value of $Z = 1.96$ has 0.025 in the two tails so that the sum is 0.05. This can be searched from the standard normal table. As the Z-value increases from $Z = 1.96$, the area on the tails becomes smaller, it ultimately becomes zero for larger values of Z. Since the larger the Z, the smaller the p-value, then smaller p-values are associated with lower levels of significance.

In particular, a p-value of $p = 0.05$ and $p = 0.01$ means the statistic is significant at 5 percent and 1 percent level of significance respectively. A p-value of $p = 0.03$ is significant at 5 percent level of significance but not at 1 percent. Therefore, in making the decision, the null hypothesis is accepted if $p > 0.05$ and rejected if $p < 0.01$ using a two-critical point rule.

13.4 Testing of hypothesis in large samples

The testing of hypotheses is divided into two categories; large samples and small samples. Two main reasons stand out. First, the reader will recall that with large samples, most variables can be approximated by the normal distribution. As such, we are less worried of the specific distribution of the variable because the large samples will allow the approximation of whatever distribution with the normal distribution. In small samples, however, we must make specific assumptions of the distribution because they cannot be approximated by the normal. Second, we have stated in the earlier sections that often, true parameters are unknown. They are estimated from sample data. While we may be confident of closing in on the true parameter with large samples, the same cannot be said about small samples. Thus, it is important that the test of hypothesis is discussed depending on whether the underlying parameters are estimated based on large enough samples or small samples. This subsection deals with testing of hypothesis in large samples while the latter is dealt with in the succeeding subsection.

In general, a test of hypothesis is conducted in four steps: the statement of the hypotheses, determining the decision rule, the calculation of the test statistic and finally the decision. We discuss these below.

13.4.1 Statement of Hypotheses

Every test must be clear on what is to be tested. There should be a definite statement of possible outcomes from the test. This is the statement of hypotheses. To illustrate, let us carry along the following scenario. Assume that a fertiliser manufacturing company is accused of selling underweight bags. The bags are said to be underweight because they are considered to fall below the declared weight of say 50 kg. The Bureau of Standards, upon receipt of the complaint, wants to investigate the matter. The procedure for conducting this enquiry up to the decision is what we describe in the hypothesis testing.

The standards agency must consider the company innocent of the allegations until sufficient evidence is adduced to the contrary. That is, they must assume that the weight on average, is as declared. This is the null hypothesis. The alternative hypothesis follows the allegations, that the bags are underweight. We have stated only two of three possible outcomes. It is also possible that the bags are greater than the declared weight. This also constitutes a deviation from the declared weight. Does it then become part of the alternative hypothesis?

To answer this question, we need to examine the allegations. The allegations in this scenario are direction specific. It is alleged that the bags are underweight. If the bags are overweight, no one would complain. It would be a bonus to the buyers. So being overweight should be considered part of the null hypothesis. This forms a one sided claim on the deviation of the variable and tested using a one sided or *one tailed* test. If the claim was that the bags are not as declared, either they are overweight or underweight, it would be a two sided or *two tailed* test. We shall see this scenario later. The reader will now understand that the earlier scenario regarding moisture content of maize was also a one sided test.

Taking μ as the true mean weight of the bags, we state the hypotheses as follows.

$$H_0: \mu \geq 50$$
$$H_a: \mu < 50$$

The two hypotheses must be both mutually exclusive and exhaustive. That is, they should not have any region in common nor should any region be unrepresented.

Had the above test been two tailed, the hypotheses would have been

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

The probability of Type I error or level of significance must now be equally apportioned on the two tails. There will be half of the level of significance on each tail. This will result in two critical points, the lower and upper critical points. Because of the symmetrical nature of the standard normal distribution, the two critical points are always of the same absolute value.

13.4.2 Decision Rule

Almost in every sphere, the rules of a game must be stated before the game is played. In hypothesis testing as well, the rule for the decision must be stated before any statistic is calculated. This ensures that rules are not altered in favour of a particular outcome. The decision rule is a rule that will define when the null hypothesis is rejected and when we fail to reject.

In the case of underweight bags, we know that the null is not rejected if the sample evidence shows that it is equal to 50, the declared weight. But will the null be rejected if the sample mean $\bar{x} = 49$? Well, maybe not. The deviation may only be due to sample variation, the true mean still being 50. But it could also be true that the true mean is actually as established from the sample. So the rule must be clear in terms of setting out the rejection and acceptance regions on the scale. Since an acceptable deviation of the sample mean from the declared mean depends on the standard deviations of sample means, the rule must be stated in terms of the standard normal distribution. The standard score Z is given by

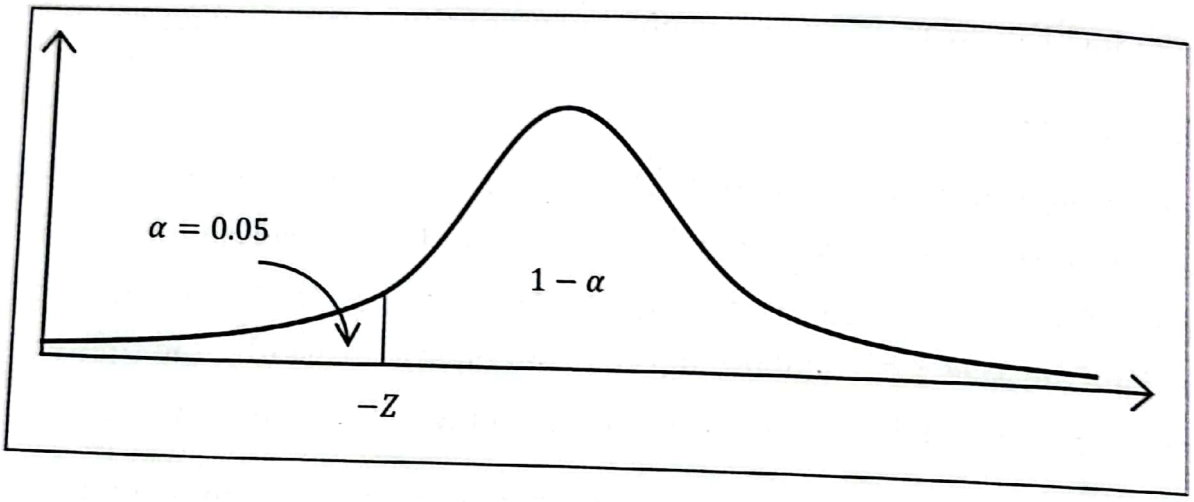
$$Z = \frac{\bar{x} - \mu_0}{se(\bar{x})}$$

Where μ_0 is the presumed true mean. This will be the Z-value observation from the sample. To distinguish it from other Z-values, it is called the observed Z and denoted by

$$Z_{obs}$$

Assuming the level of significance $\alpha = 0.05$, we look for the Z value that has the probability of Type I error equal to the level of significance. Since this is a one sided test, the whole $\alpha = 0.05$ will fall on one tail, particularly the left tail.

Figure 13.3. Decision rule in hypothesis testing



In the figure above, the standard normal value with 5 percent on the lower tail is $Z = -1.645$. This forms the critical or rejection point. It demarcates the rejection and acceptance regions.

The decision rule is: Reject H_0 if $Z_{obs} < -1.645$, otherwise fail to reject it. The next step is to make observations (calculation) which will ultimately lead to a decision depending on how the observed value will compare with the stated rule.

13.4.3 Calculating the test statistic.

Under the decision rule, we stated that the test will be based on a Z-score which will be calculated based on the observed sample mean and standard deviation. At this state, the enquirer can then proceed to make observations. If it was possible to examine all the bags (population) the true parameter would be known and the hypothesis testing rendered irrelevant.

Because this is complex, a decision or conclusion has to be made on the basis of a sample. So, a sample is selected, presumably a large one. Since the sample is large, we do not have to worry whether the standard deviation is population or sample based. Both permit the use of the standard normal distribution.

On the basis of the sample of size $n = 100$, suppose the following statistics are observed: mean $x = 48.2kg$ and standard deviations $= 16$kg. it was stated already that

$$Z_{obs} = \frac{\bar{x} - \mu_0}{se(\bar{x})} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

$$Z_{obs} = \frac{48.2 - 50}{16/\sqrt{100}}$$

$$Z_{obs} = \frac{-1.8}{1.6} = -1.125$$

Since the observed Z-value falls in the acceptance region, we fail to reject the null hypothesis. We conclude that the evidence gathered is not strong enough to conclude that the true weight is actually below 50 kg. Despite the observed sample mean falling below the declared weight, it was not significantly low. We therefore accept the company's claim that the average weight of fertilizer bags is 50 kg.

Example 13.1

A quarrying company supplies aggregates to a road constructing company. The two have agreed that the aggregates be supplied in 25 tonne loads. Due to a busy schedule, the company is not able to weigh each load but relies on the accuracy of the loader. It is therefore believed that on average, each load is 25 tonnes. The quarrying company is worried of any deviation in weight. If the mean weight exceeds 25 tonnes, the company would be losing revenue through oversupply. If the average is found to be lower than 25

tonnes, the company may face litigation. In order to alley these fears, the company hires a quality assurance consultant.

The consultant then reroutes every 10^{th} load to the scale for weighing. The standard deviation of weight is known to be $\sigma = 1.4$. A sample of 7 loads give a mean weight of $\bar{x} = 26.6$. Does the evidence suggest any deviation from 25 tonnes average weight.

The null hypotheses in this case are

$$H_0: \mu = 25$$
$$H_a: \mu \neq 25$$

The test statistic is $Z_{obs} = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$

This is a two tailed test. The critical values at 5 and 1 percent levels of significance are:

$$Z_{0.025} = 1.96$$
$$Z_{0.005} = 2.57$$

We accept the null hypothesis if $Z_{obs} < 1.96$ and reject it if $Z_{obs} > 2.57$. We fail to conclude if the observed value falls between the two critical values.

$$Z_{obs} = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{26.6-25}{1.4/\sqrt{7}} = \frac{1.6}{0.529} = 3.02$$

We conclude that there is strong evidence against the null. Specifically, the trucks are on average overloaded. The loader must be cautioned to pay attention to weight.

When dealing with the difference of two means μ_1 and μ_2 with respective variances ${\sigma_1}^2$ and ${\sigma_2}^2$, the standard error of the difference in respective sample means $d = \mu_1 - \mu_2$ is

$$se(d) = \sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}$$

where n_1 and n_2 are the respective sample sizes. If the two variances are equal, the above expression reduces to

$$se(d) = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Example 13.2

A chicken retailer runs two branches in two towns of Zambia. Each branch is serviced by different chicken producers. The retailer suspects that the chickens from the two suppliers are of different weights. The weight of dressed chickens is known to have a standard deviation of $\sigma = 0.5$. A random sample of 34 chickens from supplier A gives a mean of $\bar{x} = 1.7$ and a random sample of 42 chickens from supplier B has a mean weight of $\bar{x} = 1.9$. Is there evidence from the data that the weight of chickens from the two suppliers is different?

The null hypothesis is that the true means are not different and the alternative alleges that the two are in fact different.

$$H_0: \mu_A - \mu_B = 0$$
$$H_a: \mu_A - \mu_B \neq 0$$

This is a two sided alternative hypothesis and therefore, a two sided test. The critical values are constructed on both tails, each having half the level of significance in the tail. At 5 percent level of significance, the critical value is $Z_{0.025} = 1.96$. On the basis of a symmetrical normal distribution, we reject the null hypothesis is if $Z_{obs} > 1.96$ or $Z_{obs} < -1.96$. We fail to reject the null if the observed statistic falls between the two critical values.

With a common variance in both samples, the test statistic is

$$Z = \frac{\bar{x}_A - \bar{x}_B}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$= \frac{-0.2}{0.5\sqrt{\frac{1}{34}+\frac{1}{42}}}$$
$$= \frac{-0.2}{0.5(0.23)}$$
$$= -1.73$$

Since the observed statistic $Z_{obs} = -1.75$ is neither above $Z_{0.025} = 1.96$ nor below $Z_{0.025} = -1.96$, the null hypothesis is accepted. The conclusion is that there is no sufficient evidence that the true mean weight of the chickens from the two suppliers are different.

The *Z*-tests can also be used to test significance of difference in proportions. A proportion is a relative measure based on a numerator and denominator. It can be expressed as a number between 0 and 1 or in percentage terms. Examples of proportions include the proportion of men or women in a certain grouping, proportion of animals infected with a particular disease, the proportion of output that is defective.

Like in the preceding set of tests, the test of proportions can take two forms: first testing whether an observed proportion is significantly different from a preconceived level and second, testing the significance of difference between two proportions. Let us start with the first case.

In the earlier application of the *Z*-test, we looked at a case where a company has the declared or true mean weight of output. The tasks was to examine whether any observed difference from the held truth is significant to cause change in the way we perceive the commodity. In proportions, we deal with the accuracy of a machine. Virtually all machines are prone to error. They are never 100 percent perfect or safe. What differs, however, is the chance or probability of failure. How often the machine fails or how often it produces defective output. Defining a variable X as the number of particular outcomes in n trials, the proportion of the particular outcomes is given by

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

Using p to denote the true proportion and $\hat{p}$ for the observed or estimated proportion, the test employs the statistic,

$$Z = \frac{\hat{p} - p}{se(\hat{p})}$$

The standard error $se(\hat{p})$ is given by

$$se(\hat{p}) = \sigma = \sqrt{\frac{p(1-p)}{n}}$$

Therefore,

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

If both the numerator and denominator are multiplied by the sample size n, the statistic can be expressed as,

$$Z = \frac{x - np}{\sqrt{np(1-p)}}$$

The statistic is compared to the critical values in order to arrive at a conclusion; to accept the null hypothesis, reject is or fail to make any conclusion.

Example 13.3

A manufacturer of a patent medicine claims that the medicine is 90 percent effective in relieving pain for a period of 5 hours. In a sample 320 people who had pain, the medicine relieved pain for the stated period in 272 people. Determine whether the manufacturer's claim is legitimate.

The hypotheses are:

$$Ho: p = 0.9$$
$$Ha: p < 0.9$$

The critical values of Z statistic at 5 and 1 percent level of significance are

$$Z_{0.05} = 1.645, and\ Z_{0.01} = 2.326$$

The null hypothesis is accepted if $|Z_{obs}| < 1.645$ and rejected if $|Z_{obs}| > 2.326$. We fail to make a conclusion if the observed value falls between the two critical values.

$$Z = \frac{x - np}{\sqrt{np(1-p)}}$$

$$Z = \frac{272 - 320(0.9)}{\sqrt{320 \times 0.9 \times 0.1}}$$

$$Z = \frac{-16}{\sqrt{28.8}} = -2.98$$

The absolute value of the observed statistic is greater than the highest of the two critical points. The deviation is significant even as 1 percent level of significance. There is sufficient evidence to reject the null hypothesis. We conclude that the efficacy of the drug is not as high as claimed.

Example 13.4

An immunisation programme has claimed that it has immunised 70 percent of eligible children with a particular antigen. A random sample of 72 eligible children however shows that only 58 percent are immunised. Is there evidence to suggest that the coverage rate is actually below the claimed 70 percent?

The hypotheses are

$$Ho: p = 0.7$$
$$Ha: p < 0.7$$

As in the preceding example, the two critical values are

$$Z_{0.05} = 1.645, and\ Z_{0.01} = 2.326$$

The null hypothesis will be accepted if $|Z_{obs}| < 1.645$ and rejected if $|Z_{obs}| > 2.326$, otherwise we fail to make a conclusion. Using the alternative expression of Z, we have

$$Z = \frac{0.58 - 0.7}{\sqrt{\frac{0.7(0.3)}{72}}}$$

$$Z = \frac{-0.12}{0.054} = 2.22$$

The observed statistic falls between the two critical values. The available evidence is in the border region. We can neither accept nor reject the null hypothesis. Instead, a further investigation on an expanded sample would be advisable.

The Z test can also be used to test the significance of differences in proportions. Let $\hat{p}_1$ and $\hat{p}_2$ be the sample proportions obtained in large samples of size n_1 and n_2 respectively drawn from respective populations with true proportions p_1 and p_2. Supposed based on sample evidence, we wish to know if two samples with proportions $\hat{p}_1$ and $\hat{p}_2$ come from the same population or two populations of equal proportions, that is, $p_1 = p_2$. The null hypothesis is always on the basis of no difference. The standard error of the difference of two proportions is

$$se(p_1 - p_2) = \sigma_{p_1 - p_2} = \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Where p is the overall proportion or weighted average of the two proportions,

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

Example 13.5

A class of 160 pupils is divided into two equal streams A and B, each handled by a different teacher. A common test is administered to all the pupils to determine whether one stream is better than another.

In the test, 60 pupils from stream A pass the while 52 make it from stream B. Test whether there is a significance difference in the performance of the two streams.

The null hypothesis is that the true pass rates are the same for the two streams. The alternative is that there is a difference, with no specific direction.

$$H_0: p_A = p_B$$
$$H_a: p_A \neq p_B$$

This is a two directional or two tailed test. The test statistic is

$$Z = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

The critical points are

$$Z_{0.025} = 1.96, \text{and} Z_{0.005} = 2.576$$

Given the numbers, the respective proportions are: $p_A = 0.75$ and$p_B = 0.65$. The overall pass rate is the total number of pupils that passed divided the combined total

$$p = \frac{60 + 52}{160} = 0.7$$

Then

$$Z = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$Z = \frac{0.75 - 0.65}{\sqrt{0.7(0.3)\left(\frac{1}{80} + \frac{1}{80}\right)}}$$

$$Z = \frac{0.1}{\sqrt{0.7(0.3)\left(\frac{1}{80} + \frac{1}{80}\right)}}$$

$$Z = \frac{0.1}{0.0724} = 1.38$$

The observed statistic is less than the lower of the two critical points. We therefore accept the null hypothesis and conclude that the performance of the two streams is the same.

13.5 Testing of hypothesis in small samples

When the sample is not large enough, statisticians must worry about the specific distribution of the variable. Often, this is covered by making assumption of the unknown distribution. If the variable is known or assumed to be normally distributed, the size of the sample fails to make a fair estimation of the necessary parameters, the variance. This is dealt with by using the student *t*-distribution as opposed to the standard normal distribution in the tests. Particularly, the *t*-test is used when the variance to be used is based on a small sample. Two cases are worth noting. First, if the collected sample is small but the population variance is available, the Z—test is used. Second, if the population variance is not known but the sample size is large enough, the *Z*-test should be used.

All the steps of testing the hypothesis remain the same as in the *Z*-test.

Example 13.6

A consumer protection unit has complained to a bureau of standards that a named bakery sells underweight bread. In order to investigate the claim, the bureau weighs 15 random samples of bread. The data is given in the table below. Is there evidence that the true mean of the bread is below the labelled weight of $700g$?

650 678 704 688 700 712 687 668 689 715 675 682 710 708 691

The hypotheses for this test are

$$H_0: \mu = 700$$
$$H_a: \mu < 700$$

The test statistic is

$$t = \frac{\bar{x} - \mu}{se(\bar{x})}$$

For a one tailed t-test with 14 degrees of freedom, accept the null if $|t_{obs}| < t_{14,\ 0.05} = 1.761$ and reject it if $|t_{obs}| > t_{14,\ 0.01} = 2.624$.

Based on the weights given above, the reader will confirm that

$$n = 15, \qquad \bar{x} = 690.47g \text{ and } s = 18.24$$

Therefore,

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$= \frac{690.47 - 700}{18.24/\sqrt{15}}$$

$$= \frac{-9.53}{4.71} = -2.02$$

The absolute value of the observed statistic falls between the two critical values. Though there is evidence to doubt the null hypothesis, the evidence isn't strong enough to warrant its rejection. The null is neither accepted nor rejected. In this case, the analyst must recommend a further investigation on the matter as the available evidence is inconclusive.

Example 13.7

A block making company has been operating with one machine. It has now bought a new machine, making identical blocks. However, customers have complained on differences in the weight (and therefore strength) of the blocks from the two machines. The company hires the services of a statistician to investigate this claim. The statistician measures the weights of 25 randomly picked blocks from each machine's output. The following data is generated.

Machine 1: $n = 25,\ \bar{x} = 11.9,\ s = 0.3$

Machine 2: $n = 25,\ \bar{x} = 12.2,\ s = 0.2$

Do the results suggest differences in the strength of blocks from the two machines?

The company as well as the customers have been working on assumptions that the machines are producing identical blocks, and hence equal weight. This forms the null hypothesis. For the alternative, the allegation of difference in weight is not specific on which machine is producing lighter or heavier blocks. It simply says the two are not equal.

Hypotheses

$$H_0: \mu_1 = \mu_2$$
$$H_a: \mu_1 \neq \mu_2$$

Alternatively, the hypotheses can be stated in terms of the difference in mean weights $d = \mu_1 - \mu_2$. In this format

$$H_0: d = 0$$
$$H_a: d \neq 0$$

On the assumption of equal means,

$$d \sim N(0, (\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}))$$

The test statistic is still based on the standard normal distribution. Since the variance is based on estimates from small sample, this is *t*-distribution based test. The test statistic is

$$t = \frac{d - 0}{se(d)} = d \Big/ \sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}$$

At 95 percent ($\alpha = 0.05$) level of significance, we look for the critical value of t with $n_1 + n_2 - 2$ degrees of freedom which have $\frac{0.05}{2}$ probabilties in the upper tail. From the *t*-table,

$$t_{48,\ 0.025} \cong 2.009$$

Corollary, for $\alpha = 0.01$,

$$t_{48,\ 0.005} \cong 2.678$$

With the notion of two critical points, accept the null if $|t_{obs}| < 2.009$ and reject it if $|t_{obs}| > 2.678$. If the observed value falls between the two critical values, recommend for further investigation.

$$t_{obs} = \frac{\mu_1 - \mu_2}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}} = \frac{-0.3}{\sqrt{\frac{0.09}{25} + \frac{0.04}{25}}}$$

$$t_{obs} = \frac{-0.3}{\sqrt{\frac{0.13}{25}}}$$

$$= \frac{-0.3}{0.072} = -4.17$$

Since the absolute value of the observed t is greater than the critical value, we conclude that the two machines produce blocks of different weights. The observed difference in sample estimates are highly significant and are pointing to differences in true means.

Example 13.8

A vocational training regulator wishes to assess the effectiveness of short-courses being advertised by a training institution. The training institution has claimed that the courses enhances the productivity of the trainees, enabling them to more work per unit of time. In order to make the assessment, the regulator selects a random sample of 20 bricklayers and measures the number of blocks each can lay per 8 working hour day. The bricklayers are then administered the short training and the number of blocks each can lay is measured again. The data is given in the table below. You are asked as a statistician to analyse the data and determine whether the training is effective in enhancing productivity.

Id	Pre-training	Post-training	Diff ($\bar{d}$)	$\bar{d}^2$
1	162	198	36	1296
2	189	225	36	1296
3	144	153	9	81
4	198	216	18	324
5	171	144	-27	729
6	216	261	45	2025
7	153	180	27	729
8	189	207	18	324
9	207	171	-36	1296
10	162	180	18	324
11	126	135	9	81
12	144	135	-9	81
13	144	162	18	324
14	171	234	63	3969
15	162	162	0	0
16	180	216	36	1296
17	108	162	54	2916
18	198	225	27	729
19	135	171	36	1296
20	153	144	-9	81
Sum			369	19197

In paired observations, the interest in not in the change in overall mean parse but in the change for each subject. Instead of comparing the two means, we are interested in the difference for each candidate. The task is to test whether the mean difference, $\bar{d}$, are significantly greater than zero. This is a confirmation that the test is effective and that the candidates' outputs increase.

The hypotheses are

$$H_0: \bar{d} = 0$$
$$H_a: \bar{d} > 0$$

The test statistic is:

$$t_{obs} = \frac{\bar{d}}{se(\bar{d})}$$

The critical values for this one-tailed test at 19 degrees of freedom are $t_{19,\ 0.05} = 1.729$ and $t_{19,\ 0.01} = 2.539$. The null hypothesis is accepted if $t_{obs} < 1.729$ and rejected if $t_{obs} > 2.539$. A decision is not reached if the statistic falls between the two critical values.

From the data in the table

$$\bar{d} = \frac{\sum d_i}{n}$$
$$= \frac{369}{20}$$
$$= 18.45$$

$$s_d^2 = \frac{\sum {d_i}^2 - n \cdot \bar{d}^2}{n-1}$$
$$= \frac{19197 - 20(340.4025)}{19}$$
$$= \frac{12388.95}{19} = 652.05$$

$$s = \sqrt{652.05} = 25.535$$

Based on the mean and standard deviation, the t-statistics is

$$t_{obs} = \frac{\bar{d}}{se(\bar{d})} = \frac{18.45}{25.535/\sqrt{20}} = 3.231$$

The t-statistic is highly significant. The null hypothesis is rejected at 1 percent level of significance and we conclude that the short training is effective in increasing the pace of building.

13.6 Other Tests: Chi-square test of goodness of fit,

Hypothesis testing can also be used to test association between variables which may not be quantitative. For instance, Zambia's education system is

divided between public (including mission) and private run schools. This is the case for both secondary schools and tertiary institutions like universities. Suppose the ministry wants to ascertain whether pupils that go to private secondary schools also end in private universities and similarly for public school taught pupils, they end in public university. This may be important in trying to understand the drivers of demand for education.

For this kind of variables, the outcomes are binary or categorical. One either belongs to this group or the other. The question of interest is whether most private school pupils also go to private universities and public school pupils also go to public universities.

Information of this nature is usually presented in contingency tables or cross tabulations. Consider the following table

		University		
		Public	Private	Total
Secondary School	Public	A	B	$A+B$
	Private	C	D	$C+D$
	Total	$A+C$	$B+D$	$TOTAL$

The table shows two variates, nature of secondary school and university, whether they went to a private or government (including mission) owned. Each letter shows the number of persons in each intersection. That is, A, is the number of persons whose secondary and university were both public owned. Letter B is the number that went to public secondary schools but switched to private university. Similarly, C and D represent private school pupils that went to public and private universities respectively.

If the choice of secondary and university are related, we would expect to see some pattern in the numbers. Particularly, numbers A and D would be larger, relative to B and C, as more public school pupils also get into public

university and private school pupils getting into private universities. Alternatively, B and C would be large relative to the other two if private school pupils get into public universities and public school pupils into private universities. This forms expected frequencies of each cell, based on the observed totals for each category. That is, given the numbers in the two categories of school and university, a relationship in the selection of the two will provide a hint on the numbers of each intersection cell. This is referred to as *expected frequency*.

The measure of such relationships is known as *goodness of fit*. The test measures the degree or extent to which observed frequencies fit with the expected frequencies. The test is based on Pearson's chi-square test which uses the chi-square distribution.

We generate a statistic χ which follows a chi-square distribution with $d = (r-1)(c-1)$ degrees of freedom where r and c represent the number of rows and columns respectively.

$$\chi^2 = \sum_{j=1}^{c}\sum_{i=1}^{r}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

In the statistic, $E_{ij} = \frac{Ri \times Cj}{Total}$ (where Ri is the total of the i^{th} row and Cj is the total of the j^{th} column) and O_{ij} are the expected and observed frequencies respectively for the cell in the ith row and jth column

Example 13.9

An Agricultural Extension Officer wants to investigate whether women in his catchment area are more inclined to belong to a cooperative than men. She then randomly interviews 100 people on whether they belonged to a cooperative or not. The data is shown in the table below.

	Membership to a Cooperative			
Gender		Member	Not Member	Total
	Male	23	24	47
	Female	28	25	53
	Total	51	49	100

Is there enough evidence from the data to suggest that women are more likely to belong to associations than men?

On merely looking at the data, more than half the sampled females belong to cooperative while for males, less than half are members. One may be tempted to conclude that yes, more females belong to cooperatives than males. We put this to a test using the chi-square test of goodness of fit. The first step is to calculate the expected values, which are given in parentheses in the table below.

The null hypothesis is that there is no such pattern or trend. Both men and women are equally likely to belong to cooperatives. The alternative hypothesis states that women are more likely to belong to cooperatives than men. As we start the calculations, we bear in mind that the 5 percent critical value of the chi-square distribution with 1 degree of freedom is

$$\chi^2_{1,\ 0.05} = 3.84$$

	Membership to a Cooperative			
Gender		Member	Not Member	Total
	Male	23 (25.5)	27 (24.5)	50
	Female	28 (25.5)	22 (24.5)	50
	Total	51	49	100

The statistic is calculated using the following table

Cell	O_{ij}	E_{ij}	$O_{ij} - E_{ij}$	$(O_{ij} - E_{ij})^2$	$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
A	23	25.5	-2.5	6.25	0.245098
B	27	24.5	2.5	6.25	0.255102
C	28	25.5	2.5	6.25	0.245098
D	22	24.5	-2.5	6.25	0.255102
				Total	1.0004

Since the observed statistic is less than the critical value, we fail to reject the null hypothesis that there is no evidence to suggest that women are more inclined to belong to cooperative than men.

Example 13.10

A study seeks to determine whether alcoholic drivers are more likely to be involved in a road traffic accident or not. A total of 300 randomly selected persons are asked on the state of alcoholism and whether they have been involved in an accident or not. The data is presented in the contingency table below.

	Accident	No-accident	Total
Alcoholic	85	95	180
Non-alcoholic	25	95	120
Total	110	190	300

The null hypothesis is that the two variables are not associated. That is, being alcoholic is not associated with high involvement in road traffic accident. The alternative hypothesis is that the two are associated.

The test for significance of association involves a chi-square test. The table has two rows and two columns, therefore, degrees of freedom $d = 1$. The associated critical value is $\chi^2_{1,\ 0.05} = 3.84$ and $\chi^2_{1,\ 0.01} = 6.635$

The null hypothesis is rejected in favour of the alternative hypothesis if the observed chi-square value is greater than the critical value, $\chi^2_{obs} > 6.635$. Accept the null if $\chi^2_{obs} < 3.84$

Cell	O_{ij}	E_{ij}	$O_{ij} - E_{ij}$	$(O_{ij} - E_{ij})^2$	$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
A	85	66	19	361	5.47
B	95	114	-19	361	3.17
C	25	44	-19	361	8.20
D	95	76	19	361	4.75
	Total				21.59

The observed statistic $\chi^2_{obs} = 21.59$ is far much greater than the critical value. There is overwhelming evidence to reject the null hypothesis and conclude that alcoholism and road accidents are not independent.

13.7 Other Tests: The *F*-test

In the standard normal (and *t*-test), we looked at the significance of difference of mean from an assumed value or a difference of one mean from another. The test is limited to testing the significance of one difference only and as a consequence, it can only be used for no more than two means. The chi-square test also was used to test association between two variates presented in contingency tables. This test is limited to discrete data only, and becomes complex for variables with many values. The two weaknesses in the two tests are overcome by the use of the *F-test*, named after a twentieth century English mathematician Ronald A. Fischer. This is a test based on the *F*-distribution.

The *F*-test is therefore mainly used for two tests: first, it is used to test the equality of means of many groups. Like in the standard normal test for difference of means, the *F*-test can be used to test whether there is a significant difference in means but among many groups. It tests the null hypothesis that all the k-means are not significantly different. That is,

$$\mu_1 = \mu_2 = \cdots = \mu_k$$

Second, the *F*-test is used to test the significance of the regression models, the extent to which they fit the data well. Given a general regression model

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + e_i$$

The test is based on testing the significance of any of the regression coefficients. The null hypothesis holds that all the coefficients are not significantly different from zero. That is,

$$\beta_1 = \beta_2 = \cdots = \beta_k = 0$$

In general, the *F*-test is based on comparing variances. It uses the ratio of two variances and is thus often referred to as analysis of variance (ANOVA).

Consider the data set $x_{ij}, i = 1,2, \ldots n_j,\ j = 1,2 \ldots k$. The second subscript j represents the group where the observation is. Since there are k groups, j will run from 1 to k. The i is the position of the observation within a group. Since the groups are not always of the same size, we allow the highest value of i to be dependent on the group. This is written as n_j, to emphasise that it will vary from group to group.

From the data, means are calculated for each group and denoted by $\bar{x}_j$ and the grand mean denoted by $\bar{x}$. Then the deviation of each observation from the grand mean can be broken into two; its deviation from the respective group mean and the deviation of the group mean from the grand mean. That is,

$$x_{ij} - \bar{x} = (x_{ij} - \bar{x}_j) + (\bar{x}_j - \bar{x})$$

Algebraically, the group mean is included on the right hand side twice, as a positive and negative entrant so that it can effectively cancel out.

Then square and sum both sides of the equation. Since the summation will be in two dimensions, within each group and then summing the group total, it is shown using double summation symbols.

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{x})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2 + \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\bar{x}_j - \bar{x})^2$$

The left hand side of the equation shows the *total* sum of squared deviation from the mean (SST). On the right, the sum is broken down into two. The first shows square differences being taken from the respective group means. This is known as the sum of square differences *within* groups and often denoted by SSW. It shows variations within groups. The second is the square of differences of each group mean from the overall mean. This shows variation across or *between* groups. It is referred to as sum of square differences between groups and denoted by SSB.

Therefore

$$SST = SSW + SSB$$

Where

$$SST = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{x})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}{}^2 - n\bar{x}^2$$

$$SSW = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2$$

$$SSB = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\bar{x}_j - \bar{x})^2 = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{x})^2$$

The SSB will increase with the differences in group means. Thus, a larger SSB is an indication that the differences among or between groups are large and may be significant. However, large SSW points to the fact that large differences may be due to huge variations in the observations. Therefore, the large SSW weakens the significance of the difference of group means. Therefore, the ratio of appropriately adjusted sum of squared differences between and within groups can provide a test for the significance of the differences. This is known as the *F*-test.

The adjustment of the SSB and SSW makes use of the degrees of freedom. These are defined as $k - 1$ and $n - k$ for SSB and SSW respectively where k is the number of groups and n is the overall sample size. Therefore, the statistic is

$$F = \frac{SSB/k-1}{SSW/n-k}$$

This is said to follow an F-distribution with $d_1 = k-1$ and $d_2 = n-k$ degrees of freedom. This is often summarized in an ANOVA table below.

Source	Sum of Squares	Deg of freedom	Mean squares	F-statistic
Between	SSB	k-1	$MSB = SSB/k-1$	$F = \frac{MSB}{MSW}$
Within	SSW	n-k	$MSW = SSW/n-k$	
Total	SST	n-1		

If the observed statistic is greater than the critical value, $F_{d_1,d_2,\alpha}$, we reject the null hypothesis of no difference in means. If the observed value is less than the critical value, then the null hypothesis will hold true.

Example 13.11

A District education authority wants to determine whether there are differences in the standards of teaching at three schools measured by pupil performance. A mock examination is administered to 25 randomly selected Grade 12 pupils in each of the three schools; A, B and C. The following information emerges

$$\sum_{i=1}^{50} x_{iA}{}^2 = 90467, \qquad \bar{x}_A = 59\%,$$

$$\sum_{i=1}^{50} x_{iB}{}^2 = 82400, \qquad \bar{x}_B = 56\%$$

$$\sum_{i=1}^{50} x_{iC}{}^2 = 117481, \qquad \bar{x}_C = 67\%$$

Do the results support the notion that pupil's performances at the three schools are different?

In this problem, the null hypothesis is that there is no difference. That is, all the means are the same. In the alternative hypothesis, we hold the view that not all the means are the same.

The gathered data is organised in three groups or schools ($k = 3$) and there is a total of 75 observations ($n = 75$). Using the 5 percent and 1 percent levels of significance, the critical value is

$$F_{2,\ 72,\ 0.05} = 3.12$$
$$F_{2,\ 72,\ 0.01} = 4.91$$

In order to calculate the observed statistic, we have to evaluate the sum of squares. The reader will know that the overall mean is the average (weighted) of the three group averages. Therefore $\bar{x} = 60.67$.

$$SST = \sum_{j=1}^{k}\left(\sum_{i=1}^{n_j} {x_{ij}}^2\right) - n\bar{x}^2$$
$$= 90467 + 82400 + 117481 - 75(60.67)^2$$
$$= 290348 - 276033.3$$
$$= 14314.67$$

$$SSB = \sum_{j=1}^{k} n_j\left(\bar{x}_j - \bar{x}\right)^2$$
$$= 25[(59 - 60.67)^2 + (56 - 60.76)^2 + (67 - 60.67)^2]$$
$$= 25[(59 - 60.67)^2 + (56 - 60.76)^2 + (67 - 60.67)^2]$$
$$= 25[64.67] = 1616.67$$

$$SSW = SST - SSB$$
$$= 14314.67 - 1616.67 = 12698$$

This completes all the information required for the ANOVA table.

Source	Sum of Squares	Degrees of freedom	Mean squares	F-statistic
Between	1616.67	2	808.3	4.58
Within	12698	72	176.4	
Total	14314.67	74		

The observed statistic $F_{obs} = 4.58$ falls between the two critical points. There is sufficient evidence to reject the notion that all the means are the same at 5 percent level fo significance. However, the evidence is insufficient at 1 percent level of significance. We therefore fail to arrive at a conclusion or the evidence is not conclusive.

Example 13.12

A company runs four branches in different malls in the city. It intends to assess the performance of the branch managers by looking at the volume of sales per day. Data is collected on the volume of sales for a selected week as follows.

Day	1	2	3	4	5	Mean
A	12	11	9	11	10	10
B	11	11	13	14	12	12.2
C	10	9	9	11	10	9.8
D	10	10	12	11	13	11.2

The null hypothesis is that the true mean sales is the same for all the branches. That is

$$H_0: \mu_A = \mu_B = \mu_C = \mu_D$$

The alternative hypothesis is the opposite of the null, that is, not all the means are equal.

The 5 percent and 1 percent levels of significance critical values are

$$F_{3,\ 16,\ 0.05} = 3.24$$

$$F_{3,\ 16,\ 0.01} = 5.29$$

The null hypothesis will be accepted if $F_{obs} < 3.24$ and rejected if $F_{obs} > 5.29$. The evidence will be inconclusive if $F_{obs} \in (3.24,5.29)$.

For the statistic, we need to evaluate the SSB and SSW. The overall mean sales is the average of the four branch sales

$$\bar{x} = \frac{\sum \bar{x}_j}{k} = \frac{43.2}{4} = 10.8$$

For SSB

$$SSB = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{x})^2$$

$$= (10 - 10.8)^2 + (12.2 - 10.8)^2 + (9.8 - 10.8)^2 + (11.2 - 10.8)^2$$

$$= (-0.8)^2 + (1.4)^2 + (-1.0)^2 + (0.4)^2$$

$$= 3.76$$

For SSW

Day	1	2	3	4	5	Sum
$(x_{iA} - \bar{x}_A)^2$	4	1	4	1	0	10
$(x_{iB} - \bar{x}_B)^2$	1.44	1.44	0.64	3.24	0.04	6.8
$(x_{iC} - \bar{x}_C)^2$	0.04	0.64	0.64	1.44	0.04	2.8
$(x_{iD} - \bar{x}_D)^2$	1.44	1.44	0.64	0.04	3.24	6.8

$$SSW = 10 + 6.8 + 2.8 + 6.8$$

$$= 26.4$$

The anova table will then be.

Source	Sum of Squares	Degrees of freedom	Mean squares	F-statistic
Between	3.76	3	1.2533	0.76
Within	26.4	16	1.65	
Total	30.16	19		

The observed statistic $F_{obs} = 0.76$ falls in the acceptance region. We therefore accept the null hypothesis and conclude that there is no

sufficient evidence to suggest that the true mean sales for each of the four branches are different.

As mentioned already, the *F*-test can also be used to test the significance of the overall regression model. In this usage, k is the number of parameters in the model including the constant and n remains the number of observation. Assume a general regression model given by,

$$y_i = \alpha + \beta x_i + e_i$$

Where α and β are regression parameters and e_i is the error term.

We stated under regression analysis that the deviation of the dependent variable from its average can be divided into two.

$$\begin{aligned} y_i - \bar{y} &= \beta(x_i - \bar{x}) + e_i \\ &= (\hat{y}_i - \bar{y}) + e_i \end{aligned}$$

If we square on both sides and sum, we get

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} {e_i}^2$$

This assumes independence of the error term from the explained variable so that $\sum(\hat{y}_i - \bar{y})e_i = 0$. This is rewritten as

$$SST = ESS + RSS$$

The ESS is the explained sum of square deviations. This is the variation in the dependent variable that is explained by the model. RSS is the residual sum of squared deviations or the sum of variations not explained by the model. It is attributed to random variation in the variable. The statistic is then defined as follows.

$$F = \frac{ESS/k-1}{RSS/n-k}$$

This follows an *F*-distribution with $d_1 = k - 1$ and $d_2 = n - k$ numerator and denominator degrees of freedom respectively.

Example 13.13

Gloria tries to estimate the effect of education (measured by number of years of schooling completed) and experience (measured by years in employment) on the level of earnings. She then runs the following regression on 30 observations.

$$y_i = \alpha + \beta x_{1i} + \delta x_{2i} + e_i$$

The statistical package used for the analysis produces an F-value of $F_{obs} = 2.83$. How should Gloria interpret the results?

The null hypothesis in this test states that all the model coefficients are not significantly different from zero. The alternative is that not all coefficients are identically zero.

The model has 3 parameters and based on a sample of 30 observations. Hence, $k = 3$ and $n = 30$. The critical value at 5 percent and 1 percent level of significance with 2 and 27 degrees of freedom are

$$F_{2,\ 27,\ 0.05} = 3.35$$
$$F_{2,\ 27,\ 0.01} = 5.48$$

Since observed value falls in the acceptance, we accept the null hypothesis that all the parameter of the model are zero. Therefore, the model is insignificant and will explain little variation in the dependent variable.

At times, we are interested to test whether two samples are drawn from the same distribution. This is made on the basis of equality of variances. Two samples are drawn from populations with variances ${\sigma_1}^2$ and ${\sigma_2}^2$ respectively. The true variances are not known but estates are available from the sample. That is, the samples have variances ${s_1}^2$ and ${s_2}^2$ respectively. The question then is, 'are the true variances different?' this is referred to as difference in variances. Given the two variances from two samples of sizes $n_1 - 1$ and $n_2 - 1$ respectively, the statistic $F = \frac{{s_1}^2}{{s_2}^2}$ follows an F-distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom. That is

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

Using the above theory, critical points can be formulated to test the equality of the two variances. If the two variances are equal, their ratio is one. Therefore, the deviation of the ratio from one is a measure of the difference between the two. The null hypothesis takes the form

$${\sigma_1}^2 = {\sigma_2}^2$$

The test can be two tailed or one tailed depending on how the alternative hypothesis is stated. If the alternative hypothesis only emphasises on the difference, the test is two tailed. Such case apply if all we are interested is to see if the variances are significantly different. If, however, the alternative hypothesis is specific on the direction of the difference, then a one tailed test is used. This applies when the test is to confirm that one particular sample has less variation than another. The reader will note that in a one tailed test, the test that:

$${\sigma_1}^2 > {\sigma_2}^2 \ which\ implies\ \frac{{\sigma_1}^2}{{\sigma_2}^2} > 1$$

Is the same as the test that

$${\sigma_2}^2 < {\sigma_1}^2 \ which\ implies\ \frac{{\sigma_2}^2}{{\sigma_1}^2} < 1$$

The two above tests are the same and will generate the same conclusion as long as the order of the degrees of freedom as swapped accordingly.

Example 13.14

A processing company is contemplating on buying one of the two brands of packaging machines being promoted at a trade fair. The company is satisfied that though there are variations in actual quantities for the two machines, the true means for both machines is always to the set level. That is, is set to measure 50kgs, the true means will be 50 kgs. The company is worried that huge variations in actual quantities may dent the reputation of the company. It therefore decides to go for the machine with the lowest variance. A

sample of 32 units of output from machine A has a variance of $s_A{}^2 = 12.4$ and a sample of 29 units of output from machine B has a variance of $s_B{}^2 = 10.3$. You are requested to certify whether the second machine has less variability in its output.

In claim in this enquiry is one sided, that the variance of machine B is less than the variance of machine A. this gives the hypotheses

$$H_0 : \sigma_A{}^2 = \sigma_B{}^2$$
$$H_a : \sigma_A{}^2 > \sigma_B{}^2$$

Alternatively, the hypotheses are stated as

$$H_0 : \frac{\sigma_A{}^2}{\sigma_B{}^2} = 1$$
$$H_a : \frac{\sigma_A{}^2}{\sigma_B{}^2} > 1$$

The latter format is in line with the test statistic. $F = \frac{s_A{}^2}{s_B{}^2}$ which on the assumptions of the null hypothesis, follows an F-distribution. The critical values for the F-distribution with $n_A - 1 = 31$ and $n_B - 1 = 28$ degrees freedom at 5 percent and 1 percent level of significance are:

$$F_{0.05,\ 31,\ 28} = 1.86$$

$$F_{0.01,31,\ 28} = 2.43$$

The null hypothesis will be accepted if $F_{obs} < 1.86$ and rejected if $F_{obs} > 2.43$. No decision will be made if $1.86 \leq F_{obs} \leq 2.43$.

The observed statistic then is

$$F_{obs} = \frac{s_A{}^2}{s_B{}^2} = \frac{12.4}{10.3} = 1.2$$

Since the observed statistic $F_{obs} < 1.86$, we accept the null hypothesis that there is no difference in the variations caused by the two machines.

CHAPTER 14

14 NON-PARAMETRIC TESTS

These are techniques that do not rely on data belonging to any particular distribution. They are 'distribution-free' methods. Even when parametric methods can be used, non-parametric methods may be used because of their simplicity and robustness. But this is at a cost: where parametric methods are appropriate, non-parametric methods have less *power*. A large sample would be required to draw conclusions with the same level of confidence. There is a large number of non-parametric tests that are available. They include: Wilcoxon test; the Mann-Whitney-Wilcoxon U test, Anderson-Darling test, Kruskall-Wallis Anova by ranks test, Kendall's coefficient of concordanceand the famous Spearman's rank correlation. To every parametric test, there are at least one corresponding non-parametric test. As much as possible, we must use parametric test.

Suppose the distribution is non-normal. Sometimes normality could be achieved through data manipulations. This can be done through, for example: deleting outliers, winsorising data or trimming the data. We explain the three methods below:

14.1.1 Deleting outliers:

This method relies on ability to identify observations that make the data non-normal which are called outliers. Outliers make the distribution skewed in the direction of outliers. If these are identified and deleted, the remaining data will be closer to normal. For instance, consider the following data set $X = 1,2,3,3,4,4,4,5,6,18$ which has an average of $\bar{x} = 5$. The distribution is clearly skewed because of one outlier value. If the 18 is deleted, the remaining distribution $X = 1,2,3,3,4,4,4,5,6$ will be near-normal with mean $\bar{x} = 3.55$. Caution however must be exercised that the number of deleted 'outliers' do not amount to 5 percent of the observations.

14.1.2 Trimming data

An alternative to deleting outliers only is to chop the tails of the distribution. This minimises the shift in the central measure in highly skewed distribution. For instance, given the same observations $X = 1,2,3,3,4,4,4,5,6,18$, trimming involves cutting off the lowest and highest. With this, we remain with the variable $X = 2,3,3,4,4,4,5,6$ with a mean of $\bar{x} = 3.87$

14.1.3 Winsorising data

Instead of trimming and losing the outliers, winsorising involves recomputing them so that the number of observations remains unchanged. Winsorising, named after a twentieth century biostatistician Charles P. Winsor, involves changing the values of the outliers as opposed to ridding them. For instance, a 90 person winsorisation will maintain the middle 90 percent of the observations but change all the lower 5 percent to the 5^{th} percentile value and the upper 5 percent to the 95^{th} percentile value. Given any dataset, a desirable winsorising percent is chosen which determines the boundaries of the resultant dataset. Then outliers falling below are equated to the new minimum and outliers falling above, to the new maximum. Consider the data set $X = 1,2,3,3,4,4,4,5,6,18$. There are 10 observations. The 90 percent winsorisation will maintain the mid 9 observations, that is 2,3,3,4,4,4,5,6. The outlier on the left is then recomputed to 2 and the outlier on the right recomputed to 6. The Winsorised data is $X = 2,2,3,3,4,4,4,5,6,6$ with mean $\bar{x} = 3.9$.

14.2 Use of Non-parametric tests.

While parametric test remain useful and appealing, their use also has limitations. In these cases, non-parametric tests prove to be useful alternatives. In particular, non-parametric test are particularly advantageous in the following situations:

- Sample size is very small
- Data are inherently in ranks
- Data are simply classificatory or categorical

- Samples are made up of observations from several different populations.

14.3 Application of Non-parametric tests:

14.3.1 Wilcoxon Signed-Rank Test

Suppose there are n pairs of observations (x_i, y_i) and we wish to test the hypothesis that the two variables, x and y, have the same distribution. This implies having the same parameters, the mean and variance. That is $E(x) = E(y) = \mu$ and $Var(x) = Var(y) = \sigma^2$. In this test, the null hypothesis is states that the two variables come from the same population and hence have the same distribution and that the differences follow a symmetric distribution around the zero mean. That alternative is that the differences do not have a zero mean.

We then define two variables. The first is the absolute value of the difference between the two. That is, let D_i be

$$D_i = |x_i - y_i|$$

The second is the actual sign of the difference. Let the sign S_i

$$S_i = sign(x_i - y_i)$$

where $sign$ is the sign function which assigns 1 for a positive difference and -1 for a negative difference. No sign will be assigned for zero difference. Such pairs can as well be deleted since they are immaterial in the test. This reduces the number of pairs to say n^* The two steps have split the actual difference into the absolute number D_i and the sign S_i.

Then rank the data according to the ascending order of the absolute difference D_i and assign ranks R_i, starting with the smallest. If there is a tie, an average of the ranks involved is assigned to the tied values. The next step involves attaching the signs to the ranks. This is the basis of the name of the test, the signed rank test because the ranks are given the sign of the difference. In other words, the ranks which are ordinarily positive, get signed to positive or negative. The test statistic W, is the sum of the signed ranks.

$$W = \sum_{i=1}^{n^*} S_i R_i$$

Where n^* is the number of remaining pairs after removing pairs with zero differences. The statistic is compared to the critical values from the Wilcoxon Signed Rank Test table for a decision. The null hypothesis is rejected if $W > W_\alpha$, otherwise we fail to reject.

Example 14.1

A group of 30 factory workers are asked to express their concern for job security and job pay on a rating scale from 1 (no concern) to 10 (ultimate concern). The results are shown in the table below.

Id	Security (x)	Pay (y)	Diff D	Sign S	Rank R	SR
1	3	4	1	-1	5.5	-5.5
2	1	9	8	-1	25	-25
3	6	8	2	-1	14	-14
4	7	5	2	1	14	14
5	6	4	2	1	14	14
6	5	6	1	-1	5.5	-5.5
7	3	6	3	-1	19.5	-19.5
9	8	7	1	1	5.5	5.5
10	3	5	2	-1	14	-14
11	4	6	2	-1	14	-14
13	6	7	1	-1	5.5	-5.5
14	3	6	3	-1	19.5	-19.5
16	5	6	1	-1	5.5	-5.5
17	4	5	1	-1	5.5	-5.5
18	4	3	1	1	5.5	5.5
19	6	5	1	1	5.5	5.5
20	6	7	1	-1	5.5	-5.5
21	1	6	5	-1	24	-24
22	7	3	4	1	22.5	22.5
23	4	6	2	-1	14	-14
24	5	6	1	-1	5.5	-5.5
26	2	5	3	-1	19.5	-19.5
28	4	6	2	-1	14	-14
29	3	6	3	-1	19.5	-19.5
30	2	6	4	-1	22.5	-22.5
Sum						**-191**

On the basis of this data, can we claim that workers are more concerned with one over the other?

The first step to calculating the Wilcoxon statistic is to calculate the absolute differences and signs. Then delete pairs with zero

differences and arrange the remaining in ascending order. The above distribution has five (5) zero differences. We delete these and remain with 25 observations, $n^* = 25$. With this number of observations, the critical value of the statistic at 5 and 1 percent level of significance are: $W_{25,\ 0.05} = 89$ and $W_{25,\ 0.01} = 68$.

Id	Security (x)	Pay (y)	Diff D	Sign S	Rank R	SR
1	3	4	1	-1	5.5	-5.5
6	5	6	1	-1	5.5	-5.5
9	8	7	1	1	5.5	5.5
13	6	7	1	-1	5.5	-5.5
16	5	6	1	-1	5.5	-5.5
17	4	5	1	-1	5.5	-5.5
18	4	3	1	1	5.5	5.5
19	6	5	1	1	5.5	5.5
20	6	7	1	-1	5.5	-5.5
24	5	6	1	-1	5.5	-5.5
3	6	8	2	-1	14	-14
4	7	5	2	1	14	14
5	6	4	2	1	14	14
10	3	5	2	-1	14	-14
11	4	6	2	-1	14	-14
23	4	6	2	-1	14	-14
28	4	6	2	-1	14	-14
7	3	6	3	-1	19.5	-19.5
14	3	6	3	-1	19.5	-19.5
26	2	5	3	-1	19.5	-19.5
29	3	6	3	-1	19.5	-19.5
22	7	3	4	1	22.5	22.5
30	2	6	4	-1	22.5	-22.5
21	1	6	5	-1	24	-24
2	1	9	8	-1	25	-25
Sum						**-191**

In the table above, the observed statistic is $W_{obs} = -191$. The absolute value of the observed statistic is greater than the critical value even at 1 percent level of significant. Therefore, there is

overwhelming evidence from the data that the workers are not equally concerned between job security and pay. In particular, the evidence is suggesting that the workers are less concerned about job security but more concerned about the pay.

14.3.2 Mann-Whitney-Wilcoxon test

Take two independent samples A and B of sizes n_A and n_B so that there is a total of $n = n_A + n_B$ observations. We wish to test whether the two samples are from the same distribution. The null hypothesisholds that the two samples have the same distribution. The alternative will argue that the two samples are not of the same distribution. Here are the steps for the test.

Whilst keeping the identity of the observations as to which sample they originated, rank all the n observations in an ascending order. For each value from sample B, count the number of observations from sample A that precede it. That is, for each observation b_i, how many a_i's come before it. Then sum the numbers and denote this as U_A. It is based on the number of observations of A that precede each observation in B. Repeat the process with the two swapped. Then we have the number of observation from B that precede each observation from A and the sum is denoted by U_B.

The test statistic U is the minimum between U_A and U_B. That is,

$$U = \min(U_A, U_B)$$

The null hypothesis is rejected, in favour of the alternative, if the observed U_{obs} is less than the critical value at a given level of significance. Notice that the value of the statistic reduces with strong evidence against the null hypothesis. It is smaller values to point to the rejection of the null and larger values point to acceptance of the null hypothesis.

Example 14.2

The bacteria count per unit volume for two types of cultures A and B are shown below.

A	B
27	32
31	29
26	35
25	28

This gives 8 observations. When ranked in ascending order, we get the following:

Value	25	26	27	28	29	31	32	35
Sample	A	A	A	B	B	A	B	B

For each observation from B, the number of observations from A preceding it are as shown in the first row of the following table. The number of observation from B preceding each observation from A are shown in the second row. Each row has the sum in the last column.

U_A	3	3	4	4	=14
U_B	0	0	0	2	=2

Then

$$U = \min(U_A, U_B) = 2$$

14.3.3 Kruskall-Wallis test

The Kruskall-Wallis test by ranks is an extension of the Mann-Whitney-Wilcoxon test in many samples. The test is named after its pioneers, the two American statisticians William Henry Kruskall and Wilson Allen Wallis. It is used to test differences in medians of many samples which may or may not be of the same size. Its parametric equivalent is the analysis of variance based on sample means. In this test, the null hypothesis holds that all the samples are drawn from the same population and therefore have the same true median. This is equivalent to assuming equality in means under the analysis of variance. The alternative then states that at least one sample median is different from median of another sample. Note that the

alternative does not imply differences in all the sample medians but there is at least a pair with different medians.

In order to use the Kruskall-Wallis K test, the some assumptions about the data are made. That is:

- All the samples used are random and independent
- There are 5 or more observations in each sample
- The observations can be ranked

To illustrate the test, assume there are h-samples, each of size $n_j, j = 1,2, \ldots h$. There is a total of N observations, where

$$N = \sum_{j=1}^{h} n_j$$

Put all the samples together, noting the sample they come from. Arrange the grand sample in order of magnitude and assign ranks. Thus, r_{ij} is the rank of the ith observation from the jth sample. If there is a tie, an average of ranks the observations would have been assigned if not for a tie is assigned to all the tied observations. Then for each sample, find the average ranks of the members. That is

$$\bar{r}_j = \frac{\sum_{i=1}^{n_j} r_{ij}}{n_j}$$

This is the mean rank for each sample. The average of all the ranks $\bar{r}$ is given by

$$\bar{r} = \frac{\sum_{j=1}^{N} \sum_{i=1}^{n_j} r_{ij}}{N} = \frac{N+1}{2}$$

The test statistic k is then defined as

$$K = \frac{(N-1)\sum n_j (\bar{r}_j - \bar{r})^2}{\sum\sum (r_{ij} - \bar{r})^2}$$

The formula can also be expressed as

$$K = (N-1)\frac{\sum n_j \bar{r}_j^2 - N\bar{r}^2}{\sum\sum r_{ij}^2 - N\bar{r}^2}$$

The statistic K is assumed to follow a chi-square distribution with $(h-1)$ degrees of freedom. Depending on how the statistic compares with the critical value $\chi^2_{h-1,\ \alpha}$, we can reject or fail to reject the null hypothesis. Alternatively, the p-value can be computed based on

$$p = \Pr(\chi^2_{h-1} > K)$$

Example 14.3

An agriculturalist wants test whether there is difference in the performance of three maize seed varieties in a particular geographical location. She then enumerates beneficiaries of the farmer input support programme in which each farmer receives a voucher worth four bags of fertilizer and a 10kg maize seed of one's choice. There are only three maize varieties on the market and each farmer must use all the fertilizer on the 10kg seed. The agriculturist has a list of 15 farmers and the data on yield is given below.

Variety	Yield				
A	60	70	55	88	101
B	53	43	65	65	63
C	75	85	84	80	93

Use the Kruskall-Wallis test to determine whether there is a significant difference in the performance of the three varieties of seed.

The first step in the Kruskall-Wallis test is to assign ranks to observations. In the table below, each observation is replaced by its rank and the corresponding group average ranks calculated.

Variety	Rank r_{ij}					Mean $\bar{r}_j$
A	4	8	3	13	15	8.6
B	2	1	6.5	6.5	5	4.2
C	9	12	11	10	14	11.2

The overall average rank is

$$\bar{r} = \frac{N+1}{2} = \frac{15+1}{2} = 8$$

The K-statistic is

$$K = (N-1)\frac{\sum n_j \bar{r}_j^2 - N\bar{r}^2}{\sum\sum r_{ij}^2 - N\bar{r}^2}$$

$$= (15-1)\frac{[(5*8.6^2)+(5*4.2^2)+(5*11.2^2)] - 15*8^2}{1239.5 - 15*8^2}$$

$$= 14\frac{125.2}{279.5}$$

$$K = 6.27$$

The chi-square critical value at 5 percent levels of significance and 14 degrees of freedom is 23.69. The observed K-statistic is way below the critical value. Therefore, the null hypothesis is accepted. There is no evidence to suggest that the performance of the varieties are different.

14.3.4 Kendall's Coefficient of Concordance

Kendall's W also known as Kendall's coefficient of concordance is based on comparing rankings of many sets. At an exhibition such as the Agriculture and Commercial Show of Zambia in Lusaka or the Trade fair in Ndola, a lot of exhibitors will compete for various prices. Other application are in selecting say the best performer in a game. All these are based on subjective perceptions of those chosen as judges. Because the choice can be subjective, more than one judge is often preferred. Assume now there are n exhibitors

or competitors and the winder is being decided by a panels of m judges. Each judge will assign a rank between 1 and n to each participant.

The level of confidence in the selection of winners will depend on how the choices of individual judges are in consonance. There should be some resemblance, abet not perfect, in the way the judges will assign ranks to different participants. For instance, participants will have a high level of trust in the selection if all judges give same ranking to each participant. In this case, there is perfect concordance or agreement in ranking. If however, the rankings are completely different, very independent, the competence of judges is then brought into question as well as the credibility of the rankings. In reality however, the rankings will not be at any of the two extreme cases. There is often some agreement on some rankings but not perfectly so. At what point then does one conclude that there is sufficient concordance to accept the results or when are results discredited because of lack of concordance?

The Kendall's W test is a test of concordance or agreement in the ranks. Suppose now that r_{ij} is the rank assigned to the ith participant by the jth judge. With n participants and m judges, there is a total of $g = mn$ rankings. Define R_i as the total rank or summation of ranks assigned to the ith participant.

$$R_i = \sum_{j=1}^{m} r_{ij}$$

There will be n such total ranks, one corresponding to each participant. The average across participants is then given as

$$\bar{R} = \frac{1}{n}\sum_{i=1}^{n} R_i$$

When there is a high level of agreement in the rankings, the variations in the total ranks will also be high. That is, because of consistence in the assignment of ranks, the lowly ranked will have a lower total because the summation will involve roughly all small numbers. The corollary applies to

the poorly or 'highly' ranked. They have big numbers leading to a high sum. However, when there is complete lack of consonance in the assignment of ranks, participants will get low ranks from some judges and high ranks from some. At the end, the sum of one's ranks, which we are calling the total rank, will not be very different implying 'almost-same-average'. Therefore, the variance of the total sum will indicate the level of concordance or agreement in the assignment of ranks. A high variance points to consistence in ranks while a lower variance is indicative of independence or lack of concordance in the assignment of ranks.

Kendall's W statistic uses the variance in total ranks to measure the level of concordance in ranks. The statistic is defined as

$$W = \frac{12}{m^2(n^2-1)} \times \frac{1}{n}\sum_{i=1}^{n}(R_i - \bar{R})^2$$

The reader will note that formula is basically in two parts. The part on the right is the variance in total ranks. As mentioned already, the variance will be higher with higher levels of concordance and vice-versa. The weakness with the variance only is the lack of benchmarks. It is not possible to tell when it is high or when it is low because there is nothing to compare with.

The part on the left is added so that the product will range between zero and one. That is $0 \leq W \leq 1$. The value of $W = 1$ means perfect consistence in the way ranks are assigned. It means all the judges will have been unanimous in the assignment of ranks to participants. A value of $W = 0$ on the other hand indicate a complete lack of semblance in ranks. It may suggest that ranks we assigned randomly and not guided by the performance of participants. Typically, the value of W will lie between the two extreme cases. With the maximum and minimum known, a researcher will find it easy to comment on whether the statistic is too low, low, fair, high or very high. This reader should note that this is not a test, but a relative measure of agreement in the assignment of ranks.

Example:

During the annual Trade fair in Ndola, six exhibitors are competing in a particular categories. There are three judges, each required to assign ranks, one for the best exhibitor, two for the second best and so on to the sixth. The exhibitor with the lowest total score wins the prize. After a tour by judges, the ranks are as below.

Exhibitor	judge 1	judge 2	judge 3
A	1	3	4
B	5	4	3
C	4	1	5
D	3	2	1
E	6	6	2
F	2	5	6

Determine the level of concordance.

To determine the level of concordance among judges, we use Kendall's W test which requires calculating the total ranks. In the table below, the total ranks are added in the last column.

Exhibitor	judge 1	judge 2	judge 3	R_j	$(R_j - \bar{R})^2$
A	1	3	4	8	6.25
B	5	4	3	12	2.25
C	4	1	5	10	0.25
D	3	2	1	6	20.25
E	6	6	2	14	12.25
F	2	5	6	13	6.25
Sum				63	47.5

Based on the above ranking, exhibitors D and A scoop the first and second positions respectively, while C and B follow at third and fourth positions respectively.

The mean ranks

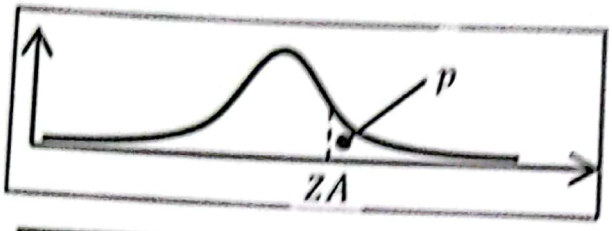
$$\bar{R} = \frac{1}{n}\sum_{i=1}^{n} R_i = \frac{63}{6} = 10.5$$

The W-statistic is

$$W = \frac{12}{m^2(n^2-1)} \times \frac{1}{n}\sum_{i=1}^{n}(R_i - \bar{R})^2 = \frac{12}{9(36-1)}\frac{47.5}{6} = \frac{570}{1890}$$
$$= 0.30$$

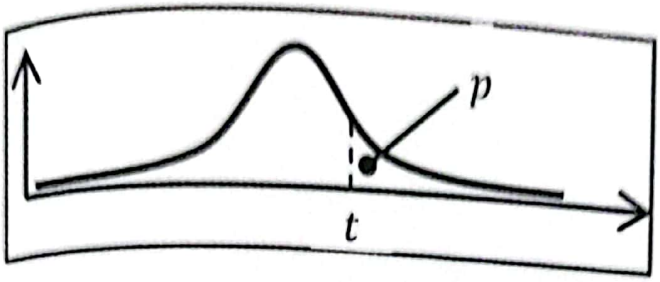
Even though a decision on the winner is arrived at, there is need to consider the degree of unanimity in the ranking. The value of $W = 0.3$ is in the lower half. There was a lesser degree of unanimity in the way ranks were assigned.

Standard Normal Distribution



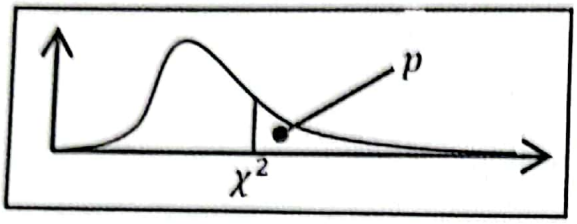
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003

Critical values of the Student t-distribution



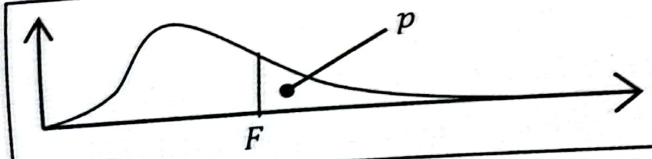
	Upper-tail probability p								
df	0.2500	0.1000	0.0500	0.0250	0.0100	0.0050	0.0025	0.0010	0.0005
1	1.000	3.078	6.314	12.706	31.821	63.66	127.3	318.3	636.6
2	0.816	1.886	2.920	4.303	6.965	9.925	14.09	22.33	31.60
3	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.21	12.92
4	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24	0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
60	0.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
90	0.677	1.291	1.662	1.987	2.368	2.632	2.878	3.183	3.402
120	0.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	0.675	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

Critical values of the Chi-square distribution



χ^2	Upper-tail probability p										
df	0.995	0.990	0.975	0.950	0.900	0.500	0.100	0.050	0.025	0.010	0.005
1	0.000	0.000	0.001	0.004	0.016	0.455	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	1.386	4.605	5.991	7.378	9.210	10.60
3	0.072	0.115	0.216	0.352	0.584	2.366	6.251	7.815	9.348	11.34	12.84
4	0.207	0.297	0.484	0.711	1.064	3.357	7.78	9.49	11.14	13.28	14.86
5	0.412	0.554	0.831	1.145	1.610	4.351	9.24	11.07	12.83	15.09	16.75
6	0.676	0.872	1.237	1.635	2.204	5.348	10.64	12.59	14.45	16.81	18.55
7	0.989	1.239	1.690	2.167	2.833	6.346	12.02	14.07	16.01	18.48	20.28
8	1.344	1.646	2.180	2.733	3.490	7.344	13.36	15.51	17.53	20.09	21.95
9	1.735	2.088	2.700	3.325	4.168	8.343	14.68	16.92	19.02	21.67	23.59
10	2.156	2.558	3.247	3.940	4.865	9.342	15.99	18.31	20.48	23.21	25.19
11	2.603	3.053	3.816	4.575	5.578	10.34	17.28	19.68	21.92	24.72	26.76
12	3.074	3.571	4.404	5.226	6.304	11.34	18.55	21.03	23.34	26.22	28.30
13	3.565	4.107	5.009	5.892	7.042	12.34	19.81	22.36	24.74	27.69	29.82
14	4.075	4.660	5.629	6.571	7.790	13.34	21.06	23.68	26.12	29.14	31.32
15	4.601	5.229	6.262	7.261	8.547	14.34	22.31	25.00	27.49	30.58	32.80
16	5.142	5.812	6.908	7.962	9.312	15.34	23.54	26.30	28.85	32.00	34.27
17	5.697	6.408	7.564	8.672	10.09	16.34	24.77	27.59	30.19	33.41	35.72
18	6.265	7.015	8.231	9.390	10.86	17.34	25.99	28.87	31.53	34.81	37.16
19	6.844	7.633	8.907	10.12	11.65	18.34	27.20	30.14	32.85	36.19	38.58
20	7.434	8.260	9.591	10.85	12.44	19.34	28.41	31.41	34.17	37.57	40.00
21	8.034	8.897	10.28	11.59	13.24	20.34	29.62	32.67	35.48	38.93	41.40
22	8.643	9.542	10.98	12.34	14.04	21.34	30.81	33.92	36.78	40.29	42.80
23	9.260	10.20	11.69	13.09	14.85	22.34	32.01	35.17	38.08	41.64	44.18
24	9.886	10.86	12.40	13.85	15.66	23.34	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	24.34	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	25.34	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	26.34	36.74	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	16.93	18.94	27.34	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	28.34	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	29.34	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	39.34	51.81	55.76	59.34	63.69	66.77
90	59.20	61.75	65.65	69.13	73.29	89.33	107.6	113.1	118.1	124.1	128.3
120	83.85	86.92	91.57	95.70	100.6	119.3	140.2	146.6	152.2	159.0	163.6
∞	888.6	898.9	914.3	927.6	943.1	999.3	1058	1075	1090	1107	1119

Critical values of the F-distribution



The top, middle and bottom numbers in each cell correspond to 0.1, 0.05 and 0.01 levels of significance respectively

Upper Percentage Points of the *F* distribution

Denominator df		Numerator df																			
	p	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40	50	60	100	∞
1	0.10	39.9	49.5	53.6	55.8	57.2	58.2	58.9	59.4	59.9	60.2	60.7	61.2	61.7	62.1	62.3	62.5	62.7	62.8	63	63.3
	0.05	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	252	253	254
	0.01	4052	5000	5403	5625	5764	5859	5928	5981	6022	6056	6106	6157	6209	6240	6261	6287	6303	6313	6334	6363
2	0.10	8.53	9	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.47	9.48	9.49
	0.05	18.5	19	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5	19.5
	0.01	98.5	99	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5	99.5	99.5
3	0.10	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.2	5.18	5.17	5.17	5.16	5.15	5.15	5.14	5.13
	0.05	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.7	8.66	8.63	8.62	8.59	8.58	8.57	8.55	8.53
	0.01	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	26.9	26.7	26.6	26.5	26.4	26.4	26.3	26.2	26.1

Upper Percentage Points of the *F* distribution

Denominator df		Numerator df																			
	p	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40	50	60	100	∞
4	0.10	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.9	3.87	3.84	3.83	3.82	3.8	3.8	3.79	3.78	3.76
	0.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6	5.96	5.91	5.86	5.8	5.77	5.75	5.72	5.7	5.69	5.66	5.63
	0.01	21.2	18	16.7	16	15.5	15.2	15	14.8	14.7	14.5	14.4	14.2	14	13.9	13.8	13.7	13.7	13.7	13.6	13.5
5	0.10	4.06	3.78	3.62	3.52	3.45	3.4	3.37	3.34	3.32	3.3	3.27	3.24	3.21	3.19	3.17	3.16	3.15	3.14	3.13	3.11
	0.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.52	4.5	4.46	4.44	4.43	4.41	4.37
	0.01	16.3	13.3	12.1	11.4	11	10.7	10.5	10.3	10.2	10.1	9.89	9.72	9.55	9.45	9.38	9.29	9.24	9.2	9.13	9.03
6	0.10	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.9	2.87	2.84	2.81	2.8	2.78	2.77	2.76	2.75	2.72
	0.05	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.1	4.06	4	3.94	3.87	3.83	3.81	3.77	3.75	3.74	3.71	3.67
	0.01	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.1	7.98	7.87	7.72	7.56	7.4	7.3	7.23	7.14	7.09	7.06	6.99	6.89
7	0.10	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.7	2.67	2.63	2.59	2.57	2.56	2.54	2.52	2.51	2.5	2.47
	0.05	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.4	3.38	3.34	3.32	3.3	3.27	3.23
	0.01	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.06	5.99	5.91	5.86	5.82	5.75	5.66
8	0.10	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.5	2.46	2.42	2.4	2.38	2.36	2.35	2.34	2.32	2.3
	0.05	5.32	4.46	4.07	3.84	3.69	3.58	3.5	3.44	3.39	3.35	3.28	3.22	3.15	3.11	3.08	3.04	3.02	3.01	2.97	2.93
	0.01	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.26	5.2	5.12	5.07	5.03	4.96	4.87

Upper Percentage Points of the F distribution

Denominator df	p	Numerator df																			
		1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40	50	60	100	∞
9	0.10	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.34	2.3	2.27	2.25	2.23	2.22	2.21	2.19	2.16
	0.05	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.89	2.86	2.83	2.8	2.79	2.76	2.71
	0.01	10.6	8.02	6.99	6.42	6.06	5.8	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.71	4.65	4.57	4.52	4.48	4.41	4.32
10	0.10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28	2.24	2.2	2.17	2.16	2.13	2.12	2.11	2.09	2.06
	0.05	4.96	4.1	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.73	2.7	2.66	2.64	2.62	2.59	2.54
	0.01	10	7.56	6.55	5.99	5.64	5.39	5.2	5.06	4.94	4.85	4.71	4.56	4.41	4.31	4.25	4.17	4.12	4.08	4.01	3.92
12	0.10	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.1	2.06	2.03	2.01	1.99	1.97	1.96	1.94	1.91
	0.05	4.75	3.89	3.49	3.26	3.11	3	2.91	2.85	2.8	2.75	2.69	2.62	2.54	2.5	2.47	2.43	2.4	2.38	2.35	2.3
	0.01	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.5	4.39	4.3	4.16	4.01	3.86	3.76	3.7	3.62	3.57	3.54	3.47	3.37
15	0.10	3.07	2.7	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.02	1.97	1.92	1.89	1.87	1.85	1.83	1.82	1.79	1.76
	0.05	4.54	3.68	3.29	3.06	2.9	2.79	2.71	2.64	2.59	2.54	2.48	2.4	2.33	2.28	2.25	2.2	2.18	2.16	2.12	2.07
	0.01	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.8	3.67	3.52	3.37	3.28	3.21	3.13	3.08	3.05	2.98	2.88
20	0.10	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.89	1.84	1.79	1.76	1.74	1.71	1.69	1.68	1.65	1.61
	0.05	4.35	3.49	3.1	2.87	2.71	2.6	2.51	2.45	2.39	2.35	2.28	2.2	2.12	2.07	2.04	1.99	1.97	1.95	1.91	1.85
	0.01	8.1	5.85	4.94	4.43	4.1	3.87	3.7	3.56	3.46	3.37	3.23	3.09	2.94	2.84	2.78	2.69	2.64	2.61	2.54	2.43

Upper Percentage Points of the *F* distribution

Denominator df		Numerator df																			
	p	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40	50	60	100	∞
24	0.10	2.93	2.54	2.33	2.19	2.1	2.04	1.98	1.94	1.91	1.88	1.83	1.78	1.73	1.7	1.67	1.64	1.62	1.61	1.58	1.54
	0.05	4.26	3.4	3.01	2.78	2.62	2.51	2.42	2.36	2.3	2.25	2.18	2.11	2.03	1.97	1.94	1.89	1.86	1.84	1.8	1.74
	0.01	7.82	5.61	4.72	4.22	3.9	3.67	3.5	3.36	3.26	3.17	3.03	2.89	2.74	2.64	2.58	2.49	2.44	2.4	2.33	2.22
30	0.10	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.77	1.72	1.67	1.63	1.61	1.57	1.55	1.54	1.51	1.46
	0.05	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.88	1.84	1.79	1.76	1.74	1.7	1.63
	0.01	7.56	5.39	4.51	4.02	3.7	3.47	3.3	3.17	3.07	2.98	2.84	2.7	2.55	2.45	2.39	2.3	2.25	2.21	2.13	2.02
40	0.10	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.71	1.66	1.61	1.57	1.54	1.51	1.48	1.47	1.43	1.38
	0.05	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.78	1.74	1.69	1.66	1.64	1.59	1.52
	0.01	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.8	2.66	2.52	2.37	2.27	2.2	2.11	2.06	2.02	1.94	1.82
50	0.10	2.81	2.41	2.2	2.06	1.97	1.9	1.84	1.8	1.76	1.73	1.68	1.63	1.57	1.53	1.5	1.46	1.44	1.42	1.39	1.33
	0.05	4.03	3.18	2.79	2.56	2.4	2.29	2.2	2.13	2.07	2.03	1.95	1.87	1.78	1.73	1.69	1.63	1.6	1.58	1.52	1.45
	0.01	7.17	5.06	4.2	3.72	3.41	3.19	3.02	2.89	2.78	2.7	2.56	2.42	2.27	2.17	2.1	2.01	1.95	1.91	1.82	1.7
60	0.10	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.66	1.6	1.54	1.5	1.48	1.44	1.41	1.4	1.36	1.3
	0.05	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.1	2.04	1.99	1.92	1.84	1.75	1.69	1.65	1.59	1.56	1.53	1.48	1.4
	0.01	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.5	2.35	2.2	2.1	2.03	1.94	1.88	1.84	1.75	1.62

Upper Percentage Points of the *F* distribution

Denominator df		Numerator df																			
	p	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40	50	60	100	∞
90	0.10	2.76	2.36	2.15	2.01	1.91	1.84	1.78	1.74	1.7	1.67	1.62	1.56	1.5	1.46	1.43	1.39	1.36	1.35	1.3	1.24
	0.05	3.95	3.1	2.71	2.47	2.32	2.2	2.11	2.04	1.99	1.94	1.86	1.78	1.69	1.63	1.59	1.53	1.49	1.46	1.41	1.31
	0.01	6.93	4.85	4.01	3.53	3.23	3.01	2.84	2.72	2.61	2.52	2.39	2.24	2.09	1.99	1.92	1.82	1.76	1.72	1.62	1.48
120	0.10	2.75	2.35	2.13	1.99	1.9	1.82	1.77	1.72	1.68	1.65	1.6	1.55	1.48	1.44	1.41	1.37	1.34	1.32	1.28	1.2
	0.05	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.6	1.55	1.5	1.46	1.43	1.37	1.27
	0.01	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.93	1.86	1.76	1.7	1.66	1.56	1.4
∞	0.10	2.71	2.31	2.09	1.95	1.85	1.78	1.72	1.68	1.64	1.61	1.55	1.49	1.43	1.38	1.35	1.3	1.27	1.25	1.2	1.08
	0.05	3.85	3	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84	1.76	1.68	1.58	1.52	1.47	1.41	1.36	1.33	1.26	1.11
	0.01	6.66	4.63	3.8	3.34	3.04	2.82	2.66	2.53	2.43	2.34	2.2	2.06	1.9	1.79	1.72	1.61	1.54	1.5	1.38	1.16