BASIC STATISTICS FOR ADVANCED LEVEL MATHEMATICS

By

Kalani Jonathan, M.Sc., B.Ed., Dip. Ed.

ABOUT THE AUTHOR

The author was born to Pastor and Mrs. Yowasi Mukirania in December 1965. He studied at Mitandi Primary school starting Primary one in 1973. He joined Mitandi Secondary school, Saad Secondary school and later Nyakasura School for his A levels. He later enrolled for a Diploma in Education at the Institute of Teacher Education (now Kyambogo University) , a B.Ed degree(1997) and M.Sc (Mathematics) degree in 0ctober 2005 respectively, at Makerere University.

He has taught at Rwenzori High School and Kilembe Secondary School between 1990 and now. He is a former examiner of Uganda National Examinations Board (marked for thirteen years). He has held various responsibilities some of which are being Games Master in charge of Net ball, patron of the Mathematics and Wildlife Clubs, Heading the Mathematics Department and caretaking the school administration as Head teacher and Deputy Head teacher.

He is a member of the Uganda Mathematical Society and has always presented his students for the National Mathematics Contest at O and A level since 1999. As a student, he participated in the contest at primary level in 1979 and university level in 1996 where he achieved the first position. He has consistently attended the annual conferences of the Uganda Mathematical Society at Makerere University.

Currently, he is training secondary school teachers in Rwanda.

ACKNOWLEDGEMENTS

I want to thank the members of the Mathematics department of Kilembe Secondary school especially Mr. Ndeghe Joshua for encouraging me in my endeavour to write this book and all my former students on whom the material was tested.

I want to give special thanks to Mr. M. Magino (formerly, Chair of the Secondary School Mathematics panel, National Curriculum Development Center, Uganda) for accepting to proof read my work and also find publishers for this book. However, all errors and omissions are entirely mine. I will be very happy if the users respond to anomalies through the publishers.

I am grateful to Mr. Herbert Oundo who accepted to typeset my work without which I would have taken much longer to produce the manuscript.

Lastly I want to thank my wife Christine and children Musoki, Biira and Kambale for the patience whenever I missed their company while working on the text. I did not get leave to author this book and had to utilize long hours at night and during weekends hence minimum socialization with the family.

Finally, I want to thank the publishers for accepting to publish my work.

J. Kalani

PREFACE

I have written this book purposely to meet the requirements for those students offering Mathematics at advanced level through out the commonwealth countries. This book could be useful to university students who are doing social sciences, business administration or education with a requirement of doing probability and statistics which they utilize in research or measurement and evaluation.

I was prompted to write this text upon discovering that there were few text books that broadly covered the requirements for A Level statistics section of Mathematics as examined by Uganda National Examinations Board and since these change periodically (as syllabi are reviewed quite often), I decided to write for all possible syllabi for Advanced Level.

My intention was to write a book for personal and class use. One of the assumptions is that students are conversant with the necessary pure mathematics. Numerous examples have been given because the author believes that this promotes comprehension and thus encourages the reader to attempt most if not all of the exercises.

Having taught A level Mathematics for twenty years and currently being a trainer of secondary mathematics teachers, I have found out that there is need for a book which approaches the subject in an elementary way and assumes that the reader has had little or no exposure to probability and statistics; which need this book ably satisfies.

Finally I must thank my wife, Christine, for all her help and encouragement, . for reading the manuscript and checking the arithmetic and Herbert Oundo for Typesetting. Any errors and omissions are entirely mine and I hereby welcome comments and criticisms from my fellow teachers.

Contents

Chapter 1

DESCRIPTIVE STATISTICS

1.1 Introduction

- Statistics is a branch of mathematics. It deals with collection, interpretation and analysis of data. Statistics is so important because it is used in almost all spheres of of life ranging from governments, non-governmental organisations, individual businesses, e.t.c. It is a subject that cannot be avoided while plannning and making decisions.
- Statistical data is either qualitative or quantitive.
  Attributes like colour, sex altitude, opinion are measured using qualitative data while numerical quantities like distance, mass height, time e.t.c. are measured using quantitative data.
- Data may be discrete or continous. Dicrete data is collected by counting and can only take on integral values while continouous data can take on any value, integral or not.

1.2 Definition of common terms

1.2.1 Population:

This is a collection of items about which information is required. If there is something about which information is needed, for instance, ten year olds in a certain town, all ten year olds comprise the population.

1.2.2 Sample:

This is a finite sub-set of the population. It may not be possible to examine all the ten year olds in a given town but a certain number of these ten year olds may be examined and these form the sample of the population.

In many cases, we rarely deal with population, but deal with a section of the population which we have called the sample. An appropriate sample must have all the attributes of the population about which information is required.

1.2.3 Variable:

This is the observed item and varies between members of the population. A variable is sometimes called a variate. A variate can be qualitative or quantitative and may be defined as discrete or continous.

1.2.4 Statistic:

This is a number which characterises the distribution of the variate in the sample. A statistic represents the parameter. A parameter characterises the distribution of the variate in the population.
If the value of the statistic is near the value of the parameter, we say that the sample represents the population. In most cases, the parameter may never be known if the population is infinite and we are content with the value of the statistic. For instance if the mean of the population is μ, and the mean of the sample is $\bar{x}, \bar{x}$ is utilised in place of the true paramenter value μ. It is said to be a good estimator of the population mean μ.

1.3 Data Representation:

Data may be presented diagramatically or visually by use of bar graphs, histograms, frequency polygon, Oigve or Pie-chart. These visual diagrams give a visual impression to the statistician who now goes ahead to analyse and make conclusions about the data.

1.3.1 Bar Graph

This is at times called a bar chart. Class frequencies are plotted against class limits. Since consecutive classes can never have common limits, the bars have spaces between them when plotted.

1.3.2 Histogram:

This is a graph where the class frequencies are plotted against class boundaries. The example below illustrates the difference between a bar graph and a histogram.

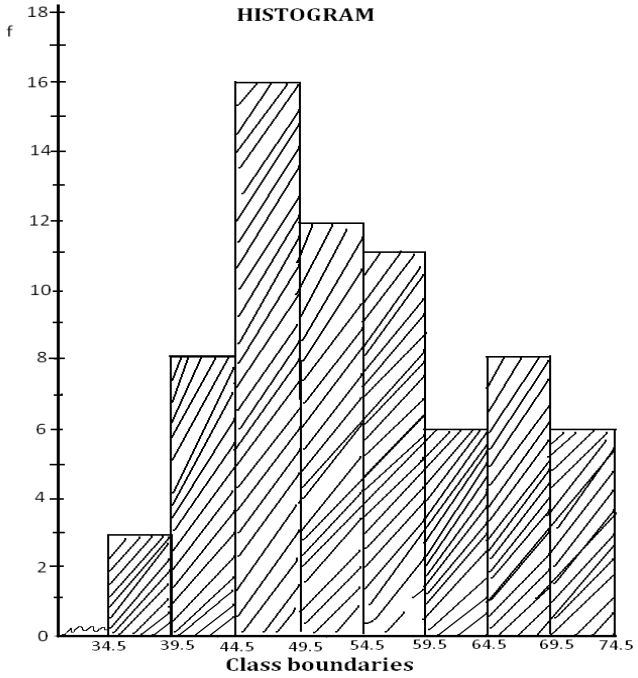
Example 1.3.1

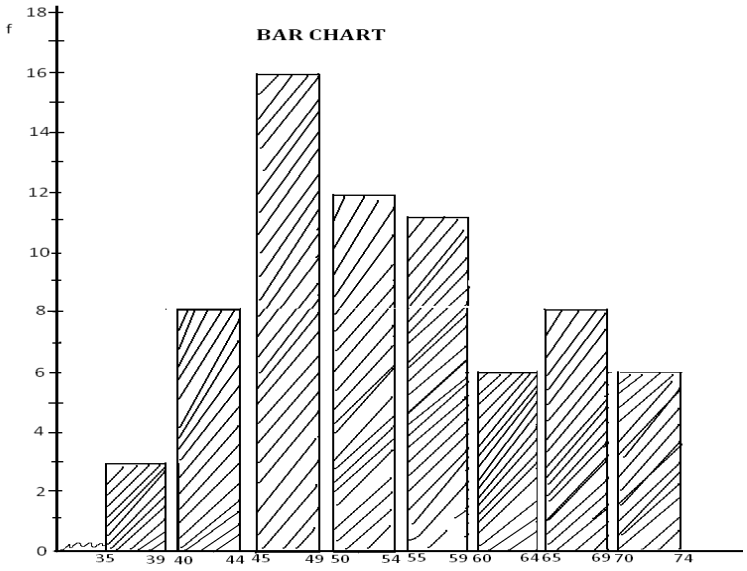
The table below gives marks obtained in a test given to Sophomores in the 2009/2010 academic year.

Marks	Number of students
35 - 39	3
40 - 44	8
45 - 49	16
50 - 54	12
55 - 59	11
60 - 64	6
65 - 69	8
70 - 74	6

Use the given data to plot a bar graph and a histogram
Solution:

Class limits	Class boundaries	frequency
35 - 39	34.5 - 39.5	3
40 - 44	39.5 - 44.5	8
45 - 49	44.5 - 49.5	16
50 - 54	49.5 - 54.5	12
55 - 59	54.5 - 59.5	11
60 - 64	59.5 - 64.5	6
65 - 69	64.5 - 69.5	8
70- 74	69.5- 74.5	6





You may be asked to draw a histogram even when the class width is not uniform. This is done in any of the two ways which utilize frequency density plotted against class boundaries or standard frequency plotted against class boundaries.
Frequency density is obtained if frequencies of the classes are divided by their widths while standard frequency is obtained by using the common class width is as the standard one and then dividing the frequency values by numbers of times the class is the standard one.

Example 1.3.2

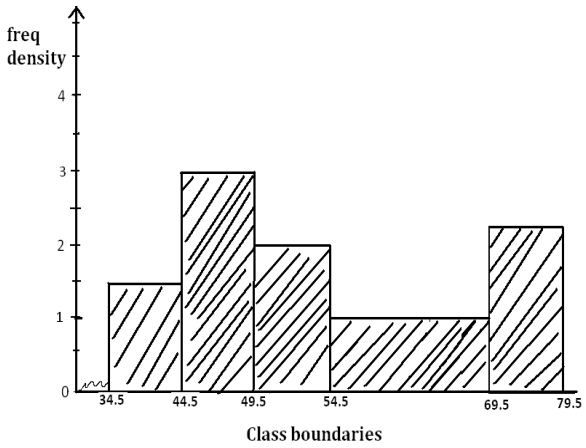
From the information given below, construct a histogram using the two options

Class boundaries	Frequency
34.5 - 44.5	14
44.5 - 49.5	15
49.5 - 54.5	10
54.5 - 69.5	15
69.5 - 79.4	16

Solution:

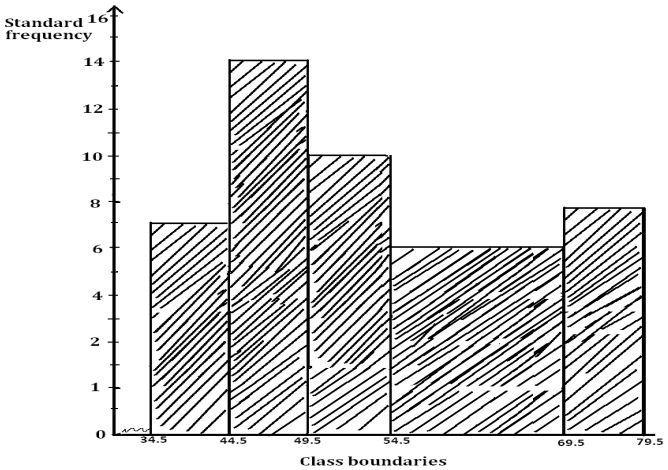
(i)

Class boundaries	Frequency	Frequency density
34.5 - 44.5	14	1.4
44.5 - 49.5	15	3.0
49.5 - 54.5	10	2.0
54.5 - 69.5	15	1.0
69.5 - 79.5	16	1.6



The common classes are 44 - 49 and 50 - 54.

(ii)



The two diagrams above give the same visual impression. Whenever the class width is not uniform, then the standard frequency or frequency density should be plotted against the class boundaries while constructing a histogram.
Sometimes the data is given as a collection in terms of inequalities. histograms for such do not require one to find class boundaries since they are already given as integral values as the next example shows.

Example 1.3.3

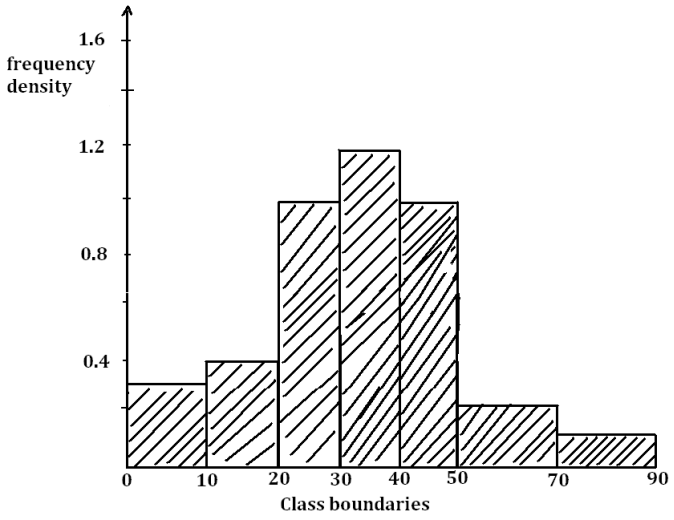
The table below shows the population in thousands of different ages in some selected subcounties of Kasese district in Uganda

Age group	Population in thousands
Below 10	3
10 and under 20	4
20 and under 30	10
30 and under 40	12
40 and under 50	10
50 and under 70	4
70 and under 90	2

Construct a histogram representing the above data.
Solution:

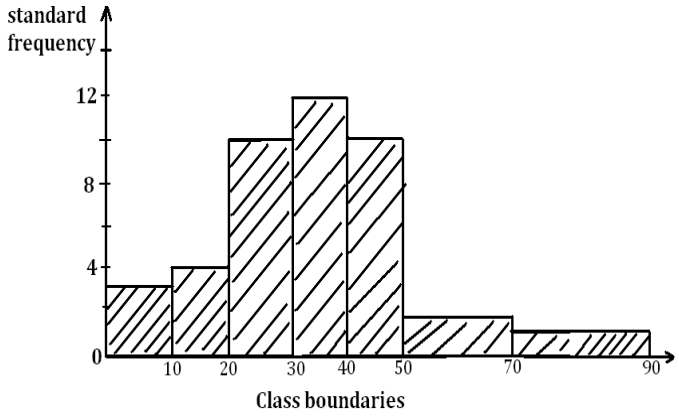
Class	Class width	Frequency	Frequency density
0-< 10	10	3	0.3
10- < 20	10	4	0.4
20- < 30	10	10	1.0
30- < 40	10	12	1.2
40- < 50	10	10	1.0
50- < 70	20	4	0.2
70- < 90	20	2	0.1

The histogram is thus constructed.



Using standard frequency, the frequency table is

Class limits	No. of times	frequency	standard frequency
0- < 10	1	3	3
10- < 20	1	4	4
20- < 30	1	10	10
30- < 40	1	12	12
40-<50	1	10	1.0
50- < 70	2	4	2
70- < 90	2	2	1



Once the class widths differ, the histogram should be drawn using either the frequency density or the standard frequency.

1.3.3 Frequency Polygons

When class frequencies are plotted against class marks, the figure obtained is called a frequency polygon. This polygon is obtained when consecutive points are joined by a straight line.

Example 1.3.4

The frequency distribution below shows the weights to the nearest kilogram of students at Kavumu College of Education.

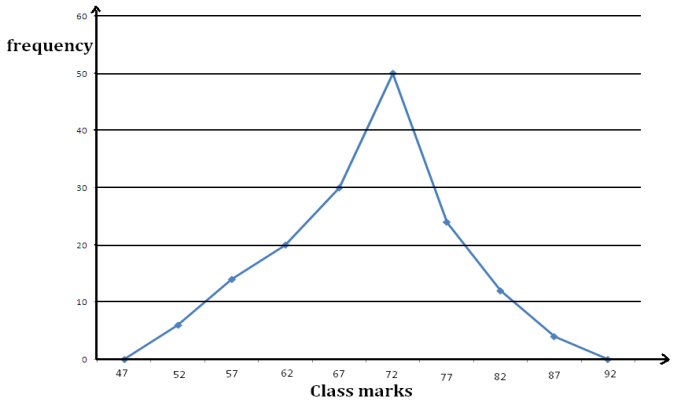
Mass (kg)	frequency
50 - 54	6
55 - 59	14
60 - 64	20
65 - 69	30
70 - 74	50
75 - 79	24
80 - 84	12
85 - 89	4

Construct a frequency polygon for the above data.

Solution:

To make a good polygon, we shall assume there is a class before the 50 - 54 class that has frequency zero.

Class limits	class mark	frequency
additional class	47	0
50 - 54	52	6
55 - 59	57	14
60 - 64	62	20
65 - 69	67	30
70 - 74	72	50
75 - 79	77	24
80 - 84	82	12
85 - 89	87	4
90 - 94	92	0



It may be required that a frequency polygon and a histogram be plotted on the same graph. This is said to impose a frequency polygon on a histogram.

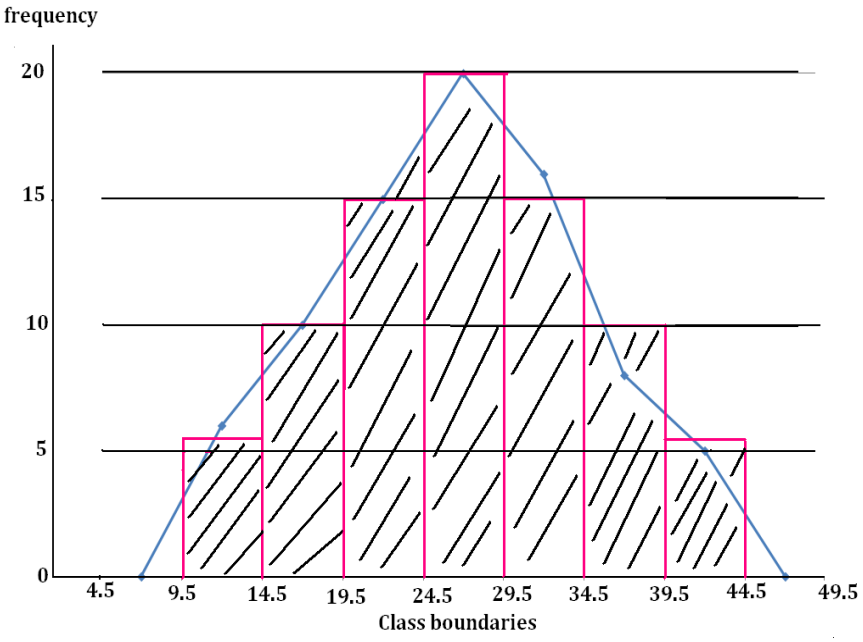
Example 1.3.5

Using the data below, draw and construct a histogram and on it, super impose a frequency polygon.

Class limits	frequency
10 - 14	6
15 - 19	10
20 - 24	15
25 - 29	20
30 - 34	16
35 - 39	8
40 - 44	5

Solution:

Class limits	Class boundaries	frequency
Class before	4.5 - 9.5	0
10 - 14	9.5 - 14.5	6
15 - 19	14.5 - 19.5	10
20 - 24	19.5 - 24.5	15
25 - 29	24.5 - 29.5	20
30 - 34	29.5 - 34.5	16
35 - 39	34.5 - 39.5	8
40 - 44	39.5 - 44.5	5
45 - 49	44.5 - 49.5	0



Note that after constructing a histogram, the midpoints of the tips of the bars are connected by straight lines.

1.3.4 The Cummulative Frequency Curve

This is also called an Ogive. It is obtained by plotting the cummulative frequency curve against the Class boundaries.

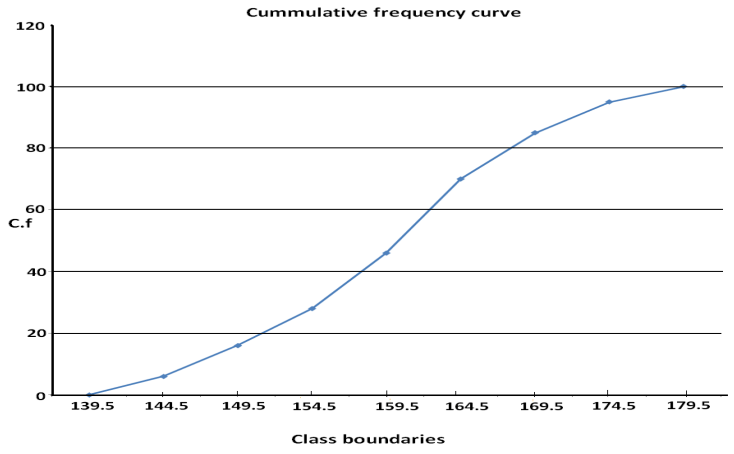
Example 1.3.6

The table below shows the heights of children measured to the nearest centimetre.

Height (cm)	frequency
140 - 144	6
145 - 149	10
150 - 154	12
155 - 159	18
160 - 169	24
165 - 169	15
170 - 174	10
175 - 179	5

Draw a cummulative frequency curve for the data.
Solution:

Classes	Class boundaries	frequency	cum freq
140 - 144	139.5 - 144.5	6	6
145 - 149	144.5 - 149.5	10	16
150 - 154	149.5 - 154.5	12	28
155 - 159	154.5 - 159.5	18	46
160 - 164	159.5 - 164.5	24	70
165 - 169	164.5 - 169.5	15	85
170 - 174	169.5 - 174.5	10	95
175 - 179	174.5 - 179.5	5	100



Example 1.3.7

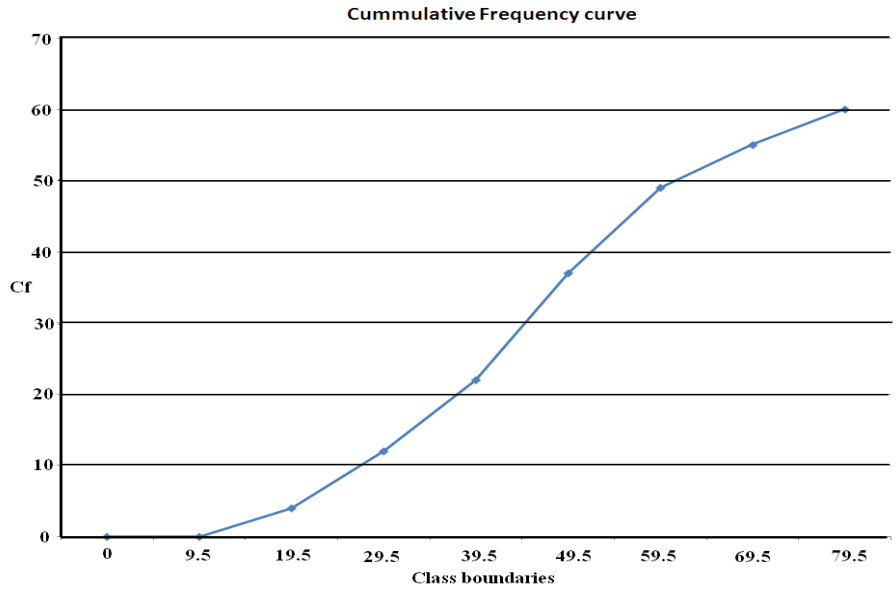
The frequency distribution below shows the ages in months of goats in the school farm

Classes	frequency
10 - 19	4
20 - 29	8
30 - 39	10
40 - 49	15
50 - 59	12
60 - 69	6
70 - 79	5

Construct a cummulative frequency curve for the data.

Solution:

Class limits	Class boundaries	frequency	cum freq
10 - 19	9.5 - 19.5	4	4
20 - 29	19.5 - 29.5	8	12
30 - 39	29.5 - 39.5	10	22
40 - 49	39.5 - 49.5	15	37
50 - 59	49.5 - 59.5	12	49
60 - 69	59.5 - 69.5	6	55
70 - 79	69.5 - 79.5	5	60



1.4 Measures of Central Tendency

These include the mean, median and mode. These values locate the average value of a variable in a specific position of the number line with respect to the data.

1.4.1 Mean

This is at times called arithmetic mean. Mean is the average value of the observations, that is, it is the sum of the observations divided by the number of items observed. The mean of the n values $x_1, x_2, x_3, \ldots, x_n$ is

$$x = \frac{(x_1 + x_2 + \ldots + x_n)}{n} = \frac{\sum x_i}{n}$$

Some values may occur in the data more than once. If there are n values with respective frequencies $f_1, f_2, \ldots, f_n$, then the mean is given by

$$\bar{x} = \frac{(f_1 x_1 + f_2 x_2 + \ldots + f_n x_n)}{(f_1 + f_2 + \ldots + f_n)} = \frac{\sum fx}{\sum f}$$

Example 1.4.1

Determine the mean mark of a class test using the data $28, 35, 18, 40, 62, 50$ and 70.
Solution:

$$\text{Mean} = \frac{28 + 35 + 18 + 40 + 62 + 50 + 70}{7} = 43.3$$

Example 1.4.2

The data below gives the heights of some students at Kavumu College of Education. Find the mean height.

Height	150	152	160	162	170	175	180
Frequency	5	9	7	6	5	4	2

Solution:

Height (x)	**Frequency** (f)	(fx)
150	5	750
152	9	1368
160	7	1120
162	6	972
170	5	850
175	4	700
180	2	360
Total	38	6120

$$\begin{aligned} \text{Mean} &= \frac{\sum fx}{\sum f} \\ &= \frac{6120}{38} \\ &= 161.0526316 \\ &\simeq 161 \end{aligned}$$

Example 1.4.3

The table below gives the weights of cows in kg that were slaughtered at the city butchery during the last Christmas season

Weight	200	250	300	450	600	700	900
Frequency	3	4	2	9	10	6	2

Find the average weight of the cows.
Solution:

Weight (kg) (x)	frequency (f)	(fx)
200	3	600
250	4	1000
300	2	600
450	9	4050
600	10	6000
700	6	4200
900	2	1800
Total	36	18250

$$\begin{aligned} \text{Mean} &= \frac{\sum fx}{\sum f} \\ &= \frac{18250}{36} \\ &= 506.94 \\ &\simeq 507kg \end{aligned}$$

There may be need to handle data that has so many items such that it has to be grouped so that it is handled appropriately in a relatively short time without tedious work.

Example 1.4.4

The table below shows the ages of people who attended a rally last week.

20	21	20	22	24	26	30	35
31	32	23	27	28	34	40	42
37	45	50	54	49	60	64	63
54	47	46	49	62	61	57	25
26	29	44	53	36	48	63	50

Beginning with the 20 - 24 class construct a frequency table and calculate the mean age.

Solution:

The figures are tallied to find the frequency for each class.

Classes	Tallies	f	x	fx
20-24	𝍸/	6	22	132
25-29	𝍸/	6	27	162
30-34	////	4	32	128
35-39	///	3	37	111
40-44	///	3	42	126
45-49	𝍸/	6	47	282
50-54	𝍸	5	52	260
55-59	/	1	57	57
60-64	𝍸/	6	62	372
		∑f=40		1630

$$\begin{aligned} \text{Mean} &= \frac{\sum fx}{\sum f} \\ &= \frac{1630}{40} \\ &= 40.75 \end{aligned}$$

Sometimes a working mean is used to solve for the mean, Any x value (midpoint of a class) can be assumed to be the mean and then the mean is calculated basing on that. Other x values are handled as deviations from the assumed mean.

Example 1.4.5

Use frequency table in the solution to example 1.4.4 to find the mean using 2 as the working (assumed) mean.
Solution:

Classes	f	x	d	fd
20 - 24	6	22	-20	-120
25 - 29	6	27	-15	-90
30 - 34	4	32	-10	-40
35 - 39	3	37	-5	-15
40 - 44	3	42	0	0
45 - 49	6	47	5	30
50 - 54	5	52	10	50
55 - 59	1	57	15	15
60 - 64	6	62	20	120
	40			-50

If assumed mean $= \bar{x}$, then $d = x - \bar{x}$

$$\begin{aligned} \text{Mean} &= \bar{x} + \frac{\sum fd}{\sum f} \\ &= 42 - \frac{50}{40} \\ &= 40.75 \end{aligned}$$

just as before.
If assumed mean is A then $d = x - A$

$$\text{mean} = A + \frac{\sum fd}{\sum f}$$

$$\bar{x} = A + \frac{\sum fd}{\sum f}$$

1.4.2 Median

The median is that value of the variable which divides the distribution into two equal parts. If the variable is plotted on a number line, the median is the value in the middle.

Example 1.4.6

Determine the median of

(i) $8, 6, 10, 9, 9, 7, 5$

(ii) $12, 15, 10, 11, 16, 18, 14$

(iii) $3, 7, 9, 10, 13, 12, 8$

Solution:
The data is arranged in either ascending or descending order

(i) $5, 6, 7, 8, 9, 9, 10$
The middle value is 8
Median = 8

(ii) $10, 11, 12, 14, 15, 10, 18$
The middle value is 14
Median = 14

(iii) $3, 7, 8, 9, 10, 12, 13, 14$
The middle values are 9 and 10

$$\text{Median} = \frac{9+10}{2}$$

$$= 9.5$$

Median for grouped data is found using a frequency table with the formula

$\text{Median} = L_1 + \left(\frac{\frac{N}{2} - Cf_b}{fm}\right) \cdot i$

where L_1 = Lower class boundary of the modal class

N = Total of all class frequencies (total no, of observations)
Cf_b = Cummulative frequency before the median class
fm =Frequency of the median class.
i = Class width (interval).

Example 1.4.7

Utilise the table in example 1.4.1 to calculate for the median.
Solution:

Classes	Class boundaries	f	C.f
35 - 39	34.5 - 39.5	3	3
40 - 44	39.5 - 44.5	8	11
45 - 49	44.5 - 49.5	16	27
50 - 54	49.5 - 54.5	12	39
55 - 59	54.5 - 59.5	11	50
60 - 64	59.5 - 64.5	6	56
65 - 69	64.5 - 69.5	8	64
70 - 74	69.5 - 74.5	6	70
		70	

The median class is $49.5 - 54.5, (50 - 54)$
$L_1 = 49.5, N = 70, C.f_b = 27, fm = 12, i = 5$
Therefore

$$\begin{aligned}
\text{Median} &= L_1 + \frac{\left(\frac{N}{2} - C.f_b\right)}{fm} \times i \\
&= 49.5 + \frac{(35 - 27)}{12} \times 5 \\
&= 49.5 + \frac{8}{12} \times 5 \\
&= 49.5 + 3.33 \\
&= 52.8333 \simeq 52.8
\end{aligned}$$

Example 1.4.8

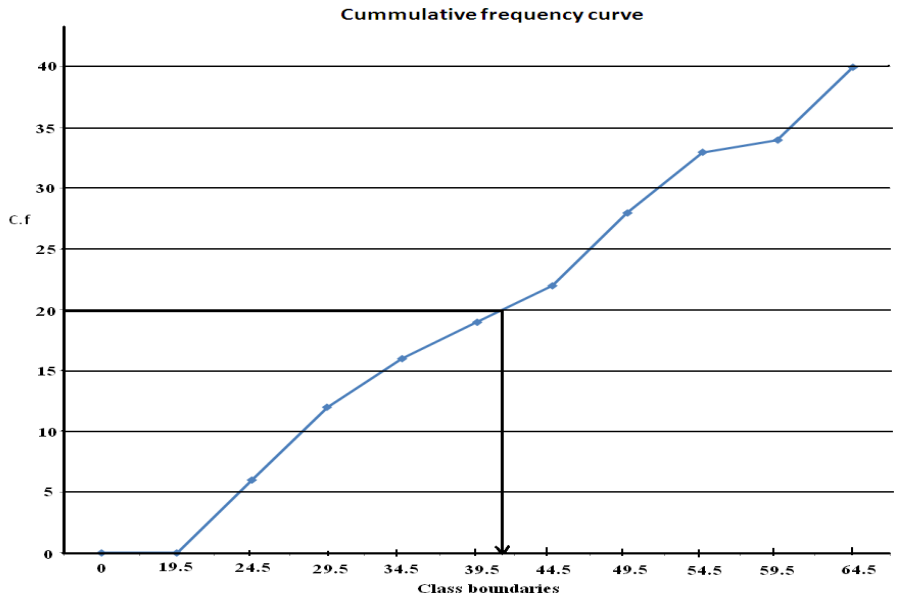
Use the data in example 1.4.4 to find the median.
Solution:

Classes	**Class boundaries**	f	C.f
20 - 24	19.5 - 24.5	6	6
25 - 29	24.5 - 29.5	6	12
30 - 34	29.5 - 34.5	4	16
35 - 39	34.5 - 39.5	3	19
40 - 44	39.5 - 44.5	3	22
45 - 49	44.5 - 49.5	6	28
50 - 54	49.5 - 54.5	5	33
55 - 59	54.5 - 59.5	1	34
60 - 64	59.5 - 64.5	6	40
		40	

In this case $L_1 = 39.5, N = 40, C.f_b = 19, fm = 3, i = 5$
Therefore median is

$$\begin{aligned} \text{Median} &= L_i + \frac{\frac{N}{2} - C.f_b)}{fm} \times i \\ &= 39.5 + \frac{(20-19)}{3} \times 5 \\ &= 39.5 + \frac{5}{3} \\ &= 41.16666 \\ &\simeq 41.2 \end{aligned}$$

Median can be estimated from the cummulative frequency curve. This is done by reading off a value half way the total frequency on the vertical axis which should tally with a value on the horizontal axis. The value read off from the horizontal axis gives the median.
Constructing a cummulative frequency curve for data in this example,



The arrow points at the median value, which is 41 according to the graph. This is a good estimate of 41.2 which we got by calculation.

1.4.3 Mode

This is a value which occurs most frequently. This is got by observation in case of ungrouped data or by calculation in case of grouped data.

Example 1.4.9

State the mode of the set of data below

(i) $(2, 3, 4, 4, 4, 5, 6, 6, 7, 8, 9)$

(ii) $1, 1, 2, 3, 3, 3, 3, 7, 7, 8, 9, 9$

Solution:

(i) Mode is 4

(ii) Mode is 3

In case of grouped data, the mode is given as

$$\text{Mode} = L_1 + \left(\frac{d_1}{d_1 + d_2}\right) \times i$$

Example 1.4.10

Use the data of example 1.4.1 to find the mode.
Solution:

Classes	**Class boundaries**	f
35 - 39	34.5 - 39.5	3
40 - 44	39.5 - 44.5	8
45 - 49	44.5 - 49.5	16
50 - 54	49.5 - 54.5	12
55 - 59	54.5 - 59.5	11
60 - 64	59.5 - 64.5	6
65 - 69	64.5 - 69.5	8
70 - 74	69.5 - 74.5	6

$\text{Mode} = L_i + \left(\frac{d_1}{d_1+d_2}\right) \times i$

where
L_i = lower boundary of modal class
d_1 = difference between modal frequency and frequency of the class before it.
d_2 = difference between modal frequency and frequency of class after it
i = class interval.
Therefore

$$\begin{aligned} \text{Mode} &= 44.5 + \left[\frac{16-8}{(16-8)+(16-12)}\right] \times 5 \\ &= 44.5 + \frac{8}{8+4} \times 5 \\ &= 44.5 + \frac{40}{12} \\ &= 47.83. \end{aligned}$$

Example 1.4.11

The data below shows marks obtained in an aptitude test by candidates who wished to get admitted to University through Mature Age entry scheme.

72	43	36	57	47	68	75	79	82	31
52	47	74	52	29	72	57	72	87	73
32	52	62	55	42	47	37	57	22	81
27	53	37	64	62	32	47	37	52	88
55	25	30	67	70	52	67	36	38	76

Beginning with the 20 - 29 class, construct a frequency table for the data. Using the frequency table

(a) Calculate the mean, median and mode.

(b) construct a cummulative frequency curve and use it to estimate the median

(c) Construct a histogram and use it to estimate the mode.

Classes	Tallies	f	x	fx	Cf
20-29	////	4	24.5	98	4
30-39	///// ////	10	34.5	345	14
40-49	///// /	6	44.5	267	20
50-59	///// /////	11	54.5	599.5	31
60-69	///// /	6	64.5	387	37
70-79	///// ////	9	74.5	670.5	46
80-89	////	4	84.5	338	50
		50		2705	

(a) (i)

$$\begin{aligned} \text{Mean} &= \frac{\sum fx}{\sum f} \\ &= \frac{2705}{50} \\ &= 54.1 \end{aligned}$$

(ii)

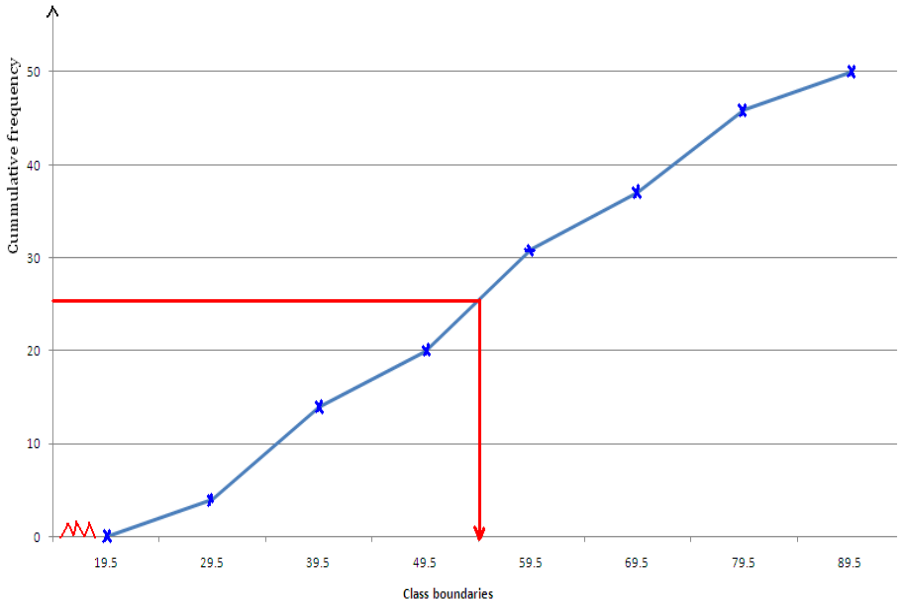
$$\begin{aligned} \text{Median} &= L_1 + \frac{\left(\frac{N}{2} - Cf_b\right) \times i}{fm} \\ &= 49.5 + \left(\frac{25-20}{11}\right) \times 10 \\ &= 49.5 + \frac{50}{11} \\ &= 54.05 \end{aligned}$$

(iii)

$$\begin{aligned} \text{Mode} &= L_1 + \frac{d_1}{d_1 + d_2} \times i \\ &= 49.5 + \frac{5}{5+5} \times 10 \\ &= 54.5 \end{aligned}$$

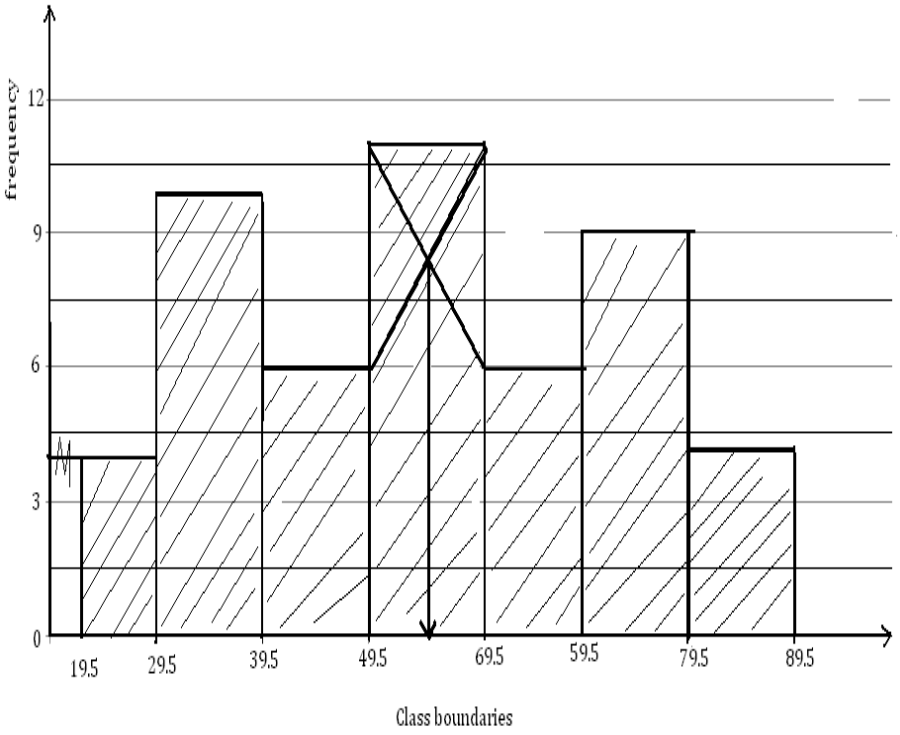
Note that in this particular case, the modal class is the same as the median class.

(b)



From the graph, median is 54

(c)



From the graph, mode = 54.5

1.4.4 Index Numbers

If a set of data is reduced to relative values by comparing it with a fixed number (base), the relative values are called index numbers or percentage relatives. If they refer to wages they are called wage relatives. If they refer to prices, they are called price relatives. The period against which the corresponding values are compared is called the base period. Examples of index numbers include price index, wage index, quantity index, e.t.c. A sample quantity index is given by

$$\frac{q_n}{q_0} \times 100$$

where q_n is the quantity in the current period and
q_0 is the quantity in the base period.

Example 1.4.12

One kilogram of maize cost shs.800 in 2006 and sh.1200 in 2009. Taking 2006 as the base year, find the price index in 2009.
Solution:

$$\text{Price Index} = \frac{1200}{800} \times 100 = 150$$

This implies that from 2006 to 2009, the price of maize has gone up by 50%

Example 1.4.13

Broilers on Fred's farm consumed 2000kg of feeds in January and consumed 3200kg of feeds in March. Using January as the base period, calculate the quantity index for March.
Solution:

$$\begin{aligned} \text{Quantity index} &= \frac{3200}{2000} \times 100 \\ &= 160 \end{aligned}$$

The quantity of feeds consumed in March was 60% above that of January.

Example 1.4.14

The wage of a porter per day was 1000 in 1999 but in 2009 it was 2500. Using 1999 as the base year, calculate the wage index for 2009.
Solution:

$$\text{Wage index} = \frac{2500}{1000} \times 100 = 250$$

The wage increased by 150% in 2009.

Note 1.4.1

Items are often grouped so that the total price for that group is compared with the total for that group in a previous (base) period. The index from such a consideration is called the simple aggregate price index.
This is given by

$$\frac{\sum P_n}{\sum P_0}$$

where P_n is the period under consideration and P_0 is the base period price.

Example 1.4.15

Musoki spent on the following items per month in the years 2005 and 2008.

Item	**2005 amount in shs**	**2008 amount in shs**
Rent	144,000	180,000
Clothing	45,000	65,000
Power	35,000	48,000
Water	8000	9000
Food	150,000	170,000
Transport	60,000	75,000
	442000	547,000

Using 2005 as the base year, calculate the simple aggregate experience index for 2008.

Solution:

The totals for 2005 and 2008 are 442000 and 547000 respectively.
The simple aggregate expenditure index is

$$\begin{aligned} &= \frac{547,000}{442,000} \\ &= 1.237556561 \\ &\simeq 1.2376 \end{aligned}$$

The expenditure rose by 23.76% in 2008.

Note 1.4.2

There are situations which have many contributory factors which warrant more complicated index numbers. These are found by use of weighted averages of the percentage relatives of the contributory factors. If the percentage relatives are $r_1, r_2, r_3 \ldots r_n$ having respective weights $w_1, w_2, w_3, \ldots r_n$, then the weighted index is $\frac{\sum r_n w_n}{\sum w_n}$

Example 1.4.16

What is the cost of living based on the data below:

Item	**Price index**	**Weight**
Rent	130	180
Clothing	129	150
Power	140	130
Water	115	210
Food	110	190
Transport	150	200

Solution:

Item	Price index (A)	Weight(W)	WA
Rent	130	180	23400
Clothing	129	150	19350
Power	140	130	18200
Water	115	210	24150
Food	110	190	20900
Transport	150	200	30000
Total		1060	136000

Cost of living index

$$\begin{aligned} &= \frac{136000}{1060} \\ &= 128.3018868 \\ &\simeq 128.3 \end{aligned}$$

This is a composite index number because we need price indices and their weights to arrive at one resultant figure.
Economists and government planners normally use composite index numbers when assessing the cost of living and other issues which are influenced by many contributing factors.

Example 1.4.17

In a manufacturing process, five different raw materials are used . The masses required are in the ratio $1 : 2 : 2 : 5 : 6$. The table below shows the cost per unit of these materials in 2001 and in 2009. Calculate the price index for the cost of the process in 2009 taking 2001 as the base year.
Solution:
Let the raw materials be A,B,C,D and E

Material	A	B	C	D	E
Cost in 2001	2000	3500	1500	1000	800
Cost in 2009	2400	4300	1800	1400	1100

We first find the weighted average of the cost per unit for each year.

$$\begin{aligned}
\text{Weighted average for 2001} &= \frac{1\times 2000 + 2\times 3500 + 2\times 1500 + 5\times 1000 + 6\times 800}{1+2+2+5+6} \\
&= \frac{21800}{16} \\
&= \frac{21,800}{16} \\
\text{Weighted average for 2009} &= \frac{1\times 2400 + 2\times 4300 + 2\times 1800 + 5\times 1400 + 6\times 1100}{1+2+2+5+6} \\
&= \frac{28200}{16} \\
\text{Therefore, price index for 2009} &= \frac{28200/16}{21800/16}\times 100 \\
&= 129.3577982 \\
&\simeq 129.4
\end{aligned}$$

NB: Composite index = weighted aggregate index.

Example 1.4.18

A building contractor bought various materials for building two houses of the same size and quality in 2004 and 2009. The corresponding prices with their weights are given in the table below. Using 2004 as the base year, calculate the weighted average relative price index.

Item	Weight	Price 2004	Price 2009
Cement	130	14,000	24000
Bricks	80	50	80
Iron bars	110	8000	16000
Iron sheets	120	10000	15000
Timber	125	1200	2000
Aggregate	90	9000	15000
Murram	70	6000	8000
Nails	115	2000	2300
Sand	140	9000	11000

Solution:

Item	Weight	Price 2004	Price 2009	$\frac{P_1}{P_0}$	$\frac{P_1}{P_2} \times w$
Cement	130	14,000	24000	$\frac{12}{7}$	222.86
Bricks	80	50	80	$\frac{8}{5}$	128.0
Iron bars	110	8000	16000	2	220.0
Iron sheets	120	10000	15000	$\frac{3}{2}$	180.0
Timber	125	1200	2000	$\frac{10}{6}$	208.33
Aggregate	90	9000	15000	$\frac{5}{3}$	150.0
Murram	70	6000	8000	$\frac{4}{3}$	93.33
Nails	115	2000	2300	$\frac{23}{20}$	132.25
Sand	140	9000	11000	$\frac{11}{9}$	171.11
Total	980				1505.88

Weighted average relative price index

$$\begin{aligned} &= \frac{1505.88}{980} \times 100 \\ &= 152.6612245 \\ &\simeq 153.66 \end{aligned}$$

The commodity prices increased by 53.66%

1.4.5 Moving Averages

For a set of numbers $x_1, x_2, \ldots$, the moving average of order n is given by the following set of arithmetic means

$$\frac{x_1 + x_2 + \ldots + x_n}{n}, \quad \frac{x_2 + x_3 + \ldots + x_{n+1}}{n}, \quad \frac{x_3 + x_4 \ldots x_{n+2}}{n}, \ldots$$

If the data is quarterly, then the average is n quarterly moving average. If it is monthly average, the average is known as a monthly moving average.
Since moving averages are found over a period of time, they are greatly utilised in time series analysis.

Example 1.4.19

For the numbers $8, 12, 14, 15, 16, 19, 21, 23, 25,$ find the moving averages of order 4 and 6.

Solution:
For order 4:

$$\frac{8+12+14+15}{4}, \frac{12+14+15+16}{4}, \frac{14+15+16+19}{4},$$
$$\frac{15+16+19+25}{4}, \frac{16+19+21+23}{4}, \frac{19+21+23+25}{4}$$

$= 12.25, 14.25, 16, 17.75, 19.75$ and 22

For order 6:

$$\frac{8+12+14+15+16+19}{6}, \frac{12+14+15+16+19+21}{6},$$
$$\frac{14+15+16+19+21+23}{6} \frac{15+16+19+21+23+25}{6}$$

$= 14, 16.5, 18$ and 19.83

Example 1.4.20

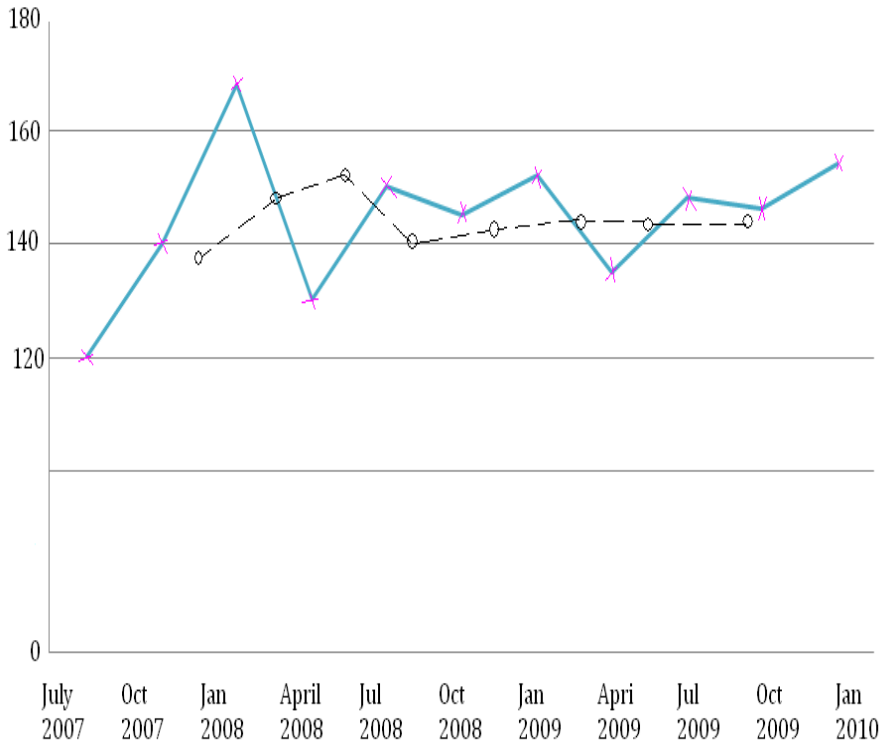
University ABC made quarterly admission for evening computer courses as shown in the table below:

Year	Month	Number admitted
2007	July	120
	October	140
2008	January	168
	April	130
	July	150
	October	145
2009	January	152
	April	135
	July	148
	October	146
2010	January	154

Plot the row data and four quarterly moving averages on the same graph.

Solution:

	Month	Admitted	Difference	$\frac{D}{4}$	Moving Average
2007	July	120			
	October	140	10	2.5	139.5
2008	January	168	10	2.5	147
	April	130	−23	−5.75	148.25
	July	150	22	5.5	144.25
	October	145	-15	-3.75	145.25
2009	January	152	3	0.75	145
	April	135	-6	-1.5	145.25
	July	148	19	4.55	145.75
	October	146			
2010	January	154			



The line connecting 4 point moving averages is the trendline. The discrepancies between the trend line and the individual points enable us to estimate the seasonal effects. The trendline and the estimates of the seasonal effects considered together make it possible to predict future figures.

Example 1.4.21

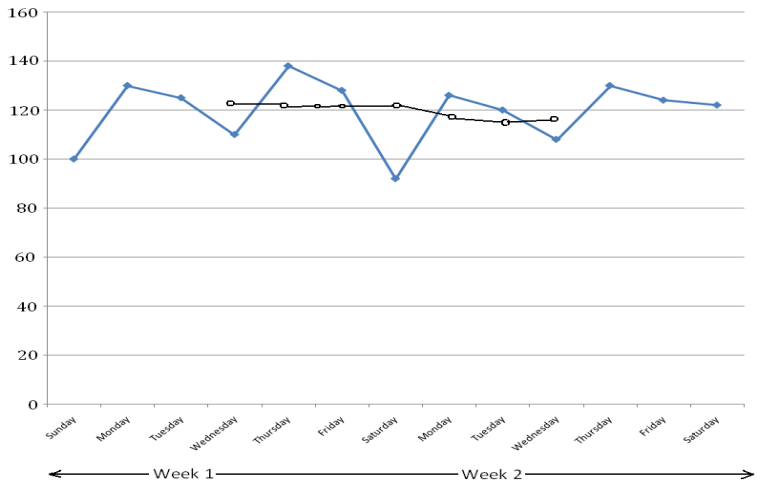
The out patient department of a hospital, open seven days a week attended to the following numbers of patients

	Sun	Mon	Tue	Wed	Thur	Fri	Sat
Week 1	100	130	125	110	138	130	128
Week 2	92	126	120	108	130	124	122

Calculate seven day moving averages for the out patient attendances. Plot both on the same graph. Use your graph to estimate attendance on sunday of the third week.

Solution:

Week	Day	No. Admitted	7-Day Moving Average
1	Sunday	100	
	Monday	130	
	Tuesday	125	
	Wednesday	110	123
	Thursday	138	121.86
	Friday	130	121.29
2	Saturday	128	120.57
	Sunday	92	120.86
	Monday	126	119.14
	Tuesday	120	118.29
	Wednesday	108	117.43
	Thursday	130	
	Friday	124	
	Saturday	122	



Attendence on Sunday of week 3 will be 82 patients.

1.5 Measures of Dispersion

These measures are used to find the spread of the observation from the mean or about the mean. They include range, mean deviation, quartiles, percentiles, deciles, variance and standard deviation.

1.5.1 Range

This is the difference between the highest and the lowest value in the set.

Example 1.5.1

For the set S below, find the range

(i) $S = \{12, 17, 21, 14, 23, 19\}$

(ii) $S = \{43, 50, 64, 74, 85, 67, 79, 38\}$
Solution:

(i) The range of S is $23 - 12 = 11$

(ii) The range of S is $85 - 38 = 47$

1.5.2 Mean Deviation

For a set of numbers, the mean deviation is given by

$$\frac{\sum |x_i - M|}{n}$$

where M is the mean

Example 1.5.2

Find the mean deviation for the set of numbers $82, 93, 99, 108, 112$.
Solution:

$$\begin{aligned} \text{Mean} &= \frac{\sum x}{n} \\ &= \frac{82 + 93 + 99 + 108 + 112}{5} \\ &= 98.8 \end{aligned}$$

$$\text{Mean deviation} = \frac{|82 - 98.8| + |93 - 98.8| + |99 - 98.8| + |108 - 98.8| + |112 - 98.8|}{5}$$

$$= \frac{(16.8 + 5.8 + 0.2 + 9.2 + 13.2)}{5}$$

$$= 9.04$$

Example 1.5.3

Using the frequency table displayed below, calculate the mean deviation

Classes	$10-14$	$15-19$	$20-24$	$25-29$	$30-34$
freq	2	3	6	4	2

Solution:

Classes	f	x	fx	$\lvert x-\bar{x}\rvert$	$f\lvert x-\bar{x}\rvert$
10 - 14	2	12	24	10.3	20.6
15 - 19	3	17	51	5.3	15.9
20 - 24	6	22	132	0.3	1.8
25 - 29	4	27	108	4.7	18.8
30 - 34	2	32	64	9.7	19.4
	17		379		76.5

$$\text{Mean } \bar{x} = \frac{\sum fx}{\sum f}$$

$$= \frac{379}{17}$$

$$= 22.29411765$$

$$\simeq 22.3$$

$$\text{Mean deviation} = \frac{\sum f|x-\bar{x}|}{\sum f}$$

$$= \frac{76.5}{17} = 4.5$$

1.5.3 Variance and Standard Deviation

Variance is the sum of the squares of the mean deviation divided by the numbers of observations, and standard deviation is the positive square root of the variance.

For ungrouped data.

$$\text{variance} \quad = \quad \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

where $\bar{x}$ is the mean of the given set of numbers $x_1, x_2, \ldots x_n$

Example 1.5.4

Find the variance of $43, 46, 50, 53, 57, 61$.
Solution:

x	$(x - \bar{x})$	$(x - \bar{x})^2$
43	-8.66	74.9956
46	-5.66	32.0356
50	-1.66	2.7556
53	1.34	1.7956
57	5.34	28.5156
61	9.34	87.2356
		227.3336

$$\begin{aligned} \text{Mean } \bar{x} &= \frac{43 + 46 + 50 + 53 + 57 + 61}{6} \\ &= \frac{310}{6} \\ &= 51.66 \end{aligned}$$

Therefore variance $= \frac{227.3336}{6} = 37.8893333$

It therefore follows that the standard deviation is $\sqrt{37.8893333} = 6.1553395$

Note 1.5.1

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

which is a simpler formula used to find the variance, using $\frac{\sum x_i^2}{n} - \bar{x}^2$ we get

x	x^2
43	1849
46	2116
50	2500
53	2809
57	3249
61	3721
	16244

$$\text{Variance} = \frac{16244}{6} - \left(\frac{310}{6}\right)^2$$

$$= 37.888889333 \text{ as before.}$$

For Grouped data.

$$\text{Variance} = \frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2$$

$$\text{or} \qquad \frac{\sum fd^2}{\sum f} - \left(\frac{\sum fd}{\sum f}\right)^2$$

where $d = X - A$ and A is the assumed (working) mean.

Example 1.5.5

For the distribution given below, find the standard deviation.

Classes	$20-29$	$30-39$	$40-49$	$50-59$	$60-69$	$70-79$
f	3	5	8	6	4	2

Solution:

Classes	f	x	x^2	fx	fx^2
$20-29$	3	24.5	600.25	73.5	1800.75
$30-39$	5	34.5	1190.25	172.5	5951.25
$40-49$	8	44.5	1980.25	356	15842
$50-59$	6	54.5	2970.25	327	17821.5
$60-69$	4	64.5	4160.25	258	16641
$70-79$	2	74.5	5550.25	149	11100.5
	28			1336	69157

$$\begin{aligned}
\text{Variance} &= \frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2 \\
&= \frac{69157}{28} - \left(\frac{1336}{28}\right)^2 \\
&= 193\frac{47}{196} \text{ or } 193.2397959 \\
\text{Standard deviation} &= \sqrt{variance} \\
&= \sqrt{193\frac{47}{196}} \text{ or} \sqrt{193.2397959} \\
&= 13.90107175 \\
&\simeq 13.9
\end{aligned}$$

We can still find the standard deviation by using a working mean. Let Assumed/working mean be $A = 44.5$

Classes	f	x	$d = (x - A)$	$d^2 = (x - A)^2$	fd	fd^2
$20 - 29$	3	24.5	-20	400	-60	1200
$30 - 39$	5	34.5	-10	100	-50	500
$40 - 49$	8	44.5	0	0	0	0
$50 - 59$	6	54.5	10	100	60	600
$60 - 69$	4	64.5	20	400	80	1600
$70 - 79$	2	74.5	30	900	60	1800
	28				90	5700

$$\begin{aligned}
\text{Variance} &= \frac{\sum fd^2}{\sum f} - \left(\frac{\sum fd}{\sum f}\right)^2 \\
&= \frac{5700}{28} - \left(\frac{90}{28}\right)^2 \\
&= 193.239759 \\
\text{Standard deviation} &= \sqrt{193.239759} \\
&= 13.90107175 \\
&= 13.9
\end{aligned}$$

Example 1.5.6

Use the data of example 1.3.4 to find the standard deviation.
Solution:

Classes	f	x	fx	x^2	fx^2
$50-54$	6	52	312	2704	16224
$55-59$	14	57	798	3249	45486
$60-64$	20	62	1240	3844	76880
$65-69$	30	67	2010	4489	134670
$70-74$	50	72	3600	5184	259200
$75-79$	24	77	1848	5929	142296
$80-84$	12	82	984	6724	80688
$85-89$	4	87	348	7569	30276
	160		11140		$785,720$

$$\begin{aligned}
\text{Variance} &= \frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2 \\
&= \frac{785720}{160} - \left(\frac{11140}{160}\right)^2 \\
&= 63.109375 \\
\therefore \text{S.d} &= \sqrt{63.109375} \\
&= 7.944149923 \\
&\simeq 7.944.
\end{aligned}$$

We get the same answer when we use the method of working mean. For instance let the working mean be 67.

Classes	f	x	d	fd	d^2	fd^2
$50-54$	6	52	-15	-90	225	1350
$55-59$	14	57	-10	-140	100	1400
$60-64$	20	62	-5	-100	25	500
$65-69$	30	67	0	0	0	0
$70-74$	50	72	5	250	25	1250
$75-79$	24	77	10	240	100	2400
$80-84$	12	82	15	180	225	2700
$85-89$	4	87	20	80	400	1600
	160			420		11200

$$\begin{aligned}
\text{Variance} &= \frac{\sum fd^2}{\sum f} - \left(\frac{\sum fd}{\sum f}\right)^2 \\
&= \frac{11200}{1600} - \left(\frac{420}{160}\right)^2 \\
&= 63.109375 \\
\therefore \text{Standard deviation} &= \sqrt{63.109375} \\
&= 7.944140923 \\
&\simeq \underline{7.944}
\end{aligned}$$

Variance for a sample

For a sample that has n items, variance is denoted by

$$S^2 = \left(\frac{n}{n-1}\right)\left\{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2\right\}$$

Example 1.5.7

Find the variance of a sample that has the following data: 12, 15, 17, 13, 20, 21, 25, 18, 14.

Solution:

x	x^2
12	144
15	225
17	289
13	169
20	400
21	441
25	625
18	324
14	196
155	2813

$$\begin{aligned}
n &= 9 \\
S^2 &= \frac{9}{8}\left\{\frac{2813}{9} - \left(\frac{155}{9}\right)^2\right\}
\end{aligned}$$

$$= \quad 17\frac{17}{18} \text{ or } \quad 17.94444$$

Example 1.5.8

For a sample of heights of students of year I in cm, find the standard deviation
$171, 180, 160, 156, 158, 161, 165, 154, 174, 176, 178$.
Solution:

x	x^2
171	29241
180	32400
160	25600
156	24336
158	24964
161	25921
165	27225
154	23716
174	30276
176	30976
178	31684
1833	306339

$$\begin{aligned} S^2 &= \frac{n}{n-1}\left\{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2\right\} \\ &= \frac{11}{10}\left\{\frac{306339}{11} - \left(\frac{1833}{11}\right)^2\right\} \\ &= 89.4545454545 \\ &\simeq 89.5 \\ \text{Standard deviation } S &= \sqrt{89.454545} \\ &= 9.458041319 \\ &\simeq 9.458 \end{aligned}$$

1.5.4 Percentiles, Deciles and Quartiles

- Percentiles are values which devide the data into 100 equal parts. A value P_1 has one percent of the data falling below it, P_{13} has 13% of the data falling below it and P_n has n% of the data falling below it.

Example 1.5.9

If the number of observations is 120, then

$$P_{25} = \frac{25}{100} \times 120 = 30$$

So 30 observations are below the value of P_{25} and

$$P_{15} = \frac{15}{100} \times 120 = 18$$

So 18 observations fall below P_{15}
In the earlier case P_{25} has a quarter of the observations below it. So $P_{25} = Q_1 =$ lower quartile.

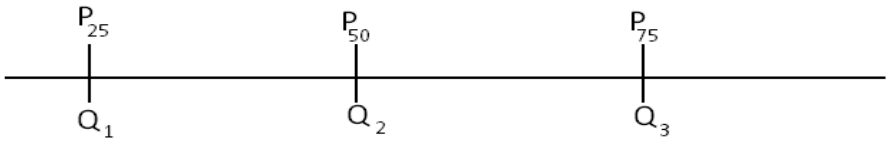
- Deciles devide a set of values into ten equal parts. These are values D_1,D_2, . . . ,D_9.

Example 1.5.10

Let the observations be 200. Find D_3 and D_7.
Solution:

$$D_3 = \frac{3}{10} \times 200 \qquad D_7 = \frac{7}{10} \times 200$$
$$= 60 \qquad = 140$$

Note that $P_{25} = Q_1 =$ lower Quartile
$P_{50} = Q_2 =$ Median (middle quartile)
$P_{75} = Q_3 =$ Upper quartile
If the data is arranged on a number line, then



The Interquartile Range

This is at times called quartile range. It is given by Upper quartile - Lower quartile

$$= Q_3 - Q_1$$

Upper quartile given by $L_3 + \frac{\left(\frac{3N}{4} - Cf_b\right)}{fq_3} \times i$
Where L_3 = Lower class boundary of upper quartile class
N = Total frequency (no. of observation)
Cf_b = Cummulative frequency before upper quartile class
f$_{q3}$ = frequency of the upper quartile class
i = Class interval (width)

The lower quartile is given by $L_1 + \left(\frac{\frac{N}{4} - Cfb}{f_{q1}}\right) \cdot i$

Where L$_1$ = Lower class boundary of lower quartile class
N = Total frequency (no.of observation)
C$_{fb}$ =Cummulative frequency before lower quartile class
f$_{q1}$ = frequency of the lower quartile class
i = class interval (width).
The semi-interquartile range = $\frac{1}{2}(Q_3 - Q_1)$.

Example 1.5.11

The data below is weights of students in lower secondary school at school ABC.

Mass	frequency
$40 - 44$	8
$45 - 49$	24
$50 - 54$	37
$55 - 59$	30
$60 - 64$	26
$65 - 69$	18
$70 - 74$	10
$75 - 79$	7

Find the

(i) Median

(ii) Interquartile range

Solution:

Class	f	Cf
$40-44$	8	8
$45-49$	24	32
$50-54$	37	69
$55-59$	30	99
$60-64$	26	125
$65-69$	18	143
$70-74$	10	153
$75-79$	7	160
	160	

(i)

$$\begin{aligned}
\text{Median} &= L_2 + \left(\frac{\frac{N}{2} - Cf_b}{fm}\right) \cdot i \\
&= 54.5 + \frac{(80-69)}{30} \times 5 \\
&= 54.5 + \frac{11 \times 5}{30} \\
&= 56.33333 \simeq 56.3
\end{aligned}$$

(ii)

$$\begin{aligned}
Q_3 &= L_3 + \left(\frac{\frac{3}{4}N - Cf_b}{f_{l3}}\right) i \\
&= 59.5 + \left(\frac{120-99}{26}\right) \times 5 \\
&= 59.5 + \frac{21}{26} \times 5 \\
&= 63.53846154 \simeq 63.54 \\
Q_1 &= L_1 + \frac{\left(\frac{N}{4} - Cf_b\right)}{f_{q1}} i \\
&= 49.5 + \frac{(40-32)}{37} \times 5 \\
&= 49.5 + \frac{8}{37} \times 5
\end{aligned}$$

$$= \quad 50.58108 \ \simeq 50.58.$$

$$\text{Interquartile range} = Q_3 - Q_1$$
$$= 63.54 - 50.58$$
$$= 12.96$$

Example 1.5.12

For the data given in the table below,

$26-50$	$51-75$	$76-100$	$101-125$	$126-150$	$151-175$	$176-200$
4	18	17	28	22	16	15

(a) Draw an ogive and use it to estimate

(i) the median

(ii) interquartile range

(iii) 20 - 90 Percentile range

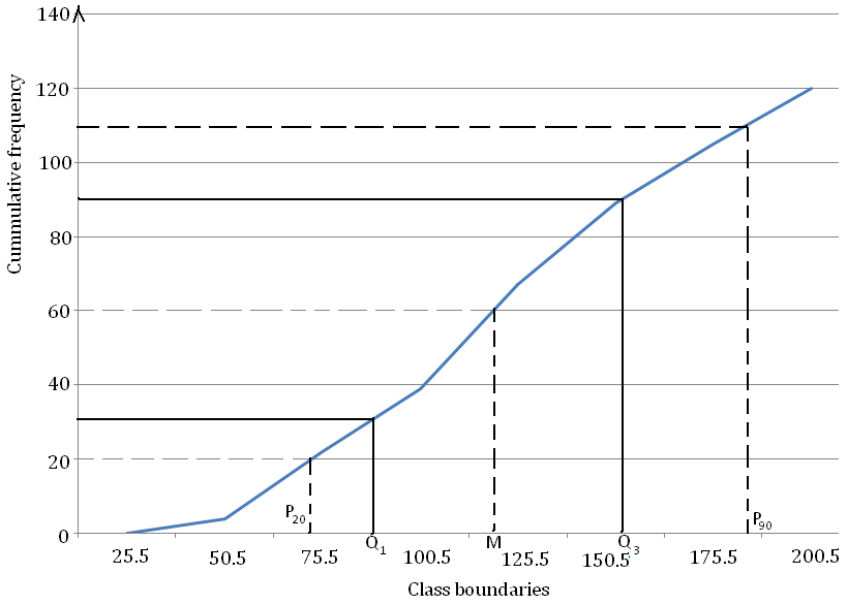
(b) Find by calculation

(i) the median

(ii) interquartile range

(c) Compare (i) and (ii) of (a) and (i) and (ii) of (b) and comment on your results.
Solution:

Class	f	Cf
$26-50$	4	4
$51-75$	18	22
$76-100$	17	39
$101-125$	28	67
$126-150$	22	89
$151-175$	16	105
$176-200$	15	120
	120	



(b) (i)

$$\begin{aligned} \text{Median} &= 100.5 + \frac{(60-39)}{28} \times 25 \\ &= 119.25 \end{aligned}$$

(ii)

$$\begin{aligned} Q_3 &= 150.5 + \frac{(90-89)}{16} \times 25 \\ &= 152.0625 \\ Q_1 &= 75.5 + \frac{(30-22)}{17} \times 25 \\ &= 87.26470588 \simeq 87.2647 \\ \text{Interquartile range} &= Q_3 - Q_1 \\ &= 152.0625 - 87.2647 \\ &= 64.7978 \end{aligned}$$

(c) The figures are nearly the same. The discrepancy is due to the scale (the decimal places are not easy to plot if they are very small) and the innacurracy of the hand drawn ogive.

Exercise 1

1. The table below shows the heights of sticks to the nearest centimeter

Height	Frequency
$10-19$	6
$20-29$	8
$30-39$	14
$40-49$	20
$50-59$	22
$60-69$	12
$70-79$	10
$80-89$	8

   (a) Draw a cummulative frequency curve and use it to estimate the median

   (b) Construct a histogram and use it to estimate the median.

2. The table below gives the weights in grams of babies delivered in a hospital on New years day.

Weight	700	850	950	1000	1100	1300
Frequency	3	4	2	9	10	6

   Find the average weight of the babies.

3. The table below shows the marks obtained in a test

15	16	15	17	19	21	25	30
26	28	18	22	23	29	35	37
32	40	45	49	44	55	59	58
49	42	41	44	57	56	62	20
21	24	39	48	31	43	58	45

   (i) Form a frequency table with a class interval of 5, the lowest class limit being 15

   (ii) Using a working mean of 32, find the mean.

   (iii) Construct a cummulative frequency curve and use it to estimate the median (36.17)

4. The distribution table below shows the ages of patients admitted at a clinic

AGE	NUMBER
$0- < 5$	10
$5- < 15$	16
$15- < 30$	25
$30- < 50$	29
$50- < 70$	19
$70- < 90$	11

   Determine the mean and median age of the patients

5. Find the variance of the sample data below

   (i) 43, 47, 56, 66, 78, 88, 95, 101, 105 and 110

   (ii) 21, 30, 38, 45, 50, 56, 71, 82 and 87.

6. Find the standard deviation of the population data below

   (i) 60, 63, 69, 71, 78, 80 and 81

   (ii) 15, 19, 28, 32, 33 and 35.

7. The table below shows the mistakes per page typed by a learner typist who typed 60 pages.

Class	$5-9$	$10-14$	$15-19$	$20-24$	$25-29$
f	13	15	18	8	6

   Calculate the mean, median and mode of the mistakes made.

8. Use the data of question 7 to find the standard deviation using a working mean of 17.

9. Kambale's family spent the following money per month on the items shown in the years 2007 and 2008.

Item	2007 sh.	2008 sh.
Rent	130, 000	150, 000
Clothing	90, 000	110, 000
Power	60,000	68,000
Water	12,000	13,000
Food	200,000	230,000
Transport	80,000	90,000
Medical	70,000	100,000

Using 2007 as the base year, find the simple aggregate experience index for 2008.

10. Find the cost of living index based on the data below:

Item	Price Index	Weight
Rent	130	180
Clothing	120	165
Power	160	130
Water	110	190
Food	125	175
Transport	140	200
Medical	105	135

11. The figures below are the yields to the nearest kilogram of cassava obtained from plots of equal size.

22	20	24	22	23	24	23	21	42	43
18	21	27	19	20	21	25	21	35	29
26	23	22	25	21	19	21	23	44	39
20	22	24	26	18	25	19	27	38	44
32	36	31	15	15	39	36	40	17	42

Beginning with the 15 -19 class, construct a frequency table. Using an assumed mean estimate the mean yield and standard deviation.

12. The following table gives a summary of weights of patients suffering from High Blood Pressure.

Mass (kg)	Frequency
$60-64$	7
$65-69$	16
$70-74$	14
$75-79$	66
$80-84$	35
$85-89$	29
$90-94$	20
$95-99$	13

Find the

(i) median

(ii) interquartile range

(iii) mean

(iv) standard deviation

13. The table below gives the quarterly electricity costs for my family during three successive years. Plot these results on a graph. Calculate the four quarterly moving averages and plot them on the same graph. Draw straight lines to fit these averages as closely as possible and use the graph to estimate the rate at which the costs increasing over the three years.

	1st quarter	2nd quarter	3rd quarter	4th quarter
2007	110	80	70	100
2008	125	92	78	105
2009	130	98	82	120

14. The table below shows the number of births at a clinic across all the 12 months of last year

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sept.	Oct	Nov.	Dec
64	61	77	83	79	74	71	70	71	73	71	69

Find the five-monthly moving averages. Plot both the original figures. Estimate the number of births in the month of January of the following year.

15. A factory uses six raw materials A,B,C,D,E,F to manufacture a toy in the ratios $1:2:4:3:2:2:5$ respectively. The prices of the materials in shs per tonne in the years 2008 and 2009 are given as

Raw material	A	B	C	D	E	F
2008	40000	30000	10000	20000	50000	12000
2009	90000	50000	15000	35000	80000	32000

Taking 2008 as the base year, calculate an index number for the total cost of the raw materials used in the manufacture of the toy in 2009.

16. The table below gives quarterly sales in millions of shillings at a whole sale store for the period 2008 to 2010.

	2008	2009	2010
1st Quarter	170	174	182
2nd Quarter	184	191	194
3rd Quarter	191	196	202
4th Quarter	212	216	222

calculate the 4-point moving averages. Plot on one graph both the original figures and the moving average values. Estimate the sales in the first quarter of 2011.

Chapter 2

PROBABILITY

2.1 Introduction

Probability theory is the foundation of inferential statistics. It originated in games of chance. In real life we can never be sure of certain events hence the need for the utilisation of the theory of probability. We are sure of events like, the sun setting tomorrow, dying at an age less than 150 years, being hungry if one does not eat within 12 hours, e.t.c. On the other hand, we are not sure if it will rain tomorrow, if it rains, we are not sure that there will not be lightning and thunder, we are not sure if the pregnant woman will give birth to a baby boy, and many other similar illustrations. An experiment has a definite number of possible outcomes which comprise the outcome set S. The set S is the possibility space or sample space.

2.2 Theoretical Probability

It is not possible to derive all values from an experiment. For instance, when a card is picked from a well shuffled pack of cards, the probability that it is a hearts is

$$\text{P(hearts)} = \frac{\text{number of ways getting a heart}}{\text{total number of possible outcomes}}$$

$$= \frac{13}{52} = \frac{1}{4}$$

This does not mean that for every four cards we pick, there is one heart, but if we pick a card many times with replacement, the chance that it is a hearts is $\frac{1}{4}$.
An event E of an experiment is a subset of the outcome set S. Let E be the event that "John was born on Monday. There are seven days in a week and Monday is only one

of those days.

$$\text{Therefore } P(E) = \frac{1}{7}$$

The outcome set is generated using a table of outcomes, tree diagram or permutations and combinations.

Example 2.2.1

Give the probabilities of

(i) getting a tail when a fair coin is tossed

(ii) picking a diamond from a pack of 52 cards

(iii) picking a black card from a pack of 52 cards

(iv) tossing an even number with a die

(v) being born in the month of March
Solution:

(i) There are only two possibilities i.e a head and a tail. Picking a tail is one of the two probabilities. Therefore $P(T) = \frac{1}{2}$

(ii) There are 13 diamonds in a pack of cards

$$P(D) = \frac{13}{52} = \frac{1}{4}$$

(iii) There are 26 black cards in a pack

$$\text{P(Black card)} = \frac{26}{52} = \frac{1}{2}$$

(iv) The sample space is $S = \{1, 2, 3, 4, 5, 6\}$ There are three even numbers

$$\therefore \text{ P(Even number)} = \frac{3}{6} = \frac{1}{2}$$

(v) There are twelve months in a year. Therefore

$$\text{P(born in march)} = \frac{1}{12}$$

Example 2.2.2

What is the probability of throwing a total score of 8 with two dice
Solution:
We find the points of the sample space using a table.

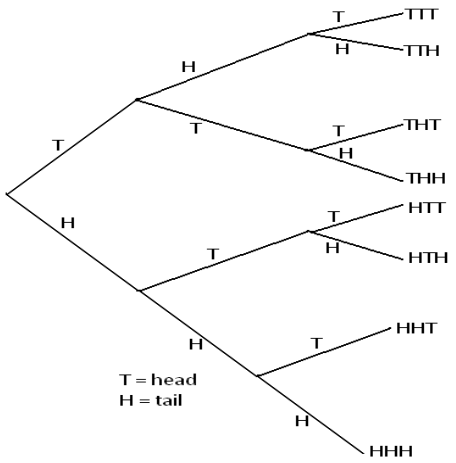
	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

All the points in the sample space have equal probability

$$\text{P(total score } 8) = \frac{5}{36}$$

Example 2.2.3

What is the probability of getting two tails and one head when a coin is tossed three times?
Solution:
We draw a tree diagram as follows.



The eight equally likely outcomes are
S = {TTT,TTH,THT,THH,HTT,HTH,HHT,HHH}
S = { outcomes of throwing a coin three times}
$n(S) = 8$
Let $E = \{\text{two tails and one head}\}$

$$\begin{aligned} n(E) &= 3 \\ P(E) &= \frac{n(E)}{n(S)} = \frac{3}{8} \end{aligned}$$

Note 2.2.1

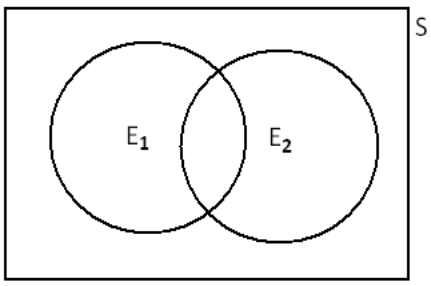
If an experiment has $n(S)$ equally likely outcomes where $n(E)$ of them are the event E, then theoretically, the probability of event E occuring is

$$\begin{aligned} P(E) &= \frac{n(E)}{n(S)} \\ \text{Always} \quad & 0 \leq P(E) \leq 1 \\ P(E) &= 0 \text{ for an impossible event.} \\ P(E) &= 1 \text{ for a sure event} \end{aligned}$$

2.3 Addition Rule

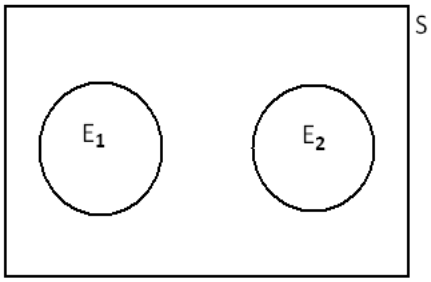
Let E_1 and E_2 be two events from the same experiment. Then, the probability of E_1 or E_2 or both occuring is given by

$$\begin{aligned} P(E_1 \text{ or } E_2) &= P(E_1) + P(E_2) - P(E_1 \text{ and } E_2) \\ \Rightarrow P(E_1 \cup E_2) &= P(E_1) + P(E_2) - P(E_1 \cap E_2) \end{aligned}$$



If the events E_1 and E_2 are mutually exclusive, (have nothing in common) then $P(E_1 \cap E_2) = 0$
So that

$$P(\ E_1 \text{ or} E_2) = P(E_1 \cap E_2) = P(E_1) + P(E_2)$$



Example 2.3.1

What is the probability of drawing a diamond or a six from a pack of cards?
Solution:

$$\begin{aligned}
S &= \{\text{ pack of cards}\},\ n(S) = 52 \\
E_1 &= \{\text{ diamonds},\} \\
n(E_1) &= 13, \\
P(E_1) &= \frac{13}{52} \\
E_2 &= \{\text{sixes}\}, \\
n(E_2) &= 4, P(E_2) = \frac{4}{52} \\
E_1 \text{ and } E_2 &= \{\text{diamonds and sixes}\} \\
&= \{\text{six of diamonds}\} \\
n(E_1 \text{ and } E_2) &= 1, \\
P(E_1 \ \cap \ E_2) &= \frac{1}{52} \\
\Rightarrow P(E_1 \text{ or} E_2) &= P(E_1) + P(E_2) - P(E_1 \text{ and } E_2) \\
&= \frac{13}{52} + \frac{4}{52} - \frac{1}{52}
\end{aligned}$$

$$= \frac{16}{52}$$

$$= \frac{4}{13}$$

Example 2.3.2

Two dice are tossed. What is the probability of scoring either a double or a sum less than four.

Solution: We use a table of outcomes to generate the sample space.

Second die \ First die						
	1,1	1,2	1,3	1,4	1,5	1,6
	2,1	2,2	2,3	2,4	2,5	2,6
	3,1	3,2	3,3	3,4	3,5	3,6
	4,1	4,2	4,3	4,4	4,5	4,6
	5,1	5,2	5,3	5,4	5,5	5,6
	6,1	6,2	6,3	6,4	6,5	,6,6

Table of Sums

2	3	4	5	6	7
3	4	5	6	7	8
4	5	6	7	8	9
5	6	7	8	9	10
6	7	8	9	10	11
7	8	9	10	11	12

Let E_1 represent a double and E_2 a sum less than four.

$$\begin{aligned} E_1 &= \{(1,1),(2,2),(3,3),(4,4),(5,5),(6,6)\} \\ E_2 &= \{(1,1),(2,1),(1,2)\} \\ n(E_1) &= 6,\ n(E_2) = 3 \\ \text{and } n(E_1 \cap E_2) &= 1 \\ \text{From } P(E_1 \cup E_2) &= P(E_1) + P(E_2) - P(E_1 \cap E_2) \end{aligned}$$

$$P(E_1 \cup E_2) = \frac{6}{36} + \frac{3}{36} - \frac{1}{36}$$

$$= \frac{8}{36}$$

$$= \frac{2}{9}$$

$$\Rightarrow P(\text{Double or sum less than four}) = \frac{2}{9}$$

Example 2.3.3

If three coins are tossed once, what is the probability of obtaining three heads or one head?

Solution:

Tossing three coins at once is equivalent to tossing one coin three times. From our earlier example

$$\begin{aligned} S &= \{TTT, TTH, THT, THH, HTT, HTH, HHT, HHH\} \\ n(S) &= 8 \end{aligned}$$

Let $E_1 = \{\text{three heads are obtained}\}$
$E_2 = \{\text{one head is obtained}\}$
E_1 and E_2 are exclusive events so that

$$\begin{aligned} P(E_1 \cap E_2) &= 0 \\ \Rightarrow P(E_1 \cap E_2) &= P(E_1) + P(E_2) \\ \text{But } P(E_1) &= \frac{1}{8} \text{ and } P(E_2) = \frac{3}{8} \\ \therefore P(E_1 \cap E_2) &= \frac{1}{8} + \frac{3}{8} \\ &= \frac{4}{8} \\ &= \frac{1}{2} \end{aligned}$$

For n mutually exclusive events

$$P(E_1 \cap E_2 \cup E_n) = P(E_1) + P(E_2) + ... + P(E_n)$$

This is the addition law for mutually exclusive events.

Example 2.3.4

If $P(A) = 0.5, P(B) = 0.4$ and $P(A \cap B) = 0.2$, find

(i) $P(A \cup B)$

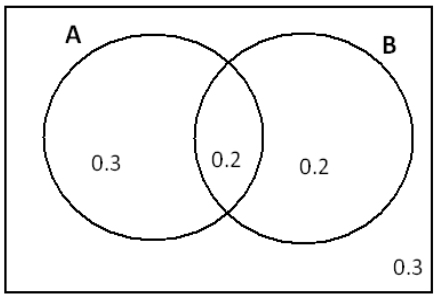
(ii) $P[(A \cup B)']$

(iii) $P(A \cup B')$

(iv) $P(A' \cap B)$.

Solution:

We draw a venn diagram to have a visual impression of the situation, though it is not a necessary thing to do in this case.



(i) From the addition law

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= 0.5 + 0.4 - 0.2 \\ &= 0.7 \end{aligned}$$

(ii)

$$\begin{aligned} P[(A \cup B)'] &= 1 - P(A \cup B) \\ &= 1 - 0.7 \\ &= 0.3 \end{aligned}$$

(iii)

$$\begin{aligned} P(A \cup B') &= P(A) + P(B') - P(A \cap B') \\ &= 0.5 + 0.6 - 0.3 \\ &= 0.8 \end{aligned}$$

(iv) From the diagram $P(A' \cap B) = 0.2$

Note 2.3.1

If A, B and C are events with respective probabilities $P(A), P(B)$ and $P(C)$, the respective probabilities of their not occuring is $P(A'), P(B')$ and $P(C')$. And also

$$\begin{aligned} P(A) + P(A') &= 1, \\ P(B) + P(B') &= 1 \\ \text{and } P(C) + P(C') &= 1. \end{aligned}$$

2.4 Mutually Exclusive Events

Two or more events of the same experiment are mutually exclusive if they cannot occur simultaneously.
For instance if a coin is tosses three times, Let $A =$ { Outcomes with three tails } and $B =$ { Outcomes with one tail }. The events A and B are mutually exclusive. Both cannot occur at the same time.

Example 2.4.1

If a coin is tossed three times, find the probability of obtaining three tails or one tail.
Solution:
The sample space is

$$\begin{aligned} S &= \{TTT,\ TTH,\ HTH,\ HTT,\ THT,\ THH,\ HHT,\ HHH\} \\ \text{Let } A &= \{ \text{ outcomes with three tails}\} \\ B &= \{\text{outcomes with one tail}\} \end{aligned}$$

$$\begin{aligned} P(\text{three tails or one tail}) &= \frac{n\{A \cup B\}}{n(S)} \\ &= \frac{n(A) + n(B)}{n(S)} \\ &= \frac{n(A)}{n(S)} + \frac{n(B)}{n(S)} \\ &= P(A) + P(B) \\ &= \frac{1}{8} + \frac{3}{8} \\ &= \frac{4}{8} \\ &= \frac{1}{2} \end{aligned}$$

This confirms that for mutually exclusive events (no intersection)

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

Example 2.4.2

Events A and B are such that $P(A) = \frac{1}{10}$ and $P(B) = \frac{1}{5}$ and $P(A \cup B) = \frac{3}{10}$. Find out whether A and B are mutually exclusive events.
Solution:

$$\begin{aligned} \text{Generally } P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ \Rightarrow \frac{3}{10} &= \frac{1}{10} + \frac{1}{5} - P(A \cap B) \\ \Rightarrow P(A \cap B) &= \frac{1}{10} + \frac{1}{5} - \frac{3}{10} \\ &= \frac{1+2-3}{10} \\ &= 0 \end{aligned}$$

Since $P(A \cap B) = 0$, A and B are mutually exclusive events.

Note 2.4.1

Two or more events are exhaustive if atleast one of them must happen. If these events are $A, B, C, \ldots K$, then $P(A \cap B \cap C \cap ... \cap K) = 1$

Example 2.4.3

E_1 and E_2 are two events such that $P(E_1) = 0.4, P(E_2) = 0.8$ and $P(E_1 \cap E_2) = 0.2$.
Show that E_1 and E_2 are exhaustive events.
Solution:

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ \text{So } P(E_1 \cup E_2) &= P(E_1) + P(E_2) - P(E_1 \cap E_2) \\ &= 0.4 + 0.8 - 0.2 \\ &= 1 \end{aligned}$$

Since $P(E_1 \cap E_2) = 1$, the events E_1 and E_2 are exhaustive.

Example 2.4.4

A and B are two events such that $P(A) = 0.7, P(B) = 0.4$ and $P(A \cup B) = 0.9$. Find

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

(ii) $P(A' \cap B)$

(iii) $P(A$ or B but not both occur)

(iv) $P(A' \cap B')$.

Solution:

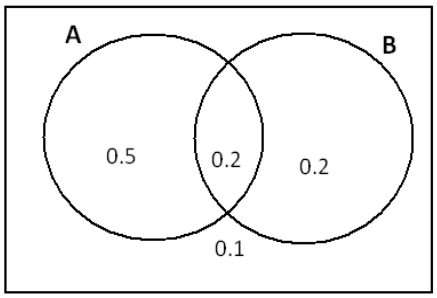
(i)

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ 0.9 &= 0.7 + 0.4 - P(A \cap B) \\ P(A \cap B) &= 0.2 \end{aligned}$$

(ii)

$$\begin{aligned} P(A' \cap B) &= P(B) - P(A \cap B) \\ &= 0.4 - 0.2 \\ &= 0.2 \end{aligned}$$

(iii)



From the diagram

$$\begin{aligned} \text{P(A or B but not both)} &= 0.5 + 0.2 \\ &= 0.7 \end{aligned}$$

(iv)

$$\begin{aligned} P(A' \cap B') &= 1 - P(A \cup B) \\ &= 1 - 0.9 \\ &= 0.1 \end{aligned}$$

2.5 Conditional Probability

If A and B are two events (they dont have to be from the same experiment), then the conditional probability that A will occur given that B has occured is

$$\begin{aligned} P(A/B) &= \frac{P(A\text{and } B)}{P(B)} \\ &= \frac{n(A \cap B)}{n(B)} \end{aligned}$$

Example 2.5.1

If A and B are two events such that $P(A) = \frac{1}{4},\ P(B) = \frac{1}{2}$ and $P(A/B) = \frac{2}{5}$, find

(i) $P(A \cap B)$ (ii) $P(A \cup B)$ (iii) $P(B/A')$

Solution:

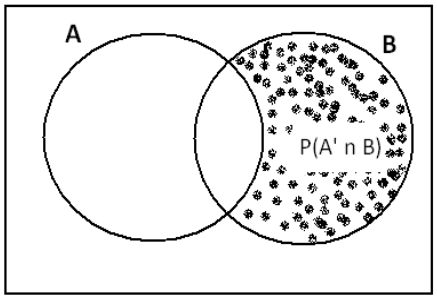
(i)

$$\begin{aligned} P(A/B) &= \frac{P(A \cap B)}{P(B)} \\ \Rightarrow \frac{2}{5} &= \frac{P(A \cap B)}{\frac{1}{2}} \\ P(A \cap B) &= \frac{1}{5} \end{aligned}$$

(ii)

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= \frac{1}{4} + \frac{1}{2} - \frac{1}{5} \\ &= \frac{11}{20} \end{aligned}$$

(iii)



$$\begin{aligned} P(B/A') &= \frac{P(A' \cap B)}{P(A')} \\ &= \frac{P(B) - P(A \cap B)}{P(A')} \\ &= \frac{1/2 - 1/5}{3/4} \\ P(B/A') &= \frac{2}{5} \end{aligned}$$

Example 2.5.2

A and B are events such that $P(B) = \frac{2}{5}$, $P(A \text{ and } B) = \frac{1}{4}$ and $P(B/A) = \frac{1}{3}$. Find

(i) $P(A)$ (ii) $P(A/B)$ (iii) $P(B/A')$ (iv) $P(A/B')$

Solution:

(i)

$$\begin{aligned} P(B/A) &= \frac{P(A \cap B)}{P(A)} \\ \frac{1}{3} &= \frac{1/4}{P(A)} \\ P(A) &= \frac{3}{4}. \end{aligned}$$

(ii)

$$\begin{aligned} P(A/B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{1/4}{2/5} \\ P(A/B) &= \frac{5}{8} \end{aligned}$$

(iii)

$$\begin{aligned} P(B/A') &= \frac{P(B \cap A')}{P(A')} \\ &= \frac{P(B) - P(A \cap B)}{P(A')} \\ &= \frac{2/5 - 1/4}{1/4} \\ P(B/A') &= \frac{3}{5}. \end{aligned}$$

(iv)

$$\begin{aligned} P(A/B') &= \frac{P(A \cap B')}{P(B')} \\ &= \frac{P(A) - P(A \cap B)}{P(B')} \\ &= \frac{3/4 - 1/4}{3/5} \\ P(A/B') &= \frac{5}{6} \end{aligned}$$

Example 2.5.3

Two cards are dealt from a well shuffled pack of 52 cards without replacement. If the first card is a heart, what is the probability that the second card is also a heart?
Solution:
Let $A_1 = \{$first card drawn is a heart$\}$
$A_2 = \{$second card drawn is a heart$\}$

$$P(A_1) = \frac{13}{52} = \frac{1}{4}, \quad P(A_2/A_1) = \frac{12}{51} = \frac{4}{17}$$

Example 2.5.4

A bag contains six white and 14 black beads. What is the probability that if three beads are chosen randomly, they are all black.
Solution:
Let A_n be the event "the n^{th} ball drawn is black
We require $A_1 \cap A_2 \cap A_3$.

$$\begin{aligned} P(A_1 \cap A_2 \cap A_3) &= P(A_1) \cdot P(A_2/A_1) \cdot P(A_3/A_1 \cap A_2) \\ &= \frac{14}{20} \times \frac{13}{19} \times \frac{12}{18} \\ &= \frac{91}{285} \end{aligned}$$

Example 2.5.5

Two cards are drawn from a well shuffled pack of 52 cards without replacement. What is the probability that

(i) they are both Qs

(ii) neither is a Q

(iii) atleast one is a Q,

(iv) either one but not both is a Q.
Solution:

1. (i) Let $A_1 = \{$ first card is a Q$\}$, $A_2 = \{$ second card is a Q $\}$

$$\begin{aligned} P(A_1) &= \frac{4}{52} \text{ and } P(A_2/A_1) = \frac{3}{51} \\ \Rightarrow P(A_2/A_1) &= \frac{P(A_1 \cap A_2)}{P(A_1)} \end{aligned}$$

$$\begin{aligned} P(A_1) \cdot P(A_2/A_1) &= P(A_1 \cap A_2) \\ \frac{4}{52} \times \frac{3}{51} &= P(A_1 \cap A_2) \\ P(A_1 \cap A_2) &= \frac{4}{52} \times \frac{3}{51} \\ &= \frac{1}{221} \end{aligned}$$

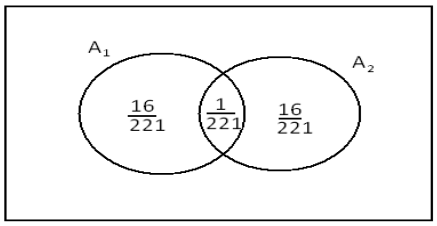
(ii)

$$\begin{aligned} P(A_1' \cap A_2') &= P(A_1') \cdot P(A_2'/A_1') \\ &= \frac{48}{52} \times \frac{47}{51} \\ &= \frac{188}{221} \end{aligned}$$

(iii) P(atleast one is a Q) $= 1 - P$(neither is a Q)

$$\begin{aligned} &= 1 - \frac{188}{221} \\ &= \frac{33}{221} \end{aligned}$$

(iv)



A_2 can occur with or without A_1 occuring.

$$P(A_2) = P(A_1)P(A_2) + P(A_1')P(A_2)$$

$$
\begin{aligned}
&= \frac{4}{52} \cdot \frac{3}{51} + \frac{48}{52} \cdot \frac{4}{51} \\
&= \frac{17}{221} \\
\Rightarrow P(A_2 \cap A_1') &= \frac{17}{24} - \frac{1}{221} \\
&= \frac{16}{221} \\
\text{Therefore } P(A_1 \ or A_2 \text{but not both}) &= \frac{16}{221} + \frac{16}{221} \\
&= \frac{32}{221}
\end{aligned}
$$

Example 2.5.6

A disease affects 2% of the population. A trainee laboratory technician can detect the disease if it is present in the person with a probability of $\frac{9}{10}$ and if the person does not suffer from the disease there is a probability of $\frac{1}{20}$ that he will still say the person has the disease. Find the probability that

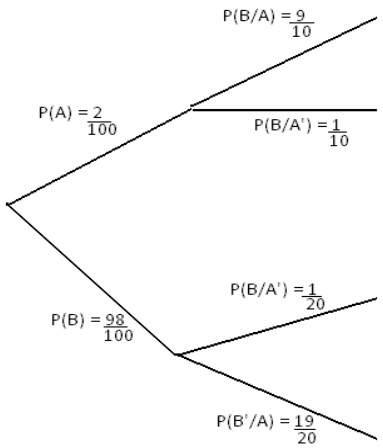
(i) a person has the disease and the technician correctly detects it.

(ii) the test indicates the person has the disease

(iii) the person has the disease given that the test indicates so.

Solution:

Let $A = \{$ a person has the disease $\}$ B = $\{$ a person tests positive $\}$



(i) We need

$$\begin{aligned} P(A \cap B) &= P(A) \times P(B/A) \\ &= \frac{2}{100} \times \frac{9}{10} \\ &= \frac{9}{500} \end{aligned}$$

(ii) Either a person has the disease or he does not have it. A and B are mutually exclusive events. We need to find P(B).

$$\begin{aligned} P(B) &= P(A \cap B) + P(A' \cap B) \\ &= P(A) \cdot P(B/A) + P(A') \cdot P(B/A') \\ &= \frac{2}{100} \times \frac{9}{10} + \frac{98}{100} \times \frac{1}{20} \\ &= \frac{18}{1000} + \frac{98}{2000} \\ &= \frac{67}{1000} \end{aligned}$$

(iii)

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

Utilising (i) and (ii),

$$\begin{aligned} P(A \cap B) &= \frac{9}{500} \\ \text{and } P(B) &= \frac{67}{1000} \\ \Rightarrow P(A/B) &= \frac{9/500}{67/1000} \\ &= \frac{18}{67} \end{aligned}$$

The probability that a person has no disease when the technician says he has it is $1 - \frac{18}{67} = \frac{49}{67}$.
The trainees findings cant be relied on. He needs a lot more training in laboratory work.

2.6 Independent Events

Events are independent if the occurence of one event does not influence or affect the occurance of another. If the events are A and B

$$\begin{aligned} P(A) &= P(A/B) \text{ and} \\ P(B) &= P(B/A). \end{aligned}$$

In that case $P(A \cap B) = P(A) \cdot P(B)$.

Example 2.6.1

A coin is tossed three times and a die is tossed once. What is probability of obtaining one head and an even number.
Solution:

$$\begin{aligned} \text{P(one head)} &= \frac{3}{8} \\ \text{P(an even number)} &= \frac{3}{6} \\ &= \frac{1}{2} \\ \Rightarrow \text{P(one head and an even number)} &= \frac{3}{8} \times \frac{1}{2} \\ &= \frac{3}{16} \end{aligned}$$

Example 2.6.2

Alfred, Bright and Charles are each given one shot. Their probabilities of hitting the target are $\frac{2}{5}, \frac{4}{5}$ and $\frac{1}{4}$ respectively. Find the probability that if they all fire at the target, only one shot will hit the target.
Solution:
Let A = {Alfred hits the target}
B = { Bright hits the target}

C = {Charles hits the target}
D = {only one shot hits the target

$$\begin{aligned}
\text{Then } P(A) &= \frac{2}{5} \text{ and } P(A') = \frac{3}{5} \\
P(B) &= \frac{4}{5} \text{ and } P(B') = \frac{1}{5} \\
P(C) &= \frac{1}{4} \text{ and } P(C') = \frac{3}{4}
\end{aligned}$$

They hit or miss the target independent of each other.
If only one is to hit the target, we need P(D)

$$\begin{aligned}
P(D) &= P(A \cap B' \cap C') + P(A' \cap B \cap C') + P(A' \cap B' \cap C) \\
&= \frac{2}{5} \times \frac{1}{5} \times \frac{3}{4} + \frac{3}{5} \times \frac{4}{5} \times \frac{3}{4} + \frac{3}{5} \times \frac{1}{5} \times \frac{1}{4} \\
&= \frac{6}{100} + \frac{36}{100} + \frac{3}{100} \\
P(D) &= \frac{45}{100} = \frac{9}{20}
\end{aligned}$$

Example 2.6.3

A box contains 6 red, 4 white and 5 black balls. Three balls have to be removed randomly, with replacement. What is the probability that they have the same colour?
Solution:
There are three possible joint events

$$\begin{aligned}
\text{P(all red)} &= \frac{6}{15} \times \frac{6}{15} \times \frac{6}{15} \\
&= \frac{216}{3375} \\
\text{P(all white)} &= \frac{4}{15} \times \frac{4}{15} \times \frac{4}{15} \\
&= \frac{64}{3375} \\
\text{P(all black)} &= \frac{5}{15} \times \frac{5}{15} \times \frac{5}{15}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{125}{3375} \\
\text{P(all are the same colour)} &= \frac{216}{3375} + \frac{64}{3375} + \frac{125}{3375} \\
&= \frac{3}{25} \\
&= 0.12
\end{aligned}
$$

Example 2.6.4

Alfred has a probability of $\frac{1}{2}$ of solving a mathematics problem while Ben has a probability of $\frac{1}{3}$ to do so.
Find the probability that the problem will be solved if both Alfred and Ben solve it independently.
Solution:
Let A = {Alfred solves the problem}
B = {Ben solves the problem }
and P(A) = $\frac{1}{2}$, $P(B) = \frac{1}{3}$
We require $P(A \cup B)$.

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

But for independent events

$$
\begin{aligned}
P(A \cap B) &= P(A) \cdot P(B). \\
\Rightarrow P(A \cup B) &= P(A) + P(B) - P(A) \cdot P(B) \\
&= \frac{1}{2} + \frac{1}{3} - \frac{1}{2} \cdot \frac{1}{3} \\
&= \frac{5}{6} - \frac{1}{6} \\
&= \frac{2}{3}
\end{aligned}
$$

Example 2.6.5

If A and B are independent, show that A and B' are also independent.
Solution:
For independent events

$$
\begin{aligned}
P(A \cap B) &= P(A) \cdot P(B) \\
\text{But } P(A) &= P(A \cap B) + P(A \cap B')
\end{aligned}
$$

$$\begin{aligned} \text{So that } P(A \cap B') &= P(A) - P(A \cap B) \\ P(A \cap B) &= P(A) - P(A) \cdot P(B) \\ &= P(A)[1 - P(B)] \\ &= P(A)P(B') \end{aligned}$$

Hence $P(A \cap B') = P(A)P(B')$. A and B' are also independent events.

Example 2.6.6

Events A, B and C are such that $P(A) = \frac{1}{2}$, $P(B) = \frac{1}{3}$ and $P(C) = \frac{2}{5}$ and $P(B \cap C) = \frac{1}{5}$. If A and B are independent, A and C are mutually exclusive find

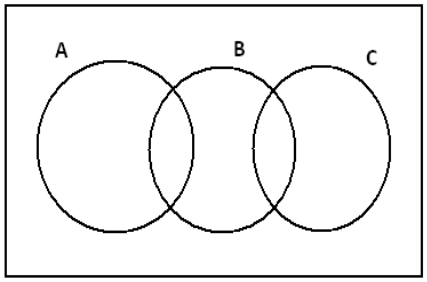
(i) $P(A \cup B)$

(ii) $P(C/B)$

(iii) $P(B/A \cup C)$

(iv) $P(B/C)$.

Solution:



(i)

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= \frac{1}{2} + \frac{1}{3} - \frac{1}{2} \cdot \frac{1}{3} \\ (P(A \cap B) &= P(A) \cdot P(B) \text{ for independence}) \\ &= \frac{5}{6} - \frac{1}{6} \\ &= \frac{2}{3} \end{aligned}$$

(ii)

$$\begin{aligned} P(C/B) &= \frac{P(B \cap C)}{P(B)} \\ &= \frac{1/5}{1/3} \\ &= \frac{3}{5} \end{aligned}$$

(iii)

$$\begin{aligned} P(B/A \cup C) &= \frac{P(B \cap (A \cup C))}{P(A \cup C)} \\ &= \frac{P\{(B \cap A) \cup (B \cap C)\}}{P(A) + P(C)} \end{aligned}$$

Since A and C are mutually exclusive.

$$\begin{aligned} \text{but } P(B \cap A) &= \frac{1}{6} \\ \text{and } P(B \cap C) &= \frac{1}{5} \\ \Rightarrow P(B/A \cup C) &= \frac{P(B \cap A) + P(B \cap C)}{P(A) + P(C)} \\ &= \frac{1/6 + 1/5}{1/2 + 2/5} \\ &= \frac{11/30}{9/10} \\ P(B/A \cup C) &= \frac{11}{27} \end{aligned}$$

(iv)

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

$$= \frac{1/5}{2/5}$$

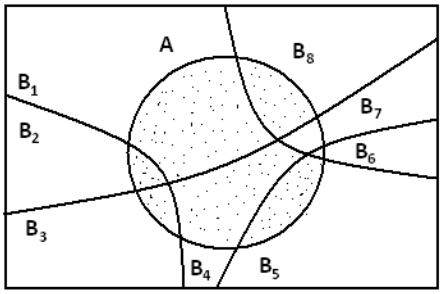
$$= \frac{1}{2}$$

2.7 Baye's Theorem

This theorem is a result of the extension of the theory of conditional probability. If an outcome set S is composed of subsets $B_1, B_2, \ldots, B_n$ and A is also a subset of S with intersections with any of $B_1, B_2, \cdot, B_n$. Then

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

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



The probability of event A is given by

$$P(A) = P(B_1 \cap A) + P(B_2 \cap A) + \ldots + P(B_n \cap A)$$

$$\text{So that } P(B_k/A) = \frac{P(B_k \cap A)}{P(A)}$$

$$= \frac{P(B_k)P(A/B_k)}{P(B_1 \cap A) + P(B_2 \cap A) + \ldots + P(B_n \cap A)}$$

$$P(B_k/A) = \frac{P(B_k)P(A/B_k)}{\sum_{k=1}^{n} P(B_n)P(A/B_n)} \text{ for } k = 1, 2, \ldots n$$

Example 2.7.1

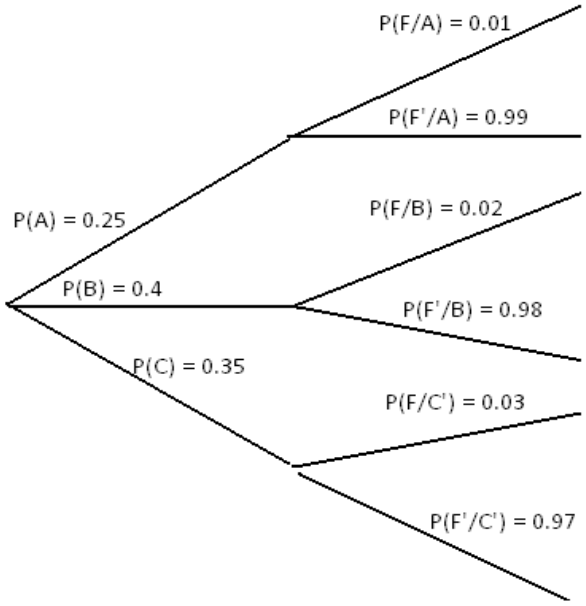
A machine operator buys a particular size of ball bearing from three manufactures A, B and C. He buys 25% of the ball bearings from A, 40% from B and 35% from C. Earlier on he had found that 1% of As ball bearings are faulty whereas 2% of Bs and 3% of Cs are. If he chooses a bearing and finds it faulty, what is the probability that it was one of Cs ball bearings?

Solution:

From the informatioon $P(A) = 0.25, P(B) = 0.4$ and $P(C) = 0.35$. The probability will be greater than P(C) = 0.35 since C produces a greater proportion of faulty ball bearings than A and B. Let F be the event that the ball bearing is faulty. We require P(C/F).

We also have $P(F/A) = 0.01, P(F/B) = 0.02$ and $P(F/C) = 0.03$

Drawing a probability tree diagram



$$P(F \cap C) = P(F/C)P(C) = P(C/F) \cdot P(F)$$

Rearranging, the probability we require is

$$\begin{aligned} P(C/F) &= \frac{P(F/C) \cdot P(C)}{P(F)} \\ \text{But } P(F) &= P(F \cap A) + P(F \cap B) + P(F \cap C) \\ &= P(F/A) \cdot P(A) + P(F/B) \cdot P(B) + P(F/C) \cdot P(C) \\ &= 0.01 \times 0.25 + 0.02 \times 0.4 + 0.03 \times 0.35 \end{aligned}$$

$$
\begin{aligned}
&= 0.0025 + 0.008 + 0.0105 \\
&= 0.021 \\
P(C/F) &= \frac{0.03 \times 0.35}{0.021} \\
&= 0.5
\end{aligned}
$$

The probability that a faulty ball bearing was produced by C is 0.5.
Using a similar argument, the probability that a faulty ball bearing was supplied by A is given by

$$
\begin{aligned}
P(A/F) &= \frac{P(F/A) \cdot P(A)}{P(F/A) \cdot P(A) + P(F/B) \cdot P(B) + P(F/C) \cdot P(C)} \\
&= \frac{0.01 \times 0.25}{0.01 \times 0.25 + 0.02 \times 0.4 + 0.03 \times 0.35} \\
&= \frac{0.0025}{0.0025 + 0.008 + 0.0105} \\
&= 0.119047619 \\
&\simeq 0.119.
\end{aligned}
$$

Example 2.7.2

Three Cooks Annet, Brenda and Cathy sort rice for cooking. Annet sorts 45% of, Brenda sorts 25% and Cathy sorts 30%. The probability that what Annet sorted contains stones is 0.5 and the respective probabilities for Brenda and Cathy are 0.3 and 0.2. What is the probability that a rice container with stones found by a quality assurance person was sorted by Annet?

Solution:

Let P(A), P(B) and P(C) denote the probabilities of quantities sorted by Annet, Brenda and Cathy respectively. Then $P(A) = 0.45, P(B) = 0.25$ and $P(C) = 0.3$
If F is the event "a container of rice has stones
Then P(F/A) = 0.5, P(F/B) = 0.3, P(F/C) = 0.2
We require P(A/F).

$$
\begin{aligned}
\text{Now } P(A/F) &= \frac{P(F/A) \cdot P(A)}{P(F/A) \cdot P(A) + P(F/B) \cdot P(B) + P(F/C) \cdot P(C)} \\
&= \frac{0.5 \times 0.45}{0.5 \times 0.45 + 0.3 \times 0.25 + 0.2 \times 0.3} \\
&= \frac{0.225}{0.36}
\end{aligned}
$$

$$= \quad 0.625$$

2.8 Permutations and Combinations

Assume that we have boys A,B,C,D and E, how many groups of two can we have? These are AB, AC, AD, AE, BC, BD, BE, CD, CE, DE. There are ten possible groups. If we are concerned with the arrangements of the five boys taken two at a time, this can be done in twenty different ways i.e
AB, AC, AD, AE, BC, BD, BE, CD, CE, DE
BA, CA, DA, EA, CB, DB, EB, DC, EC, ED
Each selection is a combination and each arrangement is a permutation. The number of permutations which can be made from n unlike objects taken r at a time is given by

$$nP_r = n(n-1)(n-2)\dots(n-r+1)$$

In our illustration of 5 objects taken 2 at time $nP_r = 5P_2 = 5(4) = 20$
Similarly, the number of combinations which can be made from n unlike objects taken r at a time is nC_r.
Conventionally it is taken that

$$\begin{aligned} n! &= n(n-1)(n-2)\dots 3\times 2\times 1 \\ 6! &= 6\times 5\times 4\times 3\times 2\times 1 \\ 1! &= 1 \\ 0! &= 1 \\ \text{and } nC_r \times r! &= nP_r \\ &= n(n-1)(n-2)\dots(n-r+1) \\ \text{Therefore } nC_r &= \frac{n(n-1)(n-2)\dots(n-r+1)}{r!} \\ \text{and } \ nC_r &= \frac{n!}{r!(n-r)!} \end{aligned}$$

This also implies that $nC_r = nC_{n-r}$

Example 2.8.1

How many numbers can be formed by using three out of the seven digits $1, 2, 3, 4, 5, 6, 7$?
Solution:
We need the number of permutations of seven things taken three at a time

$$7P_3 = 7\times 6\times 5 = 210$$

Example 2.8.2

In how many ways can a committee of five people be selected from 10 people?
Solution:
We need $10C_5$

$$= \frac{10!}{5!5!}$$

$$= 252$$

Example 2.8.3

How many different arrangments of letters can be made by using all the letters of the word "appropriate?
Solution:
We have eleven letters including three P's, two a's and two r's. The required number of arrangements is

$$\frac{11!}{3!2!2!} = \frac{11 \times 10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{3 \times 2 \times 1 \times 2 \times 1 \times 2 \times 1}$$

$$= 1663200.$$

Example 2.8.4

Find the probability that a hand of 7 cards dealt from a pack of playing cards contains only black cards.
Solution:
The total number of possible hands is $52C_7$ and the total number of black cards in a hand is $26C_7$ because the selections of the seven black cards have to be made from 26 black cards.
The required probability is

$$\frac{26C_7}{52C_7} = \frac{657800}{133784560}$$

$$= 0.004916860361$$

$$\simeq 0.004917$$

Example 2.8.5

In how many ways can a netball team of seven girls be picked from 18 possible players?
Solution:
We need

$$18C_7 = \frac{18!}{7!11!}$$

$$= 31824$$

Example 2.8.6

From a group of 8 men and 5 women, a committee of three is selected at random.

(a) What is the probability that all the committeee members are men?

(b) What is the probability that there is one woman on the committe.
Solution:

(a) The total number of people is 13.A committee of three people can be selected in $13C_3$ ways.
If the three committee members are all men, then the three men are chosen from among the men in $8C_3$ ways

$$\begin{aligned} \frac{8C_3}{13C_3} &= \frac{56}{286} = \frac{28}{143} \\ \therefore \text{P(all three are men)} &= \frac{8C_3}{13C_3} = 0.195804195 \\ &\simeq 0.1958 \end{aligned}$$

(b) If there is one woman on the committee, then there are two men on the committe. These are respectively selected in $5C_1$ and $8C_2$ ways. The probability of a woman on the committee is

$$\begin{aligned} \frac{5C_1 \times 8C_2}{13C_3} = \frac{5 \times 28}{286} = \frac{140}{286} = \frac{70}{143} & \\ = 0.489510489 & \\ \simeq 0.4895 & \end{aligned}$$

Example 2.8.7

A box contains 10 red, 6 blue and 4 green balls.
If 5 balls are picked at random, determine the probability that

(i) all balls are red

(ii) exactly three are blue

(iii) atleast one red ball

(iv) no green ball
Solution:

(i)

$$\begin{aligned} \text{P(all balls are red)} &= \frac{10C_5 \times 6C_0 \times 4C_0}{20C_5} \\ &= \frac{21}{1292} \end{aligned}$$

(ii) If three are blue, then two are not blue

$$\begin{aligned} \text{P(three blue balls)} &= \frac{6C_3 \times 14C_2}{20C_5} \\ &= \frac{455}{3876} \end{aligned}$$

(iii)

$$\begin{aligned} \text{P(atleast one red ball)} &= 1 - \text{P(no red ball)} \\ &= 1 - \frac{10C_0 \times 10C_5}{20C_5} \\ &= 1 - \frac{252}{15504} \\ &= \frac{1271}{1292} \end{aligned}$$

(iv)

$$\begin{aligned} \text{P(no green ball)} &= \frac{4C_0 \times 16C_5}{20C_5} \\ &= \frac{4368}{15504} \\ &= \frac{91}{323} \end{aligned}$$

Example 2.8.8

A box contains 3 white, 7 red and 5 blue balls. If three balls are selected at random, find the probabililty that they are

(i) all blue

(ii) of different colours

(iii) of one colour

Solution:

(i)

$$\begin{aligned} \text{P(all blue)} &= \frac{5C_3}{15C_3} \\ &= \frac{10}{455} = \frac{2}{91} \end{aligned}$$

(ii)

$$\begin{aligned} \text{P(different colours)} &= \frac{\begin{pmatrix} 3 \\ 1 \end{pmatrix}\begin{pmatrix} 7 \\ 1 \end{pmatrix}\begin{pmatrix} 5 \\ 1 \end{pmatrix}}{\begin{pmatrix} 15 \\ 3 \end{pmatrix}} \\ &= \frac{3 \times 7 \times 5}{455} = \frac{105}{455} \\ &= \frac{3}{13} \end{aligned}$$

(iii)

$$\begin{aligned} \text{P(all balls are one colour)} &= \frac{3C_3 + 7C_3 + 5C_3}{15C_3} \\ &= \frac{1 + 35 + 10}{455} \\ &= \frac{46}{455} \end{aligned}$$

Note that $nC_r = \begin{pmatrix} n \\ r \end{pmatrix} = \frac{n!}{r!(n-r)!}$

Example 2.8.9

A committe of 4 people is to be chosen from a group of 18 men and 12 women. What is the probability that

(i) all 4 are men

(ii) there are 2 men and 2 women.
Solution:

(i)

$$\begin{aligned} \text{P(all men)} &= \frac{\binom{18}{4}\binom{12}{0}}{\binom{30}{4}} \\ &= \frac{68}{609} \end{aligned}$$

(ii)

$$\begin{aligned} \text{P(2 men and 2 women)} &= \frac{\binom{18}{2}\binom{12}{2}}{\binom{30}{4}} \\ &= \frac{374}{1015} \end{aligned}$$

Exercise 2

1. Given that A and B are independent events such that $P(A) = \frac{3}{5}$ and $P(A \cup B) = \frac{7}{10}$, find
(i) P(B) (ii) P($A' \cup B'$)

2. In Kampala, the probability that a person owns an autombile is $\frac{1}{30}$. Given that the probability that a person who owns an automobile is Hiv positive is $\frac{1}{4}$, find the probability that a person selected at random owns an automobile and is HIV positive.

3. If a die is tossed once, what is the probability of a number less or equal to two showing up?

4. Given that A and B are two events such that $P(A) = \frac{6}{10}, P(B) = \frac{7}{10}$ and $P(A \cup B) = \frac{4}{5}$, find

   (i) $P(A \cap B)$ (ii) $P(A \cap B')$

5. A and B are mutually exclusive events such that $P(A) = \frac{2}{5}$ and $P(A \cup B) = \frac{17}{20}$. Find

(i) P(A′ ∪ B) (ii) $P(A' \cap B')$

6. Two dice are thrown together. Find the probability of scoring a double or a sum greater than 6.

7. Four balls are drawn at random one after the other without replacement from a bag containing 15 white, 10 blue, 8 green and 7 yellow balls. Determine the probability that

   (i) the first is white, the second is blue, the third is green and the fourth is yellow

   (ii) there are two white balls

   (iii) there is one ball of each colour

   (iv) all balls are of one colour.

8. Three men A,B and C have probabilities $\frac{2}{5}, \frac{1}{3}$ and $\frac{3}{10}$ of hitting a target. If each of them is allowed one shot at the target, find the probability that

   (i) one and only one shot will be on target

   (ii) the target will be hit

   (iii) all will miss target.

9. A and B are events such that $P(A) = \frac{1}{5}, P(A \cap B) = \frac{1}{7}$ and $P(A/B) = \frac{1}{3}$. Find (i) P(B) (ii) P(B/A) (iii) P(A/B′)

10. Show that $P(C/A) + P(C'/A) = 1$.

11. A box contains 6 white marbles and five green marbles. Two marbles are drawn at random one at a time without replacement. Find the probability that

    (i) all are white

    (ii) the second marble is white

    (iii) the first marble is white given that the second is white

    (iv) none of the marbles are white.

Chapter 3

DISCRETE PROBABILITY DISTRIBUTIONS

3.1 Introduction

A probability density function (pdf), $P(X = x)$ is discrete if its domain is countable. If the outcome set S of an experiment is devided into n mutually exclusive and exhaustive events $E_1, E_2, \ldots, E_n$, then a variable, X which can take (assume) exactly n numerical values each of which corresponds to only one of events is called a random variable.
The pdf of a discrete random variable X is a function which allocates probabilities to all distinct values that X can take. If $P(X = x)$ is the pdf for a random variable X defined for the discrete values of X, then $\sum P(X = x) = 1$.

Example 3.1.1

A bag contains 10 blue and 6 red marbles. Three marbles are drawn at random without replacement. Find the probability distribution for the number of red marbles drawn.
Solution:
Let the random variable X be "the number of red marbles drawn"

$$\begin{aligned} P(X=0) &= (\text{no red marbles}) = P(B_1 \cdot B_2 \cdot B_3) \\ P(x=0) &= P(B_1) \times P(B_2/B_1) P(B_3/B_2 \text{ and } B_1) \\ &= \frac{10}{16} \times \frac{9}{15} \times \frac{8}{14} \\ &= \frac{3}{14} \\ P(X=1) &= P(R_1 \cdot B_2 \cdot B_3 \text{ or } B_1 \cdot R_2 \cdot B_3 \text{ or } B_1 \cdot B_2 \cdot R_3) \\ &= \frac{6}{10} \times \frac{10}{15} \times \frac{9}{14} + \frac{10}{16} \times \frac{6}{15} \times \frac{9}{14} + \frac{10}{16} \times \frac{9}{15} \times \frac{6}{14} \end{aligned}$$

$$
\begin{aligned}
&= \frac{9}{56} + \frac{9}{56} + \frac{9}{56} \\
&= \frac{27}{56} \\
P(X = 2) &= (R_1 \cdot R_2 \cdot B_3 \text{ or } R_1 \cdot B_2 \cdot R_3 \text{ or } B_1 \cdot R_2 \cdot R_3) \\
&= \frac{6}{16} \times \frac{5}{15} \times \frac{10}{14} + \frac{6}{16} \times \frac{10}{15} \times \frac{5}{14} + \frac{10}{16} \times \frac{6}{15} \times \frac{5}{14} \\
&= \frac{5}{56} + \frac{5}{56} + \frac{5}{56} \\
&= \frac{15}{56} \\
P(X = 3) &= P(R_1 \cdot R_2 \cdot R_3) \\
&= \frac{6}{16} \times \frac{5}{15} \times \frac{4}{14} \\
&= \frac{1}{28}
\end{aligned}
$$

hence the probability distribution of X is

x	0	1	2	3
P(X = x)	$\frac{3}{14}$	$\frac{27}{56}$	$\frac{15}{56}$	$\frac{1}{28}$

Note that the sum of the probabilities is 1.

Example 3.1.2

In a packet of eight bulbs, three are known to be defective. If three bulbs are chosen at random find the probabilities that

(a) (i) none
(ii) one
(iii) two
(iv) three are defective

(b) Hence give the probability distribution for defective bulbs drawn.
Solution:

(a) (i)

$$\text{P(none defective)} = \frac{\binom{3}{0}\binom{5}{3}}{\binom{8}{3}} = \frac{10}{56}$$

(ii)

$$\text{P(One defective)} = \frac{\binom{3}{1}\binom{5}{2}}{\binom{8}{3}} = \frac{30}{56}$$

(iii)

$$\text{P(two defective)} = \frac{\binom{3}{2}\binom{5}{1}}{\binom{8}{3}} = \frac{15}{56}$$

(iv)

$$\text{P(three defective)} = \frac{\binom{3}{3}\binom{5}{0}}{\binom{8}{3}} = \frac{1}{56}$$

(b)

x	0	1	2	3
$P(X = x)$	$\frac{10}{56}$	$\frac{30}{56}$	$\frac{15}{56}$	$\frac{1}{56}$

NB: The "number of defective bulbs" is variable because it can take different numerical values between zero and three inclusive; it is random because it it not easy to predict the outcome of counting the number of defective bulbs in a particular packet and it is discrete since it can take only certain specific values in a given range rather than all the values in that range.

3.2 Mean

Mean is also called expection. For a discrete random variable X with pdf $P(X = x)$, the mean of x is $E(X) = xP(X = x)$.
E(x) is the mean value μ of x. $E(x)$ has the following properties

(i) $E(a) = a$ a constant

(ii) $E(ax) = aE(x)$

(iii) $E[G(x)] = \sum G(x)P(X = x)$

(iv) If $F(x)$ and $G(x)$ are any two functions of X,

$$E[F(x) + G(x)] = E[F(x)] + E[G(x)].$$

Example 3.2.1

Use the probability distribution found in the example 3.1.2 to find (i) $E(x)$ (ii) $E(x^2)$.

Solution:

x	0	1	2	3
$P(X = x)$	$\frac{10}{56}$	$\frac{30}{56}$	$\frac{15}{56}$	$\frac{1}{56}$

(i)

$$\begin{aligned} E(x) &= \sum xP(x) \\ &= 0 \times \frac{10}{56} + 1 \times \frac{30}{56} + 2 \times \frac{15}{56} + 3 \times \frac{1}{56} \\ &= 0 + \frac{30}{56} + \frac{30}{56} + \frac{3}{56} \\ &= 1\frac{1}{8} \end{aligned}$$

(ii)

$$\begin{aligned} E(x^2) &= \sum x^2 P(X = x) \\ &= 0^2 \times \frac{10}{56} + 1^2\frac{30}{56} + 2^2 \times \frac{15}{56} + 3^2 \times \frac{1}{56} \\ &= 0 + \frac{30}{56} + \frac{60}{56} + \frac{9}{56} \\ &= 1\frac{43}{56} \end{aligned}$$

Example 3.2.2

A fair coin is tossed three times. Let x represent the number of heads which show up. Find the probability distribution of x and hence $E(x)$ and $E(x^2)$.

Solution:

The distribution of the probability is given by

$$\left(\begin{array}{c} 3 \\ x \end{array} \right) \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{n-x}, \; n = 3$$

$$\begin{aligned} P(X=0) &= \left(\begin{array}{c} 3 \\ 0 \end{array} \right) \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^3 = \frac{1}{8} \\ P(X=1) &= \left(\begin{array}{c} 3 \\ 1 \end{array} \right) \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^2 = \frac{3}{8} \\ P(X=2) &= \left(\begin{array}{c} 3 \\ 2 \end{array} \right) \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^1 = \frac{3}{8} \\ P(X=3) &= \left(\begin{array}{c} 3 \\ 3 \end{array} \right) \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^0 = \frac{1}{8} \end{aligned}$$

The probability distribution of X is

x	0	1	2	3
P(X = x)	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

$$\begin{aligned} E(x) &= \sum xP(X=x) \\ &= 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} \\ &= 1\frac{1}{2} \\ E(x^2) &= \sum x^2 P(x=x) \\ &= 0^2 \times \frac{1}{8} + 1^2 \times \frac{3}{8} + 2^2 \times \frac{3}{8} + 3^2 \times \frac{1}{8} \\ &= 3 \end{aligned}$$

Example 3.2.3

A random variable x has a probability density function shown below:

x	0	1	2	3	4
P(X = x)	0.1	0.2	0.4	0.1	0.2

Calculate

(i) $P(0 < x < 2)$

(ii) $P(X < 2)$

(iii) $E(x)$

(iv) $E(x^2)$

Solution:

(i)

$$\begin{aligned} P(0 < x < 2) &= P(X = 1) \text{ (this is discrete)} \\ &= 0.2 \end{aligned}$$

(ii)

$$\begin{aligned} P(X > 2) &= P(X = 3) + P(X = 4) \\ &= 0.1 + 0.2 \\ &= 0.3 \end{aligned}$$

(iii)

$$\begin{aligned} E(x) &= \sum xP(X = x) \\ &= 0 \times 0.1 + 1 \times 0.2 + 2 \times 0.4 + 3 \times 0.1 + 4 \times 0.2 \\ &= 0.2 + 0.8 + 0.3 + 0.8 \\ &= 2.1 \end{aligned}$$

(iv)

$$\begin{aligned} E(x^2) &= \sum x^2 P(X = x) \\ &= 0^2 \times 0.1 + 1^2 \times 0.2 + 2^2 \times 0.4 + 3^2 \times 0.1 + 4^2 \times 0.2 \\ &= 0 + 0.2 + 1.6 + 0.9 + 3.2 \\ &= 5.9. \end{aligned}$$

3.3 Variance

The variance of a probability distribution associated with a random variable X is

$$\begin{aligned} V(X) &= E[(X-\mu)^2] \text{ where } \mu = E(x). \\ V(X) &= E[(X-\mu)^2] \\ &= E[X^2 - 2\mu x + \mu^2] \\ &= E(X^2) - 2\mu E(x) + \mu^2 \\ \text{but } E(X) &= \mu \\ \Rightarrow V(X) &= E(x^2) - 2\mu^2 + \mu^2 \\ &= E(X^2) - \mu^2 \text{ or} \\ &\quad E(X^2) - [E(X)]^2 \end{aligned}$$

Example 3.3.1

The discrete random variable X has the following probability distribution.

x	0	1	2	3	4
P(X = x)	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{16}$	$\frac{1}{8}$	$\frac{3}{16}$

Find the mean and variance of X.
Solution:

$$\begin{aligned} E(x) &= \sum xP(X=x) \\ &= 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{16} + 3 \times \frac{1}{8} + 4 \times \frac{3}{16} \\ &= 0 + \frac{3}{8} + \frac{6}{16} + \frac{3}{8} + \frac{12}{16} \\ &= \frac{15}{8} \\ V(x) &= E(x^2) - [E(x)]^2 \\ &= E(X^2) - \left(\frac{15}{8}\right)^2 \\ &= \{0^2 \times \frac{1}{8} + 1^2 \times \frac{3}{8} + 2^2 \times \frac{3}{16} + 3^2 \times \frac{1}{8} + 4^2 \times \frac{3}{16}\} - \left(\frac{15}{8}\right)^2 \\ &= 5\frac{1}{4} - \left(\frac{15}{8}\right)^2 \end{aligned}$$

$$= \frac{21}{4} - \frac{225}{64}$$

$$= \frac{336 - 225}{64}$$

$$= \frac{111}{64}$$

$$= 1.734375.$$

Example 3.3.2

A random variable X has pdf given by $P(X = x) = kx, x = 1, 2, 3, 4$ and $P(X = x) = k(8 - x)$, for $x = 5, 6, 7$. Determine

(i) the constant k

(ii) Expectation of X

(iii) Variance of X.

Solution:

x	x^2	$P(X = x)$	$xP(X = x)$	$x^2P(X = x)$
1	1	k	k	k
2	4	2k	4k	8k
3	9	3k	9k	27k
4	16	4k	16k	64k
5	25	3k	15k	75k
6	36	2k	12k	72k
7	49	k	7k	49k
		16k	64k	296k

(i)

$$\sum P(X = x) = 1$$

$$16k = 1$$

$$k = \frac{1}{16}$$

(ii)

$$\begin{aligned} E(x) &= \sum xP(X=x) = 64k \\ &= 64 \times \frac{1}{16} \\ &= 4 \end{aligned}$$

(iii)

$$\begin{aligned} V(x) &= E(x^2) - (E(x))^2 \\ &= \sum x^2 P(X=x) - [E(x)]^2 \\ &= 296k - 42 \\ &= \frac{296}{16} - 16 \\ &= 2.5 \end{aligned}$$

Note that if a and b are constants

(i) $Var(a) = 0$

(ii) $Var(ax) = a^2V(X)$

(iii) $V(ax+b) = Var(ax) + V(b) = a^2V(x) + 0 = a^2V(x)$

(iv) $V(ax+by) = V(ax) + V(by) = a^2V(x) + b^2V(Y)$

(v) $V(X+Y) = V(X-Y) = V(X) + V(Y)$

Example 3.3.3

If X and Y are two independent random variable with $E(X) = 0.8$, $E(Y) = 0.9$, $V(X) = 0.4$ and $V(Y) = 0.5$ find

(i) $E[2X+3Y]$

(ii) $V(X+Y)$

Solution:

(i)

$$\begin{aligned} E(2X+2Y] &= E(2X)+E(3Y) \\ &= 2E(X)+3E(Y) \\ &= 2\times 0.8+3\times 0.9 \\ &= 1.6+2.7 \\ &= 4.3 \end{aligned}$$

(ii)

$$\begin{aligned} V(X+Y) &= V(X)+V(Y) \\ &= 0.4+0.5 \\ &= 0.9 \end{aligned}$$

3.4 The Cummulative Mass Function (cmf)

The cummulative mass function of a discrete pdf is given by $F(x) = P(X \leq x)$. The pdf of a discrete random variable is often called the probability mass function pmf.

Example 3.4.1

From the probability distribution table below, determine the cummulative mass function

x	0	1	2	3
P(X = x)	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Solution:

$$\begin{aligned} F(0) &= P(X \leq 0) = \frac{1}{8} \\ F(1) &= P(X \leq 1) \\ &= P(X=0)+(X=1) \\ &= \frac{1}{8}+\frac{3}{8} \\ &= \frac{1}{2} \\ F(2) &= P(X \leq 2) \\ &= P(X=0)+P(X=1)+P(X=1)+P(X=2) \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{8}+\frac{3}{8}+\frac{3}{8} \\
&= \frac{7}{8} \\
F(3) &= P(X \leq 3) \\
&= P(x=0)+P(x=1)+P(X=2)+P(x=3) \\
&= \frac{1}{8}+\frac{3}{8}+\frac{3}{8}+\frac{1}{8} \\
&= 1
\end{aligned}
$$

Example 3.4.2

A discrete random variable X has the cummulative mass function $F(x)=\frac{x}{k}$ for $x=1,2,3,4$.
Find the probability distribution of $X, E(X)$ and $V(X)$
Solution:

x	1	2	3	4
F(X)	$\frac{1}{k}$	$\frac{2}{k}$	$\frac{3}{k}$	$\frac{4}{k}$

$$
\begin{aligned}
\Rightarrow \quad k &= 4 \\
P(X=1) &= F(1)=\frac{1}{4} \\
P(X=2) &= F(2)=F(2)-F(1) \\
&= \frac{2}{4}-\frac{1}{4} \\
&= \frac{1}{4} \\
P(X=3) &= F(3)-F(2) \\
&= \frac{3}{4}-\frac{2}{4} \\
&= \frac{1}{4} \\
P(X=4) &= F(4)-F(3)
\end{aligned}
$$

$$= \frac{4}{4} - \frac{3}{4}$$

$$= \frac{1}{4}$$

The probability distribution is

x	1	2	3	4
P(X = x)	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

$$\begin{aligned} E(X) &= \sum xP(x) \\ &= 1 \times \frac{1}{4} + 2 \times \frac{1}{4} + 3 \times \frac{1}{4} + 4 \times \frac{1}{4} \\ &= 2.5 \\ V(x) &= E(x^2) - \mu^2 \\ &= \sum x^2 P(X = x) - \mu^2 \\ &= 1^2 \times \frac{1}{4} + 2^2 \times \frac{1}{4} + 3^2 \times \frac{1}{4} + 4^2 \times \frac{1}{4} - (2.5)^2 \\ &= 7\frac{1}{2} - \frac{25}{4} \\ &= 1\frac{1}{4} \end{aligned}$$

3.5 Median

The median of a probability distribution of a random variable X is the smallest value for which $\sum P(X = x) \geq 0.5$. If the median is $m, F(m) \geq 0.5$ or $1 - F(m-1) \geq 0.5$.

Example 3.5.1

A discrete variable X has the distribution $P(X = 1) = 0.2, P(X = 2) = 0.3$
$P(X = 3) = 0.3$ $P(X = 4) = 0.1$ $P(X = 5) = 0.1$. Find

(i) the cummulative mass function

(ii) the median

(iii) sketch the graphs of $f(x)$ and $F(x)$.
Solution:

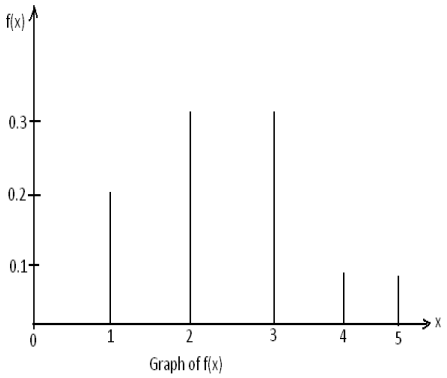
(i)

X	P(X = x)	$P(X \leq x) = F(x)$
1	0.2	0.2
2	0.3	0.5
3	0.3	0.8
4	0.1	0.9
5	0.1	1.0

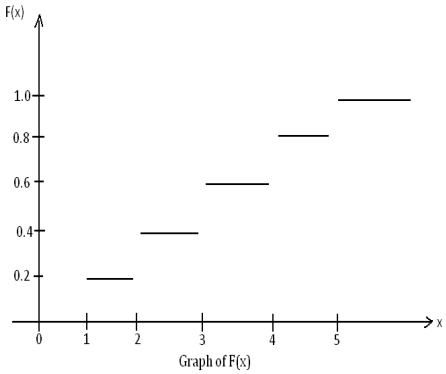
$$\begin{aligned} F(x) &= P(X \leq x) \\ F(1) &= 0.2, \quad F(2) = 0.5, \\ F(3) &= 0.8, \quad F(4) = 0.9, \\ F(5) &= 1.0 \end{aligned}$$

(ii) $F(2) = 0.5 \Rightarrow$ median $= 2$

(iii)



Graph of f(x)



Graph of F(x)

Example 3.5.2

A random variable X has the probability function

$$P(X=x)=\begin{cases} k\cdot 2^x, & x=0,1,2,3,4 \\ 0, & \text{elsewhere} \end{cases}$$

Determine

(i) the value of k

(ii) $E(x)$

(iii) $V(x)$

(iv) median of x

Solution:

(i)

$$P(X=x)=\begin{cases} k\cdot 2^x, & x=0,1,2,3,4 \\ 0, & \text{elsewhere} \end{cases}$$

$$\begin{aligned} k(2^0+2^1+2^2+2^3+2^4) &= 1 \\ K(1+2+4+8+16) &= 1 \\ k &= \frac{1}{31} \end{aligned}$$

(ii)

x	0	1	2	3	4
$P(X=x)$	$\frac{1}{31}$	$\frac{2}{31}$	$\frac{4}{31}$	$\frac{8}{31}$	$\frac{16}{31}$

$$\begin{aligned} E(X) &= \sum xP(X=x) \\ &= 0\times\frac{1}{31}+1\times\frac{2}{31}+2\times\frac{4}{31}+2\times\frac{4}{31}+3\times\frac{8}{31}+4\times\frac{16}{31} \\ &= \frac{98}{31} \end{aligned}$$

(iii)

$$\begin{aligned} V(x) &= E(x^2) - \mu^2 \\ &= \sum x^2 P(X = x) - \mu^2 \\ &= 0^2 \times \frac{1}{31} + 1^2 \times \frac{2}{31} + 2^2 \times \frac{4}{31} + 2^2 \times \frac{4}{31} + 3^2 \times \frac{8}{31} + 4^2 \times \frac{16}{31} - \left(\frac{98}{31}\right)^2 \\ &= \frac{346}{31} - \left(\frac{98}{31}\right)^2 \\ &= \frac{1122}{31} = 1\frac{161}{961} \end{aligned}$$

(iv) $F(m) \geq 0.5$ for median

$$\begin{aligned} F(0) &= \frac{1}{31} \\ F(1) &= \frac{1}{31} + \frac{2}{31} = \frac{3}{31} \\ F(2) &= \frac{3}{31} + \frac{4}{31} = \frac{7}{31} \\ F(3) &= \frac{7}{31} + \frac{8}{31} = \frac{15}{31} \\ F(4) &= \frac{15}{31} + \frac{16}{31} = 1 \\ \Rightarrow \text{ Median of } x &= 4. \end{aligned}$$

Example 3.5.3

A discrete random variable has a probability function

$$P(X = x) = \begin{cases} k\left(\frac{1}{4}\right)^x, & x = 0, 1, 2, 3, 4 \\ 0, & \text{elsewhere} \end{cases}$$

Determine

(i) the constant k

(ii) $P(x \leq 2)$

(iii) the median of x

Solution:

(i)

$$k(\frac{1}{4})^x, x = 0, 1, 2, 3, 4.$$

$$k\left[\left(\frac{1}{4}\right)^0 + \left(\frac{1}{4}\right)^1 + \left(\frac{1}{4}\right)^2 + \left(\frac{1}{4}\right)^3 + \left(\frac{1}{4}\right)^4\right] = 1$$

$$k = \frac{256}{341}$$

(ii)

$$\begin{aligned} P(X \leq 2) &= P(X = 0) + P(x = 1) + P(x = 2) \\ &= \frac{256}{341} + \frac{64}{341} + \frac{16}{341} \\ &= \frac{336}{341} \end{aligned}$$

(iii) $F(0) = \frac{256}{341}$ since this exceeds $\frac{1}{2}$ the median is 0.
NB: I have not bothered with ways of finding the mode because it is the value of x with the highest probability. If one has a probability distribution of X, then that value of x which has the highest probability is the mode.

Exercise 3

1. A discrete random variable X has the following probability distribution

x	0	1	2	3	4
$P(X = x)$	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{1}{5}$	$\frac{1}{20}$

Find the mean and variance of X

2. A random variable X has the following probability distribution

x	1	2	3	4
$P(X=x)$	0.2	0.1	0.4	0.3

Find (i) E(X) (ii) V(x) (iii) $P(x=2/x\geq 2)$

3. Musoki is given pocket money using the following condition: Her father rolls a die and gives her 400/ = (a hundred shillings) for each sport on the upper most face of the die. What is her expected pocket money?

4. A random variable X can take the values shown in the table below with the given probabilities

x	0	1	2	3	4	5
$P(X=x)$	$\frac{1}{5}$	$\frac{3}{10}$	$\frac{1}{10}$	$\frac{1}{8}$	$\frac{11}{80}$	$\frac{11}{80}$

   (i) Calculate the expected value and variance of X

   (ii) Calculate the expected value and variance of X^2.

5. A bag contains 4 green and 6 black marbles. A sample of three marbles is drawn at random from the bag without replacement.

   (i) Find the probability of obtaining exactly two green marbles in the sample.

   (ii) Find the probability of obtain exactly two black marbles in the sample

   (iii) Find the expected numbers of green marbles in the sample.

6. A discrete random variable X has the probability mass function

$$P(X=x)=\begin{cases}\frac{x}{k}, & x=0,1,2,3,4\\ 0, & \text{elsewhere}\end{cases}$$

   Find

   (i) the constant k

   (ii) the expectation of X

   (iii) median and variance of X

   (iv) $P(x=3/X\geq 3)$

7. A random variable x has the probability distribution given in the table below.

x	6	10	12	16	20
$P(X)$	$\frac{2}{15}$	$\frac{1}{5}$	$\frac{1}{3}$	$\frac{4}{15}$	$\frac{1}{15}$

(i) Find $E(X)$ and $V(X)$

(ii) If $2Y = X + 5$, find $E(Y)$ and $V(Y)$

8. If X and Y are two independent random variables with $E(X) = 0.3, E(Y) = 0.9, V(X) = 0.15$ and $V(Y) = 0.4$,
find

(i) $E(3X + 5Y)$

(ii) $V(X + Y)$

(iii) $V(X - Y)$

(iv) $V(2X + 4Y)$

9. The discrete random variable X has the following probability distribution

x	0	1	2
$P(X = x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

(i) find the mean and variance of X

(ii) If two independent random variables X_1 and X_2 have the same distribution as X, find the distribution of $X_1 - X_2$. Solve for its mean and variance.

10. (i) A couple plans to have four children. Construct the probability distribution table for the number of boys they give birth to.

(ii) Find the expected number of boys

11. A random variable X has the probability mass function

$$P(X = x) = \begin{cases} C.3^x, & x = 0, 1, 2, 3 \\ 0, & \text{elsewhere} \end{cases}$$

Find

(i) the value of the constant C

(ii) $E(x)$ and $V(x)$

(iii) $P(x \leq 2)$

12. A discrete randaom variable X has probability function

$$P(X = x) = \begin{cases} \frac{2+x}{kx}, & x = 0, 1, 2, 3, 4, 5 \\ 0, & \text{otherwise} \end{cases}$$

Determine

(i) the value of k

(ii) the expectation of X

(iii) the variance of X

(iv) $P(X \geq 3/X \leq 5)$

13. A bag contains 7 blue and 5 red marbles. Four marbles are drawn at random and not replaced.
Find the probabilities that

(i) no red marble is drawn

(ii) exactly two red marbles are drawn

(iii) three blue marbles are drawn

(iv) no blue marble is drawn

Chapter 4

THE BINOMIAL DISTRIBUTION

4.1 Introduction

A dicrete random variable X having a probability density function of the form

$$P(X = x) = \begin{pmatrix} n \\ x \end{pmatrix} p^x q^{n-x}, n = 0, 1, 2, ..., n$$

where $q = 1 - p$ is said to have a binomial distribution. The properties of a binomial distribution are that

1. A single trial has only two possible mutually exclusive and exhaustive results. These results are either a "success" or a "failure".
2. the values of p and q are constant through out all the trials
3. the result of each trial is independent of previous trials
4. the number of trials n is constant.

Example 4.1.1

The probability of winning a game is $\frac{3}{5}$. Eight games are played. What is the probability of

(i) four success

(ii) atleast two successes.
Solution:
Here $n = 8, p = \frac{3}{5}$ and $q = \frac{2}{5}$

(i)

$$\begin{aligned} P(X=4) &= \binom{8}{4}\left(\frac{3}{5}\right)^4\left(\frac{2}{5}\right)^4 \\ &= 70\times\frac{81\times 16}{5^8} \\ &= 0.2322432 \\ &\simeq 0.2322 \end{aligned}$$

(ii)

$$\begin{aligned} P(X\geq 2) &= 1-\{P(X=0)+P(X=1)\} \\ &= 1-\left\{\binom{8}{0}\left(\frac{3}{5}\right)^0\left(\frac{2}{5}\right)^8+\binom{8}{1}\left(\frac{3}{5}\right)^1\left(\frac{2}{5}\right)^7\right\} \\ &= 1-\{0.00065536+0.00786432\} \\ &= 1-0.00851968 \\ &= 0.99148032 \\ &\simeq 0.9915. \end{aligned}$$

Example 4.1.2

The probability that a patient recovers from the Ebola disease is 0.2. If 6 people are known to have contracted the disease, what is the probability that

(i) atleast 2 survive

(ii) atmost 4 will survive

(iii) between 2 and 4 survive.

Solution:

Here n = 6, p = 0.2 and q = 0.8

(i)

$$\begin{aligned} \text{P(atleast 2)} &= P(X\geq 2) \\ &= 1-\{P(X=0)+P(X=1)\} \\ &= 1-\left\{\binom{6}{0}\left(\frac{1}{5}\right)^0\left(\frac{4}{5}\right)^6+\binom{6}{1}\left(\frac{1}{5}\right)^1\left(\frac{4}{5}\right)^5\right\} \end{aligned}$$

$$
\begin{aligned}
&= 1 - \{0.262144 + 0.393216\} \\
&= 0.34464 \\
&\simeq 0.3446
\end{aligned}
$$

(ii)

$$
\begin{aligned}
\text{P(atmost } 4) &= P(X \leq 4) \\
&= 1 - P(X \geq 5) \\
&= 1 - \{P(X = 5) + P(X = 6)\} \\
&= 1 - \left\{ \begin{pmatrix} 6 \\ 5 \end{pmatrix} \left(\frac{1}{5}\right)^5 \left(\frac{4}{5}\right)^1 + \begin{pmatrix} 6 \\ 6 \end{pmatrix} \left(\frac{1}{5}\right)^6 \left(\frac{4}{5}\right)^0 \right\} \\
&= 1 - \{0.001526 + 0.000064\} \\
&= 1 - 0.0016 \\
&= 0.9984
\end{aligned}
$$

(iii)

$$
\begin{aligned}
P(2 \leq x \leq 4) &= 1 - \{P(x < 2) + P(x > 4)\} \\
&= 1 - \{P(x = 0) + P(x = 1) + P(x = 5) + P(x = 6)\} \\
&= 1 - \{0.262144 + 0.393216 + 0.001536 + 0.000064\} \\
&= 1 - 0.65696 \\
&= 0.34304.
\end{aligned}
$$

Example 4.1.3

A coin is tossed four times. Find the probability that

(i) three heads are obtained

(ii) no head is obtained

(iii) atleast one head is obtained.

Solution:

Here $P = \frac{1}{2}, q = \frac{1}{2}, n = 4$

Let X be the event "a head is obtained"

(i)

$$
P(X = 3) = \begin{pmatrix} 4 \\ 3 \end{pmatrix} \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^1
$$

$$= 4 \times \frac{1}{16}$$

$$= \frac{1}{4}.$$

(ii)

$$P(X = 0) = \begin{pmatrix} 4 \\ 0 \end{pmatrix} \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^4$$

$$= \left(\frac{1}{2}\right)^4$$

$$= \frac{1}{16}$$

(iii)

$$P(X \geq 1) = 1 - P(x = 0)$$

$$= 1 - \frac{1}{16}$$

$$= \frac{15}{16}.$$

Note 4.1.1

Values of $\begin{pmatrix} n \\ r \end{pmatrix} p^r q^{n-r}, r = 0, 1, \ldots n$

Can be obtained from tables if p is a common fraction. Try to read off the probabilities we have so far got from the binomial distribution tables at the end of the book.

Example 4.1.4

In Musoki's family, the probability of having a girl is 0.3. If there are 6 children, determine the probability that

(i) they are all boys

(ii) there is atleast two girls

(iii) there are three girls

(iv) all are girls.

Solution:
Here $n = 6, p = 0.3, q = 0.7$
Let x be the event of having a girl

(i) P(they are all boys) = P(no girl)

$$\begin{aligned} P(x = 0) &= \binom{6}{0} (0.3)^0 (0.7)^6 \\ &= 0.1176 \text{ (from tables).} \end{aligned}$$

(ii)

$$\begin{aligned} P(X \geq 2) &= 1 - \{P(x < 2)\} \\ &= 1 - \{P(x = 0) + P(x = 1)\} \\ &= 1 - \{0.1176 + 0.3025\} \text{ (from tables)} \\ &= 1 - 0.4201 \\ &= 0.5799. \end{aligned}$$

(iii)

$$\begin{aligned} P(x = 3) &= \binom{6}{3} (0.3)^3 (0.7)^3 \\ &= 0.1852. \end{aligned}$$

(iv)

$$\text{P(all girls)} = P(x = 6) = 0.0007. \text{ (from tables) or } 0.000729 \text{ (from calculator).}$$

4.2 Mean and Variance of a Binomial Distribution

For a binomial distribution of n trials and probability of success p, mean = np and variance $=\delta^2 = npq$.
To determine the mean, it is known that

$$\begin{aligned} E(X) &= \sum xP(X=x) \\ &= \sum_{x=0}^{n} x \left(\begin{array}{c} n \\ x \end{array} \right) p^x q^{n-x} \end{aligned}$$

Stating the series term by term

$$\begin{aligned} E(X) &= \mu = 0 \times q^n + 1 \times npq^{n-1} + 2 \times \frac{(n-1)}{2 \times 1} p^2 q^{n-2} \\ &\quad +3 \times \frac{n(n-1)(n-2)}{3 \times 2 \times 1} p^3 q^{n-3} + \ldots + np^n \\ &= npq^{n-1} + n(n-1)p^2 q^{n-2} + \frac{n(n-1)(n-2)}{2 \times 1} p^3 q^{n-3} + \ldots + n \times p^n \\ \mu &= np \left[q^{n-1} + (n-1)pq^{n-2} + \frac{(n-1)(n-2)}{2 \times 1} p^2 q^{n-3} + \ldots + p^{n-1} \right] \end{aligned}$$

but what is in the square bracket is the binomial expansion of $(p+q)^{n-1}$ and $p+q=1$. so that what is in the square bracket sums to 1.
Therefore $\mu = np$.
To get the variance, recall that $V(x) = E(x^2) - [E(x)]^2$ which can be expressed as

$$\delta^2 = E(x^2) - (E(x))^2 = E[x(x-1)]$$

So for the Binomial distribution

$$E[x(x-1)] = n \sum_{x=0}^{n} x(x-1) \left(\begin{array}{c} n \\ x \end{array} \right) p^x q^{n-x}$$

Stating the series term by term, first two terms are each zero. beginning with the third term

$$E[x(x.1)] = 2 \times 1 \frac{n(n-1)}{2 \times 1} p^2 q^{n-2} + 3 \times 2 \times \frac{n(n-1)(n-2)}{3 \times 2 \times 1} p^3 q^{n-3}$$

$$+4 \times 3 \times \frac{n(n-1)(n-2)(n-3)}{4 \times 3 \times 2 \times 1} p^4 q^{n-4} + \ldots + n(n-1)p^n$$

$$= n(n-1)p^2 q^{n-2} + n(n-1)(n-2)p^3 q^{n-3} + \frac{n(n-1)(n-2)(n-3)}{2 \times 1} p^4 q^{n-4}$$

$$+ \ldots + n(n-1)p^n$$

$$\Rightarrow E[x(x-1)] = n(n-1)p^2 \left[q^{n-2} + (n-2)pq^{n-3} + \frac{(n-2)(n-3)}{2 \times 1} p^2 q^{n-4} + \ldots + p^{n-2} \right]$$

(after taking out the factor $n(n-1)p^2$)
The terms in the square bracket sum to 1 since they are the expansion of
$(p+q)^{(n-2)}$ and $p+q=1$.
So $E[x(x-1)] = E(x^2) - E(x) = n(n-1)p^2$
Isolating $E(x^2)$,

$$\begin{aligned}
E(x^2) &= n(n-1)p^2 + E(x) \\
\text{but } E(x) &= np \\
\Rightarrow E(x^2) &= n(n-1)p^2 + np \\
\text{and variance} &= E(x^2) - [E(x)]^2 \\
\Rightarrow Var(X) &= n(n-1)p^2 + np - [E(x)]^2 \\
&= n(n-1)p^2 + np - (np)^2 \\
&= n^2p^2 - np^2 + np - n^2p^2 \\
&= np - np^2 \\
&= np(1-p) \text{ but } 1-p = q \\
Var(x) &= npq
\end{aligned}$$

Example 4.2.1

If the binomial distribution $B(n,p)$ has mean 12 and standard deviation 2, find n and p.
Solution:

$$\begin{aligned}
E(x) &= np = 12 \\
\sqrt{Var(x)} &= \sqrt{npq} = 2 \\
\Rightarrow npq &= 4 \\
\Rightarrow q &= \frac{1}{3} \\
\Rightarrow p &= \frac{2}{3}
\end{aligned}$$

$$\begin{aligned} np &= 12 \\ \Rightarrow n \cdot \frac{2}{3} &= 12 \\ \Rightarrow n &= 18 \\ \text{Therefore } n &= 18 \text{ and } p = \frac{2}{3}. \end{aligned}$$

Example 4.2.2

For a random variable X having a binomial distribution $B(10, \frac{1}{4})$. Determine

(i) the mean

(ii) the variance

(iii) $P(x = 4)$

(iv) $P(X > 0)$

(v) $P(2 \leq x \leq 9)$.

Solution:

(i) $n = 10, p = \frac{1}{4}, q = \frac{3}{4}$
mean $= \mu = np = 10 \times \frac{1}{4} = 2.5$

(ii) variance $= npq = 10 \times \frac{1}{4} \times \frac{3}{4} = \frac{30}{16} = 1.875$

(iii)

$$\begin{aligned} P(X = 4) &= \left(\begin{array}{c} 10 \\ 4 \end{array} \right) (\frac{1}{4})^4 (\frac{3}{4})^6 \\ &= = 0.1460 \text{ (from tables)} \end{aligned}$$

Example 4.2.3

Peter found that 25% of the people who accept invitations to a party do not come. For a party that he is going to hold next week he has 16 chairs for guests but has invited 20 people. What is the probability that there is no chair for every one who will come to the party.

Solution:

Let X be the event, "invited guest has come for the party". There will not be a chair for every guest if the guests exceed 16. So we require

$$\begin{aligned} P(X > 16) &= P(X \geq 17) \\ P(X \geq 17) &= P(x = 17) + P(x = 18) + P(x = 19) + P(x = 20) \\ &= 0.1339 + 0.0669 + 0.0211 + 0.0032 \\ &= 0.2251 \end{aligned}$$

4.3 Binomial Recurrence formula

The Binomial recurrence formula is

$$P(X = x + 1) = \frac{n - x}{x + 1} \cdot \frac{p}{1 - p} P(X = x)$$

This enables successive probabilities to be more easily calculated once the initial probability is known.

Example 4.3.1

For $B(5, \frac{2}{5})$, use the recurrence formula to solve for all the individual probabilities.
Solution:

$$P(x = 0) = \begin{pmatrix} 5 \\ 0 \end{pmatrix} (\frac{2}{5})^0 (\frac{3}{5})^5 = 0.07776$$

By the binomial recurrence formula

$$\begin{aligned}
P(X = x + 1) &= \frac{n - x}{x + 1} \cdot \frac{p}{1 - p} P(X = x) \\
P(X = 1) &= \frac{5}{1} \cdot \frac{2/5}{3/5} \times 0.0776 = 0.2592 \\
P(X = 2) &= \frac{4}{2} \cdot \frac{2/5}{3/5} \times 0.2592 = 0.3456 \\
P(X = 3) &= \frac{3}{3} \cdot \frac{2/5}{3/5} \times 0.3456 = 0.2304 \\
P(X = 4) &= \frac{1}{2} \cdot \frac{2/5}{3/5} \times 0.2304 = 0.0768 \\
P(X = 5) &= \frac{1}{5} \cdot \frac{2/5}{3/5} \times 0.0768 = 0.01024
\end{aligned}$$

Clearly, these probabilities add up to 1 since a binomial distribution has exhaustive events.

Exercise 4

1. One in five people in Kampala City is employed. What is the probability that in a random sample of 8 people

(i) none is employed

(ii) all are employed

(iii) atleast two are unemployed

(iv) atmost two are employed.

2. In a hospital ward of thirty patients, 15 are of blood group A^+. What is the probability that in a sample of 9 patients picked at random

   (i) three are of blood group A^+

   (ii) none is of blood group A^+

   (iii) all are of blood group A^+

   (iv) atleast one is of blood group A^+.

3. It is known that 80% of seeds of maize if planted in good soil will germinate

   (i) If John planted 10 seeds, what is the expected number of seeds that will germinate?

   (ii) How many seeds should he plant so that 10 of them germinate?

4. It is established that one in four men is left handed. What is the probability that in a sample of 15 men,

   (i) 5 are left handed

   (ii) none is left handed

   (iii) 13 are left handed

5. The probability that a marksman will hit a target is $\frac{1}{3}$.
   He fires 4 shots. Calculate the probability that he will hit the target

   (i) twice

   (ii) four times,

   (iii) exactly once.

   (iv) how many shots at the target should he be allowed so that his probability of hitting the target atleast once improves to $\frac{9}{10}$

6. A rifle mans probability of hitting a target is $\frac{2}{5}$

   (i) Find the probability that he will hit the target atleast once in 6 trials

   (ii) Find the probability that all shots miss the target

   (iii) Find the minimum number of shots that he must be allowed in order to have a probability of atleast 0.8 for atleast one shot to hit the target.

Chapter 5

THE POISSON DISTRIBUTION

5.1 Introduction

This distribution was first used by a French mathematician, Simeon Poisson and is used to determine the probability that a particular event will take place a certain number of times over a specified period of time or interval. For instance, the number of patients arriving per hour at a hospital is a random variable with a Poisson distribution. Other examples of random variables that exhibit a Poisson distribution are:

1. The number of days in a given month on which a worker reports late for work.
2. The number of defects detected each day by a quality control inspector of a spare parts plant.
3. The number of breakdowns per year that a bus on a given route experiences
4. The number of accidents that occur per month at a manufacturing plant.
5. Telephone calls arriving at a switch board in given time intervals
6. Insurance claims per month/year, e.t.c.

5.2 The Poisson formula

A dicrete random variable X having a probability density function of the form

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} \text{ where } x = 0, 1, 2, \ldots$$

is said to have a Poisson distribution, where X represents the discrete Poisson random variable x represents the number of rare events in a unit of time, space or volume
λ is the mean value of x

e is the base of the natural logarithm and is equal to 2.71828
For the Poisson formula to be applicable, two or more events should not occur sumultaneously, the events should be independent and the mean number of events in a given interval is constant.

Example 5.2.1

The number of lorries per hour crossing a bridge is Poisson distributed with mean 6. Find

(a) the probability that 6 lorries cross in one hour

(b) the probability that 12 lorries pass in two hours.
Solution:

(a) The mean number of lorries in an hour is 6 So that

$$\begin{aligned} P(X=6) &= \frac{6^6 e^{-6}}{6!} \\ &= 0.160623141 \\ &\simeq 0.1606 \end{aligned}$$

(b) The mean members of lorries in two hours is 12 So that

$$\begin{aligned} P(X=12) &= \frac{12^{12} e^{-12}}{12!} \\ &= 0.114367915 \\ &\simeq 0.1144. \end{aligned}$$

Example 5.2.2

Telephone calls arrive at a switch board at the rate of 10 per 5−minute period. Find the probabilities of $0, 1$ or 2 calls arriving in any 5 minute period.
Solution:

$$\begin{aligned} P(X=x) &= \frac{e^{-\lambda} \cdot \lambda^x}{x!} \text{ and } \lambda = 10 \\ P(X=0) &= \frac{e^{-10} \cdot 10^0}{0!} \\ &= 0.0000453999276 \end{aligned}$$

$$
\begin{aligned}
&\simeq 0.0000454 \\
P(X=1) &= \frac{e^{-10}\cdot 10^1}{1!} \\
&= 0.004539992976 \\
&\simeq 0.000454 \\
P(X=2) &= \frac{e^{-10}\cdot 10^2}{2!} \\
&= 0.002269996488 \\
&\simeq 0.00227
\end{aligned}
$$

NB: Tables may be used to find probabilities for given values of λ, the mean.

5.3 Mean and Variance of a Poisson distribution

Example 5.3.1

Prove that for the Poisson distribution $E(x) = V(x) = \lambda$.
Solution:
The Poisson formula is

$$
\begin{aligned}
P(X=x) &= \frac{e^{-\lambda}\lambda^x}{x!} \\
\text{Mean} &= E(x) = \sum_{x=0}^{\infty} xP(X=x) \\
&= \sum_{x=0}^{\infty} \frac{xe^{-\lambda}\lambda^x}{x!} \\
&= \frac{0\times e^{-\lambda}\lambda^0}{0!} + \frac{1\times e^{-\lambda}\lambda^2}{1!} + \frac{2e^{-\lambda}\lambda^2}{2!} + \frac{3e^{-\lambda}\lambda^2}{3!} + \ldots + \ldots + \frac{(x+1)e^{-\lambda}\lambda^{-\lambda}\lambda^{(x+1)}}{(x+1)!} \\
&= 0 + \lambda e^{-\lambda} + \frac{\lambda^2 e^{-\lambda}}{1!} + \frac{\lambda^3 e^{-\lambda}}{2!} + \ldots + \frac{\lambda^{(x+1)}e^{-\lambda}}{x!} + \ldots \\
&= \lambda\left[e^{-\lambda} + \frac{\lambda e^{-\lambda}}{1!} + \frac{\lambda^2 e^{-\lambda}}{2!} + \ldots + \frac{\lambda^x e^{-\lambda}}{x!}\right]
\end{aligned}
$$

but the bracketed terms are those of the probabilities of a Poisson distribution which sum to unity.

$$\text{Therefore } E(x) = \lambda[1] = \lambda.$$

Variance is given by $E\left\{(x-\mu)^2\right\}$ and

$$E\{(x-\mu)^2 = \sum_{x=0}^{\infty}(x-\lambda)^2 \cdot \frac{e^{-\lambda}\lambda^x}{x!} \text{ where } \mu = \lambda$$

$$= \sum_{x=0}^{\infty}\left[x(x-1) + x(1-2\lambda) + \lambda^2\right]\frac{e^{-\lambda}\lambda^x}{x!}$$

The expression $x(x-1)$ cancels with the first two terms of $x!$, so that

$$\text{Variance} = E\{(x-\mu)^2$$

$$= \sum_{x=0}^{\infty} x(x-1)\frac{e^{-\lambda}\lambda^x}{x!} + (1-2\lambda)\sum_{x=0}^{\infty}\frac{xe^{-\lambda}\lambda^x}{x!} + \lambda^2\sum_{x=0}^{\infty}\frac{e^{-\lambda}\lambda^x}{x!}$$

Of the three summations, the second summation is $E(x)$ which has been seen to be λ and the third summation is $\sum_{x=0}^{\infty} = 1$. In the first sum, the first two terms are both zero, so that

$$\sum_{x=0}^{\infty} x(x-1)\frac{e^{-\lambda}\lambda^x}{x!} = \frac{2\times 1\times e^{-\lambda}\cdot\lambda^2}{2!} + \frac{3\times 2\times e^{-\lambda}\cdot\lambda^3}{3!} + \frac{4\times 3e^{-\lambda}\cdot\lambda^4}{4!}$$

$$+ \ldots + \frac{(x+2)(x+1)e^{-\lambda}\lambda^{x+2}}{(x+2)!}$$

$$= \lambda^2 e^{-\lambda} + \lambda^3 e^{-\lambda} + \frac{\lambda^4 e^{-\lambda}}{2!} + \ldots + \frac{\lambda^{x+2}e^{-\lambda}}{x!} + \ldots$$

$$= \lambda^2\left[e^{-\lambda} + \lambda e^{-\lambda} + \frac{\lambda^2 e^{-\lambda}}{2!} + \ldots + \frac{\lambda^x e^{-\lambda}}{x!}\right]$$

$$= \lambda^2$$

Because the terms in the bracket sum to 1. Substituting for the three summations

$$\text{Variance} = E\left\{(x-\mu)^2\right\}$$

$$
\begin{aligned}
&= \lambda^2 + (1-2\lambda)\lambda + \lambda^2 \\
&= \lambda^2 + \lambda - 2\lambda^2 + \lambda^2 \\
&= \lambda.
\end{aligned}
$$

So mean equal variance.

5.4 Additive Property of the Poisson distribution

If X is $P_0(x)$ and Y is $P_0(y)$, then $X + Y = P_0(x + y)$.

Example 5.4.1

Telephone calls reach a switch board independently and randomly. External calls reach at a mean rate of 1 in any 4 minute period while internal ones reach at a mean rate of 3 in any 4 minute period. Calculate the probability that there will be more than 2 calls in any period of 2 minutes

Solution:

Let the random variable E represent the number of external calls per period of two minutes. So E is $P_0(2 \times \frac{1}{4}) = P_0(0.5)$

Let the random variable I represent the number of internal calls per period of two minutes. So I is $P_0(1.5)$

Using the additive property of the Poisson distribution

$$
\begin{aligned}
E + I \text{ is } P_0(0.5 + 1.5) &= P_0(2) \\
P(E + I > 2) &= 1 - P(X \leq 2) \\
P(E + I > 2) &= 1 - \{P(E + I = 0) + P(E + I = 1) + P(E + I = 2)\} \\
\text{But } P(E + I = 0) &= \frac{e^{-2} \cdot 2^0}{0!} \\
&= e^{-2} = 0.135335283 \\
P(E + I = 1) &= \frac{e^{-2} \cdot 2!}{1!} \\
&= 2e^{-2} = 0.270670566 \\
P(E + I = 2) &= \frac{e^{-2} \cdot 2^2}{2!} \\
&= 2e^{-2} = 0.270670566 \\
\text{So that } P(E + I > 2) &= 1 - \{0.135335283 + 0.270670566 + 0.270670566\} \\
&= 1 - 0.676676416 \\
&= 0.323323583 \\
&\simeq 0.3233
\end{aligned}
$$

Therefore P(more than 2 calls in any 2 minute period)= 0.3233

5.5 The Poisson Approximation to the Binomial Distribution

The Poisson distribution is many times used to approximate the binomial distribution in a way of avoiding the tedius calculations. This is done on condition that the number of trials n is large i.e $n > 20$ and the probability of succcess is small i.e $P \leq 0.05$. Since in a binomial distribution $\mu = \lambda = np$,

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

Example 5.5.1

The probability that a person chosen at random from a given city population is a thief is 0.05. If a sample of 25 people is taken, find the probability that there is

(i) no thief

(ii) one thief

(iii) atmost one thief

(iv) atleast one thief

Solution:

Let X be the event that "a person chosen at random is a thief"
Using binomial distribution

(i) $n = 25, P = 0.05$

$$\begin{aligned} P(X = 0) &= \binom{25}{0} (0.05)^0 (0.95)^{25} \\ &= 0.2774 \end{aligned}$$

(ii)

$$\begin{aligned} P(X = 1) &= \binom{25}{1} (0.05)^1 (0.95)^{24} \\ &= 0.365 \end{aligned}$$

(iii)

$$\begin{aligned} P(X \leq 1) &= P(X = 0) + P(X = 1) \\ &= 0.2774 + 0.365 \\ &= 0.6424 \end{aligned}$$

(iv)

$$\begin{aligned} P(X \geq 1) &= 1 - P(X = 0) \\ &= 1 - 0.2774 \\ &= 0.7226 \end{aligned}$$

And using the Poisson distribution $\lambda = np = 25 \times 0.05 = 1.25$

$$P(X = x) = \frac{e^{-np}(np)^x}{x!} = \frac{e^{-1.25}(1.25)^x}{x!}$$

(i)

$$\begin{aligned} P(X = 0) &= \frac{e^{-1.25} \times 1}{0!} \\ &= 0.2865 \end{aligned}$$

(ii)

$$\begin{aligned} P(X = 1) &= \frac{e^{-1.25} \times 1.25}{1!} \\ &= 0.358 \end{aligned}$$

(iii)

$$\begin{aligned} P(X \leq 1) &= P(x = 0) + P(x = 1) \\ &= 0.2865 + 0.358 \\ &= 0.6445 \end{aligned}$$

(iv)

$$\begin{aligned} P(X \leq 1) &= 1 - P(X = 0) \\ &= 1 - 0.2865 \\ &= 0.7135 \end{aligned}$$

Observe that the answers got due to using the Poisson distribution do not differ so much from those got using the binomial distribution. If the value of P was 0.02 or 0.01, the answers would be even closer to each other. The Poisson approximation to the binomial distribution is so handy and better as n gets larger and P tends to zero.

Exercise 5

1. The number of demands for special hire cars from a rental firm is Poisson distributed with a mean of 6 demands in 1 hour. Find the probabilities of

   (i) no call in 1 hour

   (ii) one call in one hour

   (iii) one call in 2 hours

   (iv) two calls in two hours

2. The mean number of patients arriving at a clinic in two hour intervals is 5. Calculate the probabilities of $0, 2, 4, 7$ arrivals per two hour interval.

3. The number of emergency calls at a police station each day is found to have a Poisson distribution with mean 2.5

   (i) Calculate the probability that on a particular day there will be no emergency calls

   (ii) On a given day, there are six cars available for response to emergency calls. Calculate the probability that the six cars will surfice

4. The probability that a brand of pen is faulty is 0.02. The pens are packed in boxes of 200. If a box is randomly chosen find the probability that

   (i) there are no faulty pens

   (ii) there is one faulty pen

   (iii) there are atleast three faulty pens.

5. It is known that 2% of patients who contract malaria in Uganda die if they dont reach at a hospital within one day of contracting the disease. On a particular day, 150 people contracted the disease. If these people never reached a hospital, what is the probability that

   (i) none died

   (ii) exactly two died

   (iii) atmost two died

   (iv) atleast three died.

6. The circumised men in a given district are 3%. What is the probability that in a sample of $10,000$ men

   (i) none is circumised

   (ii) 30 are circumcised

   (iii) exactly 40 men are circumcised

   (iv) exactly 15 men are circumcised.

Chapter 6

CONTINUOUS PROBABILITY DENSITY FUNCTIONS

6.1 Introduction

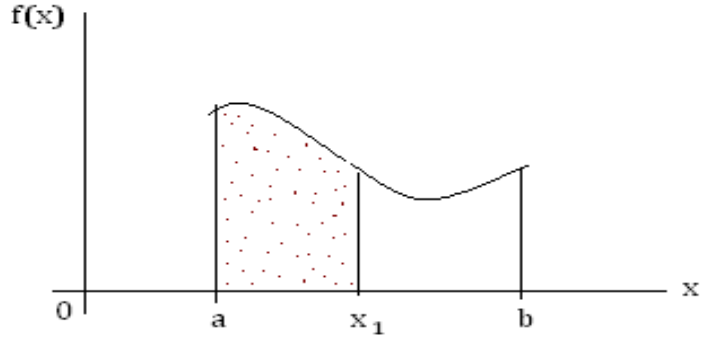
A continuous random variable theoretically represents quantities like time, weight, temperature, height, distance, mass, etc. A probability density function $f(x)$ of a random variable X is said to be continuous since it has a continuous domain. The probability density function of a continuous random variable X is a function which allocates probabilities to all values in certain intervals that X can take.

Properties of f(x):

(i) $f(x) > 0$ for all values of X.

(ii) If $a \leq x \leq b$, then $\int_a^b f(x) = 1$.

(iii) The probability of $a \leq x \leq x_1$ where $x_1 < b$ is given by $\int_a^{x_1} f(x)dx$



Example 6.1.1

A continuous random variable has a probability density function given by

$$f(x) = \begin{cases} k(1-\cos x), & 0 \le x \le \frac{\pi}{4} \\ k\sin x, & \frac{\pi}{4} \le x \le \frac{\pi}{2} \\ 0, & \text{otherwise} \end{cases}$$

Find

(a) the value of the constant k

(b) $P(\frac{\pi}{4} \le x \le \frac{\pi}{3})$

(c) $P\left(0 \le x \le \frac{\pi}{6}\right)$

Solution:

1.

$$\begin{aligned} \int_0^{\frac{\pi}{4}} k(1-\cos x)dx + \int_{\frac{\pi}{4}}^{\frac{\pi}{2}} k\sin x dx &= 1 \\ l\left[x-\sin x\right]_0^{\frac{\pi}{4}} + k\left[-\cos x\right]_{\frac{\pi}{4}}^{\frac{\pi}{2}} &= 1 \\ k\left\{\left(\frac{\pi}{4}-\frac{1}{\sqrt{2}}\right)-(0-0)\right\} - k\left\{\left(0-\frac{1}{\sqrt{2}}\right)\right\} &= 1 \\ k\left\{\frac{\pi}{4}-\frac{1}{\sqrt{2}}+\frac{1}{\sqrt{2}}\right\} &= 1 \\ \frac{\pi}{4}\,k &= 1 \\ k &= \frac{4}{\pi} \\ &= 1.273239545 \\ &\simeq 1.27324 \end{aligned}$$

(b)

$$\begin{aligned} P\left(\frac{\pi}{4} \leq x \leq \frac{\pi}{3}\right) &= \int_{\frac{\pi}{4}}^{\frac{\pi}{3}} k \sin x dx \\ &= \frac{4}{\pi}\left[-\frac{\cos x}{1}\right]_{\frac{\pi}{4}}^{\frac{\pi}{3}} \\ &= -\frac{4}{\pi}\left\{\frac{1}{2} - \frac{1}{\sqrt{2}}\right\} \\ &= -\frac{4}{\pi}\,\frac{(\sqrt{2}-2)}{2\sqrt{2}} \\ &= \frac{(\sqrt{2}-1)}{\pi} \\ &= 0.263696543 \\ &\cong 0.2637 \end{aligned}$$

(c)

$$\begin{aligned} P\left(0 \leq x \leq \frac{\pi}{6}\right) &= \int_{0}^{\frac{\pi}{6}} k(1-\cos x) dx \\ &= \left[\frac{4}{\pi}(x - \sin x)\right]_{0}^{\frac{\pi}{6}} \\ &= \frac{4}{\pi}\left[\left(\frac{\pi}{6} - \frac{1}{2}\right) - (0-0)\right] \\ &= \frac{4}{6} - \frac{4}{2\pi} \\ &= 0.030046894 \cong 0.03 \end{aligned}$$

Example 6.1.2

A continuous random variable has a probability density function where

$$f(x) = \begin{cases} kx, & 0 \leq x \leq 5 \\ k(10-x), & 5 \leq x \leq 10 \\ 0, & \text{elsewhere} \end{cases}$$

Determine

(a) the value of the constant k

(b) $P(0 \leq x \leq 3)$

(c) $P(2 \leq x \leq 4)$

(d) $P(4 \leq x \leq 8)$

Solution:

(a)

$$\begin{aligned} \int_0^5 kxdx + k\int_5^{10}(10-x)dx &= 1 \\ k\left[\frac{x^2}{2}\right]_0^5 + k\left[10x - \frac{x^2}{2}\right]_5^{10} &= 1 \\ \frac{25}{2}k + k\left[(100-50) - \left(50 - \frac{25}{2}\right)\right] &= 1 \\ \frac{25}{2}k + \frac{25}{2}k &= 1 \\ 25k &= 1 \\ k &= \frac{1}{25} \end{aligned}$$

(b)

$$P(0 \leq x \leq 3) = \int_0^3 kxdx$$

$$\frac{1}{25}\left[\frac{x^2}{2}\right] = \frac{1}{25}\cdot\frac{9}{2}$$

$$= \frac{9}{50} = 0.18$$

(c)

$$P(2 \leq x \leq 4) = \int_2^4 kx dx$$

$$\frac{1}{25}\left[\frac{x^2}{2}\right]_2^4$$

$$= \frac{1}{25}\cdot(8-2)$$

$$= \frac{6}{25} = 0.24$$

(d)

$$P(4 \leq x \leq 8) = \int_4^5 kx dx + \int_5^8 (10-x)dx$$

$$= \frac{1}{25}\left[\frac{x^2}{2}\right]_4^5 + \frac{1}{25}\left[10x - \frac{x^2}{2}\right]_5^8$$

$$= \frac{1}{25}\left\{\left(\frac{25}{2} - 8\right) + (80-32) - \left(50 - \frac{25}{2}\right)\right\}$$

$$= \frac{1}{25}\times 15$$

$$= \frac{15}{25} = \frac{3}{5} = 0.6$$

6.2 Expectation and variance

For a continuous random variable X with a probability density function $f(x)$, the expectation is given by

$$E(x) = \int_{all\ x} xf(x)dx$$

This is also called the mean value of X, μ. The properties of expectation are

(i) $E(a) = a$, a constant.

(ii) $E(ax) = aE(x)$

(iii) $E\left[G(x)\right] = \int G(x)f(x)dx$

Example 6.2.1

A probability density function of a random variable X is given by

$$f(x) = \begin{cases} \frac{4}{5}x, & 0 \leq x \leq 1 \\ \frac{2}{5}(3-x), & 1 \leq x \leq 2 \\ 0, & \text{otherwise} \end{cases}$$

Find the mean of X.
Solution:

$$\begin{aligned} E(x) &= \int_{all\ x} xf(x)dx \\ &= \int_0^1 \frac{4}{5}x^2dx + \frac{2}{5}\int_1^2 (3x - x^2)dx \\ &= \left[\frac{4x^3}{15}\right]_0^1 + \frac{2}{5}\left[\frac{3}{2}x^2 - \frac{x^3}{3}\right]_1^2 \\ &= \frac{4}{15} + \frac{2}{5}\left[\left(6 - \frac{8}{3}\right) - \left(\frac{3}{2} - \frac{1}{3}\right)\right] \\ &= 1\frac{2}{15} \end{aligned}$$

The variance of any probability distribution is given by $Var(x) = E(x^2) - [E(x)]^2$ where $E(x)$ is the expectation and $E(x^2)$ is given by

$$\int_{all\ x} x^2 f(x)dx$$

Example 6.2.2

For the probability density function below, find the variance of X.

$$f(x) = \begin{cases} \frac{k}{2}x, & 0 \leq x \leq 2 \\ k, & 2 \leq x \leq 4 \\ 0, & \text{otherwise} \end{cases}$$

Solution:
We first have to find the value of the constant k.

$$\begin{aligned}
\int f(x)dx &= 1 \\
\Rightarrow \int_0^2 \frac{k}{2}xdx + \int_2^4 kdx &= 1 \\
k\left[\frac{x^2}{4}\right]_0^2 + k\left[\frac{x}{1}\right]_2^4 &= 1 \\
k(1) + k(4-2) &= 1 \\
3k &= 1 \\
k &= \frac{1}{3} \\
Var(x) &= E(x^2) - [E(x)]^2 \\
\text{But } E(x) &= \int_{all\ x} xf(x)dx \\
&= \frac{1}{3}\int_0^2 \frac{x^2}{2}dx + \frac{1}{3}\int_2^4 xdx \\
&= \frac{1}{3}\left[\frac{x^3}{6}\right]_0^2 + \frac{1}{3}\left[\frac{x^2}{2}\right]_2^4 \\
&= \frac{1}{3}\left\{\left(\frac{8}{6} - 0\right) - (8-2)\right\} \\
&= \frac{1}{3} \times \frac{22}{3} = \frac{22}{9}
\end{aligned}$$

$$
\begin{aligned}
\text{and } E(x^2) &= \int_{all\ x} x^2 f(x)dx \\
&= \frac{k}{2}\int_0^2 x^3 dx + k\int_2^4 x^2 dx \\
&= \frac{k}{2}\left[\frac{x^4}{4}\right]_0^2 + k\left[\frac{x^3}{3}\right]_2^4 \\
&= \frac{k}{2}(4-0) + k\left(\frac{64}{3} - \frac{8}{3}\right) \\
&= \frac{1}{6}(4) + \frac{1}{3}\left(\frac{56}{3}\right) \\
\frac{4}{6} + \frac{56}{9} &= \frac{62}{9} \\
\therefore Var(x) &= \frac{62}{9} - \left(\frac{22}{9}\right)^2 \\
&= \frac{62}{9} - \frac{484}{81} \\
&= \frac{74}{81} = 0.913580246 \\
&\cong 0.9136
\end{aligned}
$$

6.2.1 The Median

The median divides the area under the curve of a probability density function into two equal parts. Therefore for a random variable X which varies over the interval $[a, b]$, that is, $a \leq x \leq b$, the median is given by

$$\int_a^m f(x)dx = \frac{1}{2}$$

Example
A continuous random variable X has a probability density function given by

$$f(x) = \begin{cases} k(x - x^3), & 0 \leq x \leq 2 \\ 0, & \text{elsewhere} \end{cases}$$

Find its median.

Solution:

We need to first find the value of the constant k. Therefore,

$$\begin{aligned}
\int_0^2 k(x-x^3)dx &= 1 \\
k\left[\frac{x^2}{} - \frac{x^4}{4}\right]_0^2 &= 1 \\
k\left[(2-4)-(0-0)\right] &= 1 \\
-2k &= 1 \\
k &= -\frac{1}{2}
\end{aligned}$$

Therefore median is given by

$$\begin{aligned}
\int_0^m -\frac{1}{2}(x-x^3)dx &= \frac{1}{2} \\
&= -\frac{1}{2}\int_0^m (x-x^3) = \frac{1}{2} \\
-\frac{1}{2}\left[\frac{x^2}{2} - \frac{x^4}{4}\right]_0^m &= \frac{1}{2} \\
\left[\frac{x^2}{2} - \frac{x^4}{4}\right]_0^m &= -1 \\
\frac{m^2}{2} - \frac{m^4}{4} &= -1 \\
m^4 - 2m^2 - 4 &= 0 \\
\text{Let } m^2 &= p. \text{ Then} \\
p^2 - 2p - 4 &= 0 \\
p &= \frac{2 \pm \sqrt{4+16}}{2} \\
p &= 1 \pm \sqrt{5}
\end{aligned}$$

But $1-\sqrt{5}$ is negative. So we take

$$p = 1+\sqrt{5} \Rightarrow m^2 = 1+\sqrt{5}$$

$$\begin{aligned} m &= \sqrt{1}+\sqrt{5} \\ &= \sqrt{3.236067977} \\ &= 1.79890744 \cong 1.799 \end{aligned}$$

So median is 1.799.

6.3 Mode

This is the maximum value of the function $f(x)$. It is the value at which maximum probability is attained. For instance, using the last example, the mode is found thus:

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

But k was found to be $\frac{1}{2}$

$$\begin{aligned} \Rightarrow f'(x) &= \frac{d}{dx}\left[-\frac{1}{2}(x-x^3)\right] \\ -\frac{1}{2}\left[1-3x^2\right] &= \\ 3x^2-1 &= 0 \\ x^2 &= \frac{1}{3} \\ x &= \pm\sqrt{\frac{1}{3}} \\ f''(x) &= -\frac{1}{2}\cdot(-6x) = 3x \end{aligned}$$

When $x = \sqrt{\frac{1}{3}}$ the value of f'' will be positive hence minimum, but when $x = -\sqrt{\frac{1}{3}}$ the value of f'' is negative and hence gives a maximum value of f(x). Hence the mode is $-\sqrt{\frac{1}{3}}$. But this value is outside the interval for which the function exists. So X has no mode.

6.3.1 The Cumulative Distribution Function

This is defined by

$$A = \int_{-\infty}^{x_1} f(x)dx$$

where $f(x)$ is the probability density function. It is denoted by $F(x)$.

$$F(x_1) = \int_{-\infty}^{x_1} f(x)dx = P(x \leq x_1)$$

$F(x)$ is the cumulative distribution function. If m is the median, then $F(m) = 0.5$.

Example 6.3.1

A continuous random variable X has probability density function

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

Find the distribution function $F(x)$.

Solution:

$$\begin{aligned} F(x) &= P(X \leq x) \\ &= \int_{-\infty}^{x} f(t)dt \\ \text{For } x < 0,\ f(t) &= 0 \\ \Rightarrow\ F(x) &= 0 \text{ and } F(0) = 0 \\ \text{For } 0 \leq x \leq 2,\ f(t) &= \frac{1}{2}t \\ F(x) &= F(0) + \int_0^x \frac{1}{2}tdt \\ &= 0 + \left[\frac{t^2}{4}\right]_0^x \\ \Rightarrow F(x) &= 0 + \frac{x^2}{4} \\ \Rightarrow F(2) &= \frac{4}{4} = 1 \end{aligned}$$

For $x > 2$, $f(t) = 0$ The function does not exist.

$$F(x) = F(2) + 0 = 1.$$

Therefore the distribution function is

$$F(x) = \begin{cases} 0, & x < 0 \\ \frac{x^2}{4}, & 0 \leq x \leq 2 \\ 1, & x \geq 2 \end{cases}$$

Example 6.3.2

Given the following probability density function, determine $F(x)$.

$$f(x) = \begin{cases} k, & 0 \leq x \leq \frac{3}{2} \\ \frac{k}{2}(2-x), & \frac{3}{2} \leq x \leq 2 \\ 0, & \text{otherwise} \end{cases}$$

Solution:
We need to first find the value of k.

$$\begin{aligned} \int_0^{\frac{3}{2}} k dx + \frac{k}{2}\int_{\frac{3}{2}}^{2}(2-x)dx &= 1 \\ \left[k\frac{x}{1}\right]_0^{\frac{3}{2}} + \frac{k}{2}\left[2x - \frac{x^2}{2}\right]_{\frac{3}{2}}^{2} &= 1 \\ \frac{3}{2}k + \frac{k}{2}\left\{(4-2) - \left(3 - \frac{9}{8}\right)\right\} &= 1 \\ \frac{3}{2}k + \frac{1}{16}k &= 1 \\ k &= \frac{16}{25} \\ F(x) &= P(X \leq x) \end{aligned}$$

$$
\begin{aligned}
&= \int_{-\infty}^{x} f(x)dx \\
\text{For } x < 0,\ f(t) &= 0 \\
\Rightarrow F(0) &= 0 \\
\text{For } 0 \le x \le \frac{3}{2},\ f(t) &= \int_0^x k dt \\
&= k\left[kt\right]_0^x = kx = \frac{16}{25}x \\
F(\frac{3}{2}) &= 0 + \frac{16}{25} \times \frac{3}{2} \\
&= \frac{24}{25}
\end{aligned}
$$

For $\frac{3}{2} \le x \le 2$,

$$
\begin{aligned}
F(x) &= F(\frac{3}{2}) + \int_{\frac{3}{2}}^{x} \frac{k}{2}(2-t)dt \\
&= \frac{24}{24} + \frac{k}{2}\left[2t - \frac{t^2}{2}\right]_{\frac{3}{2}}^{x} \\
&= \frac{24}{25} + \frac{16}{50}\left\{\left(2x - \frac{x^2}{2}\right) - \left(3 - \frac{9}{8}\right)\right\} \\
&= \frac{24}{25} + \frac{16}{50}\left\{2x - \frac{x^2}{2} - \frac{15}{8}\right\} \\
F(2) &= \frac{24}{25} + \frac{16}{50} \times \frac{1}{8} = 1 \\
\text{For } x > 2,\ f(t) &= 0 \\
&= 0 = F(2) + 0 = 1.
\end{aligned}
$$

Therefore the cumulative distribution function is given by

$$f(x) = \begin{cases} 0, & 0 \leq x \\ \frac{16}{25}x, & 0 \leq x \leq \frac{3}{2} \\ \frac{24}{25} - \frac{1}{25}(4x^2 - 16x + 15), & \frac{3}{2} \leq x \leq 2 \\ 1, & x \geq 2 \end{cases}$$

Example 6.3.3

For a random variable X with a probability density function

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

find (i) the constant a (ii) the distribution function of X.
Solution:

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

(i)

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

(ii)

$$f(x) = \begin{cases} 4x(1-x^2), & 0 \le x < 1 \\ 0, & \text{otherwise} \end{cases}$$

For $x < 0, f(t) = 0 \Rightarrow F(0) = 0$
For $0 \le x \le 1$,

$$\begin{aligned} F(1) &= F(0) + \int_0^x 4t(1-t^2)dt \\ &= 0 + 4\left[\frac{t^2}{2} - \frac{t^4}{4}\right]_0^x \\ &= 0 + (2x^2 - x^4) - 0 \\ &= 2x^2 - x^4 \\ F(1) &= 0 + 2 - 1 = 1 \end{aligned}$$

$$F(x) = \begin{cases} 0, & 0 \le 0 \\ x^2(2-x^2), & 0 \le x \le 1 \\ 1, & x \ge 1 \end{cases}$$

For $x > 1,\ f(t) = 0 \Rightarrow F(1) + 0 = 1$

Example 6.3.4

A random variable X has a probability density function given by

$$f(x) = \begin{cases} ax(4-x^2),\ 0 \le x < 2 \\ 0, & \text{otherwise} \end{cases}$$

Find

(i) the value of the constant a

(ii) the mean of X

(iii) the variance of X

(iv) the mode of X

(v) the distribution function of X, F(x).

Solution:

(i)

$$\begin{aligned}
\int_0^2 ax(4-x^2)dx &= 1 \\
a\int_0^2 (4x-x^3)dx &= a\left[2x^2-\frac{x^4}{4}\right]_0^2 = 1 \\
a(8-4) &= 1 \\
a &= \frac{1}{4}
\end{aligned}$$

(ii) Mean of X is

$$\begin{aligned}
E(x) &= \int_0^2 xf(x)dx \\
\frac{1}{4}\int_0^2 (4x^2-x^4)dx &= \frac{1}{4}\left[\frac{4}{3}x^3-\frac{x^5}{5}\right]_0^2 \\
&= \frac{1}{4}\left(\frac{32}{3}-\frac{32}{5}\right)-0 \\
&= \frac{1}{4}\times\frac{64}{15} \\
&= 1\frac{1}{15}
\end{aligned}$$

(iii) Var(x)= $E(x^2)-[E(x)]^2$
But

$$\begin{aligned}
E(x^2) &= \int x^2 f(x)dx \\
\Rightarrow E(x^2) &= \frac{1}{4}\int_0^2 (4x^3-x^5)dx
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{4}\left[x^4 - \frac{x^6}{6}\right]_0^2 \\
&= \frac{1}{4}\left(16 - \frac{64}{6}\right) - 0 \\
&= \frac{1}{4} \times \frac{16}{3} \\
&= \frac{4}{3} \\
\Rightarrow \ var(x) &= \frac{4}{3} - \left(\frac{16}{15}\right)^2 \\
&= \frac{44}{225}
\end{aligned}
$$

(iv)

$$
\begin{aligned}
f(x) &= \frac{1}{4}x(4 - x^2) \\
&= x - \frac{1}{4}x^3 \\
f'(x) &= 1 - \frac{3}{4}x^2 \\
f'(x) &= 0 \\
\Rightarrow 1 - \frac{3}{4}x^2 &= 0 \\
\Rightarrow x &= \pm\frac{2}{\sqrt{3}} \\
f''(x) &= -\frac{3}{2}x
\end{aligned}
$$

$f''(x)$ is negative when $x = \frac{2}{\sqrt{3}} = 1.154700538 \simeq 1.155$, so this value gives $f(x)$ the maximum value.
Therefore the mode of X is $\frac{2}{\sqrt{3}} = 1.154700538 \simeq 1.155$.

(v)

$$f(x) = \begin{cases} \frac{1}{4}x(4-x^2), & 0 \leq x < 2 \\ 0, & \text{otherwise} \end{cases}$$

For $x < 0, f(t) = 0 \Rightarrow F(0) = 0$.
For $0 \leq x \leq 2$

$$\begin{aligned} F(x) &= F(0) + \int_0^x \frac{1}{4}t(4-t^2)dt \\ &= 0 + \int_0^x \left(t - \frac{1}{4}t^3\right) dt \\ &= \left[\frac{t^2}{2} - \frac{t^4}{16}\right]_0^x \\ &= \frac{x^2}{2} - \frac{x^4}{16} \\ &= \frac{x^2(8-x^2)}{16} \\ F(2) &= \frac{2^2(8-2^2)}{16} = 1 \end{aligned}$$

For $x > 2, f(t)$ does not exist, i.e., $f(0) = 0 \Rightarrow F(x \geq 2) = 1$.
Therefore $F(x)$ is given by

$$F(x) = \begin{cases} 0, & x \leq 0 \\ \frac{x^2(8-x^2)}{16}, & 0 \leq x \leq 2 \\ 1, & x \geq 2 \end{cases}$$

Example 6.3.5

A random variable X takes on the values of the interval $0 \leq x \leq 2$ and has a probability density given by

$$f(x) = \begin{cases} 2c; & 0 \leq x \leq 1.5 \\ c(2-x); & 1.5 \leq x \leq 2 \\ 0, & \text{otherwise} \end{cases}$$

(i) Find the value of c

(ii) $P(1 \leq x \leq 1.8)$

(iii) the mean of X

(iv) the variance of X

(v) the cummulative distribution function, $F(x)$.

Solution:

(i)

$$\begin{aligned} \int_0^{1.5} 2cdx + \int_{1.5}^{2} c(2-x)dx &= 1 \\ \left[\frac{2cx}{1}\right]_0^{1.5} + c\left[(2x - \frac{x^2}{2})\right]_{1.5}^{2} &= 1 \\ 3c + c\left[(4-2) - (3 - \frac{9}{8})\right] &= 1 \\ 3c + c \cdot \frac{1}{8} &= 1 \\ \frac{25}{8}c &= 1 \\ c &= \frac{8}{25} \end{aligned}$$

(ii)

$$f(x) = \begin{cases} \frac{16}{25}; 0 \leq x \leq 1.5 \\ \frac{8}{25}(2-x); & 1.5 \leq x \leq 2 \\ 0, & \text{otherwise} \end{cases}$$

$$
\begin{aligned}
P(1 \leq x \leq 1.8) &= \int_1^{1.5} \frac{16}{25} dx + \int_{1.5}^{1.8} \frac{8}{25}(2-x)dx \\
&= \left[\frac{16}{25}x\right]_1^{1.5} + \frac{8}{25}\left[2x - \frac{x^2}{2}\right]_{1.5}^{1.8} \\
&= (0.96 - \frac{16}{25}) + \frac{8}{25}\left[(3.6 - 1.62) - (3 - 1.125)\right] \\
&= 0.32 + \frac{8}{25}[0.105] \\
&= 0.32 + 0.0336 \\
&= 0.3536.
\end{aligned}
$$

(iii)

$$
\begin{aligned}
\text{mean} &= E(x) = \int x f(x) dx \\
&= \int_0^{1.5} \frac{16}{25} x dx + \frac{8}{25}\int_{1.5}^{2} (2x - x^2)dx \\
&= \left[\frac{16}{50}x^2\right]_0^{1.5} + \frac{8}{25}\left[x^2 - \frac{x^3}{3}\right]_{1.5}^{2} \\
&= 0.72 + \frac{8}{25}\left[\left(4 - \frac{8}{3}\right) - \left(\frac{9}{4} - \frac{27}{24}\right)\right] \\
&= \frac{36}{50} + \frac{8}{25} \times \frac{5}{24} \\
&= \frac{36}{50} + \frac{1}{15} \\
&= \frac{59}{75}
\end{aligned}
$$

(iv)

$$
\begin{aligned}
E(x^2) &= \int x^2 f(x) dx \\
&= \int_0^{1.5} \frac{16}{25} x^2 dx + \frac{8}{25}\int_{1.5}^{2} (2x^2 - x^3)dx
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{16}{25}\left[\frac{x^3}{3}\right]_0^{1.5} + \frac{8}{25}\left[\frac{2}{3}x^3 - \frac{x^4}{4}\right]_{1.5}^{2} \\
&= \frac{16}{25}\times\frac{27}{24} - 0 + \frac{8}{25}\left[\left(\frac{16}{3}-4\right) - \left(\frac{9}{4}-\frac{81}{64}\right)\right] \\
&= \frac{18}{25} + \frac{8}{25}\times\frac{67}{192} \\
&= \frac{499}{600} \\
\text{Variance} &= E(x^2) - [E(x)]^2 \\
&= \frac{499}{600} - \left(\frac{59}{75}\right)^2 \\
&= \frac{9577}{45000} \cong 0.21822222... \\
&\cong 0.213
\end{aligned}
$$

(v)

$$
f(x) = \begin{cases} \frac{16}{25}; & 0 \le x \le 1.5 \\ \frac{8}{25}(2-x); & 1.5 \le x \le 2 \\ 0, & \text{otherwise} \end{cases}
$$

For $x < 0,\ f(0) = 0 \Rightarrow F(0) = 0.$
For $0 \le x \le 1.5$

$$
\begin{aligned}
f(t) &= \int_0^x \frac{16}{25}dt = \left[\frac{16}{25}t\right]_0^x = \frac{16}{25}x \\
F(1.5) &= F(0) + \frac{16}{25}\times 1.5 \\
&= 0 + 0.96 = 0.96
\end{aligned}
$$

For $1.5 \le x \le 2$

$$
f(t) = \frac{8}{25}\int_{1.5}^{x}(2-t)dt
$$

$$
\begin{aligned}
&= \frac{8}{25}\left[2t - \frac{t^2}{2}\right]_{1.5}^{x} \\
&= \frac{8}{25}\left\{\left(2x - \frac{x^2}{2}\right) - (3 - \frac{9}{8})\right\} \\
&= \frac{8}{25}(2x - \frac{x^2}{2} - \frac{15}{8}) \\
&= -\frac{1}{25}(2x-5)(2x-3) \\
F(2) &= F(1.5) - \frac{1}{25}(4-5)(4-3) \\
&= 0.96 + 0.04 \\
&= 1
\end{aligned}
$$

For $x > 2$, $f(x) = 0 \Rightarrow F(2) = 1$

$$
F(x) = \begin{cases} 0, & x \leq 0 \\ \frac{16}{25}x, & 0 \leq x \leq 1.5 \\ \frac{8}{25}(2 - \frac{x^2}{2} - 1.1875), & 1.5 \leq x \leq 21, \quad x \geq 2\text{otherwise} \end{cases}
$$

Example 6.3.6

A random variable X has cumulative distribution function given by

$$
F(x) = \begin{cases} 0, & x \leq -1 \\ k(x+3)(x+1), & -1 \leq x \leq 0 \\ k(3+4x), & 0 \leq x \leq 1 \\ \frac{k}{2}(10x - x^2 + 5), & 1 \leq x \leq 3 \\ 1, & x \geq 3 \end{cases}
$$

Find

(i) the value of the constant k

(ii) the probability density function, f(x)

(iii) the mean and variance of X

(iv) $P(0.5 \leq x \leq 2)$

(v) the median of X

Solution:

(i)

$$F(x) = \begin{cases} 0, & x \leq -1 \\ k(x+3)(x+1), & -1 \leq x \leq 0 \\ k(3+4x), & 0 \leq x \leq 1 \\ \frac{k}{2}(10x - x^2 + 5), & 1 \leq x \leq 3 \\ 1, & x \geq 3 \end{cases}$$

Taking

$$\begin{aligned} \frac{k}{2}(10x - x^2 + 5), & \quad 1 \leq x \leq 3 \\ \frac{k}{2}[30 - 9 + 5] &= 1 \\ \frac{k}{2}(26) &= 1 \\ k &= \frac{1}{13} \end{aligned}$$

(ii)

$$f(x) = F'(x) = \begin{cases} \frac{1}{13}(2x+4), & -1 \leq\leq 0 \\ \frac{4}{13}, & 0 \leq x \leq 1 \\ \frac{1}{13}(5-x), & 1 \leq x \leq 3 \\ 0, & elsewhere \end{cases}$$

(iii)

$$\begin{aligned}
\text{Mean} &= \int xf(x)dx \\
&= \int_{-1}^{0} \frac{1}{13}(2x^2+4x)dx + \int_{0}^{1} \frac{4}{13}xdx + \int_{0}^{3}(5x-x^2)dx \\
&= \frac{1}{13}\left[\frac{2}{3}x^3+2x^2\right]_{-1}^{0} + \frac{4}{13}\left[\frac{x^2}{2}\right]_{0}^{1} + \frac{1}{13}\left[\frac{5}{2}x^2-\frac{x^3}{3}\right]_{1}^{3} \\
&\quad \frac{1}{13}\left[0-\left(-\frac{2}{3}+2\right)\right] + \frac{4}{13}\times\frac{1}{2} + \frac{1}{13}\left[\left(\frac{45}{2}-9\right)-\left(\frac{5}{2}-\frac{1}{3}\right)\right] \\
&= -\frac{4}{39}+\frac{2}{13}+\frac{34}{39} \\
&= \frac{12}{13}
\end{aligned}$$

$$\text{Variance} = E(x^2) - [E(x)]^2$$

$$\begin{aligned}
&= \int_{-1} \frac{1}{13}(2x^3+4x^2)dx + \int_{0}^{1}\frac{4}{13}x^2dx + \frac{1}{13}\int_{1}^{3}(5x^2-x^3)dx \\
&= \frac{1}{13}\left[\frac{1}{2}x^4+\frac{4}{3}x^3\right]_{-1}^{0} + \frac{4}{13}\left[\frac{x^3}{3}\right]_{0}^{1} + \frac{1}{13}\left[\frac{5}{3}x^3-\frac{x^4}{4}\right]_{1}^{3} \\
&= \frac{1}{13}\left[0-\left(\frac{1}{2}-\frac{4}{3}\right)\right] + \frac{4}{13}\times\frac{1}{3} + \frac{1}{13}\left[\left(45-\frac{81}{4}\right)-\left(\frac{5}{3}-\frac{1}{4}\right)\right] \\
&= \frac{5}{78}+\frac{4}{39}+\frac{70}{39} \\
&= \frac{51}{26} \\
Variance &= \frac{51}{26} - \left(\frac{12}{13}\right)^2 \\
&= \frac{51}{26} - \frac{144}{169}
\end{aligned}$$

$$= \frac{375}{338}$$

$$= 1\frac{37}{338}$$

(iv)

$$\begin{aligned} P(0.5 \leq x \leq 2) &= \int_{0.5}^{1} \frac{4}{13} dx + \frac{1}{13} \int_{1}^{3} (5-x) dx \\ &= \left[\frac{4}{13}x\right]_{\frac{1}{2}}^{1} + \frac{1}{13}\left[5x - \frac{x^2}{2}\right]_{1}^{2} \\ &= \frac{4}{13} - \frac{4}{26} + \frac{1}{13}\left[(10-2) - \left(5 - \frac{1}{2}\right)\right] \\ &= \frac{2}{13} + \frac{1}{7} \\ &= \frac{11}{26} \end{aligned}$$

(v)

$$f(x) = \begin{cases} \frac{1}{13}(2x+4), & -1 \leq x \leq 0 \\ \frac{4}{13}, & 0 \leq x \leq 1 \\ \frac{1}{13}(5-x), & 1 \leq x \leq 3 \\ 0, & elsewhere \end{cases}$$

$$\begin{aligned} \frac{1}{13}\int_{-1}^{0} (2x+4) dx &= \frac{1}{13}\left[x^2 + 4x\right]_{-1}^{0} \\ &= \frac{1}{13}[0 - (1-4)] \\ &= \frac{3}{13} \end{aligned}$$

$$\begin{aligned}\int_0^1 \frac{4}{13}dx &= \left[\frac{4}{13}x\right]_0^1 \\ &= \frac{4}{13}\end{aligned}$$

Since $\frac{3}{13} + \frac{4}{13} = \frac{7}{13} > \frac{1}{2}$, the median is in the interval $0 \leq x \leq 1$. Therefore

$$\begin{aligned}\frac{1}{13}\int_{-1}^{0}(2x+4)dx + \int_0^m \frac{4}{13}dx &= \frac{1}{2} \\ \frac{3}{13} + \left[\frac{4}{13}x\right]_0^m &= \frac{1}{2} \\ \frac{3}{13} + \frac{4m}{13} &= \frac{1}{2} \\ \frac{4m}{13} &= \frac{7}{26} \\ m &= \frac{7}{8}\end{aligned}$$

Exercise 6

1. A random variable X has a probability density function given by

$$f(x) = \begin{cases} 4kx, & 0 \leq x \leq 1 \\ k(6-2x), & 1 \leq x \leq 2 \\ 0, & elsewhere \end{cases}$$

Find

(i) the value of the constant k

(ii) the mean value of X

(iii) the variance of X

(iv) the median of X

2. A continuous random variable X has a probability density function given by

$$f(x) = \begin{cases} \frac{x}{16}, & 0 \leq x \leq 4 \\ a, & 4 \leq x \leq 6 \\ 0, & elsewhere \end{cases}$$

where a is a constant. Find

(i) the value of a
(ii) the expectation of X
(iii) the median of X
(iv) the cumulative distribution function, F(x).

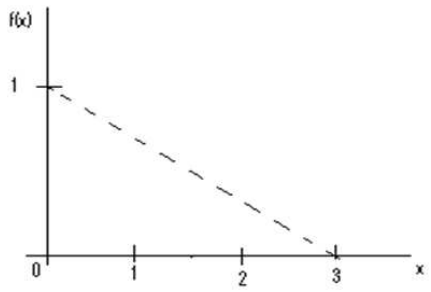
3. The probability density function of a random variable X is given by

$$f(x) = \begin{cases} \sin x, & 0 \leq x \leq \frac{\pi}{2} \\ 0, & elsewhere \end{cases}$$

Find

(i) $P(x \geq \frac{\pi}{4})$
(ii) the mean of X
(iii) the median of X
(iv) the cumulative distribution function

4. The probability density function f(x) of a random variable X takes on the form shown in the diagram below.



(i) Determine the expression for f(x). Hence find

(ii) the mean and variance of X

(iii) the cumulative density function, F(x).

5. A random variable X has probability density function

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

where k is a constant.
Determine

(i) the value of k

(ii) the mean of X

(iii) the median of X

(iv) $P(-0.5 \leq x \leq 1)$

(v) cumulative distribution function, $F(x)$.

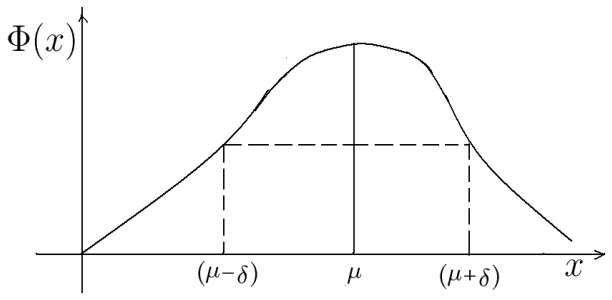
Chapter 7

THE NORMAL DISTRIBUTION

7.1 Introduction

The normal distribution is an important distribution in statistics. It arises in many situations of nature and social life. If X is a continous random variable following a normal distribution with mean μ and variance δ^2, then it can be represented by $X \sim N(\mu, \delta^2)$. The sketch of $\Phi(x)$ below is plotted from the formula

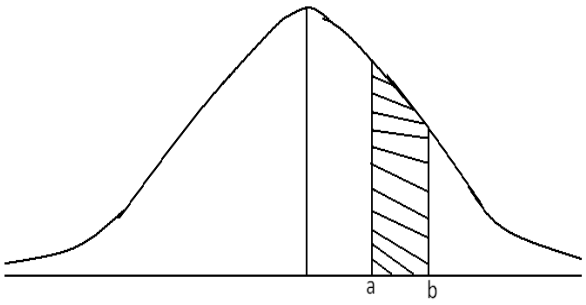
$$\Phi(x) = \frac{1}{\sqrt{2\tau} \cdot \delta} e^{-(x-\mu)^2/2\delta^2}, \quad -\infty < x < \infty$$



This curve for $\Phi(x)$ is called a normal curve. The curve has the following properties:

1. It is symmetrical about the mean $x = \mu$.
2. It never touches the x-axis but approaches it asymptotically. As $x \to \pm\infty, \Phi(x) \to 0$.
3. It has one maximum at $x = \mu$ (it is unimodal)
4. The area under the curve sums to unity

5. The mean, median and mode coincide at the maximum value of the function
6. The area under the normal curve is used to find probabilities $P(a < x < b)$.



So that

$$P(a < x < b) = \int_a^b \Phi(x)dx = \text{ Area}$$

7.2 Standardisation

$\Phi(x)$ is a probability density function which does not give the probability of x but gives the probability that x lies in a certain range. The probability that $X < x$ is given by

$$\begin{aligned} P(X < x) &= \int_{\infty}^{x} \Phi dx \\ &= \int_{-\infty}^{x} \frac{1}{\sqrt{2\tau} \cdot \delta} e^{\frac{-(x-\mu)^2}{2\delta^2}} dx \end{aligned}$$

which is evaluated using Mathematics beyond 'A' level. To avoid the complicated integration, we replace the original variable X by a standardized variable Z where

$$Z = \frac{x - \mu}{\delta}$$

So that the equation $\Phi(x) = \frac{1}{\sqrt{2\tau} \cdot \delta} e^{-(x-\mu)^2/2\delta^2}$ reduces to

$$y = \frac{1}{\sqrt{2\tau}} e^{-z^2/2}$$

The shape of the curve does not change except it becomes symmetrical about the y-axis so that the mean is 0 and standard deviation is 1.
When the variable has been standardised, tables are then used in the evaluation of areas (probabilities) between X_1 and X_2.

Example 7.2.1

The weights of a population of women who attend a maternity clinic are Normally distributed with a mean of $65kg$ and a standard deviation of $5kg$.
What is the probability that the weight of a woman chosen at random is

(i) less than $55kg$

(ii) less than $70kg$

(iii) between $55kg$ and $70kg$

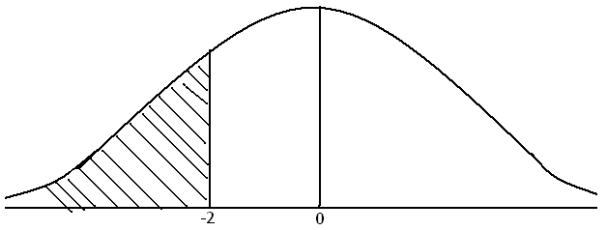
(iv) greater than $75kg$

(v) less than 60kg

Solution:

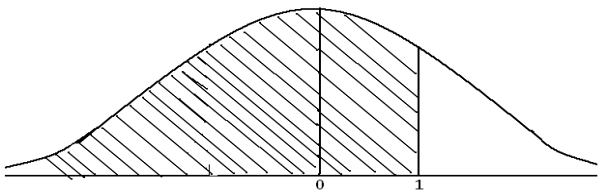
Here $\mu = 65, \delta = 5$ so that $z = \frac{x-65}{5}$

(i) $P(X < 55) = P(Z < \frac{55-65}{5}) = P(Z < -2)$

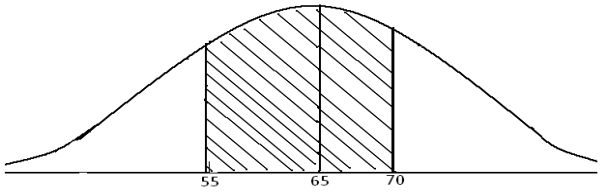


From tables, $P(Z < 2) = 0.5 - 0.4772 = 0.0228$.

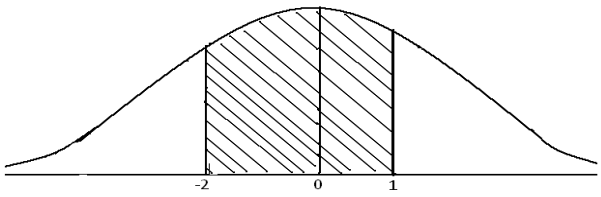
(ii) $P(X < 70) = P(Z < \frac{70-65}{5}) = P(Z < 1)$
From tables $P(Z < 1) = 0.5 + 0.3413 = 0.8413$



(iii)



$$\begin{aligned} P(55 \leq x < 70) &= P(\frac{55-65}{5} \leq Z \leq \frac{70-65}{5}) \\ &= P(-2 \leq z \leq 1) \end{aligned}$$



From tables

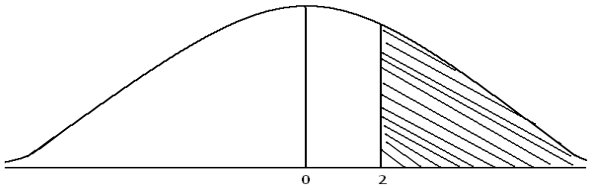
$$\begin{aligned} P(-2 \leq z \leq 1) &= 0.4772 + 0.3413 \\ &= 0.8185 \end{aligned}$$

(iv)

$$\begin{aligned} P(X > 75) &= P(Z > \frac{75-65}{5}) \\ &= P(Z > 2) \end{aligned}$$

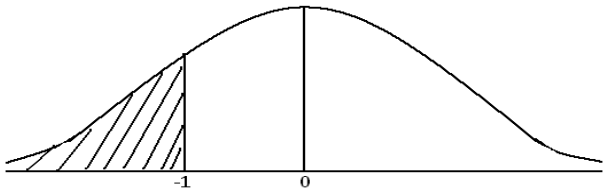
From tables

$$\begin{aligned} P(Z > 2) &= 0.5 - 0.4772 \\ &= 0.0228 \end{aligned}$$



(v)

$$\begin{aligned} P(X < 60) &= P(Z < \frac{60-65}{5}) \\ &= P(Z < -1) \end{aligned}$$



From tables, $P(Z < -1) = 0.5 - 0.3413 = 0.1587$.

Example 7.2.2

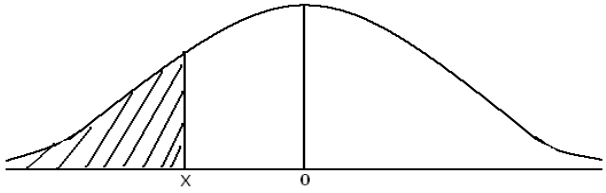
A given brand of light bulbs has a life time which is normally distributed with mean 1600 hours and standard deviation 40 hours. What should the quaranted life time of the bulbs be so that only 4% of the bulbs will have to be replaced under quarantee?

Solution:

Let X be the quaranteed life time

$$Z = \frac{X - 1600}{40}$$

The $Z-$value which leaves an area of 0.04 to the left is -1.751. Therefore.



$$-1.751 = \frac{X - 1600}{40}$$

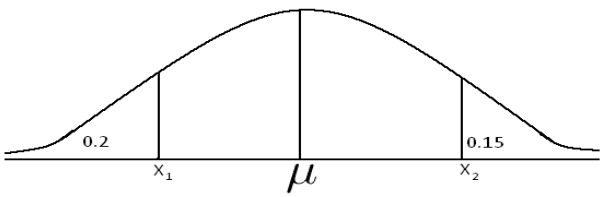
$$X = 1529.96 \simeq 1530$$

The quaranteed life time should be 1530 hours.

Example 7.2.3

In an examination 20% of the candidates fail and 15% achieve distinction. If 60 is the pass mark and the minimum mark required for a distinction is 90, assuming that the marks are normally distributed estimate the mean mark and standard deviation to the nearest whole number.

Solution:



The Z value which leaves an area of 0.2 to the left is -0.842 while that which leaves an area of 0.15 to the right is 1.036.

$$\text{Using } Z = \frac{x - \mu}{\delta},\ X_1 = 60 \text{ and } X_2 = 90$$

$$Z_1 = -0.842 = \frac{60 - \mu}{\delta}$$

$$\mu - 0.842\delta = 60 - \mu$$
$$\mu - 0.842\delta = 60 \text{.....(1)}$$

$$\text{and } Z_2 = 1.036 = \frac{90 - \mu}{\delta}$$

$$1.036\delta = 90 - \mu$$
$$\mu + 1.036\delta = 90 \text{.....(2)}$$

Solving equations (1) and (2), $\delta = 15.9744 \simeq 16$ and $\mu = 73.45047923 \simeq 73$
Therefore the mean mark is 73 and the standard deviation is 16.

7.3 Distribution of a sample mean $\bar{x}$ from a normal population

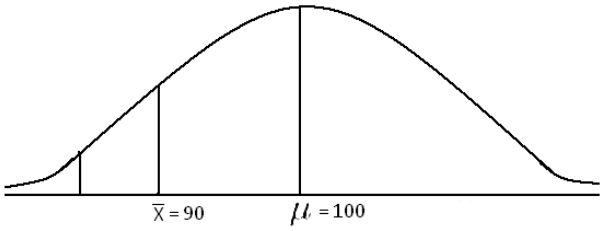
If $x_1, x_2, x_3, ..., x_n$ is a randomly chosen sample from a normal population that has mean $\bar{x}$ and standard deviation δ, then the distribution of $\bar{x}$ is also normally distributed with mean and standard deviation $\delta/\sqrt{n}$.

Example 7.3.1

A random sample of size 25 is taken from a normal population with mean 100 and standard deviation 8. Find the probaility that the mean of the sample is less than 90.

Solution:

$n = 25, \mu = 100$. Let the mean of the sample be $\bar{x}$



$$\begin{aligned} P(\bar{X} < 90) &= P(Z < \frac{90-100}{8}) \\ &= P(Z < -1.25) \end{aligned}$$

From tables,

$$\begin{aligned} P(Z < -1.25) &= 0.5 - 0.3944 \\ &= 0.1056 \end{aligned}$$

Example 7.3.2

The lengths of snakes at Nairobi snake park are normally distributed with mean 170cm and standard deviation 10cm. Calculate the probability that the mean length of a sample of 16 snakes will be between 165cm and 175cm

Solution:

$$\begin{aligned} P(165 < x < 175) &= P\left(\frac{165-170}{10/4} < z < \frac{175-170}{10/4}\right) \\ &= P(-2 < z < 2) \\ &= 0.4772 + 0.4772 \\ &= 0.9544 \end{aligned}$$

7.4 Normal Approximation to Binomial Distribution

The normal distribution is used as an approximation to the binomial distribution in cases when n is large the probability is not so far from $\frac{1}{2}$. This is done to avoid tedious calculations.
If $X \sim B(n,p)$, $E(x) = np = \mu$ and $var(x) = npq$ where $q = 1-p$. Then $X \sim N(np, npq)$, so that

$$Z = \frac{X \pm 0.5 - np}{\sqrt{npq}}.$$

Since the binomial distribution is discrete, the value ± 0.5 is used to make it continuous.

Example 7.4.1

A coin is tossed 150 times. What is the probability that

(a) there will be more than 80 heads?

(b) there will be atleast 70 and atmost 90 heads?

(c) there will be less than 60 heads?

Solution:

$$\begin{aligned} \text{Mean} &= \mu = np = 150 \times \frac{1}{2} = 75 \\ \delta^2 &= npq = 150 \times \frac{1}{2} \times \frac{1}{2} = 37.5 \\ \Rightarrow \delta &= \sqrt{37.5} \end{aligned}$$

Let X be the number of heads obtained

(a)

$$\begin{aligned} P(X > 80) &= P(X \geq 81) \\ P(X \geq 81) &= P(Z > \frac{80.5 - 75}{\sqrt{37.5}}) \\ &= P(Z > 0.898) \\ &= 0.5 - 0.3155 \\ &= 0.1845 \end{aligned}$$

(b)

$$= P(\frac{69.5-75}{\sqrt{37.5}} \leq Z \leq \frac{90.5-75}{\sqrt{37.5}})$$

$$\begin{aligned} P(70 \leq x \leq 90) &= P(69.5 \leq x \leq 90.5) \\ &= P(-0.898 \leq Z \leq 2.531) \\ &= 0.3155 + 0.4943 \\ &= 0.8098 \end{aligned}$$

(c)

$$\begin{aligned} P(X < 60) &= P(X \leq 59.5) \\ &= P(Z \leq \frac{59.5-75}{\sqrt{37.5}}) \\ &= P(Z \leq -2.531) \\ &= 0.5 - 0.4943 \\ &= 0.0057 \end{aligned}$$

Example 7.4.2

Use the normal approximation to the binomial distribution with $n = 600$ and $P = 0.5$ to find the probability of a value

(i) less than 320

(ii) greater than 280

(iv) lying between 290 and 315 inclusive

Solution:

$$\mu = np = 600 \times \ 0.5 = 300 \text{ and } \delta = \sqrt{npq} = \sqrt{600 \times 0.5 \times 0.5} = \sqrt{150}$$

(i)

$$\begin{aligned} P(X < 320) &= P(X \geq 320.5) \\ &= P(Z < \frac{320.5-300}{\sqrt{150}}) \\ &= P(Z \leq 1.674) \\ &= 0.5 + 0.4539 \\ &= 0.9539 \end{aligned}$$

(ii)

$$\begin{aligned} P(X > 280) &= P(X \geq 279.5) \\ &= P(Z \geq \frac{279.5 - 300}{\sqrt{150}}) \\ &= P(Z - 1.674) \\ &= 0.5 + 0.4539 \\ &= 0.9539 \end{aligned}$$

(iii)

$$\begin{aligned} P(290 \leq X \leq 315) &= P(289.5 \leq x \leq 315.5) \\ &= P(\frac{289.5 - 300}{\sqrt{150}} \leq Z \leq \frac{315.5 - 300}{\sqrt{150}}) \\ &= P(-0.857 \leq Z \leq 1.266) \\ &= 0.3043 + 0.3973 \\ &= 0.7016. \end{aligned}$$

7.5 Normal Approximation to Poisson distribution

The poisson probabilities may be approximated by the normal distribution when the sample size is large, i.e, greater than 30. Recall from chapter 5 that for a Poisson distribution, mean $= E(x) = \mu = \lambda$ and variance $= var(x) = \delta^2 = \lambda$.
The Z statistic is obtained by substituting for the mean and variance as follows.

$$Z = \frac{X - \lambda}{\sqrt{\lambda}}$$

So that the probability of success between a and b inclusive is given as

$$P(a \leq x \leq b) = P(\frac{a - \lambda}{\sqrt{\lambda}} \leq Z \leq \frac{b - \lambda}{\sqrt{\lambda}})$$

Example 7.5.1

The average number of customers entering a bank in a 30-minute period is 80. Find the probability that in a 30-minute period, between 60 and 95 customers inclusive, will enter the bank.

Solution:
For a Poisson distribution

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} \text{ for } x = 0, 1, 2, \ldots \text{ and } \lambda > 0$$

In this case n is large and λ is big, so that there would be tedious calculations to solve this. We are bailed out of the problem by utilising a normal approximation to the Poisson distribution

$$\begin{aligned} P(60 \leq x \leq 95) &= P(\frac{60-80}{\sqrt{80}} \leq Z \leq \frac{95-80}{\sqrt{80}}) \\ &= P(-2.236 \leq Z \leq 1.677) \\ &= 0.4873 + 0.4542 \\ &= 0.9415 \end{aligned}$$

Exercise 7

1. Use normal distribution tables to evaluate
   (a) $P(Z > 1.6)$ (b) $P(Z < -2.4)$ (c) $P(Z > 0.5)$
   (d) $P(2.1 \leq Z \leq 2.5)$ (e) $P(Z < 1.2)$

2. A random variable Z is normally distributed with mean 0 and standard deviation 1. Find the following values of c.

   (i) $P(Z < c) = 0.4$

   (ii) $P(Z > c) = 0.82$

   (iii) $P(-c \leq x \leq c) = 0.8444$

   (iv) $P(0 \leq z < c) = 0.4099$

3. The marks in an examination were normally distributed with mean μ and standard deviation δ. Solve for μ and δ, 10% of the candidates scored more than 80 and 20% scored less than 30 .16 candidates were chosen at random from those who sat for the examination. Find the probability that their average mark exceeds 55.

4. The marks of 1000 candidates in an examination are normally distributed with a mean of 60 marks and a standard deviation of 25 marks

   (i) If the pass mark is 48, estimate the number of candidates who passed the examination

   (ii) If 8% of the candidates obtained distrinction, estimate the minimum mark for a distinction

5. An unbiased coin is tossed 100 times Find

   (i) the probability of obtaining 52 heads

   (ii) the probability of obtaining more than 38 heads.

6. The life time of bulbs manufactured by an electric company is normally distributed. Out of 8000 bulbs, 400 have a life time less than 1200 hours and 350 have life time more than 1500 hours.

   (i) find the mean and standard deviation of the bulb life time

   (ii) find the percentage of the bulbs with lifetime between 1300 and 1400 hours

   (iii) If a sample of 36 bulbs is selected at random, find the probability that the mean of the lifetime exceeds 1350 hours.

7. In a crowd of people at the car park, 80% were supporters of the Democratic party while 20% were supporters of the Republic party. If 3600 of them are selected randomly, what is the probability that more than 750 were supporters of the Republican party?

8. A professor teaches statistics every year. The tests for the course are standardised so that the test scores have a normal distribution with mean 70 and a standard deviation of 10. The professor gives 12% A, 20% B, 35%C, 22% D and 11% F.

   (i) What letter grade will a student who scores 76 points on the test receive?

   (ii) What letter grade will a student who scores 60 points receive?

   (iii) How many points does a student need to score to get an A?

   (iv) What minimum points does a student need to score to avoid a failure?

9. The manages of Fina bank has found out that customers come to to cash their pay cheques on Monday. The amount of money drawn on Monday follows a normal distribution with 15 million as the mean and 3 million as the standard deviation. The manager wants to ensure that the amount of money in the bank can cover 99% of the Monday withdrawals. What is the minimum amount of of money that should be at hand to meet the demand?

10. A mobile phone hand set producer claims that the lifetimes of the hand sets follows a normal distribution with a mean of 84 months and a standard deviation of 14 months. The producer quarantees that a new hand set will last longer than 70 months or the full price will be refunded. If 1.2 million hand sets are sold, how many refunds will be claimed? The producer would like to refund not more than 5% of the hand sets sold. What should his quarantee period be?

11. A consumer protection body wants to find out whether a beverage company actually puts 300mls of soda in a can labelled 300mls. Assume the soda put in the cans follows a normal distribution with a mean of 301mls and standard deviation of 2mls

    (i) What is the probability that a certain can contains more than 300 militres of soda?

    (ii) The consumer protection body bought 196 cans of soda. What is the probability that among them, it found fewer than 50 cans that do not contain the stated amount of soda?

12. Using the normal approximation to the Poisson distribution with $\lambda = 80$, what is the probability that

    (i) there will be a value greater than 60

    (ii) there will be a value between 60 and 85 inclusive

    (iii) there will be a value less than 70.

13. A bottling company is supposed to pack 500mls of soda in a bottle. it is found that the amounts packed follow a normal distribution with a mean of 502 mililitres and standard deviation of 5 mililitres. The control procedures are designed to reject a bottle with less than 496 mililitres or more than 510 mililitres.

    (i) Find the proportion of bottles that will be rejected

    (ii) the value to which the standard deviation should be reduced, leaving the mean at 502 mililitres so that bottles rejected due to being below 496 mililitres are 5% or less for the entire production

14. A national examining body gave an examination to a large number of candidates. The marks they obtained were normally distributed. A quarter of the candidates scored less than 35 marks and half of the candidates scored more than 60 marks

    (i) Determine the mean and standard deviation of the distribution

    (ii) 18% of the candidates obtained distinction. Determine the minimum mark for a distinction

    (iii) Find the proportion of the candidates who scored between 70 and 80 marks.

Chapter 8

OTHER THEORETICAL DISTRIBUTIONS

8.1 Introduction

This chapter addresses aspects that we have not looked at before. We have looked at the normal. binomial and Poisson distribution. We shall look at the uniform distribution in its dicrete and continous, aspects, Geometric distribution, the exponential distribution and moment generating functions. Under moment generating functions, we shall revisit the binomial and Poisson distributions and then generalise for continous random variables.

8.2 Discrete uniform distribution

This distribution is utilised when handling a discrete random variable. For instance if a die is tossed, each face has a probability of $\frac{1}{6}$ of showing up. If X can takes values $1, 2, ...n$, then $P(X = x)$ is uniformly equal to $\frac{1}{n}$ because the sum of the probabilities must be 1.

$$\begin{aligned} \text{If } P(X = x) &= \frac{1}{n}, \text{ then} \\ \text{Mean} &= \mu = E(x) \\ &= \sum_{1}^{n} xP(X = x) \\ &= \sum_{1}^{n} x \cdot \frac{1}{n} \end{aligned}$$

$$= \frac{1}{n}\sum_{1}^{n} x$$

$\sum_{1}^{n} x$ gives the sum of the first n natural numbers so that

$$\begin{aligned} E(x) &= \frac{1}{n}\left[\frac{1}{2}n(n+1)\right] \\ &= \frac{1}{2}(n+1) \end{aligned}$$

Therefore, if X can take on n values and $P(X = x) = \frac{1}{n}$. Then $E(x) = \frac{1}{2}(n+1)$. Variance is given by

$$\begin{aligned} V(X) &= E(x^2) - [E(x)]^2 \\ \text{But } E(x^2) &= \sum_{1}^{n} x^2 P(X = x) \\ &= \sum_{1}^{n} x^2 \cdot \frac{1}{n} \\ &= \frac{1}{n}\sum_{1}^{n} x^2 \end{aligned}$$

The sum here is for the squares of the first n natural numbers so that

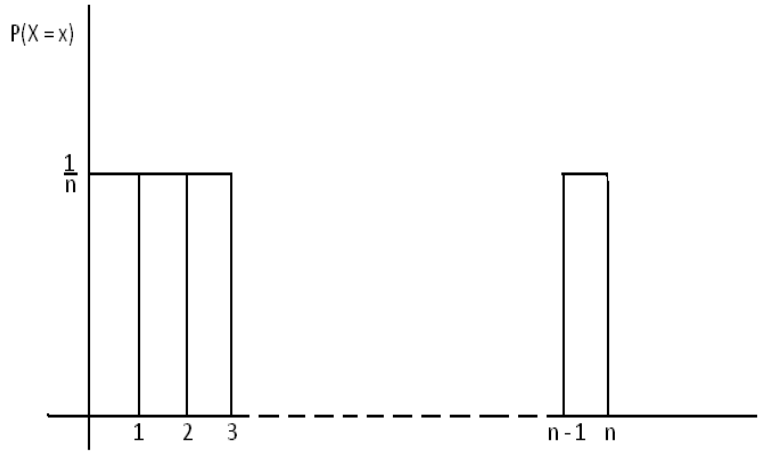
$$\begin{aligned} E(x^2) &= \frac{1}{n}\left[\frac{1}{6}n(n+1)(2n+1)\right] \\ &= \frac{1}{6}(n+1)(2n+1) \end{aligned}$$

and as we found earlier. $E(x) = \frac{1}{2}(n+1)$, so that

$$\begin{aligned} V(X) &= \frac{1}{6}(n+1)(2n+1) - \left[\frac{1}{2}(n+1)\right]^2 \\ &= \frac{1}{6}(n+1)(2n+1) - \frac{1}{4}(n+1)^2 \end{aligned}$$

$$
\begin{aligned}
&= (n+1)\left[\frac{1}{6}(2n+1) - \frac{1}{4}(n+1)\right] \\
&= (n+1)\left[\frac{2(2n+1) - 3(n+1)}{12}\right] \\
&= \frac{1}{12}(n+1)(n-1) \\
&= \frac{1}{12}(n^2 - 1)
\end{aligned}
$$

The generalised dicrete uniform distribution looks thus.

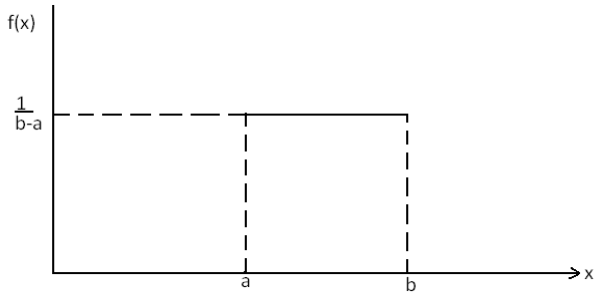


8.3 Continuous Uniform distribution

This is sometimes called the rectangular distribution. It is a continous random variable whose probability density function is defined by

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

and it is sketched thus



$$\begin{aligned}
\text{The mean } \mu &= E(x) = \int_a^b xf(x)dx \\
&= \int_a^b \frac{x}{b-a}dx \\
&= \left[\frac{x^2}{2(b-a)}\right]_a^b \\
&= \frac{b^2-a^2}{2(b-a)} \\
E(x) &= \frac{(b+a)(b-a)}{2(b-a)} \\
&= \frac{a+b}{2}
\end{aligned}$$

and the variance is given by

$$\begin{aligned}
V(X) &= E(x^2) - [E(x)]^2 \\
&= \int_a^b x^2 f(x)dx - \left(\frac{a+b}{2}\right)^2 \\
&= \left[\frac{x^3}{3(b-a)}\right]_a^b - \frac{1}{4}(a+b)^2 \\
&= \frac{b^3-a^3}{3(b-a)} - \frac{1}{(a+b)^2}
\end{aligned}$$

But $b^3 - a^3 = (b-a)^3 + 3ab(b-a)$, So that

$$\begin{aligned} V(X) &= \frac{(b-a)^3 + 3ab(b-a)}{3(b-a)} - \frac{1}{4}(a+b)^2 \\ &= \frac{(b-a)^2 + 3ab}{3} - \frac{(a+b)^2}{4} \\ &= \frac{b^2 + a^2 + ab}{3} - \frac{(a+b)^2}{4} \\ &= \frac{4(b^2 + a^2 + ab) - 3(a+b)^2}{12} \\ &= \frac{4b^2 + 4a^2 + 4ab - 3(a^2 + 2ab + b^2)}{12} \\ &= \frac{b^2 + a^2 - 2ab}{12} \\ &= \frac{(a-b)^2}{12} \end{aligned}$$

Example 8.3.1

The number of cars that are stopped at a security check point daily is uniformly distributed between 600 cars and 1100 cars.

(i) find the probability that atleast 800 cars are stopped at the check point.

(ii) what is the expected number of cars that will be stopped on any given day.
Solution:

(i) We need $P(X \geq 800)$

$$\begin{aligned} &= \int_{800}^{1100} \frac{1}{1100 - 600} dx \\ &= \left[\frac{x}{500}\right]_{800}^{1100} \\ &= \frac{1100 - 800}{500} = \frac{300}{500} = \frac{3}{5}. \end{aligned}$$

(ii)

$$\begin{aligned} E(x) &= \int_a^b xf(x)dx \\ &= \int_{600}^{1100} \frac{x}{1100-600}dx \\ &= \left[\frac{x^2}{2(500)}\right]_{600}^{1100} \\ &= \frac{1100^2-600^2}{1000} \\ &= \frac{1210000-360000}{1000} \\ &= 850 \end{aligned}$$

Example 8.3.2

Find the mean and variance of a continous random variable X which is uniformly distributed over the interval
(i) 0 to 2 (ii) 3 to k.
Solution:

(i)

$$\begin{aligned} f(x) &= \begin{cases} \frac{1}{2}, & 0 \leq x \leq 2 \\ 0, & \text{elsewhere} \end{cases} \\ E(x) &= \int_0^2 xf(x)dx \\ &= \int_0^2 \left[\frac{x}{2}\right] dx \\ &= \left[\frac{x^2}{4}\right]_0^2 = \frac{4}{4} = 1 \\ \text{Variance} &= \int_0^2 x^2 f(x)dx - [E(x)]^2 \end{aligned}$$

$$\begin{aligned}\int_0^2 \frac{x^2}{2}dx - 1^2 &= \left[\frac{x^3}{6}\right]_0^2 - 1 \\ &= \frac{8}{6} - 1 \\ &= \frac{1}{3}\end{aligned}$$

(ii)

$$f(x) = \begin{cases} \frac{1}{k-3}, & 3 \leq x \leq k \\ 0, & \text{elsewhere} \end{cases}$$

$$\begin{aligned}E(x) &= \int_3^k \frac{x}{k-3}dx \\ &= \left[\frac{x^2}{2(k-3)}\right]_3^k \\ &= \frac{k^2-3^2}{2(k-3)} = \frac{k+3}{2}.\end{aligned}$$

$$\begin{aligned}Var(x) &= \int_3^k \frac{x^2}{k-3}dx - \left(\frac{k+3}{2}\right)^2 \\ &= \left[\frac{x^3}{3(k-3)}\right]_3^k - \frac{(k+3)^2}{4} \\ &= \frac{k^3-3^3}{3(k-3)} - \frac{(k+3)^2}{4} \\ &= \frac{(k-3)^3+3(3k)(k-3)}{3(k-3)} - \frac{(k+3)^2}{4} \\ &= \frac{k^2-2(3k)+3^2}{3} - \frac{(k^2+6k+9)}{4} \\ &= \frac{1}{12}(3-k)^2 \\ &= \frac{(3-k)^2}{12}\end{aligned}$$

8.4 The Geometric distribution

This distribution has a similarity with the binomial distribution because each trial can have only two possible mutually exclusive and exhaustive outcomes with constant probabilities. The trials are also independent. The difference is that our interest is not the number of successful trials (like in the binomial distribution), but the number of trials required to achieve a success. The number of trials is not constant so it is the variable we are interested in

Example 8.4.1

A die is tossed until a 4 is obtained. Find the probability function for the number of throws required to achieve that.

Solution:

Let X be the number of throws required to obtain a 4.

$$\begin{aligned} P(X=1) &= \frac{1}{6} \\ P(X=2) &= \left(\frac{5}{6}\right)\left(\frac{1}{6}\right) \\ &= \frac{5}{36} \\ P(X=3) &= \left(\frac{5}{6}\right)^2\left(\frac{1}{6}\right) \\ &= \frac{25}{216} \end{aligned}$$

and so on.

The probability function is given by

$$P(x) = P(X=x) = \left(\frac{5}{6}\right)^{x-1}\left(\frac{1}{6}\right).$$

Generally, if the probability of success p, then

$$P(x) = P(X=x) = (1-p)^{x-1}p \text{ for } x > 0$$

The mean is given by $E(x) = \sum xP(x) = \sum xP(X=x)$

$$= \sum x(1-p)^{x-1}p$$

Putting values of $x = 1, 2, \ldots$
The mean is given by

$$\begin{aligned} & 1p + 2(1-p)p + 3(1-p)^2p + 4(1-p)^3p + \ldots \\ = \; & p\left[1 + 2(1-p) + 3(1-p)^2 + 4(1-p)^3 + \ldots\right] \end{aligned}$$

Comparing this expression with

$$(1-x)^{-2} = 1 + 2x + 3x^2 + 4x^3 + \ldots$$

$$E(x) = p \times p^{-2} = \frac{1}{p}$$

$$(\text{since } \; [1-(1-p)]^{-2} = p^{-2})$$

Example 8.4.2

Find the mean and variance of a discrete random variable X which is geometrically distributed where $P(x = q) = kq, \; q = 1, 2, \ldots n$
Solution:
The probabilities sum 1. Therefore

$$\sum_1^n P(x = q) = \sum_1^n kq = 1, \text{ so that}$$

$$k\sum_1^n q = \frac{1}{2}kn(n+1) = 1$$

Implying that $kn(n+1) = 2$
$k = \frac{2}{n(n+1)}$

$$\text{So that } P(x = q) = \frac{2q}{n(n+1)}.$$

$$\text{Now mean} = E(x) = \sum_1^n P(X = q)$$

$$= \sum_{q=1}^n \cdot \frac{2q}{n(n+1)}$$

$$= \frac{2}{n(n+1)} \sum_{q=1}^n q^2$$

This is again a sum of the squares of the first n natural numbers. So

$$\begin{aligned} E(x) &= \frac{2}{n(n+1)} \cdot \frac{1}{6} n(n+1)(2n+1) \\ &= \frac{1}{3}(2n+1) \\ \text{Variance} &= E(x^2) - [E(x)]^2 \\ \text{and } E(x^2) &= \sum_{q=1}^{1} x^2 P(X=q) \\ &= \sum_{q=1}^{n} q^2 \cdot \frac{2q}{n(n+1)} = 1 \\ &= \frac{2}{n(n+1)} \sum_{q=1}^{n} q^3 \end{aligned}$$

This is a sum of the first n cubes of the natural numbers but $\sum_{1}^{n} x^3 = \frac{1}{4}n^2(n+1)^2$, so that

$$\begin{aligned} E(x^2) &= \frac{2}{n(n+1)} \cdot \frac{1}{4} n^2(n+1)^2 \\ &= \frac{1}{2} n(n+1). \end{aligned}$$

Therefore,

$$\begin{aligned} \text{Variance} &= \frac{1}{2} n(n+1) - \left[\frac{1}{3}(2n+1)\right]^2 \\ &= \frac{1}{2} n(n+1) - \left[\frac{1}{9}(2n+1)(2n+1)\right] \\ &= \frac{1}{2}(n^2+n) - \frac{1}{9}(4n^2+4n+1) \\ &= \frac{9n^2+9n-8n^2-8n-2}{18} \end{aligned}$$

$$= \frac{n^2+n-2}{18}$$

$$= \frac{n^2+2n-n-2}{18}$$

$$= \frac{(n+2)(n-1)}{18}$$

$$= \frac{1}{18}(n+2)(n-1).$$

8.5 The Exponential Distribution

This distribution has a relationship with the Poisson distribution and is actually derived from the Poisson distribution. Consider a situation where patients arrive at a clinic, that is the number of patients in a given interval of time has a Poisson distribution. If the length of time between the arrival of patients is measured, these times form a continuous distribution. Let the average number of arrivals in unit time be λ. If time t has elapsed since a particular arrival of a patient, the probability that no further arrival is given by

$$P(X=0) = e^{-\lambda t}.$$

The probability that one call arrives in the next Δt is $P(X=1) = \lambda \Delta t e^{-\lambda \Delta t}$. These are as for Poisson distribution with mean λt and $\lambda \Delta t$, respectively. From the multiplication law for independent events, P(time elapsing between arrivals is t and $t+\Delta t$) = = P(arrivals in t) $\times$ P(one arrival in Δt)

$$\begin{aligned} &= P(X=0) \times P(X=1) \\ &= e^{-\lambda t} \times \lambda \Delta t e^{-\lambda \Delta t} \end{aligned}$$

If the change in time Δt is so small, $e^{-\lambda \Delta t}$ may be written as a series thus:

$$e^{-\lambda \Delta t} = 1 - \lambda \Delta t + \frac{(\lambda \Delta t)^2}{2!} + \ldots$$

$$\text{So that } P(X=0) \times P(x=1) = e^{-\lambda t} \times \lambda \Delta t (1 - \lambda \Delta t + \frac{(\lambda \Delta t)^2}{2!} + \ldots$$

Neglecting the terms in $(\lambda \Delta t)^2$ and higher powers because they are so small.

$$P(X=0) \times P(X=1) = e^{-\lambda t} \cdot \lambda \Delta t.$$

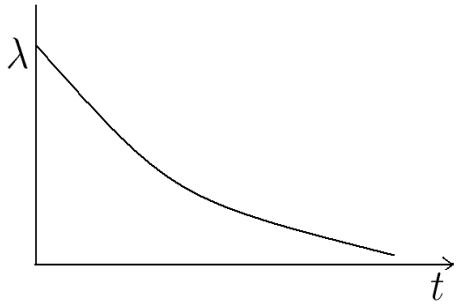
Let $f(t)$ be the probability density function for the time elapsing between arrivals of patients. Then $f(t)\Delta t$ is the probabililty that the time elapsing between arrivals is between t and $(t+\Delta t)$, so that

$$f(t) = \lambda e^{-\lambda t}.$$

So the distribution is given by

$$f(x) = \begin{cases} \lambda e^{-\lambda t}, & t \geq 0 \\ 0, & \text{elsewhere} \end{cases}$$

and is sketched thus



Example 8.5.1

Show that for the distribution

$$f(t) = \begin{cases} \lambda e^{-\lambda t}, & t \geq 0 \\ 0, & \text{elsewhere} \end{cases}$$

The mean is equal to the standard deviation which is equal to $\frac{1}{\lambda}$
Solution:

$$\begin{aligned} f(t) &= \lambda e^{-\lambda t},\ t \geq 0 \\ \text{Mean} &= \int_0^\infty t f(t) dt \\ &= \int_0^\infty \lambda t e^{-\lambda t} dt \\ &= \left[-te^{-\lambda t}\right]_0^\infty - \int -e^{-\lambda t} dt \\ &= 0 - \left[\frac{1}{\lambda} e^{-\lambda t}\right]_0^\infty \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{-\lambda}\left[e^{-\lambda t}\right]_0^{\infty} \\
&= -\frac{1}{\lambda}[0-1] \\
&= \frac{1}{\lambda} \\
\text{variance} &= \int_0^{\infty} t^2 f(t)dt - \left(\frac{1}{\lambda}\right)^2 \\
&= \int \lambda t^2 e^{-\lambda t}dt - \frac{1}{\lambda^2} \\
&= [-t^2 e^{-\lambda t}]_0^{\infty} - \frac{1}{\lambda^2} \\
&\quad [-t^2 e^{-\lambda t}]_0^{\infty} + \int 2te^{-\lambda t} - \frac{1}{\lambda^2} \\
&= \frac{2}{\lambda^2} - \frac{1}{\lambda^2} \\
&= \frac{1}{\lambda^2}
\end{aligned}
$$

but standard deviation is $\sqrt{variance}$

$\Rightarrow$ standard deviation $= \sqrt{\frac{1}{\lambda^2}} = \frac{1}{\lambda}$.

So mean = standard deviation = λ.

8.6 Moment Generating Functions

Calculations for mean and variance are made simpler when moment generating functions are used. This is done irrespective of whether the distribution is discrete or continous. This is using a sum in case of the former and using an integral in case of the latter.

8.6.1 Mean and Variance for a discrete distribution

The moment generating function for a discrete variable X is given by $M(t)$ which is defined by

$$M(t) = \sum_{all\ x} P(x)e^{xt} \text{.....(1)}$$

$$\text{but } e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \text{.....}$$

So that

$$M(t) = \sum_{all\ x} P(x)(1 + x + \frac{(xt)^2}{2!} + \frac{(xt)^3}{3!} + \frac{(xt)^4}{4!} + \text{.....(2)}$$

Differentiating equation(2) with respect to t, we get

$$\frac{dM(t)}{dt} = \sum_{all\ x} P(x)(x + x^2 + \frac{x^3t^2}{2!} + \frac{x^4t^3}{3!} + \text{.....)(3)}$$

Putting t = 0 in equation (3)

$$\left(\frac{dM(t)}{dt}\right)_{t=0} = \sum_{all\ x} xp(x) \text{.....(4)}$$

Equation (4) is an expression for the expectation (mean) of the random variable. Differentiating (3) with respect to t

$$\frac{d^2M(t)}{dt^2} = \sum_{all\ x} P(x)(x^2 + x^3t + \frac{x^4t^2}{2!} + \text{.....)(5)}$$

Putting $t = 0$ in equation (5)

$$\left(\frac{d^2M(t)}{dt^2}\right)_{t=0} = \sum_{all\ x} x^2P(x) \text{.....(6)}$$

Equation (6) gives an expression for $E(x^2)$.

$$\text{But } V(x) = E(x^2) - [E(x)]^2$$

So that the variance of x is given by

$$\begin{aligned}\text{Variance } (x) &= E(x^2) - [E(x)]^2 \\ &= \sum_{all\ x} x^2 p(x) - \left[\sum_{all\ x} xp(x)\right]^2 \\ &= \left(\frac{d^2M(t)}{dt^2}\right)_{t=0} - \left[\left(\frac{dM(t)}{dt}\right)_{t=0}\right]^2\end{aligned}$$

Which can in short be written as

$$[M''(t)]_{t=0} - [(M^{(}t))_{t=0}]^2$$

Using the moment generating functions, we can arrive at conclusions reached earlier in chapters 4 and 5 about binomial and Poisson distributions.

8.6.2 Mean and Variance for the Binomial distribution

The Binomial distribution is defined as

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

The moment generating function gives

$$\begin{aligned}M(t) &= \sum_{x=0} \binom{n}{x} e^{xt} p^x (1-p)^{n-x} \\ &= \sum_{x=0} \binom{n}{x} (pe^t\ )^x (1-p)^{n-x} \\ &= \left[Pe^t + (1-p)\right]^n\end{aligned}$$

Differentiating with respect to t

$$\begin{aligned}\frac{dM(t)}{dt} &= pe^t \cdot n(pe^t + (1-p))^{n-1} \\ \left(\frac{dM(t)}{dt}\right)_{t=0} &= = np(p+1-p)^{n-1} \\ &= np(1)^{n-1} = np.....(7)\end{aligned}$$

Equation (7) gives the mean of the binomial distribution. The second derivative gives.

$$\begin{aligned}
\frac{d^2M(t)}{dt^2} &= pe^t n(n-1)pe^t\left[pe^t+(1-p)\right]^{n-2} \\
&\quad +pe^t n\left[pe^t+(1-p)\right]^{n-1} \\
\left(\frac{d^2M(t)}{t^2}\right)_{t=0} &= pn(n-1)p[p+1-p]^{n-2}+pn[p+1-p]^{n-1} \\
&= np^2(n-1)+np
\end{aligned}$$

Therefore

$$\begin{aligned}
\text{Variance }(X) &= \left(\frac{d^2M(t)}{dt}\right)_{t=0}-\left[\left(\frac{dM(t)}{dt}\right)\right]^2 \\
&= \left[np^2(n-1)+np\right]-(np)^2 \\
&= n^2p^2-np^2+np-n^2p^2 \\
&= -np^2+np \\
&= np(-p+1) \\
&= np(1-p) \text{ but } 1-p=q \\
&= npq.....(8)
\end{aligned}$$

8.6.3 Mean and Variance for the Poisson distribution

The Poisson distribution is defined by

$$P(X=x)=\frac{e^{-\lambda}\lambda^x}{x!}\ ,x=0,1,2,\ldots$$

So that the moment generating function gives

$$\begin{aligned}
M(t) &= \sum_{x=0}\left(\frac{e^{xt}e^{-\lambda}\lambda^x}{x!}\right) \\
&= e^{-\lambda}\sum_{x=0}^{\infty}\left(\frac{(\lambda e^t\ )^x}{x!}\right) \\
M(t) &= e^{-\lambda}\sum_{x=0}^{\infty}\frac{(\lambda e^t\ )^x}{x!}
\end{aligned}$$

Introducing $e^{\lambda e^t}$ outside the summation but cancelling by putting $e^{-\lambda et}$ inside the summation

$$M(t) = e^{-\lambda} \cdot e^{\lambda e^t} \sum_{k=0}^{\infty} \frac{e^{-\lambda e^t} \cdot (\lambda e^t)^x}{x!}$$

That adjustment was done so that the terms in the expansion are those of the poisson distribution with mean λ^{e^t}. Then sum is 1 because total probability is 1. This gives.

$$M(t) = e^{-\lambda} e^{\lambda e^t}$$

$$\frac{dM(t)}{dt} = \lambda e^t e^{-\lambda} e^{\lambda e^t} \text{.....(9)}$$

$$\text{and } \frac{d^2M(t)}{dt^2} = \lambda e^t e^{-\lambda} e^{\lambda e^t} + \lambda e^t e^{-\lambda} \cdot \lambda e^t e^{\lambda e^t} \text{.....(10)}$$

Utilising equation (9)

$$\left(\frac{dM(t)}{dt}\right)_{t=0} = E(k) = \lambda$$

and utilising equation (10)

$$\left(\frac{d^2M(t)}{dt^2}\right)_{t=0} E(k^2) = \lambda + \lambda^2$$

so that the mean $= E(K) = \lambda$ and variance $= E(K^2) - [E(K)]^2$

$$= (\lambda + \lambda^2) + (\lambda)^2$$
$$= \lambda$$

In summary, for any discrete distribution with a random variable X,

$$\left[\frac{dM(t)}{dt}\right]_{t=0} = E(x) \text{ and } \left[\frac{d^2M(t)}{dt}\right]_{t=0} = E[x^2]$$

These two expressions enable us to find the mean variance.

8.6.4 Mean and Variance for continous distributions.

If X is a continous random variable with probability density function f(x), the moment generating function is defined by

$$M(t) = \int f(x)e^{xt}dt$$

and E(x) and $E(x^2)$ are defined by

$$E(x) = \left\{\frac{dM(t)}{dt}\right\}_{t=0} \text{ and}$$

$$E(x^2) = \left\{\frac{d^2M(t)}{dt^2}\right\}_{t=0}.$$

Example 8.6.1

Find the moment generating function for

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

and use it to find the mean and are variance of the distribution.
Solution:

$$\begin{aligned} M(t) &= \int_0^2 f(x)e^{xt}dx \\ &= \int_0^2 \frac{1}{2}e^{xt}dx \\ &= \left[\frac{e^{xt}}{2t}\right]_0^2 \\ &= \frac{1}{2t}(e^{2t}-1) \end{aligned}$$

Since we have t in the denominator we shall get a problem when t is equated to 0. Expanding e^{2t} as

$$M(t) = \frac{1}{2t}\left(1+2t+\frac{(2t)^2}{2!}+\frac{(2t)^3}{3!}+\frac{(2t)^4}{4!}+\ldots-1\right)$$

$$
\begin{aligned}
&= \frac{1}{2t}\left(2t + \frac{4t^2}{2} + \frac{8t^3}{6} + \frac{16t^4}{24} + \ldots\right) \\
&= 1 + t + \frac{2}{3}t^2 + \frac{t^3}{3} + \ldots \\
\frac{dM(t)}{dt} &= 1 + \frac{4}{3}t + t^2 + \ldots \\
\frac{d^2M(t)}{dt^2} &= \frac{4}{3} + 2t + \ldots \\
\text{When } t &= 0 \\
\text{Mean } \left(\frac{dM(t)}{dt}\right)_{t=0} &= 1 \\
\left(\frac{d^2M(t)}{dt^2}\right)_{t=0} &= \frac{4}{3} \\
\text{Variance} &= \left(\frac{d^2M(t)}{dt^2}\right)_{t=0} - \left[\left(\frac{dM(t)}{dt}\right)_{t=0}\right]^2 \\
&= \frac{4}{3} - 1^2 \\
&= \frac{4}{3} - 1 \\
&= \frac{1}{3}
\end{aligned}
$$

Example 8.6.2

A random variable X has the probability density given by

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

Find the moment generating function of X and use it to find the mean and variance of the distribution.

Solution:

We need to first find the constant K.

$$
\begin{aligned}
2K\int_0^\infty e^{-2x}dx &= 1 \\
&= 2K\left[\frac{-1}{2}e^{-2x}\right]_0^\infty = 1 \\
&= -k\left[e^{-2x}\right]_0^\infty = 1 \\
K &= 1 \\
\Rightarrow f(x) &= \begin{cases} 2e^{-2x}, & 0 \le x \le \infty \\ 0, & \text{elsewhere} \end{cases} \\
M(t) &= \int f(x)e^{xt}dx \\
&= \int_0^\infty 2e^{-2x}\cdot e^{xt}dx \\
&= 2\int_0^\infty e^{(t-2)x}dx \\
&= 2\times\frac{1}{t-2}\left[e^{(t-2)x}\right]_0^\infty \\
&= \frac{2}{t-2}\left[e^{(t-2)\infty}-e^0\right] \\
&= \frac{2}{t-2}\left[e^{(t-2)\infty}-1\right] \\
M(t) &= \frac{2}{t-2}\left[e^{(t-2)\infty}-1\right] \\
&= \frac{2}{2-t} \\
M(t) &= \frac{2}{t-2}\left[e^{(t-2)\infty}-1\right] \\
&= \frac{2}{2-t}
\end{aligned}
$$

$$
\begin{aligned}
M'(t) &= \frac{0-2(-1)}{(2-t)^2} \\
&= \frac{2}{(2-t)^2} \\
\text{Mean} &= [M'(t)]_{t=0} \\
&= \frac{2}{2^2} = \frac{1}{2} \\
M''(t) &= \frac{0-2[2(2-t)-1]}{(2-t)^4} \\
&= \frac{4(2-t)}{(2-t)^4} = \frac{4}{(2-t)^3} \\
M''(0) &= \frac{4}{2^3} \\
&= \frac{4}{8} = \frac{1}{2} \\
\text{Variance} &= [M''(t) - [M'(t)]]_{t=0} \\
&= \frac{1}{2} - \left(\frac{1}{2}\right)^2 \\
&= \frac{1}{2} - \frac{1}{4} \\
&= \frac{1}{4} \\
\text{Therefore , } M(t) &= \frac{2}{2-t}, \\
\text{mean} &= \frac{1}{2} \text{ and } \text{ variance} = \frac{1}{4}
\end{aligned}
$$

Exercise 8

1. The probability of Ben hitting a target is $\frac{3}{4}$. Assuming that this probability is constant and that the trials are independent, calculate the mean number of shots needed to hit the target.

2. A discrete random variable X has probability given by $P(X) = c|2 - x|$, where c is a constant for $x = 0, 1, 2, ...5$. Calculate the mean and standard deviation of X.

3. The number of times that a child will seek for permission to attend a disco before being allowed is a random variable X with probability distribution.

$$P(X = x) = c\left(\frac{1}{4}\right)^x, x = 0, 1, 2, ...$$

Find (i) the constant c (ii) the probability that he will seek for permission more than twice before he is allowed.

4. A traveller arrives late in town and has to check on various lodges until he finds accomodation. If the probability that he gets accomodation is constant and equal to $\frac{2}{5}$, what is the probability that he has to try

   (i) two lodges

   (ii) more than two lodges.

   (iii) find the expected value for the number of lodges tried and the most likely number of lodges tried.

5. Taxis arrive at stage A every twenty minutes. If Ben walks to the stage without taking note of the time to find out if he is on time for the taxi, find the mean and standard deviation of the time for which he has to wait for the taxi.

6. The random variable X has the probability density function $f(x) = 3e^{-3x}$ for $0 < x < \infty$. Find

   (i) $P(X > 30)$

   (ii) $F(x)$

7. A continous random variable has pdf f where

$$f(x) = \begin{cases} \frac{1}{4}x^3, & 0 \leq x \leq 2 \\ 0, & \text{elsewhere} \end{cases}$$

Find the

   (a) Mean

   (b) Variance

(c) Median

(d) Cummulative distribution function

(e) $P(1 \leq x \leq 2)$.

8. A continous random variable X is distributed between the values 4 and 8 and has a pdf of $\frac{8}{x^2}$. Find the mean, variance and median of x.

9. A random variable X has a probability density function given by

$$f(x) = \begin{cases} 2e^{-2x}, & 0 < x < \infty \\ 0, & \text{elsewhere} \end{cases}$$

Find the moment generating function of x and use it to find the mean and variance of the distribution.

10. A random variable X has a probaility density function

$$f(x) = \begin{cases} 4e^{-4x}, & 0 < x < \infty \\ 0, & \text{elsewhere} \end{cases}$$

Find the moment generating function of X and use it to find the mean and variance of the distribution.

11. A continuous random variable X has pdf given by

$$f(x) = \begin{cases} k(x^2 - x^3), & 0 < x < 2 \\ 0, & \text{elsewhere} \end{cases}$$

Find the constant k, mean and standard deviation of the distribution.

Chapter 9

ESTIMATION

9.1 Introduction

It is not possible in most cases to handle whole populations while analysing some statistics about those populations. Statisticians are left with the option of choosing a sample from the population which sample should represent all the characteristics of the population. The sample mean $\bar{x}$ and standard deviation S are made to stand for the population mean μ and standard deviation δ.
If samples are taken from the same population, the means of those samples $\bar{x}_1, \bar{x}_2 \ldots \bar{x}_n$ form a distribution which we call the sampling distribution of the mean. Each of the values $\bar{x}$ is an estimation of the mean.

9.2 Unbiased Estimate of the mean

A good estimator should be unbiased. This implies that the expectation of the sample mean $\bar{x}$ should be equal to the population mean μ. If a sample has size n, the sample mean $\bar{x}$ is calculated from the n observations of x. If the observations are $x_1, x_2, x_3, \ldots x_n$, then

$$\begin{aligned}
\bar{x} &= \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} \\
E(\bar{x}) &= E\left[\frac{x_1 + x_2 + \ldots + x_n}{n}\right] \\
&= E\left(\frac{x_1}{n}\right) + E\left(\frac{x_2}{n}\right) + \ldots + \frac{E(x_n)}{n} \\
&= \frac{\mu}{n} + \frac{\mu}{n} + \frac{\mu}{n} + \ldots + \frac{\mu}{n} \\
&= \mu
\end{aligned}$$

9.3 Unbiased Estimate of the variance

If all possible samples of size n are drawn from a population with replacement, and the population has mean μ and standard deviation δ, the means of the samples have a sampling distribution with mean μ and standard deviation $\frac{\delta}{\sqrt{n}}$ often called the standard error of the mean.
This value of $\frac{\delta}{\sqrt{n}}$ is derived in the following way:

Let the variance of the sampling distribution be $Var(X)$. Then

$$\begin{aligned} Var(X) &= Var\left\{\sum_1^n \frac{x_i}{n}\right\} \\ &= \frac{1}{n^2} Var\left\{\sum_1^n x_i\right\} \\ &= \frac{1}{n^2}\left\{Var(X_1) + Var(X_2) + Var(X_3) + \ldots + Var(X_n)\right\} \end{aligned}$$

but variance of $x_i = \delta^2$ so that

$$\begin{aligned} Var(X) &= \frac{1}{n^2} \cdot n\delta^2 \\ &= \frac{\delta^2}{n} \end{aligned}$$

So that standard deviation $= \sqrt{Var(x)} = \frac{\delta}{\sqrt{n}}$.
When the population mean μ is known or when we have $\bar{x}$ as the estimate of μ, we solve for the standard deviation of the sample using the formula

$$S^2 = \sum_1^n \frac{(x_i - x)^2}{n}$$

This formular gives a biased estimate of δ^2 due to the fact the sum of the squares of the deviation of the $x_i's$ from $\bar{x}$ is less than the sum of the squares of the deviations from μ. So once μ is known, we find an unbiased estimator of the variance thus:

$$\delta^2 = E\left\{\sum_1^n \frac{(x_i - \mu)^2}{n}\right\}$$

$$
\begin{aligned}
&= E\left\{\sum_1^n \frac{[(x_i-\bar{x})-(\mu-\bar{x})]^2}{n}\right\} \\
&\quad \sum\left\{\frac{\sum_1^n (x_i-\bar{x})^2 - 2(x_i-\bar{x})(\mu-\bar{x})+(\mu-\bar{x})^2}{n}\right\} \\
&= E\left\{\sum_1^n (x_i-\bar{x})^2 - 2(\mu-\bar{x})\sum_1^n (x_i-\bar{x}) + n\frac{(\mu-\bar{x})^2}{n}\right\} \\
&= E\left\{\sum_1^n \frac{(x_i-\bar{x})}{n^2}\right\} + E\left\{(\mu-\bar{x}^2)\right\}
\end{aligned}
$$

The second term is zero because $n\bar{x} = n\bar{x}$ so that $\sum_1^n \frac{(x_i-\bar{x})}{n} = 0$.

$$
\begin{aligned}
\delta^2 &= E(S^2) + E\left\{(\mu-\bar{x})^2\right\} \\
\text{since } S^2 &= \sum_1^n \frac{(x_i-\bar{x})^2}{n} \\
\text{But } E\left\{(\mu-\bar{x})\right\} &= var(\bar{x}) = \frac{\delta^2}{n} \\
&= \frac{\delta^2}{n} \\
\text{So } \delta^2 &= E(S^2) + \frac{\delta^2}{n} \\
n\delta^2 &= nE(S^2) + \delta^2 \\
n\delta^2 - \delta^2 &= nE(S^2) \\
(n-1)\delta^2 &= nE(S^2) \\
\delta^2 &= \frac{n}{n-1}E(S^2) \\
\delta^2 &= E\left[\frac{n}{n-1}S^2\right] \\
&= E\left\{\left(\frac{n}{n-1}\right)\left(\sum_1^n \frac{(x_i-\bar{x})^2}{n}\right)\right\}
\end{aligned}
$$

$$\delta^2 \;=\; E\left\{\sum_1^n \frac{(x_i-\bar{x})^2}{n-1}\right\} \simeq \hat{S}^2$$

which is an an unbiased estimator of variance. Compare this value $\hat{S}^2$ with S^2 which we stated earlier, they are related by the expression

$$\hat{S} = \sqrt{\left(\frac{n}{n-1}\right) S}$$

$n-1$ in the denominator is the number of degrees of freedom, symbolised as v. This is because if the deviations are measured from $\bar{x}$, there is one term $\sum_1^n (x_i - \bar{x}) = 0$ from $n-$observations so that we remain with $(n-1)$ terms which we can independently vary. For large n, the error is so small and negligible when we use S^2 instead of $\hat{S}^2$ to estimate the variance.

Example 9.3.1

The ages of 11 boys in the school team were recorded in years as
$16, 17.8, 19.1, 18.0, 18.6, 16.5, 19.0, 17.2, 18.8, 19.5, 20$. Determine the estimates for the mean and standard deviation of the ages of all the boys in the school if the school team can be used as a sample.

Solution:

We calculate for $\bar{x}$ and S thus

X_i	f	d	fd	fd^2
1.6	1	-2.0	-2.0	4.00
17.2	1	-0.8	-0.8	0.64
17.8	1	-0.2	-0.2	0.64
16.5	1	-1.5	-1.5	2.25
18.0	1	0	0	0
18.6	1	0.6	0.6	0.36
18.8	1	0.8	0.8	0.64
19.0	1	1.0	1.0	1
19.1	1	1.1	1.1	1.21
19.5	1	1.5	1.5	2.25
20	1	2	2	4.00
	11		2.5	16.39

Using an assumed mean $A = 18.0$

$$\begin{aligned}
\bar{x} &= A + \frac{\sum fd}{\sum f} \\
&= 18.0 + \frac{2.5}{11} \\
&= 18.2272727272727273 \\
&\simeq 18.2 \\
\text{and } S^2 &= \frac{\sum fd^2}{\sum f} - \left(\frac{\sum fd}{\sum f}\right)^2 \\
&= \frac{16.39}{11} - \left(\frac{2.5}{11}\right)^2 \\
&= \frac{180.29 - 6.25}{121} \\
&= \frac{174.04}{11} \\
&= 15.8218181818
\end{aligned}$$

The unbiased estimate for μ is $\bar{x} = 18.2$ years
The unbiased estimate for δ^2 is

$$\begin{aligned}
\hat{S}^2 &= \frac{n}{n-1} S^2 \\
&= \frac{11}{10} \times 15.82181818\ldots \\
&= 17.404 \\
S &= 4.171810159 \\
&\simeq 4.172 \text{ years.}
\end{aligned}$$

Example 9.3.2

A sample of eight measurements of heights of men aged thirty years, in metres, was as follows
$1.5, 1.7, 1.6, 1.7, 1.8, 1.5, 1.6$. Determine the estimates for the mean and standard deviation for the men aged thirty years.
Solution:

x	f	d	fd	d^2	fd^2
1.5	2	-0.1	-0.2	0.01	0.02
1.6	2	0	0	0	0
1.7	3	0.1	0.3	0.01	0.03
1.8	1	0.2	0.2	0.04	0.04
	8		0.3		0.09

Using an assumed mean $A = 1.6$

$$\begin{aligned} \bar{x} &= 1.6 + \frac{0.3}{8} \\ &= 1.6375 \end{aligned}$$

The unbiased estimate of $\mu = \bar{x} = 1.6375$ and

$$\begin{aligned} S^2 &= \frac{0.09}{8} - \left(\frac{0.3}{8}\right)^2 \\ &= 0.01125 - 0.00140625 \\ &= 0.00984375 \end{aligned}$$

The unbiased estimate of δ^2 is

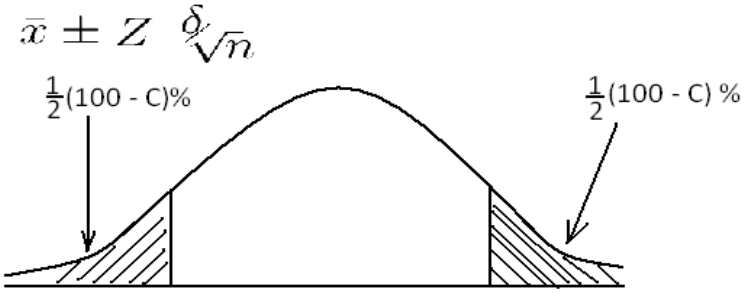
$$\begin{aligned} \hat{S}^2 &= \frac{n}{n-1} S^2 = \frac{8}{7} \times 0.00984375 \\ &= 0.01125. \end{aligned}$$

9.4 The Central Limit Theorem

As the sample size n is increased without limit, the shape of the distribution of the sample mean taken with replacement from a population (normal or not) with mean μ and standard deviation δ will approach a normal distribution. The distribution will have a mean μ and standard deviation $\delta/\sqrt{n}$.

9.5 Confidence intervals

Using data from a random sample, we find an interval within which the estimate of the parameter should lie with a certain degree of confidence (probability). The interval is called a confidence interval while its upper and lower values are called confidence limits. If $\bar{x}$ is the mean of a random sample of size n taken from a normal population with mean μ and variance δ, where δ is known, then the symmetric C% confidence interval for μ is given by



There are two cases to consider under confidence intervals. These are when δ is known or δ is not known.

Case 1 (δ known)

Example 9.5.1

It is known that the standard deviation of the ages in years of students at a polytechnic is 3 years. A sample of 49 selected students revealed a mean age of 22 years. Find the 95% confidence interval for the population mean.

Solution:

$$\begin{aligned} 100(1-2\alpha)\% &= 95 \\ 1-2\alpha &= 0.95 \\ 2\alpha &= 0.05 \\ \alpha &= 0.025 \\ Z_\alpha &= Z_{0.025} = 1.96 \end{aligned}$$

Hence the 95% confidence interval is given by

$$\begin{aligned} \bar{x} \pm \frac{Z\delta}{\sqrt{n}} &= 22 \pm \frac{1.96 \times 3}{\sqrt{49}} \\ &= 22 \pm \frac{5.88}{7} \\ &= 22 \pm 0.84 \\ &= 21.16, \quad 22.84 \\ &\Rightarrow \quad 21.16 \leq \mu \leq 22.84 \end{aligned}$$

We can say with 95% confidence that the average age of students at the polytechnic lies between 21.16 years and 22.84 years.

Example 9.5.2

A survey of 30 families revealed that the mean age of the children in the family is 5 years with a standard deviation of 0.5 years. Find the 99% confidence interval for the mean age of children in a family.

Solution:

$$\begin{aligned} 100(1-2\alpha)\% &= 99 \\ 1-2\alpha &= 0.99 \\ 2\alpha &= 0.01 \\ \alpha &= 0.005 \\ Z_\alpha &= Z_{0.005} = 2.58 \end{aligned}$$

The interval is given by

$$\begin{aligned} \bar{x} \pm Z_{\frac{\delta}{\sqrt{n}}} &= 5 \pm \frac{2.58 \times 0.5}{\sqrt{30}} \\ &= 5 \pm 0.235520699 \\ &\simeq 5 \pm 0.2355 \\ &\simeq 4.7645, \quad 5.2355 \end{aligned}$$

The interval is $4.7645 \leq \mu \leq 5.2355 \simeq 4.765 \leq \mu \leq 5.236$.

Example 9.5.3

The following data represents a sample of the assets in millions of shillings of 32 families in Kasese town. Find the 90% confidence interval for the mean value of assets.
12, 3, 13, 73, 11, 9, 8, 40, 5, 2, 17, 4, 9, 2, 7, 3,
5, 3, 1, 13, 3, 15, 22, 17, 18, 1, 15, 3, 19, 20, 50, 31.

Solution:

From the data $\bar{x} = 14.1875$ and $S = 15.60073923$
Since $n \geq 30$, we shall still use the z tables

$$\begin{aligned} 100(1-2\alpha)\% &= 90 \\ 1-2\alpha &= 0.9 \\ 2\alpha &= 0.1 \\ \alpha &= 0.05 \\ Z_\alpha &= Z_{0.05} = 1.645 \end{aligned}$$

The interval is given by

$$\bar{x} \pm Z_\alpha \frac{\delta}{\sqrt{n}} = \bar{x} \pm Z_\alpha \frac{S}{\sqrt{n}}$$

$$
\begin{aligned}
&= 14.1875 \pm \frac{1.645 \times 15.60073923}{\sqrt{32}} \\
&= 14.1875 \pm 4.536658521 \\
&= [9.65, 18.72]
\end{aligned}
$$

Example 9.5.4

A sample of the reading scores of 36 fourth graders has a mean of 80 and standard deviation 16.

(i) Find the 95% confidence interval of the mean reading scores of all fourth graders.

(ii) Find a 99% confidence interval of the mean reading scores of all fourth graders.

Solution:

$$\text{Here } n = 36,\ \mu = 80 \text{ and } \delta = 16.$$

(i) For 95% confidence interval

$$
\begin{aligned}
100(1-2\alpha)\% &= 95 \\
1-2\alpha &= 0.95 \\
2\alpha &= 0.05 \\
\alpha &= 0.025 \\
Z_\alpha &= Z_{0.025} = 1.96
\end{aligned}
$$

The interval is given by

$$
\begin{aligned}
\bar{x} \pm Z\frac{\delta}{\sqrt{n}} &= 80 \pm \frac{16 \times 1.96}{6} \\
&= 74.773,\ 85.227 \\
&\Rightarrow 74.77 \leq \mu \leq 85.23
\end{aligned}
$$

(ii) For 99% confidence interval

$$
\begin{aligned}
100(1-2\alpha)\% &= 99 \\
1-2\alpha &= 0.99 \\
2\alpha &= 0.01 \\
\alpha &= 0.005 \\
Z_\alpha &= Z_{0.005} = 2.58
\end{aligned}
$$

The interval is given by

$$\begin{aligned} \bar{x} \pm Z\frac{\delta}{\sqrt{n}} &= 80 \pm 2.58 \times \frac{16}{6} \\ & \\ & (73.12, 86.88) \\ &\Rightarrow \quad 73.12 \leq \mu \leq 86.88. \end{aligned}$$

Case 2:(δ Unknown).
When δ is known and the variable is normally distributed or when δ is unknown but $n \geq 30$, the standard normal distribution is used to find confidence intervals for the mean. But in real life the population standard deviation may not be known when the sample $n < 30$. In such a case, the sample standard deviation is used to find the confidence intervals and we use a different distribution, the student's $t-$ distribution. This distribution was discovered by W.S Gosset in 1908 while he was employed by a brewing firm. He published it under the pseudonym student; hence it is called the student's $t-$ distribution.
Like the normal distribution, the students t-distribution is bell shaped, symmetric about the mean and its mean, median and mode are equal to zero located at the centre of the distribution. It never touches the x-axis.
The difference between it and the standard normal distribution is that the variance is greater than 1, it is a family of curves based on the concept of degrees of freedom which is related to sample size and as the sample size n is increased, the t-distribution approaches the standard normal distribution.
The formula for finding a confidence interval is

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

Example 9.5.5

The masses in grams of twelve ball bearings taken from a batch at a manufacturing plant were: $20, 23, 19, 21, 24, 25, 27, 22, 25, 23, 21, 18$. Calculate a 95% confidence interval for the mean mass of the population, assumed to be normal.
Solution:
From calculator, $\bar{x} = 22.33 \quad S = 2.674$
For 95% confidence interval

$$\begin{aligned} 100(1-2\alpha) &= 95\% \\ 1-2\alpha &= 0.95 \\ 2\alpha &= 0.05 \\ \alpha &= 0.025 \\ t_\alpha &= t_{0.025} = 2.201 \quad (v = n-1 = 11) \end{aligned}$$

The interval is given by

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

$$\begin{aligned} &= 22.33 \pm \frac{2.201 \times 2.6742}{\sqrt{11}} \\ &= 22.33 \pm 1.77467 \\ &= [20.55533,\ 24.10467] \\ &\simeq 20.6 \leq \mu \leq 24.1 \end{aligned}$$

Example 9.5.6

The heights of 15 bean stalks are found to have a mean of 10cm and a standard deviation of 0.8cm. Find 99% confidence limits for the population mean.

Solution:

$$n = 15, \bar{x} = 10,\ S = 0.8, V = n - 1 = 14$$

For 99% confidence limits,

$$\begin{aligned} 100(1-2\alpha) &= 99 \\ 1 - 2\alpha &= 0.99 \\ 2\alpha &= 0.01 \\ \alpha &= 0.005 \\ t_\alpha &= t_{0.005} = 2.977 \end{aligned}$$

The limits are given by

$$\begin{aligned} \bar{x} \pm t_\alpha \frac{S}{\sqrt{n-1}} &= = 10 \pm \frac{2.977 \times 0.8}{\sqrt{14}} \\ &= 10 \pm 0.6365 \\ &\quad [9.365, 10.6365] \\ &\quad [9.4, 10.6]. \end{aligned}$$

Example 9.5.7

A random sample of egges taken from a days production of Muhindo's farm had the following masses in grams $52, 58, 53, 54, 51, 55, 57, 56, 52, 54$. Assuming the weights are normally distributed, find 98% confidence limits for the mean mass of the eggs provided that day.

Solution:

$$n = 10 \ \Rightarrow V = n - 1 = 9.$$

From calculator, $\bar{x} = 54.2$ and $S = 2.299758441$
For 98% confidence limits

$$\begin{aligned} 100(1-2\alpha) &= 98 \\ 1-2\alpha &= 0.98 \\ 2\alpha &= 0.98 \\ \alpha &= 0.01 \\ t_\alpha &= t_{0.01} = 2.821 \end{aligned}$$

The limits are given by

$$\begin{aligned} \bar{x} \pm t_\alpha \frac{S}{\sqrt{n-1}} &= \frac{54.2 \pm 2.821 \times 2.299758}{\sqrt{9}} \\ &= 54.2 \pm 2.162539 \\ &= 52.037461, 56.362539 \\ &= [53.04, 56.36] \end{aligned}$$

Confidence interval for a proportion

The probability that a certain member of the population has an attribute is p and if x is the number of members in the population who have this attribute in a sample of size n then $P = \frac{x}{n}$ so that

$$\begin{aligned} E(P) &= E(\frac{x}{n}) \\ &= \frac{1}{n}E(x) \end{aligned}$$

But one has the attributes or not. So it is a binomial situation. But for a binomial distribution B(n,p), E(x)=np and variance is npq. This shows that

$$\begin{aligned} E(P) &= \frac{1}{n} \cdot np = p \text{ and} \\ Var(P) &= Var(\frac{x}{n}) \\ &= \frac{1}{n^2}Var(x) \end{aligned}$$

but the variance of binomial distribution is npq where $q = 1 - p$

$$\begin{aligned} \Rightarrow Var(P) &= \frac{1}{n^2} \cdot np(1-p) \\ &= \frac{1}{n} p(1-p). \end{aligned}$$

When n is large the sampling distribution of x tends to a normal distribution. The sampling distribution of the proportion p tends to a normal distribution with mean p and standard deviation $\sqrt{\frac{p}{n}(1-p)}$ which is the standard error of the proportion We use the standard normal distribution Z to find a confidence interval for the proportion.

Example 9.5.8

An opinion poll taken from the electorate indicated that 40 out of 100 would vote for candidate A. What is the 98% confidence interval for the proportion of the population who will vote for candidate A. If there are only two candidates, what advice would you give candidate A?

Solution:

$$n = 100, \; P = \frac{40}{400} = 0.4$$

For 98% confidence interval

$$\begin{aligned} 100(1-2\alpha) &= 98 \\ 1 - 2\alpha &= 0.98 \\ 2\alpha &= 0.02 \\ \alpha &= 0.01 \\ Z_\alpha &= Z_{0.01} = 2.33 \end{aligned}$$

The confidence limits are given by

$$\begin{aligned} P \pm Z_\alpha \sqrt{\frac{p(1-p)}{n}} &= 0.4 \pm 2.33 \times \sqrt{\frac{0.4(0.6)}{100}} \\ &= 0.4 \pm 0.114 \\ &= [0.286, 0.514] \\ &\simeq [0.29, 0.51] \\ &= 0.29 < P < 0.51 \end{aligned}$$

Advice: A should drop out of the race because he cannot win or else change the compaign strategy.

Example 9.5.9

In a random sample of 120 airport workers, 48 have been vaccinated against swine flu. Calculated 95% confidence limits for the proportion of workers that have been vaccinated against swine flu.

Solution:

$n = 120, p = \frac{48}{120} = 0.4$. The Z-value is 1.96.
The confidence interval is given by

$$0.4 \pm 1.96 \times \sqrt{\left(\frac{0.4 \times 0.6}{120}\right)}$$

$$\begin{aligned} &= 0.4 \pm 0.08765 \\ &= (0.31235, 0.48765) \\ &\simeq (0.31, 0.49) \end{aligned}$$

Example 9.5.10

A sample of 400 nursing students included 100 men

(a) Find the 95% confidence interval for the true proportion of men who are studying nursing

(b) How large should the sample have been to reduce the confidence interval to 2%?

Solution:

(a) $n = 400, P = \frac{100}{400} = 0.25$
For 95% confidence interval, $Z = 1.96$
The interval is given by

$$\begin{aligned} P \pm Z_\alpha \sqrt{\frac{p(1-p)}{n}} &= 0.25 \pm 1.96 \sqrt{\frac{0.25 \times 0.75}{400}} \\ &= 0.25 \pm 0.0424 \\ &= [0.2076, 0.2924] \\ &\Rightarrow 0.2076 < P < 0.2924 \end{aligned}$$

(b) The confidence limits required are 0.25 ± 0.01 giving

$$\begin{aligned} 1.96 \sqrt{\frac{0.25 \times 0.75}{n}} &= 0.01 \\ 0.7203 &= 0.0001n \\ n &= 7203 \end{aligned}$$

The sample should have been 7203 nursing students.

Exercise 9

Chapter 10

SIGNIFICANCE TESTING

10.1 SETTING UP A HYPOTHESIS

In almost all situations, decisions are taken basing on the prevailing situation at the time the decision is going to be made. Hypotheses are assumptions about a parameter of the population. We test hypotheses to assess their correctness.
A cook claims she can tell whether onions or tomatoes were put in the frying pan first. She performs a series of trials to test her claim. Sauce is presented to her in two different bowls. The two bowls are randomly presented and she has to identify the bowl which has sauce where the tomatoes were put first. If she is correct four times out of five, is it right to accept her claim?
This is a binomial situation where correct identification can be said to be success and incorrect identification can be said to be "failure". In this case $n = 5, p = \frac{1}{2}, q = \frac{1}{2}$.
Hypotheses always come in pairs i.e the null hypothesis and the alternative hypothesis. In our estimation we have a binomial model. So the null hypothesis is $H_0 : p = \frac{1}{2}$ and the alternative hypothesis is $H_1 : p > \frac{1}{2}$. The hypothesis that is being tested is the null hypothesis.

Example 10.1.1

A coin is tossed fifteen times and heads showed up four times. is there evidence at 1% level of significance that the coin is biased?
Solution:
Assume that the coin is not biased. This is a binomial model.

$$H_0 : P = \frac{1}{2}$$

$$H_1 : P \neq \frac{1}{2}$$

Should the null hypothesis be true, then the number of heads X is binomially distributed as $X \sim B(15, \frac{1}{2})$. A high number or a low number of heads will lead to rejection of the

null hypothesis.
We thus solve for

$$\begin{aligned} & & P(X \leq 4 \text{ or } X \geq 11) \\ P(X \leq 4) &= 0.0592 \text{ (from tables)} \\ P(X \geq 11) &= 0.0592 \\ \Rightarrow P(X \geq 11) + P(X \leq 4) &= 0.0592 + 0.0592 \\ &= 0.1184 = 11.84\% \end{aligned}$$

Since $11.84\% > 1\%$, the result is not significant at 1% level. We retain the null hypothesis that there is no evidence at 1% level of significance that the coin is biased.

Type I and Type II errors:

When a null hypothesis is tested by application of a significance test on a sample, we may not be sure that we have made a correct conclusion. Therefore a type I error is made if a null hypothesis is retained when infact it should be rejected. The probability of making a type I error is equal to the signicance level of the test. Type II error is calculated if a particular value of the parameter is specified for the alternative hypothesis.

Tests on the mean of a normal distribution (δ known)

Example 10.1.2

A carpenter is told to produce wooden rods whose lengths are normally distributed with $\mu = 100cm$ and $\delta = 4cm$. To check on his accuracy a sample of 10 rods is taken and the mean length was found to be $101cm$. Is the carpenter accurate at 5% level of significance?

Solution:

The model is normal

$$\begin{aligned} H_0 : \mu &= 100cm,\ \delta = 4 \\ H_1 : \mu &\neq 100cm \end{aligned}$$

The sampling distribution for the sample mean $\bar{x}$ of a sample of 10 rods is normally distributed with $\mu = 100cm$, standard deviation $= \frac{\delta}{\sqrt{n}} = \frac{4}{\sqrt{10}} = 1.2649$
For the observed sample mean 101 we have

$$Z = \frac{(101 - 100)}{1.2649} = 0.790569415 \simeq 0.79$$

This is a two picked tailed. Therefore.

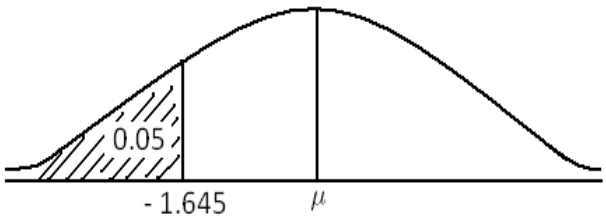
$$\begin{aligned} P(Z < -0.79) + P(Z > 0.79) &= 2P(Z > 0.79) \\ &= 2 \times\ 0.2148 \\ &= 0.4296 = 42.96\% \end{aligned}$$

Since 42.96% $>$ 5% there is no evidence that the carpenter is innacurate. We retain H_0.
NB: We could have done this using the critical region instead of calculating the probability as we shall do in the next example.

Example 10.1.3

Grain millers claim that the average weight of a bag of maize flour is 80kg. If a random sample of 100 bags had a mean of 79kg and standard deviation of 4kg test whether the average weight of the bags is less than 80kg at 5% level of significance.
Solution:
The model is normal

$$\begin{aligned} H_0 : \mu &= 80kg, \delta = 4 \\ H_1 : \mu &< 80kg \end{aligned}$$



The Z-value which leaves an are of 5% to the left is -1.645
Using the sample

$$Z = \frac{79 - 80}{4/\sqrt{100}} = -2.5$$

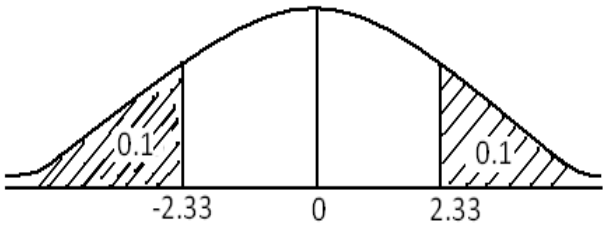
-2.5 is in the rejection region so we reject the null hypothesis, i.e. we accept the alternative hypothesis that the average weight of bags is less than 80kg.

Example 10.1.4

The average mark in an examination at kilembe secondary school is 58% with a standard deviation of 2%. Is there reason to believe that there has been a change in performance if a random sample of 40 students has an average of 60%? Test this clan at 2% level of signinance.
Solution:

$$\begin{aligned} H_0 : \mu &= 58 \\ H_1 : \mu &\neq 58 \end{aligned}$$

This is a two tailed test i.e. there is an area of 1% at either tail.



The Z value which haves an area of 1% at a either tail is 2.33.
Evaluating the sample

$$Z = \frac{(\bar{x} - \mu)}{\delta\sqrt{n)}} \Rightarrow Z = \frac{60 - 58}{2\sqrt{40}} = 6.3245$$

Since 6.3245 is greater than 2.33, we reject H_0 and conclude that there is a significant change in performance of students at 2% level of significance.

Tests for large samples (δ unknown)

Should δ be unknown when the sample is large, the test statistic Z is used with δ replaced by S so that $Z = \frac{\bar{x}-\mu}{s/\sqrt{n-1}}$

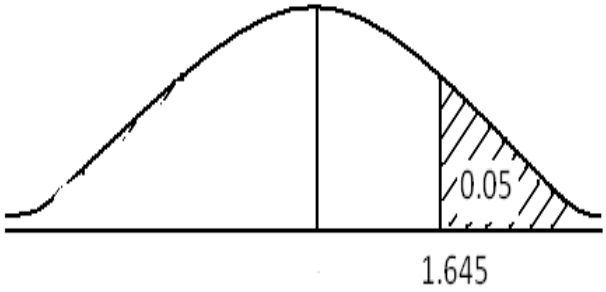
Example 10.1.5

Eighty measurements of a variate gave $\sum_{i=1}^{80} x_i = 200$ and $\sum_{i=1}^{80} x_i^2 = 625$. Test at 5% level of significance if these measurements are from a population with mean greater than 1.8.
Solution:

$$\begin{aligned}
\text{Sample mean} &= \frac{\sum x_i}{n} \\
&= \frac{200}{80} = 2.5 \\
\text{Sample s.d} &= \left\{ \frac{\sum x^2}{n} - \left(\frac{\sum x}{n} \right)^2 \right\}^{\frac{1}{2}} \\
&= \left\{ \frac{625}{80} - \left(\frac{200}{80} \right)^2) \right\}^{\frac{1}{2}}
\end{aligned}$$

$$
\begin{aligned}
&= 1.25 \\
H_0 : \mu &= 1.8 \\
H_1 : \mu &> 1.8
\end{aligned}
$$

$$
\begin{aligned}
Z &= \frac{\bar{x} - \mu}{S/\sqrt{n-1}} \\
&= \frac{2.5 - 1.8}{1.25/\sqrt{79}} \\
&= 4.98
\end{aligned}
$$



The $Z-$value which leaves an area of 0.05 to the right is 1.645.
Since $4.98 > 1.645$, we reject H_0 and conclude that the population mean is greater than 1.8 at 5% level of significance.

Example 10.1.6

The sample of marks in a test were distributed as below

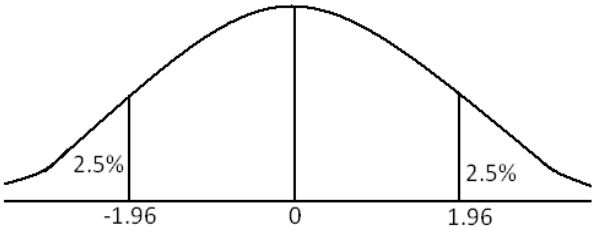
Marks	$20-24$	$25-29$	$30-34$	$35-39$	$40-44$
Frequency	8	10	18	7	12

Find the mean and variance of the marks.
The mean mark for the whole district is 30. Test whether the mean mark for the school differs significantly from the district mean at 5% level of significance.
Solution:

Classes	f	x	fx	x^2	fx^2
$20-24$	8	22	176	484	3872
$25-29$	10	27	270	729	7290
$30-34$	18	32	576	1024	18432
$35-39$	7	37	259	1369	9583
$40-44$	12	42	504	1764	21168
	55		1785		60345

$$\begin{aligned}
\text{Mean} &= \frac{\sum fx}{\sum f} = \frac{1785}{55} \\
&= 32.45 \\
\text{Standard Deviation} &= \left[\frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2\right]^{\frac{1}{2}} \\
&= \left[\frac{60345}{55} - \left(\frac{1785}{55}\right)^2\right]^{\frac{1}{2}} \\
&= \sqrt{43.88429753} \simeq 6.6245 \\
&\simeq 6.62 \\
H_0 : \mu &= 30 \text{ and } H_1 \neq 30
\end{aligned}$$

This is a two-tailed test



The $Z-$value which leaves an area of 0.025 at either tail is 1.96.

$$\text{But } Z = \frac{\bar{x} - \mu}{S/\sqrt{n-1}} = \frac{32.45 - 30}{6.62\sqrt{24}} = 0.0504$$

Since $0.0504 < 1.96$, we accept H_0 that the performance of the school does not differ from that of the distric in general.

10.2 Tests for small samples (δ unknown)

For small samples from normal populations, the test statistic is $t = \frac{\bar{x}-\mu}{S/\sqrt{n-1}}$

Example 10.2.1

A teacher claims that his students have an average I.Q of 102. To check his claim a sample of 10 students were found to have the following I.QS. 98, 95, 106, 120, 110, 105, 96, 108, 115, 90. Does this evidence support his claim? test it at 5% level of significance.

Solution:

Assume that the I.Q s are normally distributed. So

$$\begin{aligned} H_0 : \mu &= 102 \\ H_1 : \mu &\neq 102 \ . \end{aligned}$$

this is a two tailed test.
From the t-tables, the critical value $t_9 = 2.262$
From the sample $\bar{x} = 104.3$ and $s.d = 9.000555538 \simeq 9$

$$\begin{aligned} \text{So } t &= \frac{\bar{x} - \mu}{s\sqrt{n-1}} \\ &= \frac{104.3 - 102}{9/\sqrt{9}} \\ &= 0.7666666 \ldots \simeq 0.767 \end{aligned}$$

since $0.767 < 2.262$, we accept H_0, that the average I.Q of his students is 102.

Example 10.2.2

A shopkeeper claims that the average weight of a bar of soap is 700g. A sample of 8 bars gave the following weights.
$698, 701, 699, 704, 708, 695, 697, 702$. Is there evidence at 5% level of significance that the bars of soap are under weight?

Solution:

Assume that the weights are normally distributed

$$\begin{aligned} H_0 : \mu &= 700 \\ H_1 : \mu &< 700 \end{aligned}$$

This is a one tailed text.
From tables, $t = 1.895$
From the sample $\bar{x} = 700.5$ and $sd = 3.905$

$$\begin{aligned} t &= \frac{\bar{x} - \mu}{S/\sqrt{n-1}} \\ &= \frac{700.5 - 700}{3.905/\sqrt{7}} \\ &= 0.338753742 \simeq 0.339 \end{aligned}$$

Since $0.339 < 1.895$ we accept H_0 that the bars are not underweight.
Alternative answer

$$\begin{aligned} S.d &= 4.174754056 \\ t &= \frac{700.5 - 700}{4.174754056/\sqrt{7}} \\ &= 0.31687511 \\ &\simeq 0.317 \end{aligned}$$

10.2.1 Tests for the Difference between two means for large samples.

If large samples are taken from a population, the distribution of $\bar{x}$, the sample mean will be normal.
If two samples are taken from a large population and we are required to find if the means are significantly different, then the test statistic will be

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left\{\frac{\delta_1^2}{n_1} + \frac{\delta_2^2}{n_2}\right\}}}$$

If δ_1 and δ_2 are unknown and the samples are large then

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left\{\frac{\S_1^2}{n_1 - 1} + \frac{S_2^2}{n_2 - 1}\right\}}}$$

Example 10.2.3

An examination is taken by 400 men and 150 women.
The mean mark for men is 60 with a standard deviation of 3 while for women, the mean mark is 64 with a standard deviation of 5.
Test at 5% level of significance if the difference between the means is significant.
Solution:
Assume a normal distribution

$$\begin{aligned} H_0 : \mu_1 &= \mu_2 \Rightarrow \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 &\neq \mu_2 \end{aligned}$$

The test statistic is

$$\begin{aligned} Z &= \frac{\bar{x}_1 - \bar{x}_2}{\left(\frac{\delta_1^2}{n_1} + \frac{\delta_2^2}{n_2}\right)^{\frac{1}{2}}} \\ &= \frac{60 - 64}{\left(\frac{3^2}{400} + \frac{5^2}{150}\right)^{\frac{1}{2}}} \\ &= \frac{-4}{\sqrt{\left(\frac{227}{1200}\right)}} \\ &= -9.19682 \end{aligned}$$

This is a two tailed test. The $Z-$value which leaves an area of 0.025 at either tail is 1.96.
Since $9.19682 > 1.96$ we reject H_0. There is a significant difference between the means of men and women.

10.2.2 Testing if two samples are from the same population

If the two samples are from the same population $\bar{x}_1 - \bar{x}_2$ should not differ significantly from zero.
Now our null hypothesis is that $\delta_1 = \delta_2$ and the standard deviation of $\bar{x}_1 - \bar{x}_2$ then is

$$\delta \sqrt{\left\{\frac{1}{n_1} + \frac{1}{n_2}\right\}}.$$

If δ is known we use the test statistic

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\delta \left\{\frac{1}{n_1} + \frac{1}{n_2}\right\}^{\frac{1}{2}}}$$

However, if δ is unknown and S_1 and S_2 are the standard deviations of the two samples, the unbiased estimate of the standard deviation of the population is

$$\left\{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}\right\}^{\frac{1}{2}}$$

where $n_1 + n_2 - 2$ is the number of degrees of freedom.
In this case, the test statistic is

$$t_{n_1+n_2-2} = \frac{\bar{x_1} - \bar{x}_2}{\left\{\left(\frac{n_1 S_1^2 + n_2 S_2^2}{n_1+n_2-2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right\}^{\frac{1}{2}}}$$

If the populations are normal and/or the samples are large.

Example 10.2.4

A test was administered to two groups of students picked from the senior three class. The first group had 20 students who had an average of 60% with a standard deviation of 3% while the second group had 15 students had an average of 57% with a standard deviation of 2%. Test at 5% level of signicance if the two groups of students were of the same ability.

Solution:

This is a normal distribution model

$$H_0 : \mu_1 = \mu_2,\ \delta_1 = \delta_2 = \delta$$

H_1 : Samples from different ability students

$$\begin{aligned} \bar{x}_1 &= 60 \quad \bar{x}_2 = 57 \\ S_1 &= 3 \quad S_2 = 2 \\ n_1 &= 20 \quad n_2 = 15 \end{aligned}$$

The unbiased estimate of standard deviation is

$$\begin{aligned} \left\{\frac{20\times 3^2 + 15\times 2^2}{20+15-2}\right\}^{\frac{1}{2}} &= \left(\frac{240}{33}\right)^{\frac{1}{2}} \\ &= 2.69679945 \\ &\simeq 2.7 \\ t_{33} &= \frac{60-57}{2.7\left\{\frac{1}{20}+\frac{1}{15}\right\}^{\frac{1}{2}}} \\ &= 3.253 \end{aligned}$$

This is significant at 5% level so H_0 is rejected.

The students in the two groups are not of the same ability. NB: We use interpolation to read t_{33} from tables.

Example 10.2.5

Below is a list of marks obtained in a chemistry practical exam by groups A and B.
A : 50, 38, 72, 42, 28, 64, 78, 47, 37, 51, and 44
B : 59, 50, 64, 68, 58, 48, 59, 54, 46, 51, and 65
Test the teacher's claim at 5% level of significance that the students in the two groups have the same ability.

Solution:

From the data

$$\begin{aligned} \bar{x}_A &= 50.09 \quad \bar{x}_B = 56.54 \\ S_A &= 15.4 \quad S_B = 7.33 \\ n_A &= 11 \; n_B = 11 \\ H_0 : \mu_A &= \mu_B \\ H_1 : \mu_A &\neq \mu_B \end{aligned}$$

The pooled variance

$$\begin{aligned} &= S^2 = \frac{(n_A - 1)S_A^2 + (n_B - 1)S_B^2}{n_A + n_B - 2} \\ &= \frac{(11-1)(15.4)^2 + (11-1)(7.33)^2}{11 + 11 - 2} \\ &= \frac{2371.6 + 537.289}{20} \\ &= 145.4 \end{aligned}$$

The test statistic t is given by

$$\begin{aligned} t_{11+11-2} &= \frac{50.09 - 56.54}{\left\{(145.4)\left(\frac{1}{11} + \frac{1}{11}\right)\right\}^{\frac{1}{2}}} \\ &= \frac{-6.45}{5.142416305} \\ t_{20} &= -1.254274181 \simeq -1.254 \\ |t_{20}| &= 1.254 \end{aligned}$$

From tables, $t_{20} = 2.086$. Because $1.254 < 2.086$, we retain the null hypothesis that the students of the two groups have the same ability.

Tests for a proportion (large samples)

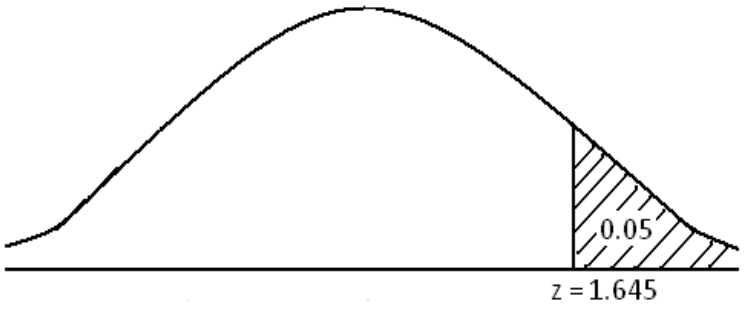
Here we consider a variable which is binomially distributed and n is large so that a normal approximation is utilised.

Example 10.2.6

Uganda National Examinations Board stated that 40% of the candidates pass Mathematics with atleast a credit. A teacher from Kilembe Secondary school claims that his students perform better than the national average because of 100 candidates he presented to UNEB, 50 got atleast a credit. Is his calim justified? Test at 5% level of significance.

Solution: The model is binomial approximated by a normal one. This is a one tailed test.

$$\begin{aligned} H_0 : p &= 0.4 \\ H_1 : p &> 0.4 \\ \mu &= np = 100 \times 0.4 = 40 \\ \delta &= \sqrt{npq} = \sqrt{100 \times 0.4 \times 0.6} = \sqrt{24} \\ Z &= \frac{x - \mu}{\delta} \\ &= \frac{50 - 40}{\sqrt{24}} \\ &= 2.041 \end{aligned}$$



The value of Z is significant at 5% level since $2.041 > 1.645$. The teachers claim is justified.

Tests for Difference between two proportions (large)

Example 10.2.7

In a sample of 1000 people from Muhanga district. There are 520 men of whom 300 are alcoholics and 480 women of whom 200 are alcoholics. Is there evidence at 5% level of significance that the men in the district are more likely to take alcohol than women?

Solution:
Let P_1 = probability that a man takes alcohol
P_2 = probability that a woman takes alcohol

$$\begin{aligned} H_0 : P_1 &= P_2 = p \\ H_1 : p_1 &> P_2 \end{aligned}$$

This is a one-tailed test. The numbers of alcoholics are binomially distributed. The best estimate for p is

$$P = \frac{\text{number of alcoholics}}{\text{number in sample}} = \frac{300 + 200}{1000} = 0.5$$

$\Rightarrow P = q = 0.5$
The observed proportion of men who take alcohol P_1 has $S.d\sqrt{(pq/n_1)}$ where n_1 is the number of men in the sample and for women P_2 has $S.d\sqrt{(pq/n_2)}$ where n_2 is the number of women in the sample.
Thus $P_1 - P_2$ has s.d δ_{1-2} given by

$$\delta_{1-2} = \left\{ pq \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right\}^{\frac{1}{2}}$$

Using an estimate of $p = 0.5$ the estimate of the pooled standard deviation is

$$\begin{aligned} S_{1-2} &= \left\{ 0.5 \times 0.5 \left(\frac{1}{520} + \frac{1}{480} \right) \right\}^{\frac{1}{2}} \\ &= 0.031648105 \\ &\simeq 0.03165 \end{aligned}$$

From H_0, the mean value of $P_1 - P_2$ is 0. The observed value is $\frac{300}{520} - \frac{200}{480} = \frac{25}{156} = 0.1602541 \simeq 0.1603$
But n is large. So we use the Z test statistic

$$\begin{aligned} Z &= \frac{0.1603 - 0}{0.03165} \\ &= 5.064770932 \simeq 5.065 \end{aligned}$$

The value is significant since $5.065 > 1.645$. So we reject H_0 and conclude that the men are more likely to take alcohol than women in the district.

10.2.3 Tests using the Poisson distribution

Example 10.2.8

The number of faults in a square metre of dyed cloth has been 8. A new dying machine has been installed and the faults per metre are now 2. Is this evidence that the new machine is better than the old one? Test at 5% level of significance.
Solution: This is a Poisson model

$$\begin{aligned} H_0 : \lambda &= 8 \\ H_1 : \lambda &< 8 \end{aligned}$$

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

$$\begin{aligned} P(X \leq 2) &= P(x = 0) + P(x = 1) + P(x = 2) \\ &= e^{-8} + \frac{8e^{-8}}{1!} + \frac{8^2 e^{-8}}{2!} \\ &= 0.00033546 + 0.002683 + 0.010734804 \\ &= 0.013753964 \simeq 1.375\% < 5\% \end{aligned}$$

The result is significant at 5% level. We reject the null hypothesis and conclude that the new machine is better that than the old one.

Exercise 10

1. Scales are set to weigh packets of wheat flour of 500g each. To find out about the accuracy of the scales, a sample of 12 packets is taken and the mean weight was found to be 502 with a standard deviation of 4kg. Assuming that the standard deviation is constant, test at the 5% level of significance if the scales are correctly set.

2. A maize miller claims that his sacks are each 60kg. A sample of 50 bags was taken and it was found to have a mean of 59kg with a standard deviation of 4kg. Is there evidence at 5% level of significance that the weights of the bags are actually less than 60kg?

3. The figures below were weighings of packets of omo in grams
   60, 58, 67, 48, 72, 55, 62, 70, 58, 74
   45, 56, 66, 47, 71, 64, 68, 52, 69, 61.
   Could they have come from a population whose mean is 60g. Test it at 5% level of signicance.

4. A bottling company sets its machine to fill bottles of 500ml of water. A sample of 100 bottles is checked and the mean quantity is found to be 496mls with a sample standard deviation of 10mls. Does this differ significantly from 500mls at 2% level of significance?

5. A manufacturer claims that his wax candles burn on average for 180 minutes. To check this claim, an officer from the bureau of standards observed the burning of 8 candles and found that they burn for the following minutes. 170, 178, 185, 182, 190, 173, 176, 181. Does this evidence support the manufactures claim at 5% level of significance?

6. The lives of light bulbs are normally distributed. If ten of them burnt for the following hours 1200, 1210, 1280, 990, 1100, 1190, 1250, 1050, 1150, 1170, estimate the population mean and show that variance is 8032 hours. The manufacture claims that his bulbs have a life span of 1200 hours. Test his claim at 5% level of significance.

7. There are 600 men and 800 women in a church. The mean number of days per year in which a man fasted was 6.5 with a standard deviation of 2.8 and for the women the corresponding figures were 7.2 and 3.1. Test at 5% level of significance if the difference between the means is significant.

8. In a random sample 600 people, there are 350 men and 250 women. 190 men are left handed while 150 women are left handed. Is there evidence at 5% level of signicance that the men in this community are more likely to be left handed than women?

9. A learner typist was making 6 mistakes per page. After a long practice, he reduces his mistakes to one per page. Is this evidence at 5% level of significance that he has improved?

Chapter 11

THE CHI-SQUARED TEST

Here we utilise the χ^2-distribution where the observed frequencies 0, are compared with the expected frequencies E. The statistic χ^2 is given by $\sum \frac{(0-E)^2}{E}$. This distribution is a function of ν, the number of degrees of freedom. For a specific value of ν, the χ^2 distribution is denoted by $\chi^2_{(\nu)}$.

The χ^2 distribution is tabulated as percentage points. A percentage point is that value of χ^2 which has a specified percentage of the distribution lying to its right.

11.1 Calculation of χ^2

As an illustration, suppose we toss three coins one hundred and twenty times. Let our interest be in the number of tails obtained. We assume that the coins are fair. This is a binomial situation with $n = 3$ and $p = \frac{1}{2}$.
The observed frequencies compared with the expected frequencies are tabulated in the table thus:

No. of tails	Observed frequency	Expected frequency
0	14	$120 \times \begin{pmatrix} 3 \\ 0 \end{pmatrix} \left(\frac{1}{2}\right)^3 = 15$
1	40	$120 \times \begin{pmatrix} 3 \\ 1 \end{pmatrix} \left(\frac{1}{2}\right) \left(\frac{1}{2}\right)^2 = 45$
2	48	$120 \times \begin{pmatrix} 3 \\ 1 \end{pmatrix} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^2 = 45$
3	18	$120 \times \begin{pmatrix} 3 \\ 3 \end{pmatrix} \left(\frac{1}{2}\right)^2 = 15$

Let the observed frequencies be O_i and the expected frequencies be E_i, the results above are put in the table below.

O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(Oi-Ei)^2}{E_i}$
14	15	-1	1	$\frac{1}{15}$
40	45	-5	25	$\frac{25}{45}$
48	45	-3	9	$\frac{9}{45}$
18	15	3	9	$\frac{9}{15}$
				$1\frac{19}{45} = 1.4222$

There are four classes so the degrees of freedom in this case is $\nu = 3$. The total frequency in this case should equal 120.

11.2 Goodness of fit

. The χ^2 distribution is used to test the goodness of fit of a given table of observed frequencies to a theoretical model.
The χ^2-test is applicable if total frequency is not less than 50 and each class has a minimum frequency of 5. In case where the class frequency is less than 5, then that class should be combined with one nearest to it.

Example 11.2.1

Four identical coins were tossed 200 times and the observed frequencies of the number of tails per toss is shown in the following table.

No. of tails	0	1	2	3	4
Observed frequency	22	50	58	42	28

Test at the 5% level if the coins are biased.
Solution:
Let $H_0 : P(T) = \frac{1}{2},\ H_1 : P(T) \neq \frac{1}{2}$

No. of tails	Observed freq	Expected freq
0	22	$200 \times \begin{pmatrix} 4 \\ 0 \end{pmatrix}^0 \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^4 = 12.5$
1	50	$200 \times \begin{pmatrix} 4 \\ 1 \end{pmatrix}^1 \left(\frac{1}{2}\right)^1 ()^3 = 50$
2	58	$200 \times \begin{pmatrix} 4 \\ 2 \end{pmatrix} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^2 = 75$
3	42	$200 \times \begin{pmatrix} 4 \\ 3 \end{pmatrix}^0 \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^1 = 50$
4	28	$200 \times \begin{pmatrix} 4 \\ 4 \end{pmatrix}^0 \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^0 = 12.5$

Then χ^2 is calculated thus.

O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
22	12.5	9.5	90.25	7.22
50	50	0	0	0
58	75	-17	289	3.853333
42	50	-8	64	1.28
28	12.5	15.5	240.25	19.22
				31.573333

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 31.57333...$$

There are five classes so that $\chi^2_{test} = 31.573$ against $\chi^2_{5\%}(4) = 9.49$
Since $\chi^2_{test} = 31.573 > 9.49$ this is significant so reject H_0, i.e. the coins are not fair. They are biased.

Example 11.2.2

For a period of thirty days 100 babies in a babies home were given a new type of food and the table below shows the recorded changes in their masses (in grams)

Change in mass	Observ freq
$-\infty < x \leq -5$	08
$-5 < x \leq 0$	06
$0 < x \leq 5$	10
$5 < x \leq 10$	29
$10 < x \leq 15$	17
$15 < x \leq 20$	08
$20 < x \leq 25$	15
$25 < x \leq \infty$	7

It is thought that these data follow a normal distribution with mean 5 and standard deviation 8. Use the χ^2 distribution at the 5% level of significance to test this hypothesis. How would the test be modified if the mean and standard deviation were unknown.
Solution:
Let the random variable X be "change in mass over thirty days".

$$H_0 : \ X \quad \text{is } N(5, 8^2)$$
$$H_1 : \ X \quad \text{is not } N(5, 8^2)$$

The expected frequencies for the given class interval is calculated in tabular form below.
If X is $N(5, 8^2)$ then the standard variable is $Z = \frac{x-5}{8}$.

Class	observed frequency	upper class Bound	standard upper class bound	Q(z)	class probability	Expected class frequency
$-2 < x \leq -5$	8	-5	-1.25	0.1056	0.1056	10.56
$-5 < x \leq 0$	6	0	-0.625	0.2660	0.1604	16.04
$0 < x \leq 5$	10	5	0	0.5000	0.2340	23.4
$5 < x \leq 10$	29	10	0.625	0.7340	0.2340	23.4
$10 < x \leq 15$	17	15	1.25	0.8944	0.1604	16.04
$15 < x \leq 20$	8	20	1.875	0.9697	0.0753	7.53
$20 < x \leq 25$	15	25	2.5	0.9938	0.0241	2.41
$25 < x \leq \infty$	7	∞	∞	1.0000	0.0062	0.62

This gives rise to the following table of observed and expected frequencies.

0	8	6	10	29	17	8	15	7
E	10.56	16.04	23.4	23.4	16.04	7.53	2.41	0.62
O-E	-2.56	-10.4	-13.4	5.6	0.96	0.47	12.59	6.38
$\frac{(O-E)^2}{E}$	0.621	6.284	7.674	1.34	0.0574	0.0293	65.771	65.652

$$\chi^2_{test} = \sum \frac{(O-E)^2}{E} = 147.4287$$

We test this against $\chi^2_{5\%}(7) = 14.07$
Since $\chi^2_{test} = 147.4287 > 14.07$, we reject H_0, i.e. the data does not follow a normal distribution with mean 5 and standard deviation 8.
If the mean and standard deviation are not known, they are estimated from the given data and then used to calculate the expected frequencies. There would be two degrees of freedom less.
If a given distribution is said or thought to be a normal, poisson or binomial, do the following to test the goodness of fit:

(i) Calculate the expected frequencies E under the null hypothesis.

(ii) Combine adjacent classes if one or some of them have frequencies less than 5 thereby also combining their frequencies.

(iii) Calculate $\frac{(0-E)^2}{E}$ for each of the classes.

(iv) Calculate the statistc $\chi^2_{test} = \sum \frac{(0-E)^2}{E}$.

(v) Determine the degrees of freedom ν. This is $n-1$ if p is known for a binomial distribution but $\nu = n-2$ if p has to be estimated using $\bar{x} = np$. The same case applies to the poisson distribution. For a normal distribution, if μ and δ are known, $\nu = n-1$ but if it has to be estimated, then $\nu = n-3$.

(vi) Find $\chi^2_{\alpha\%(\nu)}$ where $\alpha\%$ is the level of significance under which the test is being done.

(vii) Comparing χ^2_{test} with $\chi^2_{\alpha\%(\nu)}$, if $\chi^2_{test} > \chi^2_{\alpha\%}(\nu)$, reject the null hypothesis or else accept it.

Example 11.2.3

Five identical coins were tossed 210 times and the observed frequencies of the number of heads per toss was as shown in the table below:

Number of heads	0	1	2	3	4	5
Observed frequency	12	40	56	60	27	15

Test 1% level of significance if the coins are biased.

Solution:

This is a binomial situation with $n = 5$ and $p = \frac{1}{2}$. Let H$_0$:p$= \frac{1}{2}$ and H$_1$: $p \neq \frac{1}{2}$. The observed frequencies compared with the expected frequencies are tabulated below.

No. of tails	Observed frequency	Expected frequency
0	12	$210 \times \binom{5}{0} \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^5 = 6.5625$
1	40	$210 \times \binom{5}{1} \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^4 = 32.8125$
2	56	$210 \times \binom{5}{2} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^3 = 65.625$
3	60	$210 \times \binom{5}{2} \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^2 = 65.625$
4	27	$210 \times \binom{5}{4} \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^2 = 32.8125$
5	15	$210 \times \binom{5}{2} \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right)^0 = 6.5625$

Let the observed frequencies be O and the expected frequencies, E, the above results are put in the table below.

O	E	O-E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
12	6.5625	5.4375	29.56640625	4.505357143
40	32.8125	7.1875	51.66015625	1.574404762
56	65.625	−9.625	92.640625	1.411666667
60	65.625	−5.625	31.640625	0.482142857
27	32.8125	−5.8125	33.78515625	1.029642857
15	6.5625	8.4375	71.19140625	10.84821429

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 19.85142857 \cong 19.85$$

This is tested against $\chi^2_{\%}(5) = 15.09$. Since $\chi^2_{test} > \chi^2_{1\%}(5)$, we reject H_0. That is, the coins are biased so that $p \neq \frac{1}{2}$.

Exercise 11

1. For a period of one year. 100 snakes in a zoo were given a new type of diet. The table below shows the changes in mass (grams) recorded by the zoo attendant.

Change in mass(g) x	Observed frequency
$-20 \leq x \leq -15$	8
$-15 \leq x \leq -10$	15
$-10 \leq x \leq -5$	4
$-5 \leq x \leq 0$	6
$0 \leq x \leq 5$	12
$5 \leq x \leq 10$	15
$10 \leq x \leq 15$	10
$15 \leq x \leq 20$	9
$20 \leq x \leq 25$	5

It is thought that these data follow a normal distribution with mean 5 and standarddeviation 10. Use χ^2 the distribution at 10% level of significance to test this hypothesis.

2. Analysis of the goals scored per match by a certain football club over the past few years gives the following results:

No. of goals per match	0	1	2	3	4	5	6
No. of matches	15	20	30	18	7	6	4

Find the mean of this distribution and the frequencies correct to two decimal places associated with a Poisson distributin having the same mean. Use the distribution at 2.5% level of significance to determine whether or not the above distribution can reasonably be modeled by this Poisson distribution.

3. Three identical dice were tossed 80 times and the observed frequencies of the 1's showing up was as shown in the table below:

Number of 1's	0	1	2	3
Observed frequency	20	26	20	14

Test at 5% level of significance if the dice are biased.

Chapter 12

CORRELATION AND REGRESSION

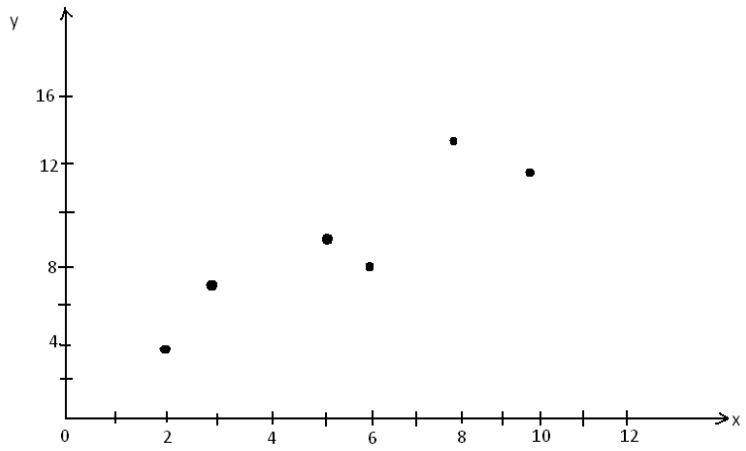
Introduction

A population may have two variables and our concern in this chapter is to find out if there is a relationship between the two variables i.e. we shall check if there is interdependence or correlation between the two variables.
If the variables are plotted in the xy-plane, we get a scatter diagram

Example 12.0.4

Draw a scatter diagram for the following data:

x	2	3	5	6	8	9	10
y	4	7	9	8	10	13	12



Each point represents the values of two variables under consideration for a particular set of sample.

Types of correlation:

- **Positive Correlation:** If y tends to increase as x increases, there is a positive correlation.
- **Negative Correlation:** If y tends to decrease as x increases, there is a negative correlation.
- **Zero Correlation:** If there is no relationship between x and y then there is zero or no correlation

On scatter diagram above shows a positive correlation.

Measurement of correlation:

The scatter diagrams give a visual impression of correlation. If we have to quantify this correlation then we have to evaluate statistically the two variables whose correlation coefficient we need to determine.
The covariance of two variables X and Y, abbreviated as Cov(X,Y) is given by the expression E(X-μ)$(Y - \mu_y)$ and if X and Y are independent Cov(X,Y)=0. If X and Y are not independent Cov(X,Y)$\neq 0$
Generally,

$$\begin{aligned} Cov(X,Y) &= E[(X-\mu_x)(Y-\mu_y)] \\ &= E[XY - \mu_x Y - \mu_x X + \mu_x\mu_y] \\ &= E(XY) - \mu_x E(Y) - \mu_x E(X) + \mu_x\mu_y \\ &= E(XY) - \mu_x\mu_y - \mu_y\mu_x + \mu_x\mu_y \\ Cov(X,Y) &= E(XY) - \mu_x\mu_y. \end{aligned}$$

Conclusively, if X and Y are independent,
$E(XY) = E(x)E(y) = \mu_x\mu_y$, so that in this case

$$\begin{aligned} Cov(X,Y) &= E(XY) - \mu_x\mu_y \\ &= \mu_x\mu_y - \mu_x\mu_y \\ &= 0 \end{aligned}$$

The unbiased estimate of Cov(X,Y) is

$$\frac{1}{n-1}\sum_{1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

and it depends on the degree of correlation and spread of the values of X and of Y.
The correlation coefficient lies between 0 and 1 for a positive correlation or between -1

and 0 for a negative correlation.
If the unbiased estimate of covariance is divided by the unbiased estimates of standard deviations of X and Y, we get the product moment correlation coefficient. (Pearsonian). Let the pearson product moment correlation coefficient be v and it is also known that the unbiased estimates for the variance of X and Y respectively are

$$\begin{aligned}
S_x &= \sqrt{\left\{\frac{1}{n-1}\sum_1^n (x_i-\bar{x})^2\right\}}, \\
S_y &= \sqrt{\left\{\frac{1}{n-1}\sum_1^n (y_i-\bar{y})^2\right\}}. \text{ Then} \\
r &= \frac{\frac{1}{n-1}\sum_1^n (x-\bar{x})(y-\bar{y})}{S_x \cdot S_y} \\
&= \frac{\frac{1}{n-1}\sum_1^n \sum_1^n (x-\bar{x})(y-\bar{y})}{\sqrt{\left\{\frac{1}{n-1}\sum_1^n (x_i-\bar{x})^2\right\}}\sqrt{\left\{\frac{1}{n-1}\sum_1^n (y_i-\bar{y})\right\}}} \\
r &= \frac{\sum_1^n (x-\bar{x})(y-\bar{y})}{\sum_1^n (x-\bar{x})^2 \cdot \sum_1^n (y-\bar{y})^2} \\
\text{But } \sum_1^n (x_i-\bar{x})^2 &= \sum_1^n x_i^2 - \frac{(\sum_1^n x_i)^2}{n} \\
\text{and } \sum_1^n (y_i-\bar{y})^2 &= \sum_1^n y_i^2 - \frac{(\sum_1^n y_i)^2}{n} \\
\text{so that } r &= \frac{\sum_1^n (x-\bar{x})(y-\bar{y})}{\sqrt{\left[\left\{\sum x_i^2 - \frac{(\sum_1^n x_i)^2}{n}\right\}\left\{\sum y_i^2 - \frac{(\sum_1^n y_i)^2}{n}\right]}}, \\
\text{And also } \sum_1^n (x-\bar{x})(y-\bar{y}) &= \sum_1^n x_i y_i - \sum_1^n \bar{x} y_i - \sum_1^n \bar{y} x_i + \sum_1^n \bar{x}\bar{y} \\
&= \sum_1^n x_i y_i - \bar{x}\sum_1^n y_i - \bar{y}\sum_1^n x_i + \sum_1^n \bar{x}\bar{x}
\end{aligned}$$

$$= \sum_1^n x_i y_i - \bar{x}\sum_1^n y_i - \bar{y}\sum_1^n x_i + n\bar{x}\bar{y}$$

$$\text{but } \sum_1^n y_i = n\bar{y} \text{ and}$$

$$\sum_1^n x_i = n\bar{x} \text{ so that}$$

$$\sum_1^n (x-\bar{x})(y-\bar{y}) = \sum_1^n x_i y_i - n\bar{x}\bar{y} + n\bar{x}\bar{y}$$

$$= \sum_1^n x_i y_i - n\bar{x}\bar{y}$$

$$\text{Hence } r = \frac{\sum_1^n x_i y_i - n\bar{x}\bar{y}}{\left[\left\{\sum_1^n x_i^2 - \frac{(\sum_1^n x_i)^2}{n}\right\}\left\{\sum_1^n y_i^2 - \frac{\left(\sum_1^n\right)^2}{n}\right\}\right]^{\frac{1}{2}}}$$

$$= \frac{\sum_1^n x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left\{n\sum_1^n x_i^2 - (\sum_1^n x_i)^2\left[n\sum_1^n y_i^2 - (\sum_1^n y_i)^2\right]\right\}}}$$

$$r = \frac{\sum_1^n x_i y_i - \left(\sum_1^n x_i \sum_1^n y_i\right)/n}{\left\{\left[n\sum_1^n x_i^2 - \left(\sum_1^n x_i\right)^2\right]\left[n\sum_1^n y_i^2 - (\sum_1^n y_i)^2\right]\right\}^{\frac{1}{2}}}$$

$$r = \frac{n\sum_1^n x_i y_i - \sum_1^n x_i \sum_1^n y_i}{\left\{\left(n\sum_1^n x_i^2 - \left(\sum_1^n x_i\right)^2\right)\left[n\sum_1^n y_i^2 - \left(\sum_1^n y_i\right)^2\right]\right\}}$$

Example 12.0.5

For the data below, calculate the product moment correlation coefficient.

x	2	3	5	6	8	10
y	4	7	9	8	10	12

Solution:

x	y	x^2	y^2	xy
2	3	4	16	6
3	7	9	49	21
5	9	25	81	45
6	8	36	64	48
8	10	64	100	80
10	12	100	144	120
34	50	238	454	320

The product moment correlation coefficient is

$$\begin{aligned} r &= \frac{n\sum xy - \sum x \sum y}{\sqrt{\{[n\sum x^2 - (\sum x^2)]\,[n\sum y^2 - (\sum y)^2]\}}} \\ &= \frac{6 \times 320 - 34 \times 50}{\sqrt{(6 \times 238 - 34^2)(6 \times 454 - 50^2)}} \\ &= \frac{1920 - 1700}{\sqrt{(1428 - 1150)(2724 - 2500)}} \\ &= \frac{232}{\sqrt{272 \times 224}} = 0.891280119 \\ &\simeq 0.8913 \end{aligned}$$

Example 12.0.6

The marks of candidates in maths and physics were given as;

Candidates	A	B	C	D	E	F	G	H	I
Maths	58	51	36	87	76	45	42	49	83
Physics	70	64	55	81	33	45	73	66	91

Calculate the product moment correlation coefficient.
Solution:

x	y	x^2	y^2	xy
58	70	3364	4900	4060
51	64	2601	4096	3264
36	55	1296	3025	1980
87	81	7569	6561	7047
76	33	5776	1089	2508
45	45	2025	2025	2025
42	73	1764	5329	3066
49	66	2401	4356	3234
83	91	6889	8281	7553
527	578	33685	39662	34737

$$\begin{aligned} r &= \frac{n\sum xy - \sum x \sum y}{\sqrt{\{[n\sum x^2 - (\sum xy^2)]\,[n\sum y^2 - (\sum y)^2]\}}} \\ &= \frac{9 \times 34737 - 527 \times 578}{\sqrt{\{[9 \times 33685 - (527)^2]\,[9 \times 39662 - (578)^2]\}}} \\ &\simeq 0.3328 \end{aligned}$$

Another way of stating the product moment correlation coefficient is

$$\begin{aligned} r &= \frac{C_{xy}}{\sqrt{C_{xx}C_{yy}}} \\ \text{where } C_{xy} &= \sum xy - n\bar{x}\bar{y} \\ C_{xx} &= \sum x^2 - n\bar{x}^2 \\ C_{yy} &= \sum y^2 - n\bar{y}^2 \end{aligned}$$

Try the example we have done using this other approach and find which will be easier for you to remember when solving such problems.

Interpretation of the magnitude correlation coefficient.

This is given in the following table.

Correlation coefficient	Interpretation
$0- < 0.2$	Very low correlation
$0.2- < 0.4$	Low correlation
$0.4- < 0.6$	Moderate correlation
$0.6- < 0.8$	High correlation
$0.8 - 1.00$	Very high correlation.

12.1 RANK CORRELATION

There are two ways of measuring correlation by ranks. These are by spearman and kendall.

12.1.1 Spearman's rank correlation ρ

In this case $\rho = 1 - \frac{6\sum d^2}{n(n^2-1)}$
where d is the difference between the rankings of a given pair of scores and n is the number of pairs.

Example 12.1.1

Two examiners X and Y marked scripts of candidates who sat for an interview. They gave the following marks.

x	62	78	59	67	42	54	71
y	60	72	66	64	56	48	67

Calculate the spearman's correlation coefficient of ρ.
Solution:

X	Y	R_x	R_y	$D=R_x-R_y$	D^2
62	60	4	5	1	1
78	72	1	1	0	0
59	66	5	3	2	4
67	64	3	4	-1	1
42	56	7	6	1	1
54	48	6	7	-1	1
71	67	2	2	0	0
					8

$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

$$
\begin{aligned}
&= 1 - \frac{6 \times 8}{7(7^2 - 1)} \\
&= 0.857142857 \\
&\simeq 0.857
\end{aligned}
$$

Example 12.1.2

Below are marks scored by 10 students in a Maths and Physics examination.

M:	50	72	43	61	80	32	66	69	54	40
P:	35	40	60	53	55	70	38	42	63	48

Calculate the spearman' s rank correlation coefficient between the performance in Maths and physics. Comment on the correlation.
Solution:

M	P	R_m	R_p	D=$R_m - R_p$	D^2
50	35	7	10	-3	9
72	40	2	8	-6	36
43	60	8	3	5	25
61	53	5	5	0	0
80	55	1	4	-3	9
32	70	10	1	9	81
66	38	4	9	-5	25
69	42	3	7	-4	16
54	63	6	2	4	16
40	48	9	6	3	9
					226

$$
\begin{aligned}
\rho &= 1 - \frac{6 \sum d^2}{n(n^2 - 1)} \\
&= 1 - \frac{6 \times 226}{10 \times 99} \\
&= -0.36969696969 \\
&\simeq -0.3697.
\end{aligned}
$$

There is a low negative correlation. Performance in one negatively affects performance in the other.

12.1.2 Kendall's Correlation Coefficient.

Kendall's rank correlation coefficient (τ) is given by

$$\tau = \frac{\text{agreements -disagrements}}{\text{total number of pairs}} = \frac{S}{\frac{1}{2}n(n-1)}$$

where S is the sum of all the scores. The scores are obtained in the following way:

- the data is arranged in two rows where the first row is in ascending order.
- the data in the second row is arranged in accordance with that of the first row
- Give ranks to each of the rows
- Compare the adjacent scores of the two rows. If there is an increases or decrease in both, allot+1. If different allot -1.

Example 12.1.3

Two examiners X and Y marked scripts of 8 candites and gave the following marks:

X	42	61	38	55	65	48	35	53
Y	45	54	46	51	72	45	49	41

Calculate Kendall's correlation coefficient for the two examiners.
Solution:
We rank each of the marks for the examiners from highest to lowest. So the ranks are:

	A	B	C	D	E	F	G	H
R_x	1	2	3	4	5	6	7	8
R_y	1	2	3	8	6.5	6.5	5	4

The pairing and scoring is done as follows

AB	AC	AD	AE	AF	AG	AH	Score
1	1	1	1	1	-1	-1	=3
	BC	BD	BE	BF	BG	BH	
	1	1	1	1	-1	-1	=2
		CD	CE	CF	CG	CN	
		1	1	1	-1	-1	=1
			DE	DF	DG	DH	
			1	1	-1	-1	=0
				EF	EG	EH	
				1	-1	-1	=-1
					FG	FH	
					-1	-1	=-2
						GH	
						1	$= \underline{1}$
							4

Total score is $4 \therefore S = 4$

$$\tau = \frac{S}{\frac{1}{2}n(n-1)} = \frac{4}{\frac{1}{2}(8)(7)} = \frac{1}{7} = 0.1429$$

Note 12.1.1

In case of the same marks the ranks are shared. In this case we had 45 and 45 so the ranks 6 and 7 are shared between them so that each has a rank of 6.5

Example 12.1.4 *In the two papers of an A-level maths examination, ten candidates gained the following marks*

Candidate	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
Paper 1	*92*	*80*	*71*	*89*	*33*	*68*	*46*	*70*		
Paper 2	*79*	*93*	*73*	*85*	*94*	*64*	*65*	*67*	*74*	

calculate

(a) the product moment correlation coefficient

(b) spearmans coefficient of rank correlation

Solution:

(a)

x	x^2	y	y^2	xy
92	8464	79	6241	7268
80	6400	93	8649	7440
71	5041	73	5329	5183
89	7921	85	7225	7565
81	6561	82	6724	6642
86	7396	94	8836	8084
33	1089	64	4096	2112
68	4624	65	4225	4420
46	2116	67	4489	3082
70	4900	74	5476	5180
716	54512	776	61290	56976

$$
\begin{aligned}
r &= \frac{n\sum xy - \sum x \sum y}{\sqrt{\{[n\sum x^2 - (\sum x)^2]\,[n\sum y^2 - (\sum y)^2]\}}} \\
&= \frac{10 \times 56976 - 716 \times 776}{\sqrt{\{[10 \times 54512 - 716^2]\,[10 \times 61290 - 776^2]\}}} \\
&= \frac{14144}{\sqrt{\{32464 \times 10724\}}} \\
&= 0.758041234 \\
&\simeq 0.758
\end{aligned}
$$

(b)

X	Y	R_x	R_y	$D=R_x-R_y$	D^2
92	79	1	5	-4	16
80	93	5	2	3	9
71	73	6	7	-1	1
89	85	2	3	-1	1
81	82	4	4	0	0
86	94	3	1	2	4
33	64	10	10	0	0
68	65	8	9	-1	1
46	67	9	8	1	1
70	74	7	6	1	1
					34

$$\begin{aligned}\rho &= 1 - \frac{6\sum D^2}{n(n^2-1)} \\ &= 1 - \frac{6 \times 34}{10 \times 99} \\ &= 0.793939393 \\ &\simeq 0.794\end{aligned}$$

Example 12.1.5

Three weighing scales from different stalls X,Y,Z in Gitarama market were used to weigh ten bags of maize A,B,C,..., J and the results in kilogrammes were as given in the table below

		A	B	C	D	E	F	G	H	I	J
Scale	X	63	66	68	61	62	60	71	73	70	76
Scale	Y	61	66	66	58	63	59	70	71	68	64
Scale	Z	61	72	76	73	62	71	77	68	65	77

Determine rank correlation coefficients for the performances of the scales.

(i) X and Y

(ii) Y and Z.
Which of the three scales X,Y and Z were in good working conditions.
Solution:
Using Spearman's coefficient.

(i)

x	y	R_x	R_y	D=R_x-R_y	D^2
63	61	7	8	-1	-1
66	66	6	4.5	1.5	2.25
68	66	5	4.5	0.5	0.25
61	58	9	10	-1	1
62	63	8	7	1	1
60	59	10	9	1	1
71	70	3	2	1	1
73	71	2	1	1	1
70	68	4	3	1	1
76	64	1	6	-5	25
					34.5

$$\begin{aligned}\rho &= 1 - \frac{6\sum D^2}{n(n^2-1)} \\ &= 1 - \frac{6 \times 34.5}{10 \times 99} \\ &= 0.791\end{aligned}$$

(ii)

Y	Z	R_y	R_z	D=R_y-R_z	D^2
61	61	8	10	-2	4
66	72	4.5	5	-0.5	0.25
66	76	4.5	3	1.5	2.25
58	73	10	4	6	36
63	62	7	9	-2	4
59	71	9	6	3	9
70	77	2	1.5	0.5	0.25
71	68	1	7	-6	36
68	65	3	8	-5	25
64	77	6	1.5	4.5	20.25
					137

$$\begin{aligned}\rho &= 1 - \frac{6\sum D^2}{n(n^2-1)} = 1 - \frac{6 \times 137}{10 \times 99} \\ &= 0.1697.\end{aligned}$$

Scales X and Y were in good working condition due to a high correlation. If you tried to check X and Z, you would get a coefficient of 0.385 which is still low.
NB: Try to do this question again using Kendall's correlation coefficient. It will be different but will give you the same conclusion. It takes a little more time.

12.2 REGRESSION

Regression is a statistical method used to find whether there is a relationship and which type of relationship between two variables. The relationships may be positive or negative, linear or non-linear.

Normally a scatter plot is made so that the general trend of the relation can be observed so that ways of measuring the relationship are utilised. After a scatter plot is drawn, the value the correlation coefficient if seen to be significant leads to the determination of the equation of the regression line. The regression line is also called the line of best fit. This line may be drawn by "eye" although the line drawn by "eye" can't be reliable enough. If the number of points above the line is equal to the number of points below the line and the sums of the squares of the vertical distances from each point to the line is at a minimum, such a line can be said to be of best fit.

12.2.1 Determination of the regression line Equation

The equation of the regression line is $y' = a + bx$ where a is the y' intercept and b is the slope of the line where

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

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

The mathematical derivation of the equation of the regression line is beyond the scope of this book.

Example 12.2.1

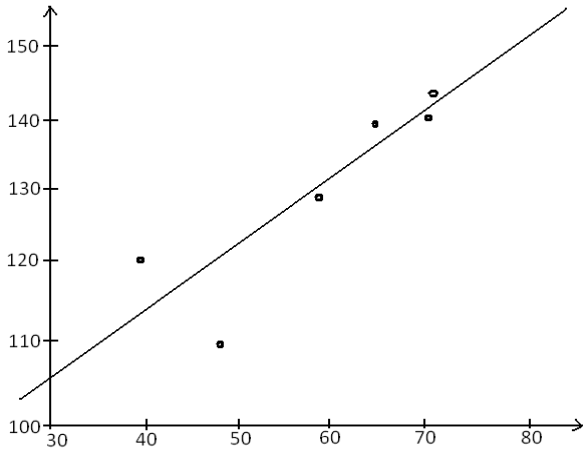
Find the equation of the regression line for the data below and graph the line on the scatter plot of the data.

x:	40	48	55	60	66	70
y:	120	110	130	140	142	150

Solution:

x	y	x^2	xy
40	120	1600	4800
48	110	2304	5280
55	130	3025	7150
60	140	3600	8400
66	142	4356	9372
70	150	4900	10,500
339	792	19785	45502

$$
\begin{aligned}
a &= \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2} \\
&= \frac{792 \times 19785 - 339 \times 45502}{6 \times 19785 - 339^2} \\
&= 64.53998416 \\
&\simeq 64.54 \\
b &= \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} \\
&= \frac{6 \times 45502 - 339 \times 792}{6 \times 19785 - 339^2} \\
&= 1.193982581 \simeq 1.194. \\
\text{The equation is } y' &= a + bx \\
\Rightarrow y &= 64.54 + 1.194x
\end{aligned}
$$



Example 12.2.2

The table below shows the percentage of sand, y in the soil at different depths x, in metres.

Soil depth (x):	4.5	7.0	5.8	3.0	5.0	8.5	9.8	10.5
% of sand (y):	80	65	75	90	68	55	64	50

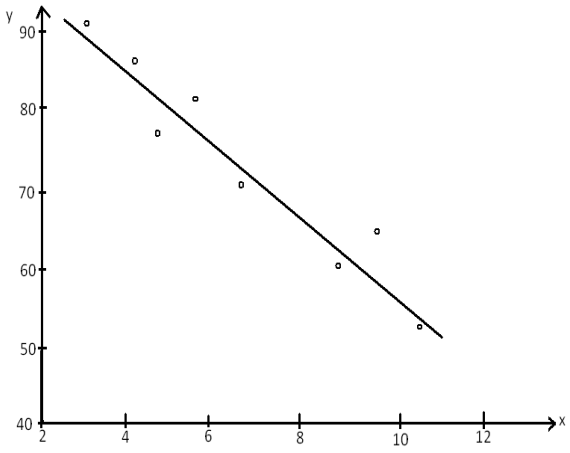
(i) Plot a scatter diagram for the data. Comment on the relationship between soil depth and the percentage of sand in the soil.

(ii) Draw a line of best fit through the points of the scatter diagram. Use your result to estimate

(a) percentage of sand at a depth of 4m

(b) depth of the soil with 70% sand.

Solution:



There is a strong negative correlation

x	y	x^2	xy
4.5	80	20.25	360
7.0	65	49.0	455
5.8	75	33.64	435
3.0	90	9.0	270
5.0	68	25.0	340
8.5	55	72.25	467.5
9.8	64	96.04	627.2
10.5	50	110.25	525
54.1	547	415.43	3479.7

Using the table

$$\begin{aligned} a &= 98.29927136 \simeq 98.3 \\ b &= -4.425030885 \simeq -4.43 \end{aligned}$$

The line of best fit is given by

$$\begin{aligned} y' &= a + bx \\ \Rightarrow y' &= 98.3 - 4.43x \end{aligned}$$

(a) when $x = 4$,y=80.58%

(b) when $y = 70$

$$\begin{aligned} 70 &= 98.3 - 4.43x \\ x &= 6.388261851 \\ &\simeq 6.39m \end{aligned}$$

Example 12.2.3

In two different competitions, schools A,B,C,. . . and H participated and their performances in points are give in the table below

		A	B	C	D	E	F	G	H
Competition	1	56	47	50	51	43	49	52	54
Competition	2	59	49	58	60	52	51	56	57

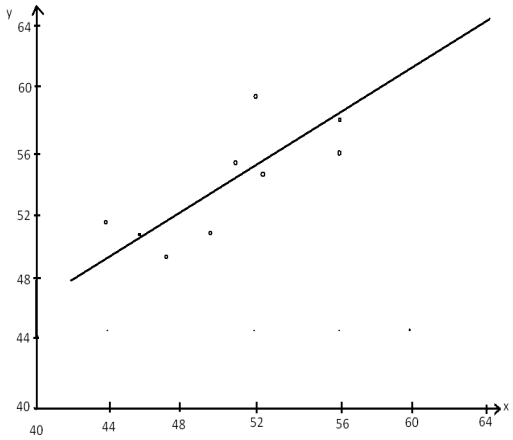
(i) Plot the points on a scatter diagram of competition 2 against competition 1.

(ii) Draw a line of best fit through the plotted points on your scatter diagram. Estimate how many points a school would have scored in competition 2 if it had 48 points in competition 1

(iii) Calculate the pearsonian product moment correlation for the data.

(iv)

Solution:



To draw a line of best fit, we look for the line $y' = a + bx$

x	y	x^2	xy	y^2
56	59	3136	3304	3481
47	49	2209	2303	2401
50	58	2500	2900	3364
51	60	2601	3060	3600
43	52	1849	2236	2704
49	51	2401	2499	2601
52	56	2704	2912	3136
54	57	2916	3078	3249
402	442	20316	22292	24536

The line of best fit is given by $y' = a + bx$ where

$$\begin{aligned} a &= \frac{(\sum y)(\sum x^2) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} \\ &= \frac{442 \times 20316 - 402 \times 442}{8 \times 201316 - (402)^2} \\ &= 19.79220779 \\ &\simeq 19.8 \\ \text{and } b &= \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} \\ &= \frac{8 \times 22292 - 402 \times 442}{8 \times 201316 - (402)^2} \\ &= 0.705627705 \\ &\simeq 0.706. \\ \Rightarrow y' &= 19.8 + 0.706x \end{aligned}$$

which line is plotted on the scatter diagram.
When $x = 48, y = 53.688 \simeq 54$

(iii) Pearsonian product moment correlation coefficent is given by

$$\begin{aligned} r &= \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2]\,[n\sum y^2 - (\sum y)^2]}} \\ &= \frac{8 \times 2292 - 402 \times 442}{[(8 \times 20316 - (402)^2]\,[8 \times 24536 - (442)^2]} = \frac{652}{\sqrt{924 \times 924}} \end{aligned}$$

$= 0.705627705$

$\simeq 0.706$

Example 12.2.4

The price of matoke is found to depend on the distance the market is away from the nearest town. The table below gives the average price for the markets around Kasese town.

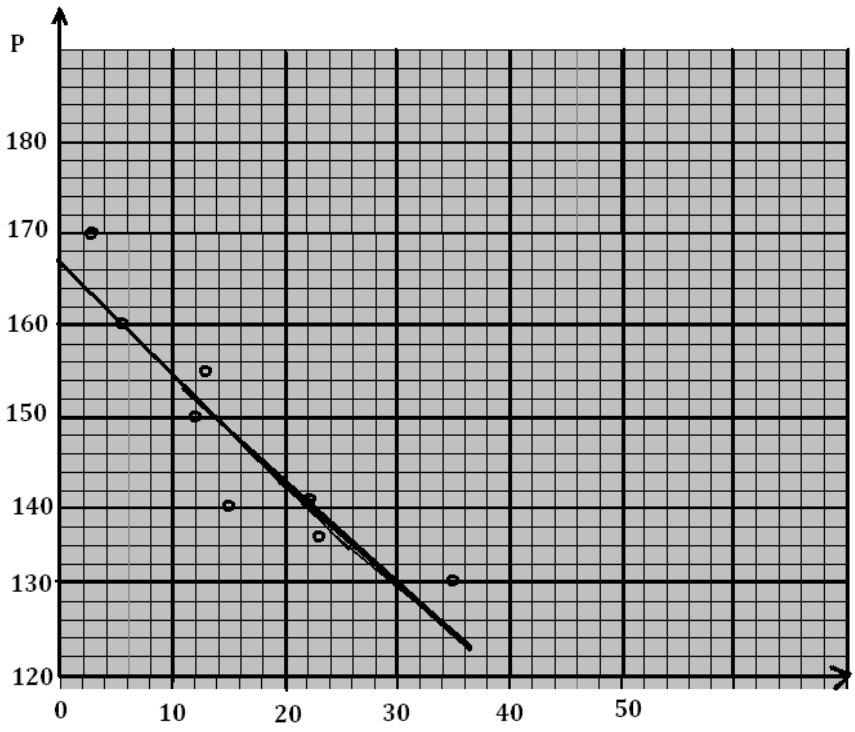
Distance dkm	35	3	12	15	19	25	5	22	11
Price P shs	130	170	150	140	145	135	160	140	155

(i) Plot these data on a scatter diagram

(ii) Draw the line of best fit on your diagram

(iii) Find the equation of your line

(iv) Estimate the price of matoke when $d = 13$



Let $d = x$ and $p = y$

x	y	x^2	xy
35	130	1225	4550
3	170	9	510
12	150	144	800
15	140	225	2100
19	145	361	2755
25	135	625	3375
5	160	25	800
22	140	484	3080
11	155	121	1705
147	1325	3219	20675

The line of best fit is $y' = a + bx$

$$\begin{aligned} a &= \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2} \\ &= \frac{1325 \times 3219 - 147 \times 20675}{9 \times 3219 - (147)^2} \\ &= 166.5240424 \\ &\simeq 166.5 \\ \text{and } b &= \frac{n(\sum y) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} \\ &= \frac{9 \times 20675 - 147 \times 1325}{9 \times 3219 - (147)^2} \\ &= -1.181744091 \\ &\simeq -1.182 \\ \Rightarrow y &= 166.5 - 1.182x \end{aligned}$$

when $d = 13$

$$\begin{aligned} y &= 166.5 - 1.182(13) \\ &= 151.134 \\ &\simeq 151 \end{aligned}$$

Exercise 12

1. For the data below, Calculate the product moment correlation coefficient

x	2	4	5	7	8	10
y	3	4	6	5	8	9

2. For the data given below calculate

   (i) the product moment correlation coefficient

   (ii) Spearman's rank correlation coefficient

x	60	72	78	59	65	70	66
y	58	74	56	64	60	65	55

3. The table below shows the marks scored by eight students in Maths and Physics.

Maths	40	48	79	26	37	60	55	70
Physics	62	68	46	39	60	52	48	32

   Calculate the Kendall's rank correlation for the students performance in the two subjects. Comment on your result.

4. The following table shows the marks scored by 10 students in English and Maths. Calculate their product moment correlation coefficient.

English	40	28	30	31	22	29	25	40	36	20
Math	45	35	31	28	27	35	24	36	30	20

5. Find the equation of the regression line for the data below:

x:	10	12	4	2	8	6	14
y:	60	55	90	89	70	80	65

6. The table below shows the heights and weights to the nearest unit of ten men.

Height (cm)	140	150	151	145	155	145	148	153	148	148
Weight (kg)	61	69	70	66	73	64	66	74	69	68

   (i) Plot the points on a scatter diagram

   (ii) Draw a line of best fit

(iii) Estimate from the graph the height of a man who weighs 67kg.

7. The following marks were scored by students in English and Mathematics examinations.

English:	73	51	42	64	38	55	47	36	51
Maths:	47	57	39	43	49	52	60	55	46

(i) Draw a scatter diagram and comment on the performance of students in the two subjects

(ii) Calculate a rank correlation and comment on your results.

8. The cost of travelling a certain distance from the city centre depends on the route and the distance a given place is away from the city centre. The table below gives the average rates of travel charged for distances travelled away from the city centre.

Distance,s(km)	5	8	10	17	20	26	29	41	42	46
Rates charged, r(shs)	500	750	900	950	1100	1000	1150	1500	1350	1750

(i) Plot the above data on a scatter diagram. Find and plot the line of best fit on the scatter diagram.

(ii) Use your result to estimate the cost of travelling a distance of 36km.

9. A speed and error typing examination was give to 9 candidates. The table below shows their speeds (y) in seconds and the number of errors in their typed scripts (x).

No. of errors (x):	10	22	18	8	30	28	26	13	15
Speed in sec(y):	120	126	114	110	143	150	145	132	135

(i) Draw a scatter diagram of these data

(ii) Calculate the equation of the regression line of y on x and draw this line on the scatter diagram

(iii) Calculate the product moment correlation coefficient.

Answers to Exercises

Exercise 1

1.

(b) 50.4

2. 1035.294

3.

(ii) 36

(iii) 36.17

4.

5. (i) (608.54444) (ii) (522.5)

6. (i) 7.70369643 (ii) (7.461009762)

7. 15.25, 15.056, 15.65

8. (6.18)

9. ($1.185358255 \simeq 1.185$)

10. (126.8)

11. $(26.6, 8.114)$

12. (i) 79.27 (ii) 11.09 (iii) 80.45 (iv) 8.55

13. 90, 93.75, 96.75, 98.75, 100, 101.25, 102.75, 103.75, 107.5

14. 72.8, 74.8, 76.8, 75.4, 73.0, 71.8, 71.2, 70.8

15. 187.5

16. 189.25, 190.25, 192, 193.25, 194.25, 196.25, 197, 198.5, 200.

Exercise 2

1. (i) $\frac{1}{4}$ (ii) $\frac{17}{20}$

2. $\frac{1}{120}$

3. $\frac{1}{3}$

4. (i) $\frac{1}{2}$ (ii) $\frac{1}{10}$

5. $\frac{3}{5}$ $\frac{3}{20}$

6. $\frac{2}{3}$

7. (i) $\frac{35}{9139}$ (ii) $\frac{3150}{9135}$ (iii) $\frac{840}{9139}$ (iv) $\frac{1680}{91390} = \frac{168}{9139}$

8. (i) $\frac{67}{150}$ (ii) $\frac{18}{25}$ (iii) $\frac{7}{25}$

9. (i) $\frac{3}{7}$ (ii) $\frac{5}{7}$ (iii) $\frac{1}{10}$

10.

11. (i) $\frac{3}{11}$ (ii) $\frac{6}{11}$ (iii) $\frac{1}{2}$ (iv) $\frac{20}{21}$

Exercise 3

1. $\frac{103}{60}$, $\frac{4451}{3600}$

2. (i) 2.8 (ii) 1.16

3. 1400/ =

4. (i) 0.1125, 2.99984 (ii) 7.4625, 77.4736

5. (i) $\frac{3}{10}$ (ii) $\frac{1}{2}$ (iii) $1\frac{1}{5}$

6. (i) 10 (ii) 3 (iii) $3, 1$ (iv) $\frac{3}{7}$

7. (i) 12.4, 13.973 (ii) $8.7, 3.4933$

8. (i) 33 (ii) 0.55 (iii) 0.55 (iv) 7.

9. (i) 1, $\frac{1}{2}$, (ii) mean = 0, variance = 1.

10. (i)

x	0	1	2	3	4
$P(x)$	$\frac{1}{16}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{1}{16}$

(ii) 2

11. (i) $\frac{1}{40} = 0.025$ (ii) $2.55, 0.5475$ (iii) $\frac{13}{40} = 0.325$

12. (i) $9\frac{17}{30}$ (ii) $2\frac{176}{287}$ (iii) 2.05599 (iv) $\frac{137}{287}$

13. (i) $\frac{7}{99}$, (ii) $\frac{14}{33}$ (ii) $\frac{35}{99}$ (iv) $\frac{1}{99}$

Exercise 4

1. (i) 0.1678 (ii) $0.00000256 \simeq 0.0000$ (iii) 0.4967 (iv) 0.7969
2. (i) 0.1641 (ii) 0.0020 (iii) 0.0020 (iv) 0.998
3. (i) 8 13
4. (i) 0.1651 (ii) 0.0134 (iii) 0.000003813766
5. (i) $\frac{8}{27}$ (ii) $\frac{1}{81}$ (iii) $\frac{65}{81}$ (iv) $\frac{80}{81}$ (v) $\frac{32}{81}$ (vi) 6
6. (i) 0.953344 (ii) 0.046656 (iii) 3.

Exercise 5

1. (i) 0.002479 (ii) 0.01487 (iii) 0.00007373 (iv) 0.0004424
2. 0.00674, 0.08422, 0.17547, 0.104445.
3. (i) 0.0821 (ii) 0.9858
4. (i) 0.0183 (ii) 0.0733 (iii) 0.7619
5. (i) 0.0498 (ii) 0.2240 (iii) 0.4232 (iv) 0.5768
6. (i) 9.356×10^{-14} (ii) 0.07263 (iii) 0.01394 (iv) 0.00103.

Exercise 6

1. (i) $\frac{1}{5}$ (ii) 1 (iii) 1 (iv) 1.0315
2. (i) $\frac{1}{4}$ (ii) $\frac{23}{6}$ (iii) 4

(iv)

$$F(x) = \begin{cases} 0, & x \leq 0 \\ \frac{x^2}{32}, & 0 \leq x \leq 4 \\ \frac{1}{4}(x-2), & 4 \leq x \leq 6 \\ 1, & x \geq 6 \end{cases}$$

3. (i) $\frac{1}{\sqrt{2}}$ (ii) 1 (iii) $\frac{\pi}{3}$

(iv)

$$F(x) = \begin{cases} 0, & x \leq 0 \\ 1 - \cos x, & 0 \leq x \leq \frac{\pi}{2} \\ 1, & x \geq \frac{\pi}{2} \end{cases}$$

4. (i)

$$f(x) = \begin{cases} \frac{2}{9}(3 - x), & 0 \leq x \leq 3 \\ 0, & elsewhere \end{cases}$$

(ii) 1 $\frac{1}{2}$

(iii)

$$F(x) = \begin{cases} 0, & 0 \leq 0 \\ \frac{1}{9}(6x - x^2), & 0 \leq x \leq 3 \\ 1, & x \geq 3 \end{cases}$$

5. (i) $k = \frac{1}{3}$, (ii) $\frac{1}{3}$ (iii) $2 - \sqrt{3}$, (iv) $\frac{3}{4}$

(v)

$$F(x) = \begin{cases} 0 & x < -1 \\ \frac{1}{3}(x+1)^2, & -1 \leq x \leq 0 \\ \frac{1}{6}(-x^2 + 4x + 2), & 0 \leq x \leq 2 \\ 1, x \geq 2 \end{cases}$$

Exercise 7

1. (a) 0.0548 (b) 0.0082 (c) 0.3085 (d) 0.0117 (e) 0.8849
2. (i) -0.253 (ii) 0.915 (iii) 1.42 (iv) 1.34

3. $49.82, 23.54,\ 0.1894$

4. (i) 684 (ii) $95.15 \simeq 95$

5. (i) 0.0736 (ii) 0.9938

6. (i) 1347.89 (ii) 41.83% (iii) 0.4199

7. 0.1018

8. (i) B (ii) D (iii) 82 (iv) 58

9. 22million

10. $190, 440, 61$ months

11. (i) 0.6915 (ii) 0.0528

12. (i) 0.9873 (ii) 0.6992 (iii) 0.1318

13. (i) 0.1699 (ii) 3.65

14. (i) $60, 37$ (ii) 94 (iii) 0.099

Exercise 8

1. $\frac{4}{3}$

2. $3, 2$

3. (i) $\frac{3}{4}$ (ii) $\frac{1}{64}$

4. (i) $\frac{6}{25}$ (ii) $\frac{9}{25}$ (iii) $2.5, 1$

5. 10 , $\frac{10}{\sqrt{3}}$

6. (i) e^{-90} (ii) $-e^{-3x}$

7. (a) 1.6 (b) $\frac{8}{3}$ (c) 1.6818 (d) $\frac{1}{16}x^4,\ 0 \leq x \leq 2,\ 1; x \geq 2$ (e) $\frac{15}{16}$

8. $8 \ln 2,\ 1.251,\ \frac{3}{16}$.

9. $M(t) = \frac{2}{2-t}, \frac{1}{2},\ \frac{1}{4}$

10. $\frac{4}{4-t}, 1, 1$

11. $\left[\frac{-3}{4}, \frac{9}{5}, \frac{1}{5}\right]$.

Exercise 9

1. (a) $9.44, 10.56$ (b) it is a lie
2. $38.8 \leq \mu \leq 58.4$
3. $[49.8 \leq \mu \leq 50.2]$
4. (i) $48.4 \leq \mu \leq 51.6$ (ii) $47.445 \leq \mu \leq 52.555$
5. $54.56 \leq \mu \leq 61.44$
6. $[102.95, 127.65]$ OR $[103.44, 127.16]$
7. $[93, 103]$
8. $[55.77, 67.73]$
9. $[12.2, 17.8]$
10. $(0.214, 0.266)$
11. $[0.124, 0.176]$ Not valid.
12. $[0.31, 0.49]$
13. $[0.723, 0.777]$

Exercise 10

1. The result is not significant. There is no evidence that the scales setting is incorrect. i.e. $1.732 < 1.96$
2. -1.768 is in the rejection region, so the weights bags are less than 60kg.
3. $0.59 < 1.96$, H_0 is retained, so the measurements could have come from a population whose mean is 60kg
4. $3.97995 > 2.33$. It differs significantly. The machine could be set again.
5. $0.2544 > 2.365$, we accept H_0. The manufacturers claim is valid.
6. $1.732 < 2.262$, accept H_0. The manufacturers claim is valid.
7. $4.42 > 1.96, Z$ is significant at 5% level, we reject H_0 and conclude thaat there is a signicant difference in means.
8. $1.392 < 1.645$ we accept H_0. Men are not more likely to be left handed than women.

9. $1.74\% < 5\%$, so he has improved actually.

Exercise 11

1. $\chi^2_{test} = 61.44 > 18.48$. We reject H_0. This data does not follow a normal distribution with mean 5 and standard deviation 10

2. $\bar{x} = 2.16$, since $\chi^2_{test} = 6.025 < 11.14$, we accept H_0. The distribution can be reasonably modeled by a Poison distribution.

3. Since $\chi^2_{test} = 554.7 > 7.815$, we reject H_0. The dice are biased

Exercise 12

1. $0.923420976 \simeq 0.923$

2. (i) $r = 0.09842$ (ii) -0.01786

3. -0.1786 low negative correlation

4. 0.7865

5. $y' = 95.86 - 2.893x$

6. (ii) $y' = -61.685 + 0.874x$ (iii) $147cm$

7. (i) There is almost no linear relationship. The trend can not be determined by the scatter plot. (ii) $e = -0.2375$

8. (i) $y' = 528.14 + 23.232x$ (ii) 1364 shs.

9. (ii) $y' = 104.4885 + 1.38x$ (ii) $v = 0.784764662$