UNIT-I

Measures of Central Tendency: In the study of a population with respect to one in which we are interested we may get a large number of observations. It is not possible to grasp any idea about the characteristic when we look at all the observations. So it is better to get one number for one group. That number must be a good representative one for all the observations to give a clear picture of that characteristic. Such representative number can be a central value for all these observations. This central value is called a measure of central tendency or an average or a measure of locations. There are five averages. Among them mean, median and mode are called simple averages and the other two averages geometric mean and harmonic mean are called special averages.

Characteristics for a good or an ideal average: The following properties should possess for an ideal average.

1. It should be rigidly defined.
2. It should be easy to understand and compute.
3. It should be based on all items in the data.
4. Its definition shall be in the form of a mathematical formula.
5. It should be capable of further algebraic treatment.
6. It should have sampling stability.
7. It should be capable of being used in further statistical computations or processing.

Arithmetic mean or mean: Arithmetic mean or simply the mean of a variable is defined as the sum of the observations divided by the number of observations. If the variable x assumes n values $x_1, x_2, ..., x_n$ then the mean, $\bar{x}$ is given by

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

This formula is for the ungrouped or raw data.

Example 1: Calculate the mean for 2, 4, 6, 8, 10

Solution: We have

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

$$\therefore \quad \bar{x} = \frac{1}{5}(2+4+6+8+10) = \frac{30}{5} = 6$$

Short-Cut method: Under this method an assumed or an arbitrary average (indicated by A) is used as the basis of calculation of deviations from individual values. The formula is

$$\bar{x} = A + \frac{\sum d}{n}$$

where, A = the assumed mean or any value in x

d = the deviation of each value from the assumed mean

Example 2: A student's marks in 5 subjects are 75, 68, 80, 92 and 56. Find his average mark.

Solution: let A=68

x	D=x-A=x-68
75	75-68=7

68	68-68=0
80	80-68=12
92	92-68=24
56	56-68=-12
Total	31

We have

$$\bar{x} = A + \frac{\sum d}{n}$$

$$\therefore \quad \bar{x} = 68 + \frac{31}{5} = 68 + 6.2 = 74.2$$

Grouped Data: The mean for grouped data is obtained from the following formula:

$$\bar{x} = \frac{\sum f\ x}{N}$$

where x = the mid-point of individual class
f = the frequency of individual class
N = the sum of the frequencies or total frequencies.

Short-cut method:

$$\bar{x} = A + \frac{\sum fd}{N} \times c$$

where $d = \frac{x - A}{c}$

A = Assumed mean or any value in x
N = total frequency
c = width of the class interval

Example 3: Given the following frequency distribution, calculate the arithmetic mean

Marks:	64	63	62	61	60	59
No. of students:	8	18	12	9	7	6

Solution: Let assumed mean=62

x	**f**	**fx**	**d=x-a=x-62**	**fd**
64	8	512	64-62=2	16
63	18	1134	63-62=1	18
62	12	744	62-62=0	0
61	9	549	61-62=-1	-9
60	7	420	60-62=-2	-14
59	6	354	59-62=-3	-18
Total	**60**	**3713**		**-7**

Direct Method:

We have

$$\bar{x} = \frac{\sum f\ x}{N}$$

$$\therefore \quad \bar{x} = \frac{3713}{60} = 61.88$$

Short-cut method:

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

Short-cut method:

$$\therefore \quad \bar{x} = 62 + \frac{-7}{60} = 62 - \frac{7}{60} = \frac{3720 - 7}{60} = \frac{3713}{60} = 61.88$$

Example 4: Following is the distribution of persons according to different income groups. Calculate arithmetic mean.

Income (Rs thousands)	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of persons	6	8	10	12	7	4	3

Solution: let A=35

Income	Number of Persons (f)	Mid value X	$d = \frac{x-A}{c} = \frac{x-35}{10}$	fd
0-10	6	5	-3	-18
10-20	8	15	-2	-16
20-30	10	25	-1	-10
30-40	12	35	0	0
40-50	7	45	1	7
50-60	4	55	2	8
60-70	3	65	3	9
Total	**50**			**-20**

$$\bar{x} = A + \frac{\sum fd}{N} \times c$$

$$\therefore \qquad \bar{x} = 35 + \frac{-20}{50} \times 10 = 35 - 4 = 31$$

Merits and demerits of Arithmetic mean:

Merits:

1. It is rigidly defined.
2. It is easy to understand and easy to calculate.
3. If the number of items is sufficiently large, it is more accurate and more reliable.
4. It is a calculated value and is not based on its position in the series.
5. It is possible to calculate even if some of the details of the data are lacking.
6. Of all averages, it is affected least by fluctuations of sampling.
7. It provides a good basis for comparison.

Demerits:

1. It cannot be obtained by inspection nor located through a frequency graph.
2. It cannot be in the study of qualitative phenomena not capable of numerical measurement i.e. Intelligence, beauty, honesty etc.
3. It can ignore any single item only at the risk of losing its accuracy.
4. It is affected very much by extreme values.
5. It cannot be calculated for open-end classes. 6. It may lead to fallacious conclusions, if the details of the data from which it is computed are not given.

MEDIAN : In the words of L.R. Conner : "The median is that value of the variable which divides the data in two equal parts, one part comprising all the values greater and the other, all values less than median." Thus, as against arithmetic mean which is based on all the items of the distribution, the median is only positional average, i.e. the value depends on the position occupied by a value in the frequency distribution.

Computation of Median

Ungrouped data: If the number of observation is odd, then the median is the middle value after the observations have been arranged in ascending or descending order of

magnitude. In case of even number of observations median is obtained as the arithmetic mean of two middle observations after they are arranged in ascending or descending order of magnitude.

Problem: The marks obtained by 12 students out of 50 are: 25, 20, 23, 32, 40, 27, 30, 25, 20, 10, 15, and 41

Solution: The values obtained by 12 students arranged in ascending order as: 10, 15, 20, 20, 23, 25, 25, 27, 30, 32, 40, 41

Here the number of items 'N' = 12, which is even.

The two middle items are 6th and 7th items

i.e., $\frac{25+25}{2} = 25$ is the median value.

Frequency (Discrete) Distribution: In case of frequency distribution where the variables take the value X_1, X_2, ..., X_n with respective frequencies f_1, f_2, ..., f_n with $N = \sum_{i=1}^{n} f_i$, median is the size of the $\frac{(N+1)}{2}th$ item or observation. In this case the use of cumulative frequency (c.f.) distribution facilitates the calculations. The steps involved are:

(i) Prepare the less than cumulative frequency (c.f.) distribution.

(ii) Find N/2.

(iii) Find the c.f. just greater than N/2.

(iv) The corresponding value gives the median.

Problem: From the following data find the value of median:

Income (Rs.)	1000	1500	800	2000	2100	1700
No. of Persons	24	26	14	10	5	28

Solution:

Income arranged in ascending order	*No. of persons (f)*	*c.f.*
800	14	14
1000	24	38
1500	26	64

1700	28	92
2000	10	102
2100	5	107

Median = Size of (N/2)th item = 107/2 = 53.5

53.5th item is consisted in the c.f. = 64. The corresponding value to this = 1500. Hence Median = Rs. 1500.

Continuous Frequency Distribution: Steps involved for its computation are:

(i) Prepare less than cumulative frequency (c.f.) distribution.

(ii) Find N/2.

(iii) Locate c.f. just greater than N/2.

(iv) The corresponding class contains the median value and is called the median class.

(v) The value of median is now obtained by using the interpolation formula:

$$mediam = L + \frac{h}{f}\left(\frac{N}{2} - C\right)$$

Where L is the lower limit or boundary of the median class;

f is the frequency of the median class;

h is the magnitude or width of class interval;

$N = \sum_{i=1}^{n} f_i$ is the total frequency; and

C is the cumulative frequency of the class preceding the median class.

Problem: The annual profits (in Rs. lacs) shown by 60 firms are given below:

Profits:	15-20	20-25	25-30	30-35	35-40	40-45	45-50	50-55	55-60	60-65
No. of firms:	4	5	11	6	5	8	9	6	4	2

Calculate the median.

Solution:

Profits	*No. of firms (f)*	*Cumulative frequency c.f.*
15-20	4	4
20-25	5	9

25-30	11	20
30-35	6	26
35-40	5	31
40-45	8	39
45-50	9	48
50-55	6	54
55-60	4	58
60-65	2	60

Median item = N/2 =60/2= 30

The cumulative frequency just greater than 30 is 31 and is corresponding class 35-40 is the median class.

$$mediam = L + \frac{h}{f}\left(\frac{N}{2} - C\right)$$

$$\therefore \qquad mediam = 35 + \frac{5}{5}\left(\frac{60}{2} - 26\right)$$

$$= 35 + 4 = 39\, marks$$

Merits and Limitations of Median

The median is superior to arithmetic mean in certain aspects. For example, it is especially useful in case of open-ended distribution and also it is not influenced by the presence of extreme values. In fact when extreme values are present in a series, the median is more satisfactory measure of central tendency than the mean.

However, since median is positional average, its value is not determined by each and every observation. Also median is not capable of algebraic treatment. For example, median cannot be used for determining the combined median of two or more groups. Furthermore, the median tends to be rather unstable value if the number of observations is small.

MODE: Mode is the value which occurs most frequently in the set of observations. It is the point of maximum frequency or the point of greatest density. In other words, the mode or modal value of the distribution is that value of the variate for which frequency is maximum.

Computation of Mode

a) In case of discrete frequency distribution, mode is the value of the variable corresponding to the maximum frequency.

But in any one (or more) of the following cases:

i) If the maximum frequency is repeated

ii) If the maximum frequency occurs in the very beginning or at the end of the distribution

iii) If there are irregularities in the distribution, the value of mode is determined by the method of grouping.

b) In case of continuous frequency distribution, mode is given by the formula :

$$Mode = l + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times h$$

Where, l is lower limit, h the width and f_m the frequency of the modal class; f_1 and f_2 are the frequencies the classes preceding and succeeding the modal class respectively.

Note:

1. For symmetric distribution mean mode and median coincide.
2. When mode is ill defined i.e., where the method of grouping also fails, its value is ascertained by the formula

Mode=3 Median – 2 Mean

Example: if seven men are receiving daily wages of Rs 150, 160, 170, 170, 170, 180 and 200. Find the modal wages.

Solution: Since 170 occur thrice and no other item occurs three times or more than three times, hence modal wages is Rs 170.

Example: Determine the mode from the following frequency distribution:

X	1	2	3	4	5	6	7	8
F	4	9	16	25	22	16	8	3

Solution: Here maximum frequency is 25and the corresponding value of X is 4. Hence mode is 4.

Example: Calculate the mode from the following frequency distribution:

Size (x)	4	5	6	7	8	9	10	11	12	13
Frequency (f)	2	5	8	9	12	14	14	15	11	13

Solution: method of grouping

Size (x)	frequency					
	I	II	III	IV	V	VI
4	2					
5	5	7		15		
6	8		13		22	
7	9	17				29
8	12		21	35		
9	14	26	28		40	
10	14			40		43
11	15	29	26		39	
12	11	24				
13	13					

In column I, original frequencies are written.

In column II, frequencies of column I are combined two by two.

In column III, leave the first frequency of column I and combine the others two by two.

In column IV, frequencies of column I are combined three by three.

In column V, leave the first frequency of column I and combine the others three by three.

In column VI, leave the first two frequencies in column I and combine the others three by three.

Now, we frame another table in which against every maximum item in column I to VI, we write down the corresponding size or sizes. The size (x) which occurs maximum number of times is the mode.

Columns	Size of item having maximum frequency
I	11
II	10,11
III	9,10
IV	10,11,12
V	8,9,10
VI	9,10,11

Since the item 10 occurs maximum number of times i.e., 5 times, hence the mode is 10.

Problem 10: find the mode of the following data :

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100

No. of candidates	3	5	7	10	12	15	12	6	2	8

Solution: Since the highest frequency is15, the modal class is 50-60.

Where, l= lower limit of modal class=50

h = width of modal class=10

f_m = frequency of the modal class=15

f_1 = frequency of the classes preceding the modal class=12

f_2 = frequency of the classes succeeding the modal class=12

We have

$$Mode = l + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times h$$

$$\therefore \quad Mode = 50 + \frac{15-12}{2(15)-12-12} \times 10 = 50 + \frac{3}{6} \times 10 = 50 + 5 = 55$$

Measures of Dispersion:

The measure of central tendency serve to locate the center of the distribution, but they do not reveal how the items are spread out on either side of the center. This characteristic of a frequency distribution is commonly referred to as dispersion. In a series all the items are not equal. There is difference or variation among the values. The degree of variation is evaluated by various measures of dispersion. Small dispersion indicates high uniformity of the items, while large dispersion indicates less uniformity. For example consider the following marks of two students.

Student I	Student II
68	85
75	90
65	80
67	25
70	65

Both have got a total of 345 and an average of 69 each. The fact is that the second student has failed in one paper. When the averages alone are considered, the two students are equal. But first student has less variation than second student. Less variation is a desirable characteristic.

Characteristics of a good measure of dispersion:

An ideal measure of dispersion is expected to possess the following properties:

1. It should be rigidly defined
2. It should be based on all the items.
3. It should not be unduly affected by extreme items.
4. It should lend itself for algebraic manipulation.
5. It should be simple to understand and easy to calculate.

Absolute and Relative Measures:

There are two kinds of measures of dispersion, namely absolute measure of dispersion and relative measure of dispersion.

Absolute measure of dispersion indicates the amount of variation in a set of values in terms of units of observations. For example, when rainfalls on different days are available in mm, any absolute measure of dispersion gives the variation in rainfall in mm. On the other hand relative measures of dispersion are free from the units of measurements of the observations. They are pure numbers. They are used to compare the variation in two or more sets, which are having different units of measurements of observations. The various absolute and relative measures of dispersion are listed below.

Absolute measure	**Relative measure**
Range	Co-efficient of Range
Quartile deviation	Co-efficient of Quartile deviation
Mean deviation	Co-efficient of Mean deviation
Standard deviation	Co-efficient of variation

Range and coefficient of Range:

Range: The range is the simplest measure of dispersion. It is the rough measure of dispersion. Its measure depends upon the extreme items and not on all the items. It is defined as the difference between the largest and smallest values of the variable.

In symbols, Range = L – S.

Where L = Largest value.

S = Smallest value.

In individual observations and discrete series, L and S are easily identified. In continuous series, the following two methods are followed.

Method 1:

L = Upper boundary of the highest class

S = Lower boundary of the lowest class.

Method 2:

L = Mid value of the highest class.

S = Mid value of the lowest class.

Co-efficient of Range:

$$\text{Co-efficient of Range} = \frac{L - S}{L + S}$$

Example1:

Find the value of range and it's co-efficient for the following data.

7, 9, 6, 8, 11, 10, 4

Solution: Here L=11 and S = 4.

We have

Range = L – S = 11- 4 = 7

Also, $\text{Co-efficient of Range} = \frac{L - S}{L + S}$

$\therefore \quad \text{Co-efficient of Range} = \frac{11 - 4}{11 + 4} = \frac{7}{15} = 0.4667$

Example 2: Calculate range and its co-efficient from the following distribution.

Size:	60-63	63-66	66-69	69-72	72-75
Number:	5	18	42	27	8

Solution:

L = Upper boundary of the highest class = 75

S = Lower boundary of the lowest class = 60

We have

Range = L – S = 75 – 60 = 15

Also, $\text{Co-efficient of Range} = \frac{L - S}{L + S}$

$\therefore \quad \text{Co-efficient of Range} = \frac{75 - 60}{75 + 60} = \frac{15}{135} = 0.1111$

Merits and Demerits of Range:

Merits:

1. It is simple to understand.
2. It is easy to calculate.
3. In certain types of problems like quality control, weather forecasts, share price analysis, etc., range is most widely used.

Demerits:

1. It is very much affected by the extreme items.
2. It is based on only two extreme observations.

3. It cannot be calculated from open-end class intervals.
4. It is not suitable for mathematical treatment.
5. It is a very rarely used measure.

Quartile deviation: The difference between the upper and lower quartiles i.e., Q_3-Q_1 is known as the inter quartile range and half of it i.e., $\frac{1}{2}(Q_3 - Q_1)$ is called the semi inter quartile range or the quartile deviation (Q.D).

$$\text{Quartile deviation} = \frac{1}{2}(Q_3 - Q_1)$$

It is better measure of dispersion than range. By eliminating the lowest 25% and highest 25% of items in a series, we are left with the central 50%, which are ordinarily free of extreme values.

$$\text{Co - efficient of Quartile deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Example: Find the Quartile Deviation and coefficient of quartile deviation for the following data:
391, 384, 591, 407, 672, 522, 777, 733, 1490, 2488

Solution:

Arrange the given values in ascending order.
384, 391, 407, 522, 591, 672, 733, 777, 1490, 2488.

We have

$$Q_1 = \left(\frac{N+1}{4}\right)th\ item$$

$$\therefore \quad Q_1 = \left(\frac{10+1}{4}\right) = 2.75th\ item$$

Q1 = 2nd value + 0.75 (3rd value – 2nd value)

= 391 + 0.75 (407 – 391)

= 391 + 0.75 × 16

= 391 + 12

= 403

We have

$$Q_3 = 3\left(\frac{N+1}{4}\right)th\ item$$

$$\therefore \quad Q_3 = 3\left(\frac{10+1}{4}\right) = 8.25th\ item$$

Q3 = 8th value + 0.25 (9th value – 8th value)

= 777 + 0.25 (1490 – 777)

= 777 + 0.25 (713)

= 777 + 178.25

= 955.25

We know that

$$\text{Quartile deviation} = \frac{1}{2}(Q_3 - Q_1)$$

$$\therefore \quad \text{Quartile deviation} = \frac{1}{2}(955.25 - 403) = \frac{552.25}{2} = 276.12$$

Also,

$$\text{Co - efficient of Quartile deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

$$\therefore \quad \text{Co - efficient of Quartile deviation} = \frac{955.25 - 403}{955.25 + 403} = \frac{552.25}{1358.25} = 0.41$$

Example: Calculated the Quartile Deviation and Coefficient of Quartile deviation from the following data:

Age in years:	20	30	40	50	60	70	80
No. of members:	3	61	132	153	140	51	3

Solution:

Age in years	No. of members	c.f
20	3	3
30	61	64
40	132	196
50	153	349
60	140	489
70	51	540
80	3	543
Total	N=543	

We have

$$Q_1 = value\ of\ \left(\frac{N+1}{4}\right)th\ item$$

$$\therefore \quad Q_1 = value\ of\left(\frac{543+1}{4}\right)th\ item = value\ of\ \frac{544}{4}th\ item = value\ of\ 136th\ item = 40\ years$$

We have

$$Q_3 = value\ of\ 3\left(\frac{N+1}{4}\right)th\ item$$

$$\therefore \quad Q_3 = 3\left(\frac{543+1}{4}\right) = value\ of\ 3\left(\frac{544}{4}\right)th\ item = value\ of\ 3(136)th\ item = value\ of\ 408th\ item = 60\ years$$

We know that

$$\text{Quartile deviation} = \frac{1}{2}(Q_3 - Q_1)$$

$$\therefore \quad \text{Quartile deviation} = \frac{1}{2}(60-40) = \frac{20}{2} = 10$$

Also,

$$\text{Co - efficient of Quartile deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

$$\therefore \quad \text{Co - efficient of Quartile deviation} = \frac{60-40}{60+40} = \frac{20}{100} = 0.20$$

Example: Calculated the Quartile Deviation and Coefficient of Quartile deviation from the following data:

Marks:	0-5	5-10	10-15	15-20	20-25	25-30
No. of students:	4	6	8	12	7	2

Solution:

Marks	No. of students	c.f
0-5	4	4
5-10	6	10
10-15	8	18
15-20	12	30
20-25	7	37
25-30	2	39
Total	N=39	

Here N=39

$$\frac{N}{4}=\frac{39}{4}=9.75$$

$$\therefore \quad Q_1 \quad class \quad is = 5-10$$

We have

$$Q_1 = l+\frac{h}{f}\left(\frac{N}{4}-C\right)=5+\frac{5}{6}(9.75-4)=5+\frac{5(5.75)}{6}=5+\frac{28.75}{6}=5+4.79=9.79$$

$$\frac{3N}{4}=\frac{3(39)}{4}=3(9.75)=29.25$$

$$\therefore \quad Q_3 \quad class \quad is = 15-20$$

We have

$$Q_3 = l+\frac{h}{f}\left(\frac{3N}{4}-C\right)=15+\frac{5}{12}(29.25-18)=15+\frac{5(11.25)}{12}=15+\frac{56.25}{12}=15+4.69=19.69$$

We know that

$$\text{Quartile deviation}=\frac{1}{2}(Q_3-Q_1)$$

$$\therefore \quad \text{Quartile deviation}=\frac{1}{2}(19.69-9.79)=\frac{9.90}{2}=4.95\,Marks$$

Also,

$$\text{Co - efficient of Quartile deviation}=\frac{Q_3-Q_1}{Q_3+Q_1}$$

$$\therefore \quad \text{Co - efficient of Quartile deviation}=\frac{19.69-9.79}{19.69+9.79}=\frac{9.90}{29.48}=0.3358 \;\; Marks$$

Merits and Demerits of Quartile Deviation

Merits:

1. It is Simple to understand and easy to calculate
2. It is not affected by extreme values.
3. It can be calculated for data with open end classes also.

Demerits:

1. It is not based on all the items. It is based on two positional values Q1 and Q3 and ignores the extreme 50% of the items
2. It is not amenable to further mathematical treatment.
3. It is affected by sampling fluctuations.

Mean Deviation or Average Deviation: The range and quartile deviation are not based on all observations. They are positional measures of dispersion. They do not show any scatter of the

observations from an average. The mean deviation is measure of dispersion based on all items in a distribution.

Mean deviation is the arithmetic mean of the deviations of a series computed from any measure of central tendency; i.e., the mean, median or mode, all the deviations are taken as positive i.e., signs are ignored. According to Clark and Schekade, "Average deviation is the average amount scatter of the items in a distribution from either the mean or the median, ignoring the signs of the deviations".

We usually compute mean deviation about any one of the three averages mean, median or mode. Sometimes mode may be ill defined and as such mean deviation is computed from mean and median. Median is preferred as a choice between mean and median. But in general practice and due to wide applications of mean, the mean deviation is generally computed from mean. M.D can be used to denote mean deviation.

If x_i , $i = 1, 2, ..., n$ be n observations, then mean deviation from the average A, (usually mean, mode or mode) is given by

$$\text{Mean deviation} = \frac{1}{n}\sum_{i=1}^{n} |(x_i - A)|$$

If $x_i \mid f_i$, $i = 1, 2, ..., n$ is the frequency distribution, then mean deviation from the average A, (usually mean, mode or mode) is given by

$$\text{Mean deviation} = \frac{1}{N}\sum_{i=1}^{n} f_i \,|(x_i - A)|$$

Where $|(x_i - A)|$ represents the modulus or absolute value of the deviation $(x_i - A)$, where the negative sign is ignored.

Since mean deviation is based on all the observations, it is better measure of dispersion than range or quartile deviation. But the step of ignoring the signs of the deviations $(x_i - A)$ creates artificiality and renders it useless for further mathematical treatment.

Coefficient of mean deviation: Mean deviation calculated by any measure of central tendency is an absolute measure. For the purpose of comparing variation among different series, a relative mean deviation is required. The relative mean deviation is obtained by dividing the mean deviation by the average used for calculating mean deviation.

$$\text{Co-efficient of mean deviation} = \frac{Mean\ deviation}{Mean\ \ or\ \ Median\ or\ \ Mode}$$

If the result is desired in percentage, then

$$\text{Co-efficient of mean deviation} = \frac{Mean\ deviation}{Mean\ \ or\ \ Median\ or\ \ Mode} \times 100$$

Example: Calculate mean deviation from mean and median for the following data:

100,150,200,250,360,490,500,600,671. Also calculate coefficients of mean deviation.

Solution:

We know that

$$\text{Mean} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\therefore \quad \text{Mean} = \frac{1}{9}(100+150+200+250+360+490+500+600+671) = \frac{3321}{9} = 369$$

Now arrange the data in ascending order

100, 150, 200, 250, 360, 490, 500, 600, 671

$$\text{Median} = \text{value of } \left(\frac{n+1}{2}\right)th \text{ item}$$

$$\therefore \quad \text{Median} = \text{value of } \left(\frac{9+1}{2}\right)th \text{ item} = \text{value of } 5th \text{ item} = 360$$

x_i	$\lvert (x_i - mean) \rvert$	$\lvert (x_i - median) \rvert$
100	$\lvert (100-369) \rvert$=269	$\lvert (100-360) \rvert$=260
150	$\lvert (150-369 \rvert$=219	$\lvert (150-360 \rvert$=210
200	$\lvert (200-369 \rvert$=169	$\lvert (200-360 \rvert$=160
250	$\lvert (250-369 \rvert$=119	$\lvert (250-360 \rvert$=110
360	$\lvert (360-369 \rvert$=9	$\lvert (360-360 \rvert$=0
490	$\lvert (490-369 \rvert$=121	$\lvert (490-360 \rvert$=130
500	$\lvert (500-369 \rvert$=131	$\lvert (500-360 \rvert$=140
600	$\lvert (600-369 \rvert$=231	$\lvert (600-360 \rvert$=240
671	$\lvert (671-369 \rvert$=302	$\lvert (671-360 \rvert$=311
	$\sum_{i=1}^{n} \lvert (x_i - mean) \rvert = 1570$	$\sum_{i=1}^{n} \lvert (x_i - median) \rvert = 1561$

$$\text{Mean deviation from mean} = \frac{1}{n}\sum_{i=1}^{n} \lvert (x_i - mean) \rvert = \frac{1570}{9} = 174.44$$

$$\text{Mean deviation from median} = \frac{1}{n}\sum_{i=1}^{n} \lvert (x_i - median) \rvert = \frac{1561}{9} = 173.44$$

$$\text{Co-efficient of mean deviation} = \frac{Mean\ deviation}{Mean} = \frac{174.44}{369} = 0.47$$

$$\text{Co-efficient of mean deviation} = \frac{\textit{Mean deviation}}{\textit{Median}} = \frac{173.44}{360} = 0.48$$

Example: Find out the mean deviation from mean and median from the following series.

Age in years	0-10	10-20	20-30	30-40	40-50
No. of persons	5	8	15	16	6

Also compute co-efficient of mean deviation.

Solution: Calculation of mean and M.D. from mean

Age in years	**Mid value** x_i	**No. of persons** f_i	$d_i = x_i - a = x_i - 25$	$f_i d_i$	$\lvert x_i - \bar{x} \rvert$	$f_i \lvert x_i - \bar{x} \rvert$
0-10	5	5	-20	-100	22	110
10-20	15	8	-10	-80	12	96
20-30	25	15	0	0	2	30
30-40	35	16	10	160	8	128
40-50	45	6	20	120	18	108
		$N = \sum_{i=1}^{n} f_i = 50$		$\sum_{i=1}^{n} f_i d_i = 100$		$\sum_{i=1}^{n} f_i \lvert x_i - \bar{x} \rvert = 472$

We know that

$$\bar{x} = a + \frac{1}{N}\sum_{i=1}^{n} f_i d_i = 25 + \frac{1}{50}(100) = 25 + 2 = 27$$

$$\text{Mean deviation} = \frac{1}{N}\sum_{i=1}^{n} f_i \lvert (x_i - \bar{x}) \rvert = \frac{1}{50}(472) = 9.44\,marks$$

$$\text{Co-efficient of mean deviation} = \frac{\textit{Mean deviation}}{\textit{Mean}} = \frac{9.44}{27} = 0.35$$

Calculation of median and M.D. from median

Age in years	**Mid value** x_i	**No. of persons** f_i	**c.f**	$\lvert x_i - median \rvert$	$f_i \lvert x_i - median \rvert$
0-10	5	5	5	23	115
10-20	15	8	13	13	104
20-30	25	15	28	3	45

30-40	35	16	44	7	112
40-50	45	6	50	17	102
		$N = \sum_{i=1}^{n} f_i = 50$			$\sum_{i=1}^{n} f_i \mid x_i - median \mid = 478$

We have

$$\frac{N}{2} = \frac{50}{2} = 25$$

The cumulative frequency just greater than 25 is 28 and is corresponding class 20-30 is the median class.

$$mediam = L + \frac{h}{f}\left(\frac{N}{2} - C\right)$$

$$\therefore \quad mediam = 20 + \frac{10}{15}\left(\frac{50}{2} - 13\right) = 20 + \frac{2}{3}(12)$$

$$= 20 + 8 = 28\, marks$$

$$\text{Mean deviation} = \frac{1}{N}\sum_{i=1}^{n} f_i \mid (x_i - median) \mid = \frac{1}{50}(478) = 9.56\, marks$$

$$\text{Co - efficient of mean deviation} = \frac{Mean\, deviation}{Median} = \frac{9.56}{28} = 0.34$$

Merits and Demerits of mean deviation:

Merits:

1. It is simple to understand and easy to compute.
2. It is rigidly defined.
3. It is based on all items of the series.
4. It is not much affected by the fluctuations of sampling.
5. It is less affected by the extreme items.
6. It is flexible, because it can be calculated from any average.
7. It is better measure of comparison.

Demerits:

1. It is not a very accurate measure of dispersion.
2. It is not suitable for further mathematical calculation.
3. It is rarely used. It is not as popular as standard deviation.
4. Algebraic positive and negative signs are ignored. It is mathematically unsound and illogical.

Standard Deviation:

Karl Pearson introduced the concept of standard deviation in 1893. It is the most important measure of dispersion and is widely used in many statistical formulae. Standard deviation is also called Root-Mean Square Deviation. The reason is that it is the square–root of the mean of the squared deviation from the arithmetic mean. It provides accurate result. Square of standard deviation is called Variance.

Definition:

It is defined as the positive square-root of the arithmetic mean of the Square of the deviations of the given observation from their arithmetic mean. The standard deviation is denoted by the Greek letter σ (sigma)

Calculation of Standard deviation-Individual Series :

There are two methods of calculating Standard deviation in an individual series.

a) Deviations taken from Actual mean

b) Deviation taken from Assumed mean (short cut method).

Deviation taken from Actual mean:

This method is adopted when the mean is a whole number.

Steps:

1. Find out the actual mean of the series $(\bar{x})$.
2. Find out the deviation of each value from the mean $(x_i - \bar{x})$
3. Square the deviations and take the total of squared deviations $\sum_{i=1}^{n}(x_i - \bar{x})^2$
4. Divide the total $\sum_{i=1}^{n}(x_i - \bar{x})^2$ by the number of observation n i.e., $\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$
5. The square root of $\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$ is standard deviation.

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

Example: Calculate the standard deviation from the following data. 14, 22, 9, 15, 20, 17, 12, 11

Solution:

x_i	$(x_i - \bar{x}) = x_i - 15$	$(x_i - \bar{x})^2$
14	-1	1
22	7	49
9	-6	36
15	0	0

20	5	25
17	2	4
12	-3	9
11	-4	16
$\sum_{i=1}^{n} x_i = 120$		$\sum_{i=1}^{n} (x_i - \bar{x})^2 = 140$

We have

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{8}(120) = 15$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}} = \sqrt{\frac{140}{8}} = \sqrt{17.5} = 4.18$$

UNIT-II

Correlation

In statistics, the word "correlation" has a very specific meaning. Statistical correlation means that, given two variables X and Y measured for each case in a sample, variation in X corresponds (or does not correspond) to variation in Y, and vice versa. That is, extreme values of X are associated with extreme values of Y, and less extreme X values with less extreme Y values. The correlation coefficient (Pearson r) measures the degree of this correspondence.

Correlation and causation

If one variable causally influences a second variable, then we would expect a strong correlation between them. However, a strong correlation could also mean, for example, that they are both causally influenced by a third variable. Therefore a strong observed correlation can suggest a causal connection, but it doesn't per se indicate the direction or nature of that causation.

$X \rightarrow Y$	$X \leftarrow Y$	$X \leftrightarrow Y$	$X \leftarrow A \rightarrow Y$
X influences Y	Y influences X	X and Y influence each other	A influences X and Y

Alternative Explanations for Strong Observed Correlation

Important: Correlation between two variables does not prove X causes Y or Y causes X.
Example: There is a statistical correlation between the temperature of sidewalks in New York City and the number of infants born there on any given day.

Pearson *r*

There is a simple and straightforward way to measure correlation between two variables. It is called the Pearson correlation coefficient (***r***) – named after Karl Pearson who invented it. It's longer name, the *Pearson product-moment correlation*, is sometimes used.

The formula for computing the Pearson *r* is as follows:

$$r = \frac{1}{n-1} \sum \frac{(x_i - \overline{X})(y_i - \overline{Y})}{s_x s_y}$$

The value of *r* ranges between +1 and -1:

- r > 0 indicates a positive relationship of *X* and *Y*: as one gets larger, the other gets larger.
- r < 0 indicates a negative relationship: as one gets larger, the other gets smaller.
- r = 0 indicates no relationship

Let's intuitively consider how this formula works. It starts by subtracting the means from *X* and *Y*, and then multiplying the results. When we subtract the mean from a variable, some of the resulting values will be positive and some negative. When we subtract the means from both *X* and *Y*, that will happen with both variables.

If there is no association between *X* and *Y*, there will be no systematic relationship between $(x_i - \overline{X})$ and $(y_i - \overline{Y})$. Therefore the positive values of one will match up with positive and negative values of the other randomly, and the same with negative values of the first variable. Therefore when we take the sum of $(x_i - \overline{X})(y_i - \overline{Y})$, all these positive and negative results will tend to cancel each other out, making *r* close to 0.

However if two variables are positively associated, then positive values of $(x_i - \overline{X})$ will match up with positive values $(y_i - \overline{Y})$, and negative values with negative values. The sum of $(x_i - \overline{X})(y_i - \overline{Y})$ will produce a positive *r*.

In a negative relationship, positive values of $(x_i - \overline{X})$ will match up with negative values of $(y_i - \overline{Y})$, and vice versa. Then the sum of $(x_i - \overline{X})(y_i - \overline{Y})$, and r, will be negative.

Note also that if we calculate the Pearson correlation of X with itself, the result will be 1:

$$r = \frac{1}{n-1}\sum \frac{(x_i - \overline{X})(x_i - \overline{X})}{s_x s_x} = \frac{\frac{\sum (x - \overline{X})^2}{n-1}}{s_x^2} = 1.$$

Computational shortcut

We can rewrite our original formula as:

$$r = \frac{1}{n-1}\sum \left[\frac{(x_i - \overline{X})}{s_x} \times \frac{(y_i - \overline{Y})}{s_y} \right]$$

Recalling the formula for a z score:

$$z = \frac{(x - \overline{X})}{s}$$

we get:

$$r = \frac{1}{n-1}\sum z_x z_y$$

Therefore we can calculate *r* by converting our original *X* and *Y* values into *z*-scores, multiplying z_x and z_y for each case, and dividing the sum of the products by $n - 1$.

Spreadsheet calculation

Pearson correlation calculator

X	Y	X-Xbar	Y-Ybar	z_x	z_y	(z_x)(z_Y)
1	1	-4.5	-4.5	-1.4863	-1.4863	2.2091
2	2	-3.5	-3.5	-1.1560	-1.1560	1.3364
3	3	-2.5	-2.5	-0.8257	-0.8257	0.6818
4	4	-1.5	-1.5	-0.4954	-0.4954	0.2455
5	5	-0.5	-0.5	-0.1651	-0.1651	0.0273
6	6	0.5	0.5	0.1651	0.1651	0.0273
7	7	1.5	1.5	0.4954	0.4954	0.2455
8	8	2.5	2.5	0.8257	0.8257	0.6818
9	9	3.5	3.5	1.1560	1.1560	1.3364
10	10	4.5	4.5	1.4863	1.4863	2.2091

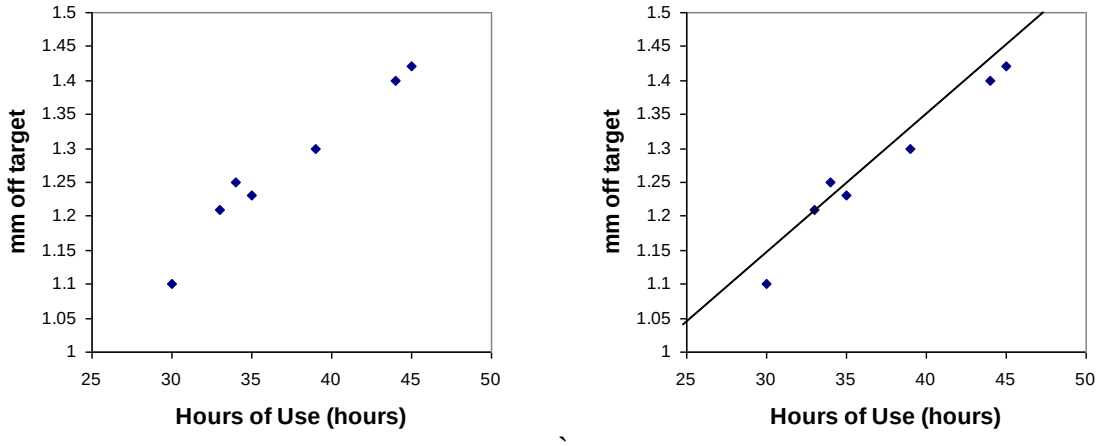
Xbar	5.5
Ybar	5.5
N	10
N-1	9
sd_s X	3.0277
sd_s Y	3.0277
r	1.00

2. Simple Linear Regression

Example

In an automated assembly line, a machine drills a hole in a certain location of each new part being made. Over time, the accuracy of the machine decreases. You have data measured at seven timepoints (hours of machine use) and degree of error (mm from target). You want to know if the data in the x–y **scatter plot** (left) can be fitted with a straight line (right).

The data (machine.xls) can be found on the class webpage.



Why do this?

– to test a hypothesis (is error a linear function of hours of machine use?)
– to predict of error for usage times not observed (interpolation and/or extrapolation)

Regression equation

At it's simplest level, linear regression is a method for fitting a straight line through an *x*-*y* scatter plot.

Recall from other math courses that a straight line is described by the following formula:

$$\hat{y} = bx + a$$

(or, equivalently, $\hat{y} = \beta_1 x + \beta_0$)

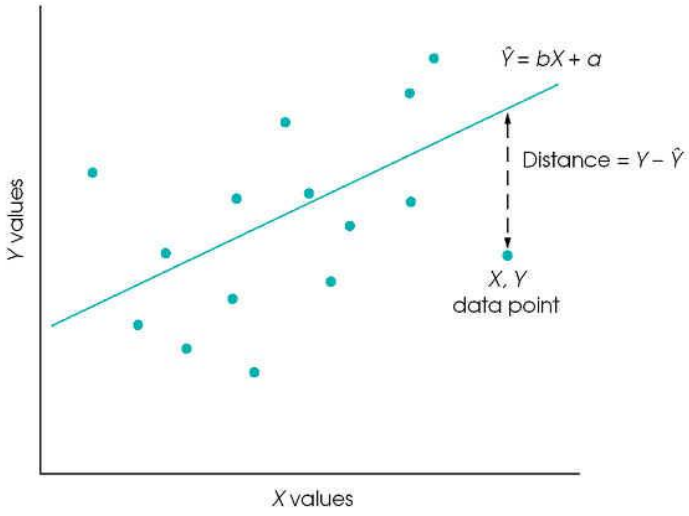
where:

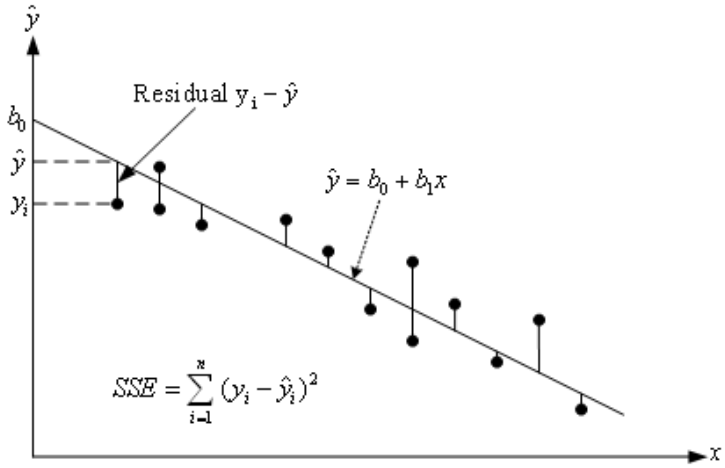
x = a value on the *x* axis
b = **slope** parameter
a = **intercept** parameter (i.e., value on *y* axis where *x* = 0 [not shown above])
$\hat{y}$ = a predicted value of *y*

We can fit infinitely many straight lines through the points. Which is the 'best fitting' line? The criterion we use is to choose those values of *a* and *b* for which our predictive errors (squared) will be minimized. In other words, we will minimize this function:

$$\text{Badness of fit} = \sum (y - \hat{y})^2$$

The difference $(y - \hat{y})$ is called a **residual**, and their sum is called the residual sum of squares or **sum of squared errors** (SSE).





In the figure above, note that instead of *a* and *b* the parameters are called b_0 and b_1.

So we have our criterion for 'best fit'. How do we estimate *a* and *b*? It turns out that we can use calculus to find the values of *a* and *b* that minimize $\sum (y - \hat{y})^2$. When we do so, we discover the following:

$$b = r \frac{s_y}{s_x}$$

where *r* is the Pearson correlation coefficient (which we calculated in the preceding lecture). Once we know *b*, we can find *a*:

$$a = \bar{y} - b\bar{x}$$

where $\bar{x}$ and $\bar{y}$ are the means of x and y, respectively.

Prediction

We now have our linear regression equation. One thing we can do with it is to predict the *y* value for some new value of *x*. For instance, in our original example, the predicated amount of drilling error for a machine after 40 hours of use is:

$$\hat{y} = b \times 40 + a$$

where *a* and *b* are the estimated regression equation coefficients.

The results are in the Parameter Estimates area:

a = Intercept
b = name of variable (e.g., lot size)

Homework

1. Calculate the Pearson *r* for X and Y. Supply <u>all</u> values indicated. Use of Excel encouraged.

X	**Y**	$\mathbf{z_x}$	$\mathbf{z_y}$	$\mathbf{z_x \times z_y}$
1	5	?	?	?
2	2	?	?	?
4	4	?	?	?
5	1	?	?	?

$\bar{X}$ = ?
$\bar{Y}$ = ?
s_x = ?
s_y = ?
n = ?
$\sum(z_x z_y)$ = ?
Pearson r = ?

2. What is the slope of the regression equation predicting Y from X?

3. If X is 6, what is the predicted value of Y? (Show formula and answer.)

How can we explore the association between two quantitative variables?

An **association** exists between two variables if a particular value of one variable is more likely to occur with certain values of the other variable.

For higher levels of energy use, does the CO_2 level in the atmosphere tend to be higher? If so, then there is an association between energy use and CO_2 level.

Positive Association: As x goes up, y tends to go up.
Negative Association: As x goes up, y tends to go down.

Correlation and Regression

How can we explore the relationship between two quantitative variables?

Graphically, we can construct a scatterplot.

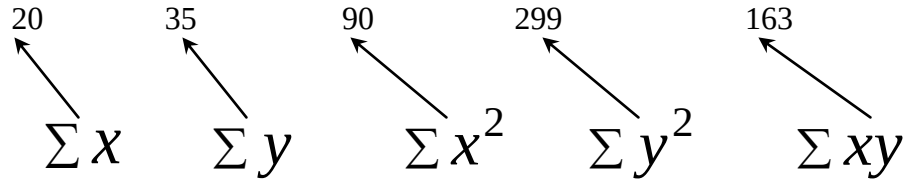
Numerically, we can calculate a correlation coefficient and a regression equation.

The Pearson correlation coefficient, *r,* measures *the **strength** and the **direction** of a straight-line relationship*.

•The ***strength*** of the relationship is determined by the *closeness of the points to a straight line.*
•The ***direction*** is determined by whether one variable generally increases or generally decreases when the other variable increases.
•***r*** is always between –1 and +1
•**magnitude** indicates the strength
•***r* = –1 or +1** indicates a perfect linear relationship
•**sign** indicates the direction
•***r* = 0** indicates no linear relationship

The following data were collected to study the relationship between the sale price, y and the total appraised value, x, of a residential property located in an upscale neighborhood.

Property	x	y	x^2	y^2	xy
1	2	2	4	4	4
2	3	5	9	25	15
3	4	7	16	49	28
4	5	10	25	100	50
5	6	11	36	121	66



Pearson correlation coefficient, r.

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

Example Among all elementary school children, the relationship between the number of cavities in a child's teeth and the size of his or her vocabulary is strong and positive.

Number of cavities and vocabulary size are both related to age.

Example Consumption of hot chocolate is negatively correlated with crime rate.

Both are responses to cold weather.

Correlation analysis

Correlation analysis measures the degree of *linear* association between two variables.

The method is based on the following relation:

$$r = \frac{n\Sigma XY - \Sigma X\Sigma Y}{\sqrt{(n\Sigma X^2 - (\Sigma X)^2)\,(n\Sigma Y^2 - (\Sigma Y)^2)}} \qquad (2.1)$$

There are a number of slightly different versions of the above formula but they all give the same result. Remember that X is the independent variable and Y is the dependent variable. $\sum$ is the summation symbol. In other words, $\sum X$ is the sum of all the values of the independent variable and ΣX^2 is the sum of all the squared values of the independent variable. n is the number of observations (the number of data points in the sample).

Application of the above formula will produce the value of *r*. *r* is known as the *correlation coefficient* and its value determines the strength and direction of linear association between the two variables under examination. In other words, the value of *r* will tell us whether there is a relationship between the two variables and how strong that relationship is. If there is a relationship, then the value of *r* will also indicate whether the value of the dependent variable increases or decreases as the value of the independent variable goes up.

The value of *r* lies between –1 and 1, with –1 indicating a *perfect negative* linear relationship and 1 indicating a *perfect positive* linear relationship. A value around zero indicates no linear relationship between the two variables. In other words, values of the correlation coefficient near –1 or 1 indicate a strong correlation between the two variables, whereas values of the correlation coefficient near zero indicate no correlation between the two variables.

Now let's use correlation analysis to examine the strength and direction of the relationship between sales and advertising cost in the advertising cost example used earlier on in this section. Table 2.1 is re-printed below. As the figures involved are very large, we have divided all sales figures by 1000 and all advertising cost figures by 100. This is a useful way to avoid dealing with calculations involving very large numbers. Although, Excel can handle data of any size, the idea here is to keep the complexity of the calculations low so that it is clear how the method works.

Table 2.2 Advertising cost data

Company	Sales (000s)	Advertising cost (00s)
A	25	] 8
B	35	12
C	29	11
D	24	5
E	38	14
F	***12***	***3***
G	18	6
H	27	8
I	17	4
J	30	9

The following table shows how the calculations have been carried out. Pay particular attention to the difference between ΣX^2 and $(\Sigma X)^2$ (which also applies to the dependent variable).

Table 2.3 Sales vs Advertising cost correlation calculations

Y	**X**	**Y^2**	**X^2**	**XY**
25	8	625	64	200
35	12	1225	144	420
29	11	841	121	319
24	5	576	25	120
38	14	1444	196	532
12	3	144	9	36
18	6	324	36	108
27	8	729	64	216
17	4	289	16	68
30	9	900	81	270
255	**80**	**7097**	**756**	**2289**

Therefore:

$n = 10$	$\Sigma X^2 = 756$
$\Sigma X = 80$	$\Sigma Y^2 = 7097$
$\Sigma Y = 255$	$(\Sigma X)^2 = (80)^2 = 6400$
$\Sigma XY = 2289$	$(\Sigma Y)^2 = (255)^2 = 65025$

Substituting the above results into relation 2.1 we will get:

$$r = \frac{10 \times 2289 - (80 \times 255)}{\sqrt{(10 \times 756 - (80)^2) \times (10 \times 7097 - (255)^2}}$$

$$= \frac{22890 - 20400}{\sqrt{(7560 - 6400) \times (70970 - 65025)}}$$

$$= \frac{2490}{\sqrt{1160 \times 5945}} = \frac{2490}{2626.06}$$

$$= 0.9482$$

Note that Excel has a very useful function which can automatically calculate the value of *r*. This together with a number of other relevant functions are listed in the last part of this section.

The value of the correlation coefficient has been found to be 0.95 (rounded up to two decimal places) indicating that there is a strong positive correlation between advertising cost and the volume of sales. This confirms the findings from the scatter diagram (graph 2.1), which indicated that the volume of sales increases linearly with advertising cost.

Always keep in mind that a low correlation coefficient value does not necessarily mean that there is no relationship between the two variables. All it says is that there is no *linear* relationship between the variables - there may be a strong relationship but of a *non-linear* kind (this will be discussed further later in this section).

Regression

We've seen how to explore the relationship between two quantitative variables graphically with a scatterplot. When the relationship has a straight-line pattern, the Pearson correlation coefficient

describes it numerically. We can analyze the data further by finding an equation for the straight line that best describes the pattern. This equation predicts the value of the response(y) variable from the value of the explanatory variable.

Much of mathematics is devoted to studying variables that are deterministically related. Saying that x and y are related in this manner means that once we are told the value of x, the value of y is completely specified. For example, suppose the cost for a small pizza at a restaurant if $10 plus $.75 per slice. If we let x= # toppings and y = price of pizza, then y=10+.75x. If we order a 3-topping pizza, then y=10+.75(3)=12.25

There are many variables x and y that would appear to be related to one another, but not in a deterministic fashion. Suppose we examine the relationship between x=high school GPA and Y=college GPA. The value of y cannot be determined just from knowledge of x, and two different students could have the same x value but have very different y values. Yet there is a tendency for those students who have high (low) high school GPAs also to have high(low) college GPAs. Knowledge of a student's high school GPA should be quite helpful in enabling us to predict how that person will do in college.

Regression analysis is the part of statistics that deals with investigation of the relationship between two or more variables related in a nondeterministic fashion.

Historical Note: The statistical use of the word regression dates back to Francis Galton, who studied heredity in the late 1800's. One of Galton's interests was whether or not a man's height as an adult could be predicted by his parents' heights. He discovered that it could, but the relationship was such that very tall parents tended to have children who were shorter than they were, and very short parents tended to have children taller than themselves. He initially described this phenomenon by saying that there was a "reversion to mediocrity" but later changed to the terminology "regression to mediocrity."

The least-squares line is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.

Equation for Least Squares (Regression) Line

$$\hat{y} = \hat{\beta}_o + \hat{\beta}_1 x$$

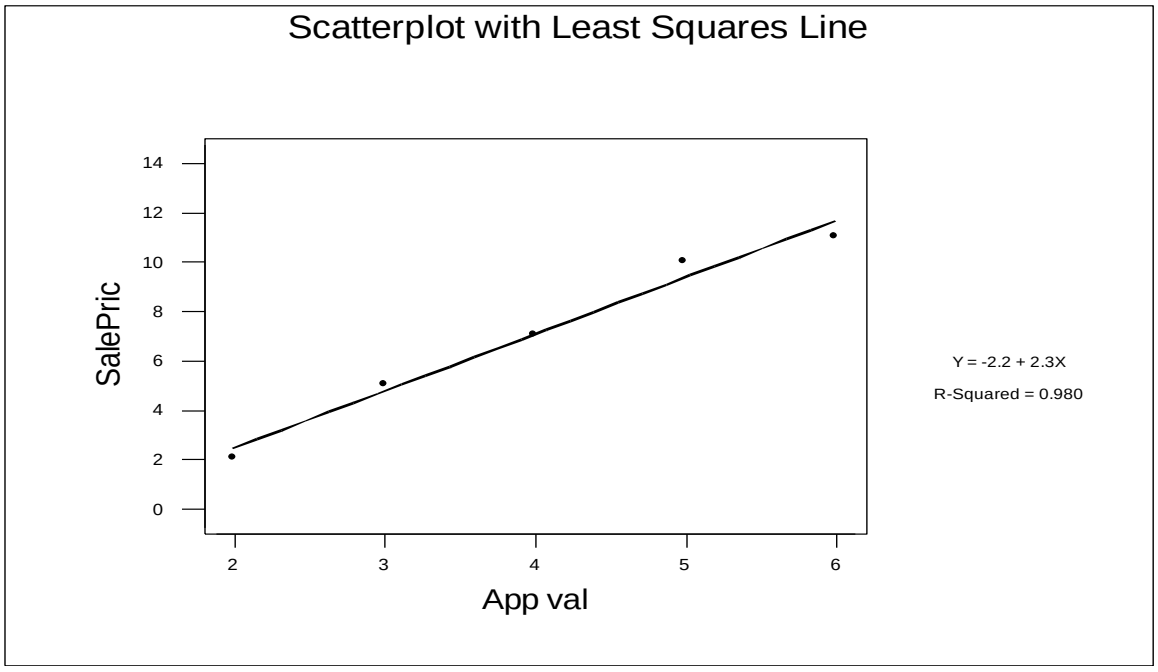
$\hat{\beta}_1$ denotes the slope. The slope in the equation equals the amount that $\hat{y}$ changes when x increases by one unit.

$$\hat{\beta}_1 = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$\hat{\beta}_0$ denotes the y-intercept. The y-intercept is the predicted value of y when x=0. The y-intercept may not have any interpretive value. If the answer to either of the two questions below is no, we do not interpret the y-intercept.

1. Is 0 a reasonable value for the explanatory variable?
2. Do any observations near x=0 exist in the data set?

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$



Equation for Least Squares Line : $\hat{y}$ = -2.2 + 2.3x

Appraisal Value, x $100,000	Sale Price, y $100,000	$\hat{y}$	(y - $\hat{y}$)	(y - $\hat{y}$)^2
2	2	2.4	-.4	.16
3	5	4.7	.3	.09
4	7	7	0	0
5	10	9.3	.7	.49
6	11	11.6	-.6	.36
				$\sum(y - \hat{y})^2$ =

1.1

****

The method of least squares chooses the prediction line $\hat{y} = \hat{B}_o + \hat{B}_1 x$ that minimizes the sum of the squared errors of prediction $\sum(y - \hat{y})^2$ for all sample points.
****

When talking about regression equations, the following are terms used for x and y
x: predictor variable, explanatory variable, or independent variable
y: response variable or dependent variable

Extrapolation is the use of the least-squares line for prediction outside the range of values of the explanatory variable x that you used to obtain the line. Extrapolation should not be done!

When the correlation coefficient indicates no linear relation between the explanatory and response variables, and the scatterplot indicates no relation at all between the variables, then we use the mean value of the response variable as the predicted value so that $\hat{y} = \bar{y}$.

Measuring the Contribution of x in Predicting y

We can consider how much the errors of prediction of y were reduced by using the information provided by x.

r^2 (Coefficient of Determination) = $\dfrac{\sum(y - \bar{y})^2 - \sum(y - \hat{y})^2}{\sum(y - \bar{y})^2}$

The coefficient of determination can also be obtained by squaring the Pearson correlation coefficient. This method works only for the linear regression model $\hat{y} = \hat{\beta}_o + \hat{\beta}_1 x$. The method does not work in general.

The coefficient of determination, r^2, represents the proportion of the total sample variation in y (measured by the sum of squares of deviations of the sample y values about their mean $\bar{y}$) that is explained by (or attributed to) the linear relationship between x and y.

Appraisal Value, x $100,000	Sale Price, y $100,000	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$	$(y - \bar{y})^2$
2	2	2.4	-.4	.16	25
3	5	4.7	.3	.09	4
4	7	7	0	0	0
5	10	9.3	.7	.49	9
6	11	11.6	-.6	.36	16
				1.1	54

r^2 (Coefficient of Determination) = $\frac{\sum(y-\bar{y})^2 - \sum(y-\hat{y})^2}{\sum(y-\bar{y})^2} = \frac{54-1.1}{54} = .98$

Interpretation: 98% of the total sample variation in y is explained by the straight-line relationship between y and x, with the total sample variation in y being measured by the sum of squares of deviations of the sample y values about their mean $\bar{y}$.

Interpretation: An r^2 of .98 means that the sum of squares of deviations of the y values about their predicted values has been reduced 98% by the use of the least squares equation $\hat{y}$ = -2.2 + 2.3x, instead of $\bar{y}$, to predict y.

The coefficient of determination is a number between 0 and 1, inclusive. That is, $0 \leq r^2 \leq 1$ If $r^2 = 0$, the least squares regression line has no explanatory value. If $r^2 = 1$, the least-squares regression line explains 100% of the variation in the response variable.

REGRESSION ANALYSIS

Introduction

Regression analysis can be defined as the process of developing a mathematical model that can be used to predict one variable by using another variable or variables. This section first covers the key concepts of two common approaches to data analysis: *graphical data analysis* and *correlation analysis* and then introduces the two main types of regression: *linear regression* and *non-linear regression*. The section also introduces a number of *data transformations* and explains how these can be used in regression analysis.

When you have worked through this section, you should be able to:

- Distinguish between a dependent variable and an independent variable and analyse data using graphical means.
- Examine possible relationships between two variables using graphical analysis and correlation analysis.
- Develop simple linear regression models and use them as a forecasting tool.
- Understand polynomial functions and use non-linear regression as a forecasting tool.
- Appreciate the importance of data transformations in regression modelling.

Advertising cost example

It is well known that some form of advertising for a particular product will be associated with and have an effect on its sales. Numerical data has been collected from ten companies on their monthly volume of sales of a particular product as well as their cost of advertising for that product. This data is shown in table 2.1. We want to develop an appropriate regression model that will be based on this data and could be used to predict the volume of sales for a particular company, given that company's advertising cost.

Table 2.1 Advertising cost data

Company	Sales	Advertising cost (£)
A	25000	800
B	35000	1200
C	29000	1100
D	24000	500
E	38000	1400
F	***12000***	***300***
G	18000	600
H	27000	800
I	17000	400
J	30000	900

In this example we have two *variables*, sales and advertising cost, and numerical data has been collected from a number of companies. The first thing to note is the distinction between this data set and the entire population of companies. What we have here is just a small *sample* taken from the entire *population* of companies selling the particular product.

The idea is to use the data from the given sample in order to develop a regression model that could then be used to predict the volume of sales for a particular company based on that company's expenditure on advertising.

The regression model to be developed will relate the volume of sales to advertising cost. As we expect the volume of sales to depend on the cost of advertising, we take sales to be the *dependent* variable and advertising cost to be the *independent* variable.

Before we start developing the regression model we should first make sure that a relationship exists between advertising cost and the volume of sales. If such a relationship does not exist between the two variables, then there is no point in developing a regression model. Although a regression model could still be developed easily, that model wouldn't be able to produce accurate forecasts and therefore make any significant contribution to decision making.

The relationship between two variables can be tested graphically using a *scatter diagram* or statistically using *correlation analysis*. The results from the analysis of data will tell us whether to use regression analysis as the forecasting tool and what type of regression model to develop.

Scatter diagrams

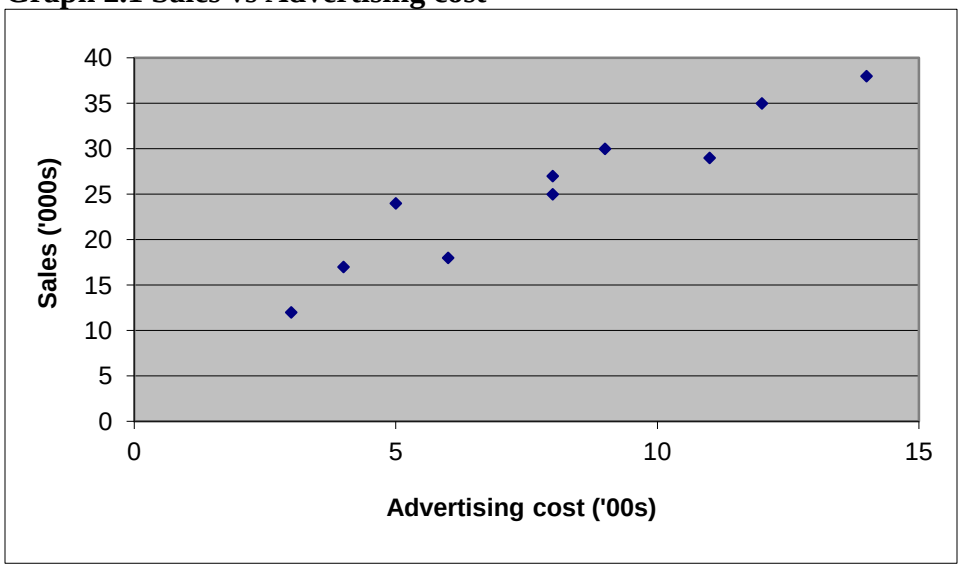
The relationship between two variables can be examined graphically using a *scatter diagram*. A scatter diagram is a simple two-dimensional graph of the values of the dependent variable and the independent variable.

The important thing to remember when drawing a scatter diagram is that the dependent variable is always drawn on the vertical axis of the diagram and that the independent variable is always drawn on the horizontal axis of the diagram. The dependent variable is usually represented by Y

and the independent variable is usually represented by X. This is the notation to be used throughout the course.

The following scatter diagram shows the volume of sales against advertising cost for the advertising cost data shown in table 2.1. The last part of this section explains how this diagram was produced on Excel.

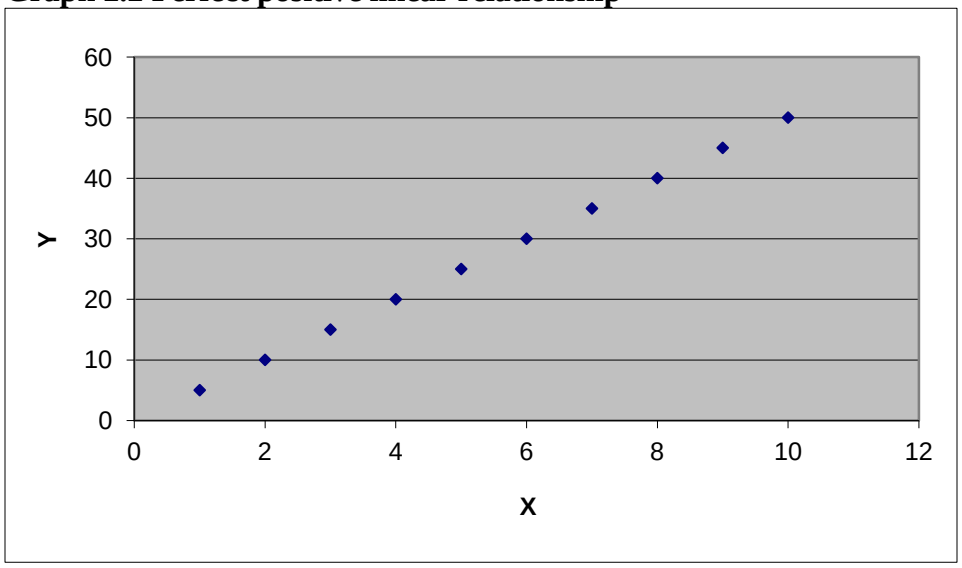
Graph 2.1 Sales vs Advertising cost



Looking at the above diagram we can see that high volumes of sales are associated with high advertising costs and that low volumes of sales are associated with low advertising costs. In other words, a relationship exists between the two variables, with the volume of sales increasing as the advertising cost goes up. As this increase is linear (i.e. the value of Y increases with the value of X in a linear way), the relationship between the two variables is a linear *relationship.*

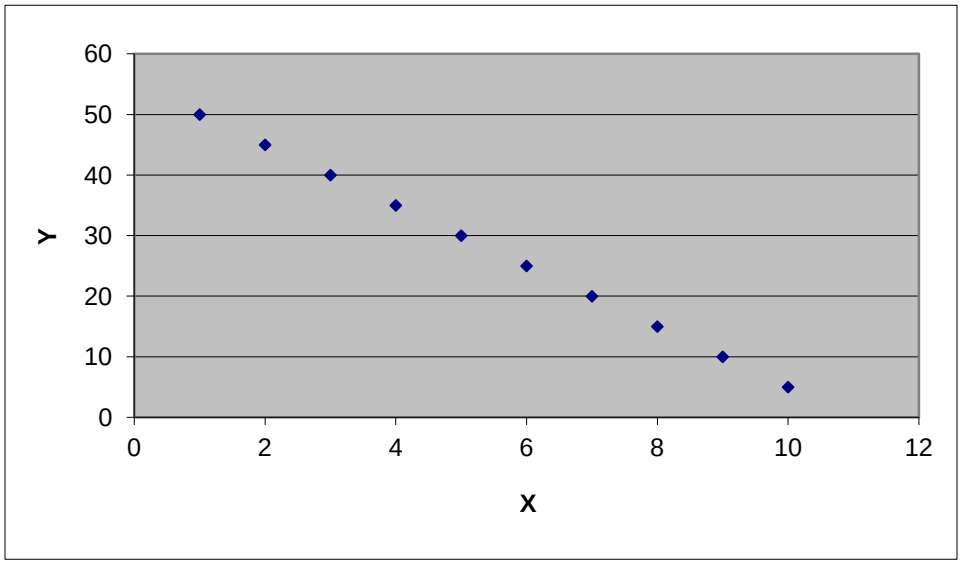
Now consider the scatter diagrams shown in graphs 2.2 and 2.3.

Graph 2.2 Perfect positive linear relationship



Graph 2.2 indicates a *perfect* linear relationship between the two variables, as all the data points on the graph have fallen exactly on a straight line. The relationship is also *positive* as the value of Y increases as the value of X goes up.

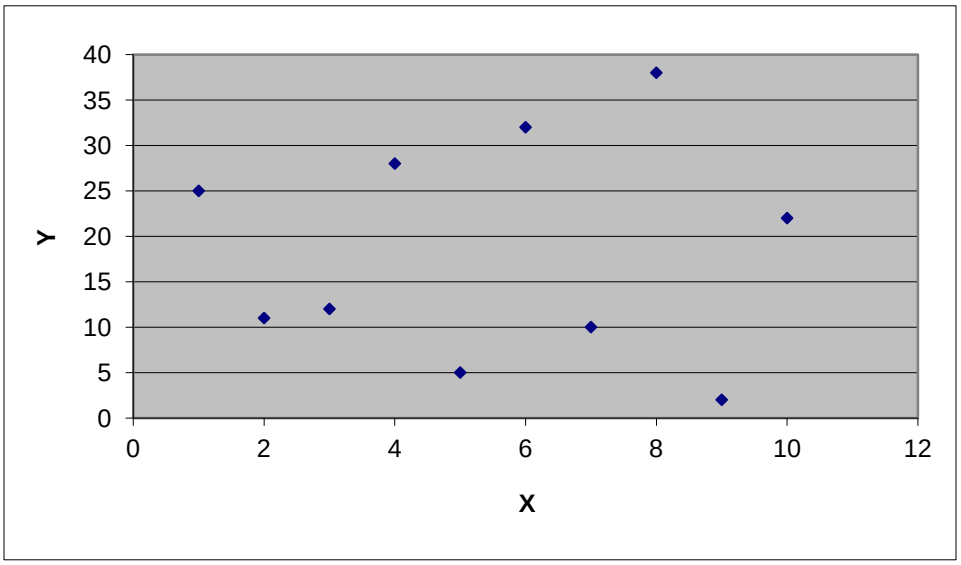
Graph 2.3 Perfect negative linear relationship



Graph 2.3 indicates another *perfect* linear relationship between the two variables, as all the data points on the graph have again fallen exactly on a straight line. This time, however, the relationship is *negative* as the value of Y decreases as the value of X goes up.

Finally, graph 2.4 shows a case of no relationship between the two variables. In such a case regression analysis would fail to produce accurate forecasts and therefore to make a contribution to decision making.

Graph 2.4 No relationship

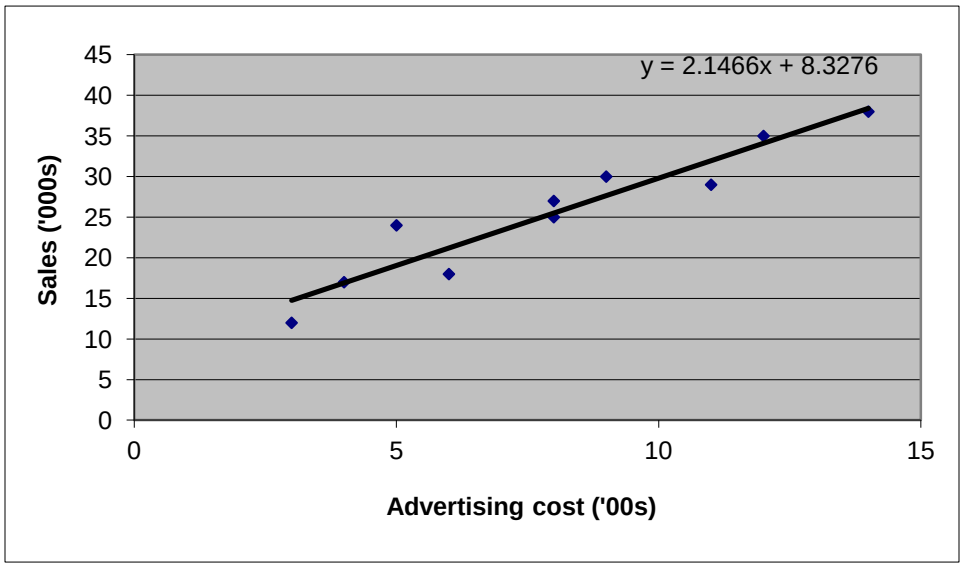


Developing a linear regression model

If we look at the advertising cost example we can see that both the scatter diagram (graph 2.1) and the value of the correlation coefficient (0.95) indicate that a strong linear relationship exists between advertising cost and sales. We can then use linear regression to describe that linear relationship.

The following graph shows the volume of sales against advertising cost with a straight line fitted on it. This line is called *regression line* and it is the result of using regression analysis. We shall now describe the process that produced this line.

Graph 2.5 Regression line for advertising cost example



As you can see from the above diagram, regression has fitted a straight line on the data. In fact that regression line has been fitted in such a way so that the sum of the distances between the data points and the line (i.e. the gaps between the data points and the line) is minimised. Because of this, the regression line is also known as the *line of best fit*.

Regression therefore aims to fit a line through the data in order to describe the relationship between two variables. If the relationship between the two variables is linear (like the one in this example), then a straight line is fitted through the data and the data points will lie very close to that line.

Obviously, we could visually draw a straight line through the data points of the scatter diagram in an attempt to fit the line to the points as closely as possible. The problem with this approach however is that, no matter how good our fit is, one could come up with a better fit. What we should do instead is to fit the regression line using a more statistical approach, which is known as the *least squares regression* method.

According to the least squares regression method, a regression line is fitted through the data in such a way so that the sum of the squares of the distances between the data points and the line is

minimised. The resulting regression line could be straight or curved depending on the type of the relationship between the two variables.

A linear regression model is based on the linear function shown in relation 2.2.

$$\text{Predicted } Y = b_0 + b_1X \qquad (2.2)$$

The parameter b_0 is called the *intercept* and the parameter b_1 is called the *slope* of the regression line. The value of the intercept determines the point where the regression line meets the Y axis of the graph. The value of the slope represents the amount of change in Y when X increases by one unit.

Another name frequently used in regression analysis to refer to the independent variable is *predictor*. Note that the above regression model uses only one predictor and is therefore called a *simple* regression model. A model which uses more than one predictor is called a *multiple* regression model. Relation 2.3 shows the general form of a multiple regression model with *k* predictors.

$$\text{Predicted } Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_kX_k \qquad (2.3)$$

The regression line which appears on graph 2.5 has been produced on Excel (the way that this is done on Excel is explained in the last part of this section). Note here that Excel displays the regression equation in the form $Y = b_1X + b_0$. In other words, the order in which the values of the intercept and the slope appear is different to the one shown in relation 2.2, and Y actually refers to the predicted value of the response variable. To avoid any confusion, in this course we will use the regression equation exactly as it is shown in relation 2.2 but we will leave the equations on the graphs exactly as Excel displays them.

In order to develop a linear regression model of the form Predicted $Y = b_0 + b_1X$ we need to calculate the values of b_0 and b_1. These values are given by the following relations:

$$b_1 = \frac{n\Sigma XY - \Sigma X\Sigma Y}{n\Sigma X^2-(\Sigma X)^2} \qquad (2.4)$$

$$b_0 = \frac{\Sigma Y}{n} - b_1\frac{\Sigma X}{n} \qquad (2.5)$$

Note that some textbooks use slightly different formulae to calculate the values of b_0 and b_1. This often happens with different textbooks but all the formulae used are mathematically equivalent to those shown above and they will give exactly the same results.
Application of the above formulae to the advertising cost data will produce the following results:

$$b_1 = \frac{10 \text{ x } 2289 - (80 \text{ x } 255)}{(10 \text{ x } 756-(80)^2)}$$

$$= \frac{22890 - 20400}{(7560 - 6400)}$$

$$= \frac{2490}{1160}$$

$$= 2.1466$$

$$b_0 = \frac{255}{10} - 2.1466 \frac{80}{10}$$

$$= 25.5 - 2.1466 \times 8$$

$$= 8.3272$$

Substituting the above values in relation 2.2 will give us the following regression equation:

$$\text{Predicted } Y = 8.3272 + 2.1466X$$

The value of b_0 is 8.3272, which means that the regression line cuts the vertical axis of the graph at that point. Similarly, the value of b_1 is 2.1466 indicating that the value of Y will increase by 2.1466 every time that the value of X increases by 1 (obviously, when X=0, Y=8.3272).

Excel can again calculate the values of b_0 and b_1 very easily and the steps are shown in the last part of this section.

Using the regression model to make predictions

Once a regression model has been developed it can then be used to predict the volume of sales for a company based on its advertising cost.

Suppose that we want to predict the volume of sales for a company which has spent £1000 on advertising. All we need to do is take this to be the value of X in the regression model and then calculate the corresponding value of Y. Note however that, as we have divided all advertising cost figures by 100 in order to make the figures more manageable, we also need to do the same with the new figure. Therefore, the value to be substituted in the regression model should be 10 (rather than 1000). The predicted volume of sales can therefore be calculated as follows:

$$\textit{Predicted } Y = 8.3272 + 2.1466\ (10) = 29.7932$$

The above result is the predicted value of sales for a company which has spent £1000 on advertising. Note that this figure should now be multiplied by 1000 in order to be converted back

to the same units as the original data (this is again because we have divided all sales figures by 1000 in order to make the figures more manageable). Thus, a company which spends £1000 on advertising for a particular product is expected to sell 29,793 units of that product.

Note that to make this prediction we used an X value (1000) from the existing range of values of the X variable (300-1200). In general, it is too risky to attempt to predict a value of Y using an X value which is outside the range of X values of the data collected. That is because the linear relationship that exists between the two variables only covers the existing data and this could change if another range of values was considered.

Also note that the above prediction is based on the regression model, which is itself based on the data obtained from the ten companies. In other words, the regression model and any forecasts produced by that model are all based on sample data. Had a different sample been used, the regression model produced would have been different. This will be discussed further in the next section.

So far we have looked at how regression analysis could be used in situations where a linear relationship exists between two variables. However, there are situations where the two variables might be related in a *non-linear* way. In other words, although the results from correlation analysis have shown that a relationship does not exist between the two variables, these variables might still be closely related (don't forget that correlation analysis measures the strength of *linear* association between the two variables). The forecaster should therefore make sure that the data is always graphed during the data analysis stage. The resulting graph will help the forecaster identify any non-linear patterns that correlation analysis has failed to spot.

If the results from data analysis show that there is a *non-linear* (also known as *curvilinear*) association between the two variables, then there is no point in developing a linear regression model. Although a linear regression model could be developed very easily, such a model would fail to produce a good fit and therefore generate accurate forecasts.

We can handle curved data in two ways: by using a *polynomial* rather than linear regression model, or by *transforming* the data and then using a linear regression model. The two methods are covered in the rest of this section.

Stopping distances example

The American National Bureau of Standards has conducted a series of tests to see how stopping distances of cars are related to automobile speed (this example has been adapted from Ryan et al (1992), *Minitab Handbook*, Duxbury Press; distance has been converted from feet to metres). The data collected is a s follows:

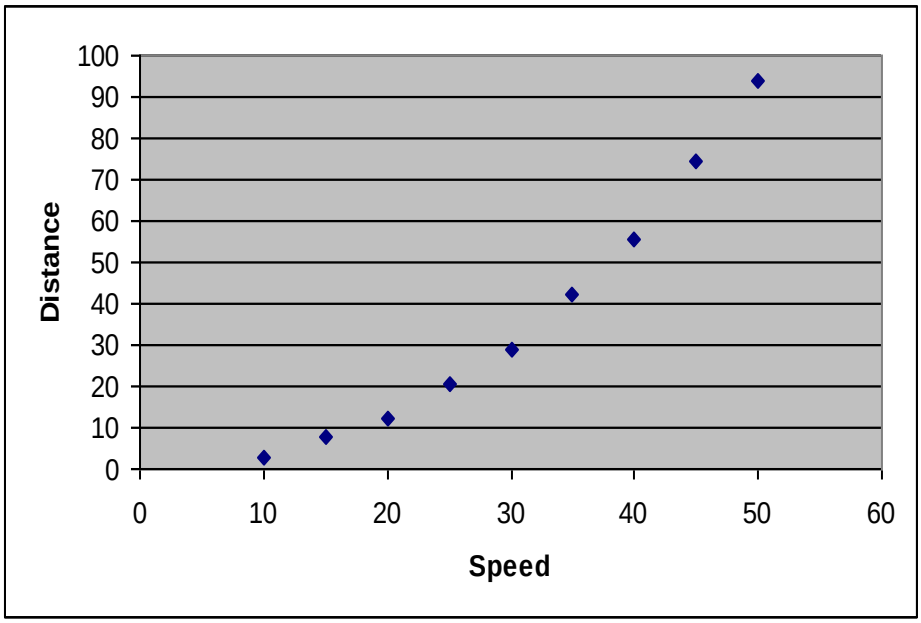
Table 2.4 Stopping distances data

Speed (m/h)	Distance (metres)
10	3.05
15	7.62

20	12.19
25	20.42
30	28.65
35	42.06
40	55.47
45	74.68
50	93.88

As we expect a car's stopping distance to be related to its speed, we take speed as the independent variable (X) and distance as the dependent variable (Y). The data is graphed on the following scatter diagram.

Graph 2.6 Speed vs Distance



Looking at the above graph we could say that the two variables are closely related in a linear way. In fact this is confirmed by correlation analysis (the value of *r* in this case has been calculated to be 0.97, indicating a very strong, almost perfect, positive correlation).

However, if we look at the above graph more carefully, we will notice that the data values appear to form a slight curve. This would make us think whether the use of non-linear regression or data transformations could be a better option. The following sections explain how these approaches can be applied in the stopping distances example.

Fitting polynomial functions

Polynomials are equations that involve powers of the independent variable. Relations (2.6) and (2.7) show a second-degree (*quadratic*) and a third-degree (*cubic*) polynomial functions.

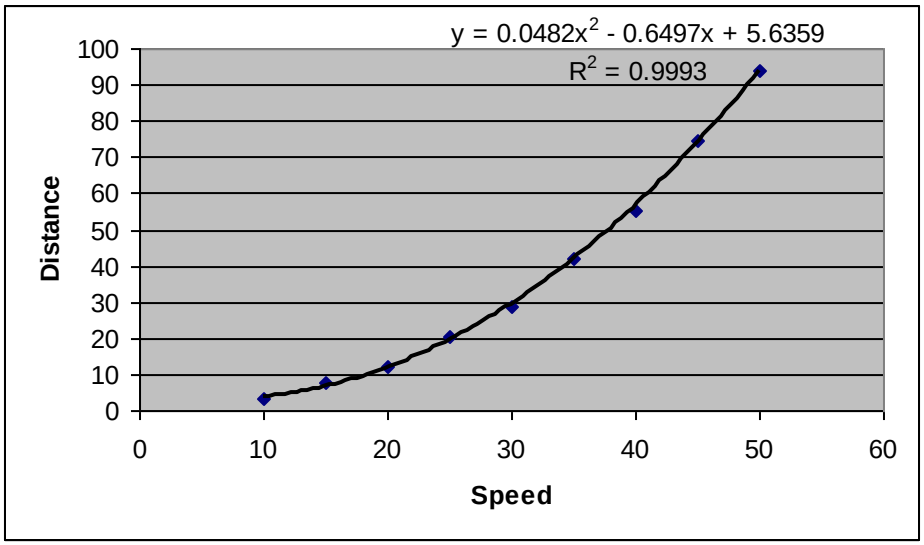
$$\text{Predicted } Y = b_0 + b_1X + b_2X^2 \qquad (2.6)$$

$$\text{Predicted } Y = b_0 + b_1X + b_2X^2 + b_3X^3 \qquad (2.7)$$

The parameter b_0 is the intercept of the regression model and the parameters b_1, b_2 and b_3 are the coefficients of the predictor (note that both models are simple regression models, as they both use only one predictor).

Graph 2.7 shows the same graph with a curve fitted on the data. That curve has been produced by a quadratic non-linear regression model based on relation 2.6. Excel has once again been used to fit the line through the data and the regression equation automatically appears on the scatter diagram (the last part of this section explains how this is done on Excel).

Graph 2.7 Regression line for stopping distances example
(2d degree non-linear regression model)



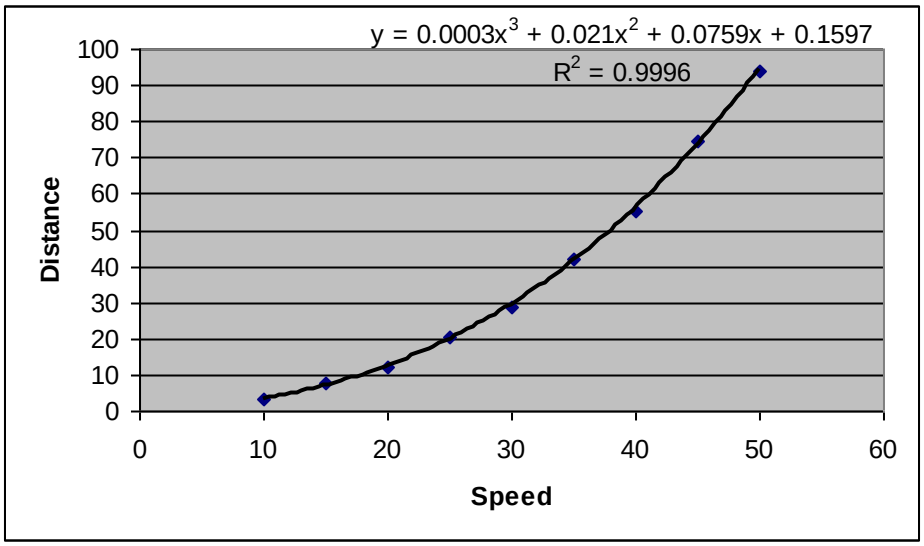
The regression model is as follows (remember that Excel displays the equation in a slightly different way):

$$\text{Predicted } Y = 5.6359 - 0.6497X + 0.0482X^2$$

Excel has also calculated the value of R^2 to be 0.9993. R^2 is a useful statistic, known as the *coefficient of determination*, that will be discussed in the next section. Basically, the nearer the value of R^2 to 1 the better the fit produced by the regression line. This therefore indicates that the above quadratic regression model has produced an excellent fit.

Graph 2.8 shows the result of a cubic non-linear regression model based on relation 2.7.

Graph 2.8 Regression line for stopping distances example
(3d degree non-linear regression model)

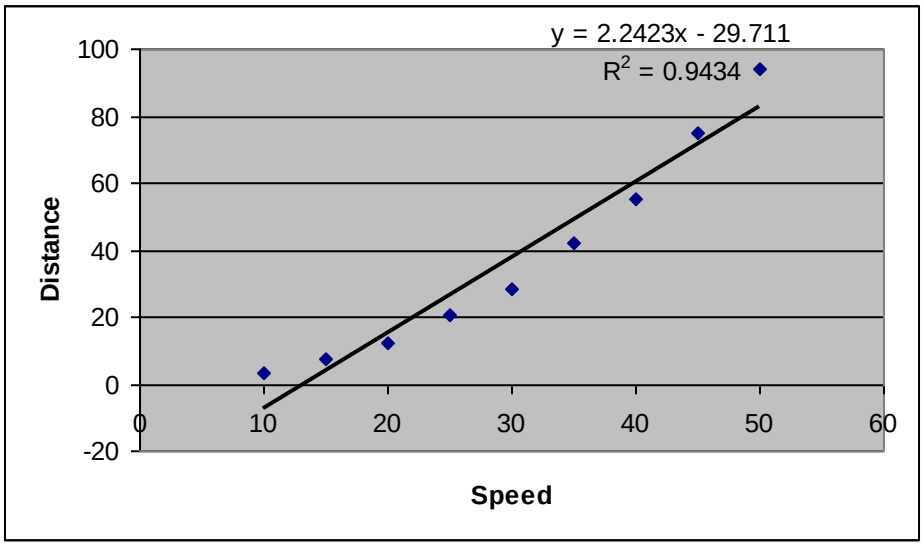


The regression model is as follows:

$$\text{Predicted } Y = 0.1597 + 0.0759X + 0.021X^2 + 0.0003X^3$$

The value of R^2 is slightly better than the one produced for the quadratic regression model, indicating that the cubic model has produced an even better fit. Had a linear regression model been used instead the results would have been as follows:

Graph 2.9 Regression line for stopping distances example (linear regression model)



The regression model is as follows:

$$\text{Predicted } Y = -29.711 + 2.2423X$$

The value of R^2 is slightly lower than the ones produced by the two non-linear regression models, indicating that non-linear regression provides better results than linear regression for the

particular case (this is the result of the slight curve in the data values – had the curve been stronger, the value of R^2 produced by the linear regression model would have been much lower).

Using data transformations

Rather than fit a polynomial to curved data, it is often preferable to try to transform the data in order to make the relationship between the two variables more linear and then use a linear regression model as the forecasting tool. Transformations aim to make a non-linear relationship between two variables more linear so that it can be described by a linear (rather than non-linear) regression model.

Of all the transformations made on data in practice, the three most popular are the *square root* ($\sqrt{X}$), the *logarithm* (logX), and the *negative reciprocal* (-1/X). The reason why we use the negative reciprocal (-1/X) rather than the reciprocal (1/X) is because we want to preserve the order of the observations. For example, if 12 is the smallest observation in the data set, then -1/12 will be the smallest observation in the transformed data set. If we used just the reciprocal, then 1/12 would be the largest observation in the transformed data set and everything would be turned around. The way that these three transformations work will be illustrated using the stopping distances example.

We normally start from the square root transformation. If this fails to straighten the curve, we can try the logarithm transformation, which is a stronger one. If this still fails to produce an acceptable outcome, we can then try the negative reciprocal transformation, which is the strongest of the three.

Now refer to the stopping distances example.

The value of the correlation coefficient is 0.97, indicating a very strong, almost perfect, positive correlation. In an attempt to make the relationship between the two variables even more linear we could use the three transformations introduced in this section.

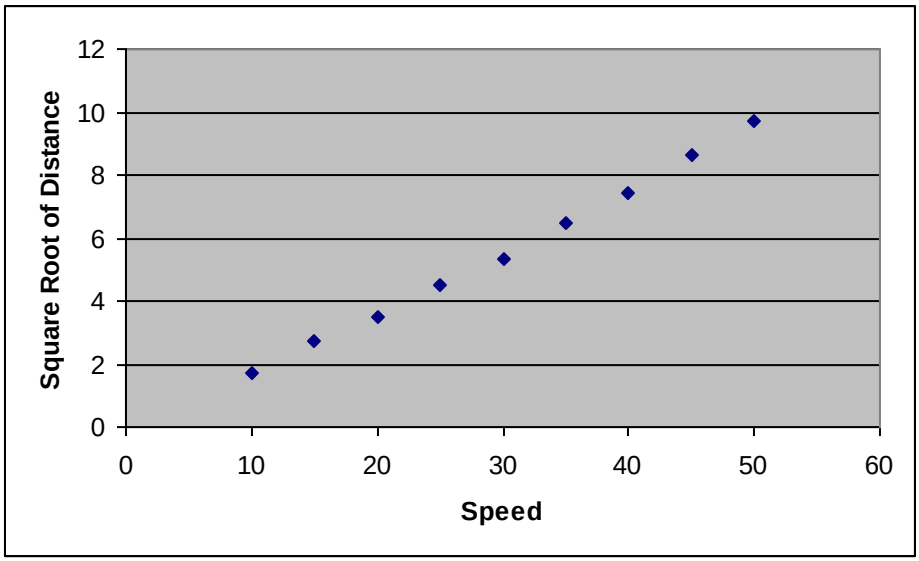
Table 2.5 shows how the square root transformation has been applied on Y (the last part of this section introduces the relevant Excel function).

Table 2.5 Square root transformation

Speed (m/h)	Square root of Distance (metres)
10	1.75
15	2.76
20	3.49
25	4.52
30	5.35
35	6.49
40	7.45
45	8.64
50	9.69

The result of the above transformation can be seen on graph 2.10.

Graph 2.10 Square Root of Distance vs Speed



The square root transformation has increased the value of the correlation coefficient from 0.97 to 0.99 (you could check this if you carry out correlation analysis using relation 2.1)

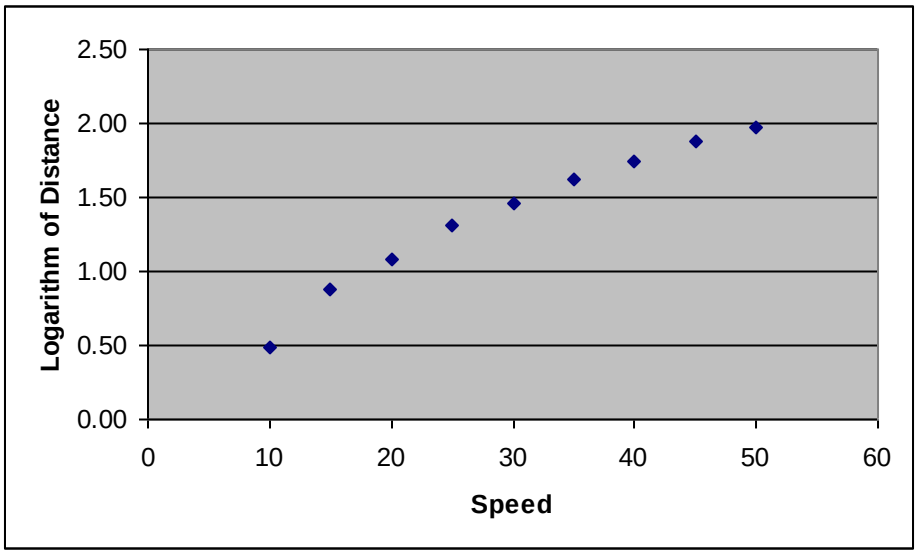
Table 2.6 shows how the logarithm transformation has been applied on Y (the last part of this section introduces the relevant Excel function).

Table 2.6 Logarithm transformation

Speed (m/h)	Logarithm of Distance (metres)
10	0.48
15	0.88
20	1.09
25	1.31
30	1.46
35	1.62
40	1.74
45	1.87
50	1.97

The result of the above transformation can be seen on graph 2.11.

Graph 2.11 Logarithm of Distance vs Speed



The logarithm transformation has increased the value of the correlation coefficient from 0.97 to 0.98. This, however, is not be as good as the result produced by the square root transformation (0.99)).

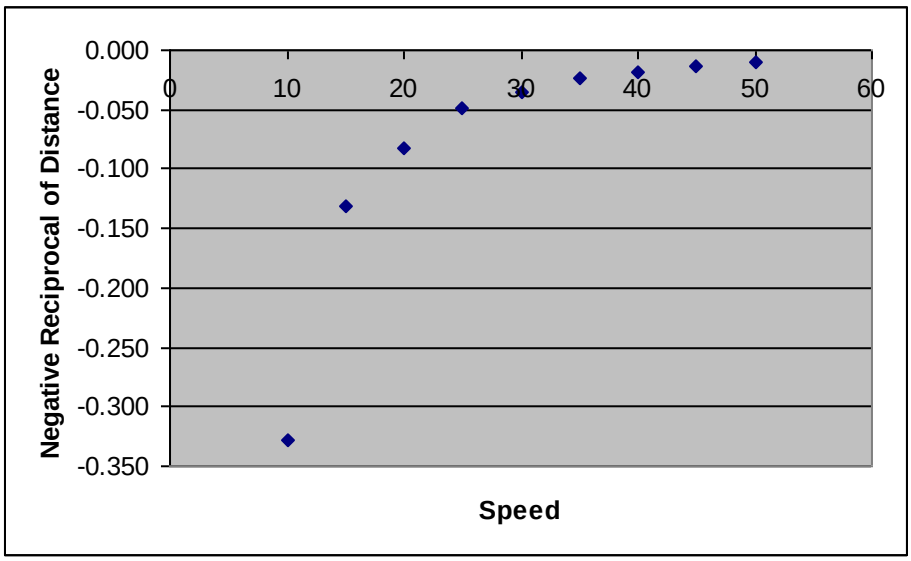
Finally, table 2.7 shows how the negative reciprocal transformation has been applied on Y (the last part of this section introduces the relevant Excel function).

Table 2.7 Negative Reciprocal transformation

Speed (m/h)	Negative Reciprocal of Distance (metres)
10	-0.328
15	-0.131
20	-0.082
25	-0.049
30	-0.035
35	-0.024
40	-0.018
45	-0.013
50	-0.011

The result of the above transformation can be seen on graph 2.12.

Graph 2.12 Negative Reciprocal of Distance vs Speed



The negative reciprocal transformation has decreased the value of the correlation coefficient from 0.97 to 0.79, indicating that this transformation would be inappropriate for this case.

Graphs 2.10-2.12 and the correlation analysis results have indicated that the square root transformation has produced the best results. The logarithm and negative reciprocal transformations have both been found to be too strong for this data set. As the data had only a slight curve, a weak data transformation produced better results than stronger transformations did.

Once the data has been transformed, a linear regression model can be used to relate the distance of a car to its speed. The resulting regression model will therefore be as follows:

$$\sqrt{\text{PredictedY}} = -0.3591 + 0.1977X$$

(this can be produced by applying relations 2.4 and 2.5).

When using the above regression model we should remember that it has been based on transformed data and that the response variable has been expressed as a square root. For example, the stopping distance of a car whose speed is 40 m/h (or 64.4 km/h) will be $\sqrt{\text{PredictedY}} = 7.5489$, that is 56.99 metres (simply square both sides of the equation to remove the square root). Similarly, a car travelling at 75 m/h (or 120.7 km/h) will have a predicted stopping distance of 209.33 metres.

In general, transformations give us some idea of a good theoretical model for the relationship between two variables. Using a polynomial function might be equally good for estimation over the range of the data, but it would probably not work very well outside the range of data.

EXCEL APPLICATIONS

This part aims to familiarise students with some of the basic functions of Excel and to explain how these can be used to carry out the work introduced in this section.

To produce a scatter diagram:

1. Select the values of the two variables.
2. Click on the chart wizart, which can be found towards the top right hand corner of the screen.
3. Choose *XY (scatter)* by clicking on that option.
4. Choose the first chart sub-type by clicking on that option. Then click on *Next*.
5. The scatter diagram appears on the screen. Click on *Next*.
6. Click on *Titles* and enter a chart title and the names of the two variables to appear on the two axes (use the mouse to move from one box to another).
7. Click on *Legend* and click on the *Show legend* option to remove the legend (this is not needed here).
8. Click on *Finish*. The scatter diagram is now complete.

Note that if there are other columns in between the two that you want to select as the two variables, you need to select the values of the first variable, press and hold the control key (at the bottom left hand corner of the keyboard) and then select the values of the second variable.

To calculate the value of the correlation coefficient for two variables (where the values of the first variable appear in cells A1 to A5 and the values of the second variable appear in cells B1 to B5):

- Click on an empty cell and type *=CORREL(A1:A5,B1:B5)*

To fit a straight line on the scatter diagram:

1. Click on any data point on the graph (notice that the colour of all data points changes).
2. Without moving the mouse right-click and choose *Add trendline* by clicking on that option.
3. Click on *Type* and choose the first type of line (linear) by clicking on that option.
4. Click on *Options* and choose *Display equation on chart* by clicking on that option.
5. Click on *OK*. A straight line has now been fitted on the scatter diagram and the regression model which has produced that line is also shown there.

To fit a curve on a scatter diagram:

1. Click on any data point on the graph (notice that the colour of all data points changes).
2. Without moving the mouse right-click and choose *Add trendline* by clicking on that option.
3. Click on *Type* and choose the third type of line (polynomial) by clicking on that option. Leave the order of the polynomial function to 2 for a quadratic regression model or increase it to 3 for a cubic regression model.

4. Click on *Options* and choose *Display equation on chart* and *Display R-squared value on chart* by clicking on the two options.
5. Click on *OK*. A curve has now been fitted on the scatter diagram and the regression model and the value of R^2 are also shown.

Finally, the three data transformations illustrated in this section can be applied as follows:

To calculate the square root of a value which appears in cell A1:

- Click on an empty cell and type *=SQRT(A1)*

To calculate the logarithm of a value which appears in cell A1:

- Click on an empty cell and type *=LOG(A1)*

To calculate the negative reciprocal of a value which appears in cell A1:

- Click on an empty cell and type *=-1/A1*

PROBLEMS

Problem 1

The management of a chain of fast food restaurants wants to investigate the relationship between the daily sales volume of a company restaurant and the number of competitor restaurants within a 1-mile radius. The following data has been collected.

Competitors	**Sales**
1	3600
1	3300
2	3100
3	2900
3	2700
5	2300
5	2000
6	1800

Draw a scatter diagram to examine whether a relationship exists between the number of competitors and the volume of sales. Once your scatter diagram has been produced it should be clearly interpreted.

Problem 2

Refer to the fast food sales data given in problem 1 and use correlation analysis to examine whether a relationship exists between the number of competitors and the volume of sales. Your results should be clearly interpreted.

Problem 3

Refer to the fast food sales data given in problem 1 and develop a linear regression model that would relate the volume of sales to the number of competitors. What is the regression model?

Problem 4

Use the regression model developed in problem 3 to predict the volume of sales if a restaurant has four competitors within a 1-mile radius. Then do the same for a restaurant that has eight competitors within a 1-mile radius. Which of your two predictions do you expect to be more accurate and why?

Problem 5

Experiments were run in several different scientific laboratories to determine the vapour pressure of cadmium as a function of temperature. The data below shows the
results from one of the laboratories:

Temper. (Kelvins)	**Pressure (Millionths of an Atmosphere)**
525	10.1000
501	2.8300
475	0.6370
452	0.1590
413	0.0086
551	31.2000
503	2.9800
488	1.3900
569	89.7000
432	0.0489

a. Plot the above data on a scatter diagram and then use correlation analysis to measure the association between the two variables.

b. Develop one linear regression model and two polynomial regression models and calculate the value of R^2 for each model.

c. Which of the three models has produced the closest fit?

Problem 6

Refer to the vapour pressure data given in problem 5 and apply the three transformations introduced in this section on the dependent variable. Then measure the correlation between the two variables after each transformation. Which transformation would you recommend for this case?

Problem 7

Using the transformed data set from the most appropriate transformation carried out in problem 6 develop a linear regression model to determine the vapour pressure of cadmium as a function of temperature. What is the regression model.

Problem 8

Use the regression model developed in problem 7 to predict the vapour pressure of cadmium if its temperature is 540 Kelvins.

UNIT-III

Introduction to Probability Distributions

Probability distributions describe the probability of observing a particular event. There are several probability distributions that are important to physicists. The binomial distribution, while not of much practical signi cance, is easy to describe, and can be used to derive the other distributions used most often by experimental physicists: the Gaussian and Poisson distributions. The Gaussian, or normal distribution, is the most important as it is most often used to describe the distribution of results for any measurement subject to small, random error. The Poisson distribution is particularly useful in describing counting experiments. A fourth distribution, the exponential distribution or interval distribution, describes the distribution of intervals between counting events. In this lab, you will investigate the Poisson and interval distributions.

Probability

In order to understand the statistical methods of dealing with random processes and how some predictability can be garnered from such chance events, we will examine some sim-ple cases involving coin tosses and dice. First we introduce three important properties of probability:

> If you consider two possible events *A* and *B* which are mutually exclusive (that is, if *A* happens *B* cannot happen and vice versa) then the probability of either *A* or *B* happening is

the sum of the probabilities of A and B: P (A or B) = P (A) + P (B). An example of two such events would be a coin toss where there are two possible events, A =heads or B =tails.

The sum of the probabilities of all possible mutually exclusive events of a trial is unity, because one of the events must happen in every trial: P (A) + P (B) + P (C) + *::* = 1. In our coin toss example, the coin must turn up either heads or tails.

The probability that two independent events will both happen is the product of the probabilities of the two single events:

P (A and B) = P (A) P (B). An example of two independent events would be two coin tosses.

From these rules we can draw the following conclusions

If a trial has n and only n possible di erent outcomes, and if you know that all of the outcomes have equal a priori probabilities of happening, then the probability of a given outcome must be equal to $1=n$.

If you classify the outcomes of a trial into di erent classes, and if the number of events belonging to one class is m, the probability that an event belonging to that class will happen is $m=n$.

We have to bear in mind that the concept of \equal probability" of events has to be derived from experience. Once we have classi ed by experience all the possible di erent and mutually exclusive events in such a manner that they have equal *a priori* probability, we can apply the rules of probabilities for detailed calculations. The key problem, therefore, is to identify which events have equal *a priori* probability. It requires considerable care to avoid mistakes. For example, if you toss two coins, you might argue that there are three possible outcomes: two heads, two tails, or one head and one tail. If you assume that each of these probabilities are equally likely then the predicted probability would be 1/3 each. Experience shows this to be wrong. The mistake is in having assumed two di erent events are only one event: heads followed by tails, and tails followed by heads. This nuance will be clari ed by working out in detail the case of tossing four coins.

Example 1: Four coins

Toss four coins. Each coin has a 50% probability of turning up heads and a 50% probability of turning up tails. (This seems logical, but it is an assumption that should be justi ed by experience.) Let p represent the probability of heads and q = 1 p that of tails: p = 0*:*5, q = 0*:*5.

The probability of no heads in a toss is the probability that all four coins turn up tails simultaneously:

(probability coin A is tails and coin B is tails and coin C is tails and coin D is tails) = (probability coin A is tails) x (probability coin B is tails) x (probability coin C is tails) x (probability coin D is tails).

Random Variable:

- A random variable *x* takes on a defined set of values with different probabilities.
    - For example, if you roll a die, the outcome is random (not fixed) and there are 6 possible outcomes, each of which occur with probability one-sixth.
    - For example, if you poll people about their voting preferences, the percentage of the sample that responds "Yes on Proposition 100" is a also a random variable (the percentage will be slightly differently every time you poll).

Random variables can be discrete or continuous

- **Discrete** random variables have a countable number of outcomes
    - Examples: Dead/alive, treatment/placebo, dice, counts, etc.
- **Continuous** random variables have an infinite continuum of possible values.
    - Examples: blood pressure, weight, the speed of a car, the real numbers from 1 to 6.

Probability functions

- A probability function maps the possible values of *x* against their respective probabilities of occurrence, *p(x)*
- *p(x)* is a number from 0 to 1.0.
- The area under a probability function is always 1.

If you toss a die, what's the probability that you roll a 3 or less?

a. 1/6

b. 1/3

c. 1/2

d. 5/6

e. 1.0

f. Two dice are rolled and the sum of the face values is six? What is the probability that at least one of the dice came up a 3?

a. 1/5

b. 2/3

c. 1/2

d. 5/6

e. 1.0

Two dice are rolled and the sum of the face values is six. What is the probability that at least one of the dice came up a 3?

a. **1/5**

b. 2/3

c. 1/2

d. 5/6

e. 1.0

The probability function that accompanies a continuous random variable is a continuous mathematical function that integrates to 1.

a. For example, recall the negative exponential function (in probability, this is called an "exponential distribution"):

b. This function integrates to 1

Expected Value and Variance:

- All probability distributions are characterized by an expected value (mean) and a variance (standard deviation squared).
- Expected value is just the average or mean (μ) of random variable *x*.
- It's sometimes called a "weighted average" because more frequent values of X are weighted more highly in the average.
- It's also how we expect X to behave on-average over the long run ("frequentist" view again).
- **Discrete case:**

$$E(X) = \sum_{\text{all } x} x_i \, p(x_i)$$

Continuous case:

$$E(X) = \int_{\text{all } x} x_i \, p(x_i) dx$$

- E(X) = μ
  - these symbols are used interchangeably

- Recall the following probability distribution of ER arrivals:

x	10	11	12	13	14
P(x)	.4	.2	.2	.1	.1

$$\sum_{i=1}^{5} x_i\, p(x) = 10(.4)+11(.2)+12(.2)+13(.1)+14(.1) = 11.3$$

Sample mean, for a sample of n subjects: = $\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \sum_{i=1}^{n} x_i(\frac{1}{n})$

- The Lottery (also known as a tax on people who are bad at math…)
- A certain lottery works by picking 6 numbers from 1 to 49. It costs $1.00 to play the lottery, and if you win, you win $2 million after taxes.
- *If you play the lottery once, what are your expected winnings or losses?*

Variance/standard deviation

σ^2=Var(x) =E(x-μ)^2

"The expected (or average) squared distance (or deviation) from the mean"

- Var(X)= σ^2
- SD(X) = σ
  - these symbols are used interchangeably
- $\sigma^2 = Var(x) = E[(x-\mu)^2] = \sum_{\text{all x}} (x_i-\mu)^2\, p(x_i)$

$$Var(X) = \sum_{\text{all x}} (x_i-\mu)^2\, p(x_i)$$

$$Var(X) = \int_{\text{all x}} (x_i-\mu)^2\, p(x_i)dx$$

$$\sigma^2 = \sum_{\text{all x}} (x_i-\mu)^2\, p(x_i)$$

$$\sigma^2 = \sum_{\text{all x}} (x_i-\mu)^2\, p(x_i) =$$
$$= (1-200{,}000)^2(.5) + (400{,}000-200{,}000)^2(.5) = 200{,}000^2$$
$$\sigma = \sqrt{200{,}000^2} = 200{,}000$$

$$= (+1 - -.053)^2 (18/38) + (-1 - -.053)^2 (20/38)$$
$$= (1.053)^2 (18/38) + (-1 + .053)^2 (20/38)$$
$$= (1.053)^2 (18/38) + (-.947)^2 (20/38)$$
$$= .997$$

$$\sigma = \sqrt{.997} = .99$$

The expected value and variance of a coin toss (H=1, T=0) are?

a. .50, .50

b. .50, .25

c. .25, .50

d. .25, .25

The expected value and variance of a coin toss are?

a. .50, .50

b. **.50, .25**

c. .25, .50

d. .25, .20

The Binomial Distribution

Now we are ready to write down an expression for the probability distribution that describes the likelihood of *r* events (e.g. heads) occurring in a total of *m* events (e.g. coin ips) where the probability of an *r*-event occurring is *p* while the probability of it not occurring is (1 *p*). Since the individual events occur independently, the probability of a subset of *r* events amongst many *m* is the product of individual probabilities. If *r* occur, then *m r* don't and the probability is $p^r(1\ p)^{m\ r}$. For the total probability of a particular event occurring (e.g. 2 heads), we multiply the probability that the event occurs by the number of ways that event can occur. The complete formula for the probability distribution is then given by

$$P_r = \frac{m!}{(m\quad r)!r!}(1\ p)^{m\ r}p^r \qquad : \qquad (1)$$

This distribution is called the *binomial distribution*. It describes the probability that *r* events occur among a total of *m* independent events. Note that it is a *discrete* distribution; it is de ned only at integral values of the variable *r*.

Problem : Work out the probability of rolling r = 0 ... 12 snake eyes and complete a table similar to the one you used in Problem 2. Plot a histogram of values. Also verify that the sum of the probabilities is unity, and that the average number of decays and the variance are reasonable.

Example : Sixteen eight-sided dice

The event of interest is again rolling a 'snake eye'.

Problem 6: What is m and p for this example? Work out P_r for r = 0 ... 16 and complete a table similar to that used in Problem 2. Plot a histogram of values. Also verify that the sum of the probabilities is unity, and that the average number of decays and the variance are reasonable.

The Poisson Distribution

The decay of radioactive atoms provides another convenient source of random events to help us explore how we can use statistics to deal with randomness. A sample of radioactive material contains a large number of atoms. Many of these atoms are unstable and will transform to another element or isotope by emitting a photon, electron or alpha particle. We will assume that, once an unstable "parent" decays, the resulting "daughter" is stable and can emit no more particles. In more complicated cases, the daughter might be unstable as well but we will not deal with that situation now.

Even though the time at which any particular atom will decay is unknown, there is some regularity in the process that we can discover by looking at the average behavior of a large number of atoms over a long time. For example, the fraction of unstable atoms that decays in a certain time period, for example one second, uctuates around a well-de ned average value.

Two characteristics are important in understanding radioactive decay. First, the proba-bility per unit time that an undecayed atom will decay within an in nitesimal time interval
t is a constant:

Binomial Probability Distribution

- A fixed number of observations (trials), n
  - e.g., 15 tosses of a coin; 20 patients; 1000 people surveyed
- A binary outcome
  - e.g., head or tail in each toss of a coin; disease or no disease
  - Generally called "success" and "failure"
  - Probability of success is p, probability of failure is 1 – p
- Constant probability for each observation
  - e.g., Probability of getting a tail is the same each time we toss the coin
- Take the example of 5 coin tosses. What's the probability that you flip exactly 3 heads in 5 coin tosses?
- *Solution:*
- One way to get exactly 3 heads: HHHTT
- What's the probability of this exact arrangement?

P(heads)xP(heads) xP(heads)xP(tails)xP(tails) $=(1/2)^3 \; x \; (1/2)^2$

Another way to get exactly 3 heads: THHHT
Probability of this exact outcome = $(1/2)^1 \; x \; (1/2)^3 \; x \; (1/2)^1 = (1/2)^3 \; x \; (1/2)^2$
In fact, $(1/2)^3 \; x \; (1/2)^2$ is the probability of each unique outcome that has exactly 3 heads and 2 tails.

So, the overall probability of 3 heads and 2 tails is:
$(1/2)^3 \; x \; (1/2)^2 + (1/2)^3 \; x \; (1/2)^2 + (1/2)^3 \; x \; (1/2)^2 +$ ….. for as many unique arrangements as there are—but how many are there??

Outcome	Probability
THHHT	$(1/2)^3 \; x \; (1/2)^2$
HHHTT	$(1/2)^3 \; x \; (1/2)^2$
TTHHH	$(1/2)^3 \; x \; (1/2)^2$
HTTHH	$(1/2)^3 \; x \; (1/2)^2$
HHTTH	$(1/2)^3 \; x \; (1/2)^2$
HTHHT	$(1/2)^3 \; x \; (1/2)^2$
THTHH	$(1/2)^3 \; x \; (1/2)^2$
HTHTH	$(1/2)^3 \; x \; (1/2)^2$
HHTHT	$(1/2)^3 \; x \; (1/2)^2$
THHTH	$(1/2)^3 \; x \; (1/2)^2$

10 arrangements $x \; (1/2)^3 \; x \; (1/2)^2$

P(3 heads and 2 tails) = ***x P(heads)^3 x P(tails)^2 = 10 x (½)^5=31.25%***

Note the general pattern emerging → if you have only two possible outcomes (call them 1/0 or yes/no or success/failure) in *n* independent trials, then the probability of exactly *X* "successes"=

$$\binom{n}{X} p^X (1-p)^{n-X}$$

- If I toss a coin 20 times, what's the probability of getting exactly 10 heads?

$$\binom{20}{10}(.5)^{10}(.5)^{10} = .176$$

- If I toss a coin 20 times, what's the probability of getting of getting 2 or fewer heads?

$$\binom{20}{0}(.5)^0(.5)^{20} = \frac{20!}{20!0!}(.5)^{20} = 9.5x10^{-7} +$$

$$\binom{20}{1}(.5)^1(.5)^{19} = \frac{20!}{19!1!}(.5)^{20} = 20x9.5x10^{-7} = 1.9x10^{-5} +$$

$$\binom{20}{2}(.5)^2(.5)^{18} = \frac{20!}{18!2!}(.5)^{20} = 190x9.5x10^{-7} = 1.8x10^{-4}$$

$$= 1.8x10^{-4}$$

All probability distributions are characterized by an expected value and a variance:

If *X* follows a binomial distribution with parameters *n* and *p*: ***X ~ Bin (n, p)***

Then:
E(X) = np
Var (X) = np(1-p)
SD (X)=sqrt(npq).

- 1. You are performing a cohort study. If the probability of developing disease in the exposed group is .05 for the study duration, then if you (randomly) sample 500 exposed people, how many do you expect to develop the disease? Give a margin of error (+/- 1 standard deviation) for your estimate.

- 2. What's the probability that **at most** 10 exposed people develop the disease?

- How many do you expect to develop the disease? Give a margin of error (+/- 1 standard deviation) for your estimate.

  X ~ binomial (500, .05)
  E(X) = 500 (.05) = 25
  Var(X) = 500 (.05) (.95) = 23.75
  StdDev(X) = square root (23.75) = 4.87
  $\therefore 25 \pm 4.87$

  What's the probability that **at most** 10 exposed subjects develop the disease?
  This is asking for a CUMULATIVE PROBABILITY: the probability of 0 getting the disease or 1 or 2 or 3 or 4 or up to 10.
  $P(X \leq 10) = P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4) + \ldots + P(X=10) =$

  Y $\binom{500}{0}(.05)^0(.95)^{500} + \binom{500}{1}(.05)^1(.95)^{499} + \binom{500}{2}(.05)^2(.95)^{498} + \ldots + \binom{500}{10}(.05)^{10}(.95)^{490} < .01$ ity of being a smoker among lung cancer cases is .6, what's the probability that in a group of 8 cases you have:

a. Less than 2 smokers?

b. More than 5?

In your case-control study of smoking and lung-cancer, 60% of cases are smokers versus only 10% of controls. What is the odds ratio between smoking and lung cancer?

a. 2.5

b. 13.5

c. 15.0

d. 6.0

e. .05

In your case-control study of smoking and lung-cancer, 60% of cases are smokers versus only 10% of controls. What is the odds ratio between smoking and lung cancer?

a. 2.5

b. **13.5**

c. 15.0

d. 6.0

e. .05

Q) What's the probability of getting exactly 5 heads in 10 coin tosses?

Q) A coin toss can be thought of as an example of a binomial distribution with N=1 and p=.5. What are the expected value and variance of a coin toss?

a. .5, .25

b. 1.0, 1.0

c. 1.5, .5

d. .25, .5

e. .5, .5

Q)If I toss a coin 10 times, what is the expected value and variance of the number of heads?

f.

g. 5, 5

h. 10, 5

i. 2.5, 5

j. 5, 2.5

k. 2.5, 10

In a randomized trial with n=150, the goal is to randomize half to treatment and half to control. The number of people randomized to treatment is a random variable X. What is the probability distribution of X?

a. X~Normal(μ=75,σ=10)

b. X~Exponential(μ=75)

c. X~Uniform

d. X~Binomial(N=150, p=.5)

e. X~Binomial(N=75, p=.5)

In a randomized trial with n=150, every subject has a 50% chance of being randomized to treatment. The number of people randomized to treatment is a random variable X. What is the probability distribution of X?

a. X~Normal(μ=75,σ=10)

b. X~Exponential(μ=75)

c. X~Uniform

d. **X~Binomial(N=150, p=.5)**

e. X~Binomial(N=75, p=.5)

The Poisson Distribution

Many experimental situations occur in which we observe the counts of events within a set unit of time, area, volume, length etc. For example,

The number of cases of a disease in different towns

The number of mutations in set sized regions of a chromosome

The number of dolphin pod sightings along a flight path through a region The number of particles emitted by a radioactive source in a given time The number of births per hour during a given day .

In such situations we are often interested in whether the events occur randomly in time or space. Consider the Babyboom dataset we saw in Lecture 2. The birth times of the babies throughout the day are shown in Figure 1. If we divide up the day into 24 hour intervals and count the number of births in each hour we can plot the counts as a histogram in Figure 2. How does this compare to the histogram of counts for a process that isn't random? Suppose the 44 birth times were distributed in time as shown in Figure 3. The histogram of these birth times per hour is shown in Figure 4. We see that the non-random clustering of events in time causes there to be more hours with zero births and more hours with large numbers of births than the real birth times histogram.

This example illustrates that the distribution of counts is useful in uncovering whether the events might occur randomly or non-randomly in time (or space). Simply looking at the histogram isn't sufficient if we want to ask the question whether the events occur randomly or not. To answer this question we need a probability model for the distribution of counts of random events that dictates the type of distributions we should expect to see.

The Poisson distribution is a discrete probability distribution for the counts of events that occur randomly in a given interval of time (or space).

If we let X = The number of events in a given interval,

Then, if the mean number of events per interval is

The probability of observing x events in a given interval is given by

$$P(X = x) = \frac{e^{x}}{x!} \qquad x = 0; 1; 2; 3; 4; : : :$$

Note e is a mathematical constant. e 2:718282. There should be a button on your calculator $\boxed{e^{x}}$ that calculates powers of e.

If the probabilities of X are distributed in this way, we write

$$X \sim Po( )$$

is the parameter of the distribution. We say X follows a Poisson distribution with parameter

Note A Poisson random variable can take on any positive integer value. In con-trast, the Binomial distribution always has a finite upper limit.

2.1 Examples

Births in a hospital occur randomly at an average rate of 1.8 births per hour.

What is the probability of observing 4 births in a given hour at the hospital?

Let X = No. of births in a given hour

1 Events occur randomly) $X \sim Po(1.8)$

2 Mean rate = 1:8

We can now use the formula to calculate the probability of observing exactly 4 births in a given hour

$$P(X = 4) = \frac{e^{1:8}\ 1:84}{4!} = 0:0723$$

What about the probability of observing more than or equal to 2 births in a given hour at the hospital?

We want $P(X \geq 2) = P(X = 2) + P(X = 3) + : : :$

i.e. an infinite number of probabilities to calculate

but

$$P(X \geq 2) = P(X = 2) + P(X = 3) + : : :$$
$$= 1 \quad P(X < 2)$$

$$= 1 - (P(X = 0) + P(X = 1))$$
$$= 1 - (e^{-1.8} 1.8^{0}/0! + e^{-1.8} 1.8^{1}/1!)$$
$$= 1 - (0.16529 + 0.29753)$$
$$= 0.537$$

- The shape of the Poisson distribution

par(mfrow = c(1, 3)) plot(0:20, dpois(0:20, 3), type = "h", ylim = c(0, 0.25), xlab = "X", main = "Po(3)", ylab = "P(X)", lwd = 3, cex.lab = 1.5, cex.axis = 2, cex.main = 2) plot(0:20, dpois(0:20, 5), type = "h", ylim = c(0, 0.25), xlab = "X", main = "Po(5)", ylab = "P(X)", lwd = 3, cex.lab = 1.5, cex.axis = 2, cex.main = 2) plot(0:20, dpois(0:20, 10),type = "h", ylim = c(0, 0.25), xlab = "X", main = "Po(10)", ylab = "P(X)", lwd = 3, cex.lab = 1.5, cex.axis = 2, cex.main = 2) Using the formula we can calculate the probabilities for a specific Poisson distribution and plot the probabilities to observe the shape of the distribution. For example, Figure 5 shows 3 different Poisson distributions. We observe that the distributions are

- unimodal
- exhibit positive skew (that decreases a increases)
- centered roughly on

(iii) the variance (spread) increases as increases

Sum of two Poisson variables

Now suppose we know that in hospital A births occur randomly at an average rate of 2.3 births per hour and in hospital B births occur randomly at an average rate of 3.1 births per hour.

What is the probability that we observe 7 births in total from the two hospitals in a given 1 hour period?

To answer this question we can use the following rule

> If $X \sim Po(\lambda_1)$ on 1 unit interval, and $Y \sim Po(\lambda_2)$ on 1 unit interval,
> then $X + Y \sim Po(\lambda_1 + \lambda_2)$ on 1 unit interval.

So if we let X = No. of births in a given hour at hospital A
and Y = No. of births in a given hour at hospital B

Then $X \sim Po(2.3)$, $Y \sim Po(3.1)$ and $X + Y \sim Po(5.4)$

$$P(X + Y = 7) = e^{-5.4} \frac{5.4^{7}}{7!} = 0.11999$$

Using the Poisson to approximate the Binomial

The Binomial and Poisson distributions are both discrete probability distributions. In some circumstances the distributions are very similar. For example, consider the Bin(100, 0.02) and Po(2) distributions shown in Figure 6. Visually these dis-tributions are identical.

In general,

> If n is large (say > 50) and p is small (say < 0.1) then a Bin(n, p) can be approximated with a Po( ) where = np

The idea of using one distribution to approximate another is widespread through-out statistics and one we will meet again. In many situations it is extremely diffi-cult to use the exact distribution and so approximations are very useful. Example Given that 5% of a population are left-handed, use the Poisson distribution to es-timate the probability that a random sample of 100 people contains 2 or more left-handed people.

X = No. of left handed people in a sample of 100

X Bin(100, 0.05)

Poisson approximation) X Po( ) with = 100 0:05 = 5

We want P (X 2)?

$$P(X \geq 2) = 1 - P(X < 2)$$

$$= 1 - P(X = 0) + P(X = 1)$$

$$= 1 - e^{-5}\frac{5^0}{0!} + e^{-5}\frac{5^1}{1!}$$

$$1 - 0{:}040428$$

$$0{:}959572$$

This consisted of 3 steps

(i) Estimating the parameters of the distribution from the data

(ii) Calculating the probability distribution

(iii) Multiplying the probability distribution by the number of observations

Once we have fitted a distribution to the data we can compare the expected fre-quencies to those we actually observed from the real Babyboom dataset. We see that the agreement is quite good.

x	0	1	2	3	4	5	6
Expected	3.837	7.035	6.448	3.941	1.806	0.662	0.271
Observed	3	8	6	4	3	0	0

When we compare the expected frequencies to those observed from the non-random clustered sequence in Section 1 we see that there is much less agreement.

x	0	1	2	3	4	5	6
Expected	3.837	7.035	6.448	3.941	1.806	0.662	0.271
Observed	12	3	0	2	2	4	1

we will see how we can formally test for a difference between the expected and observed counts. For now it is enough just to know how to fit a distribution.

NORMAL DISTRIBUTION

In probability theory, the **normal** (or **Gaussian**) **distribution** is a very common continuous probability distribution. Normal distributions are important in statistics and are often used in the natural and social sciences to represent real-valued random variables whose distributions are not known.[1][2]

The normal distribution is useful because of the central limit theorem. In its most general form, under some conditions (which include finite variance), it states that averages of random variables independently drawn from independent distributions converge in distribution to the normal, that is, become normally distributed when the number of random variables is sufficiently large. Physical quantities that are expected to be the sum of many independent processes (such as measurement errors) often have distributions that are nearly normal. Moreover, many results and methods (such as propagation of uncertainty and least squares parameter fitting) can be derived analytically in explicit form when the relevant variables are normally distributed.

The normal distribution is sometimes informally called the **bell curve**. However, many other distributions are bell-shaped (such as the Cauchy, Student's t, and logistic distributions). The terms Gaussian function and Gaussian bell curve are also ambiguous because they sometimes refer to multiples of the normal distribution that cannot be directly interpreted in terms of probabilities.

Seven features of normal distributions are listed below.

1. Normal distributions are symmetric around their mean.
2. The mean, median, and mode of a normal distribution are equal.
3. The area under the normal curve is equal to 1.0.
4. Normal distributions are denser in the center and less dense in the tails.
5. Normal distributions are defined by two parameters, the mean (μ) and the standard deviation (σ).
6. 68% of the area of a normal distribution is within one standard deviation of

the mean.

7. Approximately 95% of the area of a normal distribution is within two standard deviations of the mean.

I. We will use the normal distribution and the normal curve to do a number of statistical tasks, especially to make increasingly complex statistical assertions.

II. This kind of argumentation depends upon the normal's ability to help us see the patterns in random variation.

III. also called Bell or Gaussian curve

IV. a family of **theoretical** distributions

   a. an infinite number of curves
   b. based on rough estimations from observations and mathematical reasoning – sprang from trying to find regularities in variation
   c. DeMoivre defined the normal mathematically
   d. x-axis – all possible values of the variable
   e. y-axis (rarely drawn) – probability of that value's occurrence

Properties of the normal curve (cont'd)

I. Bell-shaped and symmetrical – recall the definition of "symmetrical"

II. For the normal, mean = mode = median

III. Asymptotic to the x-axis, i.e., the curve comes infinitely close to the x-axis without touching it. Since the y-axis of the normal indicates probability, no value in the normal is presumed to have exactly 0 probability of occurring.

IV. The curve is continuous – this is an important mathematical characteristic, but we will not explore it here

V. Inflection points of curves indicate where the slope changes from positive to negative or vice versa. For the normal curve, the inflection points are ± 1σ from the mean, □.