Course Name : BCA / BBA / BCOM/ MCA

Subject Name:

Statistics

Prepared by Assistant Professor's Team

of

Microtek College of Management & Technology

Under Guidance of

Dr. Pankaj Rajhans

An Alumni of IIT-Delhi

President & Executive Director

Microtek College of Management & Technology

Jaunpur & Varanasi (U.P)

UNIT - 1

Frequency distribution:

Graphics Representation of Frequency distribution:

There are 2 types of frequency distribution –

1. Histogram
2. Frequency Polygon

1. Histogram: In drawing the histogram of a given continuous frequency distribution we first mark off along the x-axis all the class intervals on a suitable scale. On each class interval erect rectangles with heights proportional to the frequency of the class. If, however, the classes are of unequal width then the height of the rectangle will be proportional to the ratio of the frequencies to the width of the classes. The diagram of continuous rectangles so obtained is called histogram.

2. Frequency Polygon : For an ungrouped distribution ,the frequency polygon is obtained by plotting points with abscissa as the variate values and ordinate as the corresponding frequencies and joining the plotted points by means of straight lines. For a grouped frequency distribution, the abscissa of points are mid values of the class intervals .For equal class intervals the frequency polygon can be obtained by joining the middle points of the upper sides of the adjacent rectangles of the histogram by means of straight lines. If the class intervals are of small width, the polygon can be obtained by drawing a smooth freehand curve through the vertices of the frequency polygon.

The frequency polygon so obtained should be extended to the base line (x-axis) at both the ends so that it meets the x-axis at the mid points of 2 hypothetical classes, viz, the class before the first class and the class after the last class, each assumed to have zero frequency.

…construct histograms and frequency polygons

A histogram is used to represent grouped data pictorially. Unlike a bar chart, the column widths may be different, as it is used to present data taking account of the group sizes chosen. A histogram is one form of frequency diagram.

When you group data, it is more manageable to manipulate, display and use for calculations. However, you cannot get back to original, individual readings and this means you cannot calculate accurate medians, means or modes – you have to estimate using various techniques, or else identify in broader terms what these values are. For example, at level 3 you are expected to identify the modal group – histograms can help with this.

When you construct a histogram, it is important to remember:

- The horizontal axis has a standard scale to represent the class interval.
- The vertical axis shows the frequency density, or the frequency per class of a given size.
- There are no spaces between the columns as the data represented is continuous data, i.e. along a 'continuous line' such as time or money.
- The column with the biggest area has the highest frequency:
  - if the columns are all the same width, the differences in area only depend on the height

- if the columns are different widths, you may be able to estimate if one is bigger than another by looking at the dimensions of each of them.

- You can identify the modal group by looking for the tallest column
- A frequency polygon is formed by joining the mid-points at the top of the bars of a histogram.

Worked example

We will be looking at information from a mechanic's business where the Advanced Apprentice is recording the amount of money spent by customers over a period of 6 months.

1 Group (or re-group) the data

Sensible groupings will make the histogram easier to read. Make sure that the range of frequencies of classes or groups does not vary in size by too much and that the class intervals are not too big.

Amount spent (£)	0-	20.00-	40.00-	60.00-	80.00-	100.00-	120.00-	140.00-	160.00-	180.00-200.00
Frequency	3	9	19	29	98	125	194	102	16	5

Stage 1 Decide on manageable class intervals

The range of frequencies goes from 3 to 194, so combining some classes would give a more manageable histogram. You need to try to form between 5 and 12 classes with an even spread of the frequencies across the classes. Group the data as appropriate and establish the frequency of each class

Stage 2 Group the data from the original table to calculate the new groups

Amount spent (£)	0-	80.00-	100.00-	120.00-	140.00-200.00
Frequency					

and calculate the new frequencies:

Amount spent (£)	0-	80.00-	100.00-	120.00-	140.00-200.00
Frequency	60	98	125	194	123

The range of frequencies now goes from 60 to 194, with only five classes

2 Construct the histogram accurately

The apprentice now constructs a histogram using the grouped data. First he must calculate the frequency density for each of the groups using the frequency and class interval.

Notes: The frequency of each group gives the area of bar

The class interval of each group gives the width of the bar

The frequency density of each group gives the height of the bar

Stage 1 Rearrange the formula:

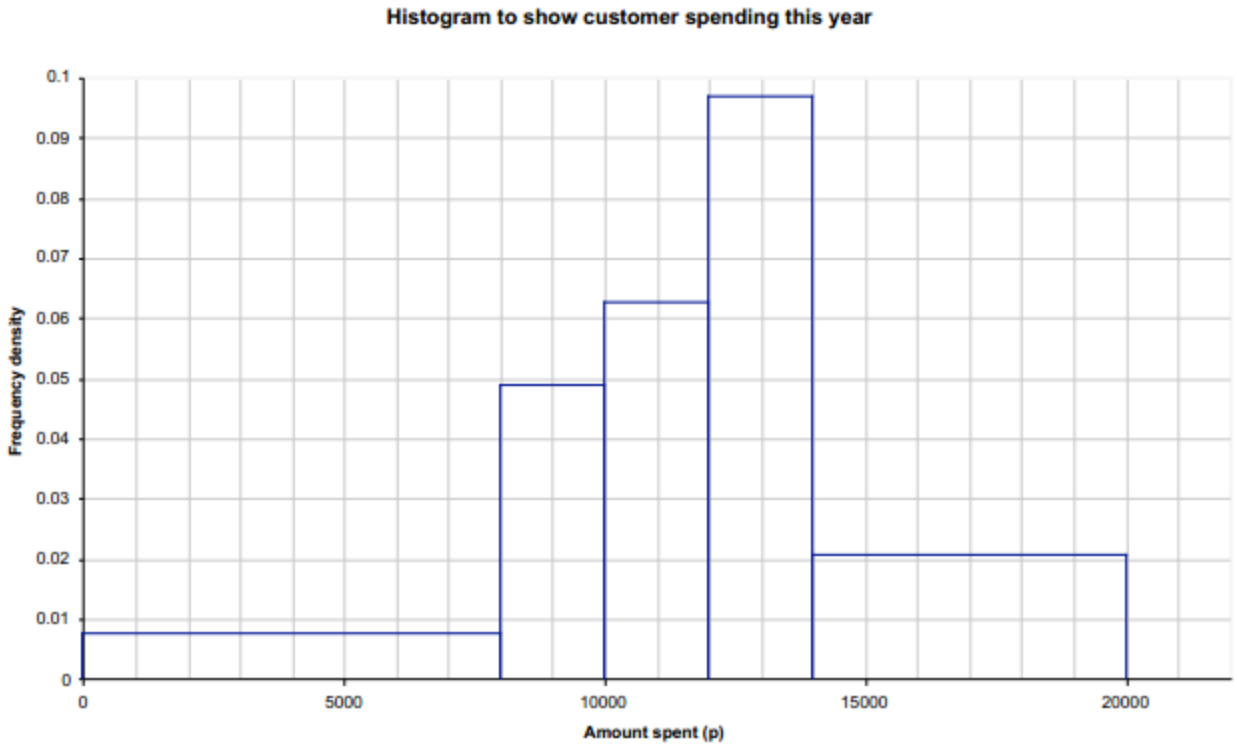
area (frequency) = **height** (frequency density) **x width** (class interval) to give:

$$\text{frequency density (height)} = \frac{\text{frequency (area)}}{\text{class interval (width)}}$$

Stage 2 Calculate the frequency density for each of the groups

Frequency	60	98	125	194	123
Amount spent (p)	0-	8 000-	10 000-	12 000-	14 000-20 000
Frequency density	0.0075	0.049	0.0625	0.097	0.0205

Stage 3 Construct the histogram and label it using accepted conventions, with the frequency density on the *y* axis and the amount spent on the *x* axis



3 Identify the modal group

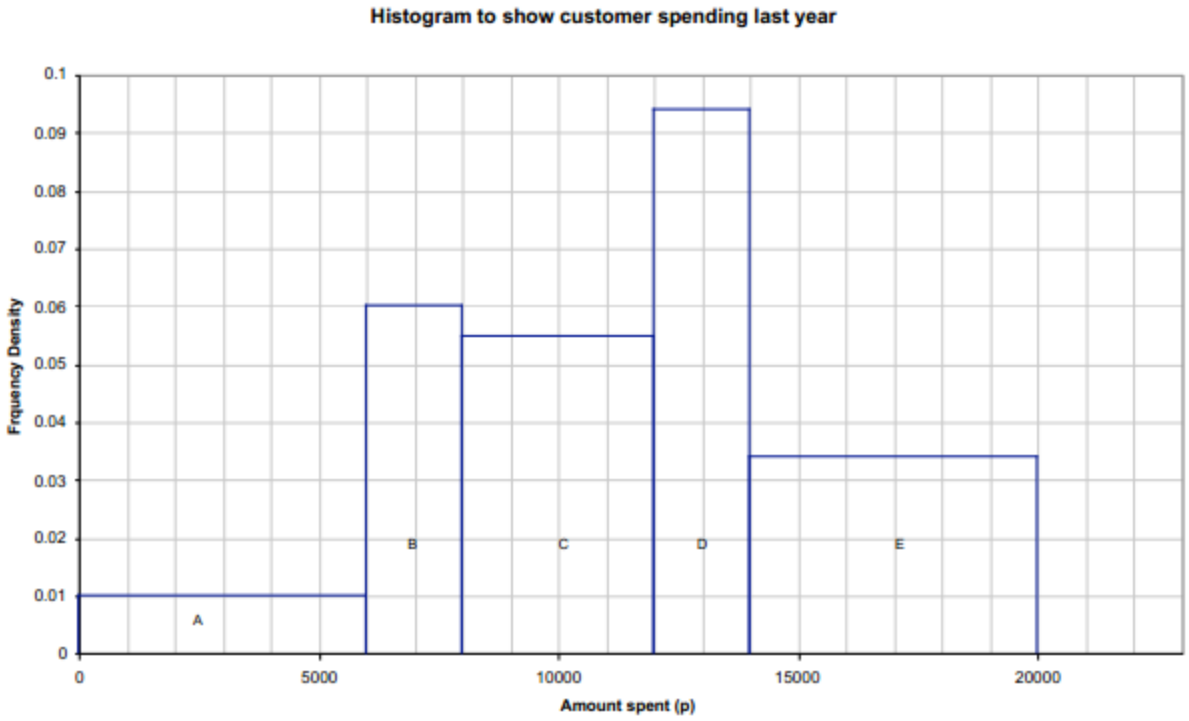
a **By looking at the grouped data to identify the modal class**

The grouping that has most people in it is £120–, so this is the modal group or class. The actual mode lies somewhere between £120 and £140.

This means that there are more customers spending £120–£140 at the garage.

b **From a histogram, the modal group is the one with the highest bar**

The apprentice has information for a similar period of time from the previous year.



The modal group is the one with the tallest bar. In this case it is the grouping £120-£140.

4 Find frequencies from histograms by calculating column areas

Look at the histogram for last year's figures (above).

We need to find the areas of each of the groups in turn to find the frequencies.

Area (frequency) = height (frequency density) x width (class interval)

Area A = 0.01 x 6 000 = 60

Area B = 0.06 x 2 000 = 120

Area C = 0.055 x 4 000 = 220

Area D = 0.094 x 2 000 = 188

Area E = 0.034 x 6 000 = 204

We can see that Area C has the biggest area and frequency. This means that the highest frequency of spending last year occurred between £80 and £120.

A **common error** here is to identify this as the modal group because it has the highest frequency.

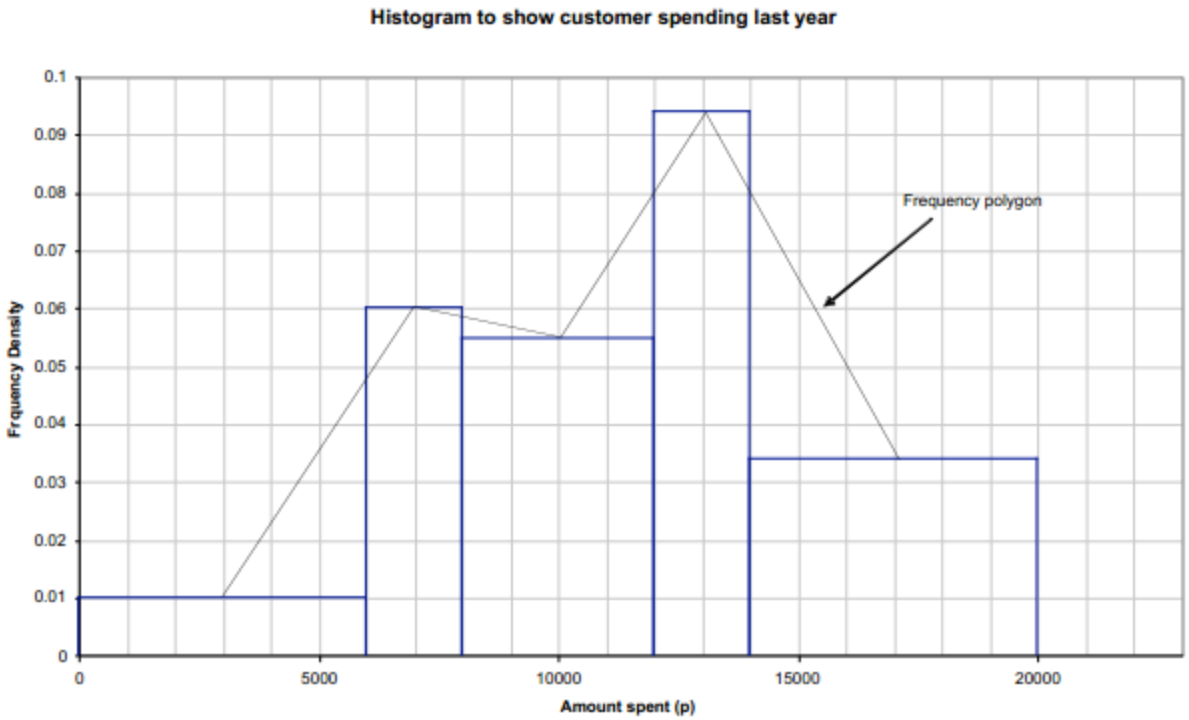
5 Forming frequency polygons

A frequency polygon is formed by joining the mid-points of the bars forming a histogram.

Breaking it down

Stage 1 Construct a histogram as shown earlier

Stage 2 Join the mid-points at the top of each bar of the histogram



Measure of Central Tendency :

According to Professor Bowely, averages are “statistical constant which enables us to comprehend in a single effort the significance of the whole.” They give us idea about the concentration of the values in the central part of the distribution.

According to Painly speaking , an average of statistical series is the value of the variable which is representative of the entire distribution.

Types of Measure of Central Tendency :

1. Arithmetic Mean or Simple Mean
2. Geometric Mean
3. Harmonic Mean
4. Median
5. Mode

Characteristics of Measure of Central Tendency:

a. It should be rigidly defined
b. It should readily comprehensible and easy to calculate
c. It should be based on all the observation
d. It should not be affected much by extreme values
e. It should be suitable for mathematical treatment

1. ARITHMETIC MEAN:

A.M mean of a set of observation is their sum divided by the number of observations.

e.g. – the arithmetic mean x of the n observation

x1, x2,x3,------,xn is given by:

$$X=(x1+\overline{x2+x3}+x4+\text{-----}+xn)/n$$

$$\overline{X} = \sum_{i=0}^{n} xi$$

In case of frequency distribution xi/fi, i-1,2,------,n , where fi is the frequency of the variable xi.

$$\overline{X} = (f1x1+f2x2+\text{-------}+fnxn)/f1+f2+\text{----------}+fn$$

$$\overline{X} = \frac{1}{N}\sum_{i=0}^{n} fixi \text{ , where } N = \sum_{i=0}^{n} fi$$

For e.g. –

- *In case of discrete data without frequency distribution-*

Find the A.M. of the following observation-

1 , 2 , 3 , 4 , 5

Solution:

$\overline{X}$= (sum of the observation)/ Total no. of the observation

$\overline{X}$= (1+2+3+4+5)/5 = 3

3 is the arithmetic mean of given observation

- *In case of discrete data with frequency distribution-*
  Find the A.M. of the following frequency distribution:

X: 1 2 3 4 5

F: 2 3 5 1 8

Solution:

X	F	xf
1	2	2
2	3	6
3	5	15
4	1	4
5	8	40
Total	19	67

$\overline{X} = \frac{1}{N}\sum_{i=0}^{n} fixi$, where $N = \sum_{i=0}^{n} fi$

$\overline{X} = 67/19$

A.M = 3.52 → Answer

- *In case of continuous data without frequency distribution-*

Find the A.M. of the following frequency distribution:

- Marks: 0-10 10-20 20-30 30-40 40-50 50-60
- frequency: 12 18 27 20 17 6

Solution:

Marks	F	x	Xf
0-10	12	5	60
10-20	18	15	270
20-30	27	25	675
30-40	20	35	700
40-50	17	45	765
50-60	6	55	330
Total	100		2800

$\overline{X} = \frac{1}{N}\sum_{i=0}^{n} fixi$, where $N = \sum_{i=0}^{n} fi$

= 2800/100

= 28 → Answer

2. GEOMETRIC MEAN:

G.M of a set of n observation is the nth root of their product.

Thus the G.M of n observation xi; i=1,2,----,n is given by—

G =(x1,x2,x3,-----xn)1/n

Taking log of both sides-

Log G = log(x1.x2.-----.xn)/n

Log G = (log x1 + log x2 + ----- + log xn)/n

Log G = $\frac{1}{n}\sum_{i=1}^{n} logxi$

G = Antilog $[\frac{1}{n}\sum_{i=1}^{n} logxi]$

In case of frequency distribution xi/fi (i = 1 ,2 ,------,n)

G = $(x_1^{f1}.x_2^{f2}.x_3^{f3}\ldots\ldots\ldots x_n^{fn})$

Taking log of both sides-

Log G = (f1.log x1 + f2.log x2 + ----- + fn.log xn)/n

Log G = $\frac{1}{n}\sum_{i=1}^{n} fi.logxi$

G = Antilog $[\frac{1}{n}\sum_{i=1}^{n} fi.logxi]$

3. HARMONIC MEAN:

H.M of a number of observations, none of which is zero , is the reciprocal of the arithmetic mean of the reciprocal of the given values. Thus, H.M(H) of n observations xi , i = 1,2,3,-----,n is given by-

$$H = 1/A.M$$

4. MEDIAN:

Median of a distribution is the value of the variable which divides into 2 equal parts, which separates the ascending or descending order.

In other words , median is the middle value that separates the higher half from lower half of the data set.

For e.g. - calculate the median of the given data-

17 , 40 , 38 , 21 , 41

Solution:

Arrange the given data in ascending order –

17 , 21 , 38 , 40 , 41

Median = 38

- If number of the given data is even then how will we calculate the median---

  Median = [(N/2) term + {(N/2) term + 1} term]/2

- If number of the given data is odd then how will we calculate the median---

  Median = [(N/2) term + 1] term

For Discrete Frequency Distribution :

In case of Discrete Frequency Distribution median is obtained by considering the cumulative frequencies. The steps for calculating median are given below :

1. Find N/2 where N = f1 + f2 + f3 + ----- + fn
2. See the (less than) cumulative frequency (c.f.) just greater than N/2.

3. The corresponding value of x is median

For e.g. - Calculate the median for the following frequency distribution :

X : 1 2 3 4 5 6 7 8 9

F: 8 10 11 16 20 25 15 9 6

Solution :

X	F	c.f.
1	8	8
2	10	18
3	11	29
4	16	45
5	20	65
6	25	90
7	15	105
8	9	114
9	6	120
Total	N= 120	

Here N/2 = 60

The cumulative frequency just greater than N/2 is 65 and the value of x corresponding to 65 is 5.

Therefore, median = 5

- For Continuous Frequency Distribution :

In case of frequency distribution, the class corresponding to the c.f. just greater than N/2 is called the median class and the value of median is obtained by the following formula:

Median = l + [(N/2 – c)h]/f

Where , l = is the lower limit of the median class ; f = is the frequency of the median class ; h = is the interval of the median class (upper limit – lower limit) ; c = is the c.f. of the class preceding the median class

And N = f1 + f2 +f3 + ------ + fn

For e.g. - Calculate the median for the following frequency distribution :

Wages	No. of employees

2000 – 3000	3
3000 – 4000	5
4000 – 5000	20
5000 – 6000	10
6000 – 7000	5

Solution :

wages	No. of employees	c.f.
2000 – 3000	3	3
3000 – 4000	5	8
4000 – 5000	20	28
5000 – 6000	10	38
6000 - 7000	5	43
Total	43	

Here N/2 = 21.5

c.f. just greater than 21.5 is 28 and the corresponding class is 4000 – 5000.

Thus the median class is 4000 – 5000

Here l = 4000 , h = 1000 , c = 8 , f = 20

Median = 4000 + [(21.5 – 8)1000]/20

Median = 4675

5. MODE: Mode is the most frequent value of the given data.
   For e.g. – 1, 2 , 2 , 6 ,4
   Solution: Mode = 2

- For Discrete Frequency Distribution :

X :	1	2	3	4	5	6	7	8
F :	4	9	16	25	22	15	7	3

Solution : Mode = 25 , the corresponding value is 4

Hence , Mode = 4 → answer

- For Continuous Frequency Distribution :

The value of mode is obtained by the following formula:

Mode = l + h (f1 – f0)/(2f1 – f0 – f2)

Where , l = is lower limit of the modal class ; h = is the interval of the modal class ;
f1 = is the frequency of the modal class ;
f0 and f2 = are frequency of the classes preceding and succeeding the modal class.
For e.g. –

Class – Interval	Frequency
0 – 10	5
10 – 20	8
20 – 30	7
30 – 40	12
40 – 50	28
50 – 60	20
60 – 70	10
70 – 80	10

Here, maximum frequency = 28
Thus the mode class is 40 – 50

Here l = 40, h =10, f1 = 28, f0 = 12, f2 = 20
By formula –
Mode = 40 + 10(28 – 12)/ (2*28 – 12 – 20)
Mode = 46.67 → answer

DISPERSION :

Averages (or the measures of central tendency) give us idea of the concentration of the observations about the central part of the distribution. If we know the average alone, we cannot form a complete idea about the distribution. If we know the average alone, we cannot form a complete idea about the distribution as will be clear from the following example.

Consider the series (i) 7,8,9,10,11 (ii) 3,6,9,12,15 (iii) 1,5,9,13,17. In all these cases we see that n, the number of observations, is 5 and the mean is 9. If we are given that the mean of 5 observations is 9.we can’nt form an idea as to whether it is the average of first series or second series or third series or of any other series of 5 observations whose sum is 45.

Thus we see that the measures of central tendency are inadequate to give us a complete idea of the distribution. They must be supported and supplemented by some other measures. One such measures is Dispersion. Literal meaning of dispersion is 'scatterdness'. We study dispersion to have an idea about the homogeneity or heterogeneity of the distribution. In the above case we say series (i) is more homogeneous(less dispersed) than the series (ii) or (iii) or we say that series (iii) is more heterogeneous (more scattered) than the series (i) or (ii).

Some important definition of dispersion is given below:

(i) "Dispersion is the measures of extent to which individual items vary."

(ii) "Measures of variation tells us how widely the data scatter about their mean"

Characteristics for Measures of central Tendency:

a) It should be rigidly defined
b) It should readily comprehensible and easy to calculate
c) It should be based on all the observation
d) It should not be affected much by extreme values
e) It should be suitable for mathematical treatment

MEASURES OF DISPERSION :

There are 4 types of measures of dispersion –

1. Range
2. Quartile Deviation
3. Mean Deviation
4. Standard Deviation

1. RANGE : the range is the difference between 2 extreme observations of the distribution. If A and B the greatest and smallest observations respectively in a distribution, then its range is given by:
   Range = A – B
2. Quartile Deviation : Q.D or semi- interquartile range Q is given by-
   Q = (Q3 – Q1)/2

Where Q1 and Q3 are the first and third quartiles of the distribution respectively.

Q.D is definitely a better measure than the range as it makes use of 50% of the data. But since it ignores the other 50% of the data, it can't be regarded as a reliable measure.

3. Mean Deviation : If xi/ fi,i=1,2,…,n is the frequency distribution, then mean deviation from the average A(usually mean median or mode)is given by:

Mean deviation from average A= $\frac{1}{N}\sum_{i=0}^{n} fi|xi - A|$, Where, N = $\sum_{i=0}^{n} fi$

and | xi-A | represents modulus or the absolute value of the deviations (xi-A), where the negative sign is ignored.

Since mean deviation is based on all the observation, it is a better measure of dispersion than rang or quartile deviation .But the step of ignoring the sighs of the deviation (xi-A) creates artificial and renders it useless for further mathematical treatment.

4.Standard Deviation and root mean square Deviation:

Standard deviation, usually denoted by the Greek letter small sigma(σ),is the positive square root of the arithmetic mean of the squares of the deviation of the given value from their arithmetic mean.

For the frequency distribution xi|fi;i=1,2,…,

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=0}^{n} fi(xi - \ddot{x})^2}$$

Where x is the arithmetic mean of the distribution and N = $\sum_{i=0}^{n} fi$

The square of standard deviation is called the variance and is given by:

$$\sigma^2 = \frac{1}{N}\sum_{i=0}^{n} fi(xi - \dddot{x})^2$$

Root mean square deviation, denoted by 's',is given by: s=

$$\sqrt{\frac{1}{N}\sum_{i=0}^{n} fi(xi - A)^2}$$

Where A is any arbitrary number .s^2 is called mean square deviation.

Example : Calculate: (i) Quartile Deviation (Q.D)

(iii) Mean Deviation (M.D) (iii) Standard Deviation (S.D)from mean , for the following data:

Marks	No. Of Student
0 – 10	6
10 – 20	5
20 – 30	8
30 – 40	15
40 – 50	7
50 – 60	6
60 – 70	3

Solution :

Marks	Mid-value (x)	No. Of students (f)	$\lvert x_i - \bar{x} \rvert$ $\bar{x} = 4.9$	$f_i \lvert x_i - \bar{x} \rvert$	$f_i (x_i - \bar{x})^2$	c.f
0 – 10	5	6	0.1	0.6	0.36	6
10 – 20	15	5	10.1	50.5	2550.25	11
20 – 30	25	8	20.1	160.8	25856.64	19
30 – 40	35	15	30.1	451.5	203852.25	34
40 – 50	45	7	40.1	280.7	78792.49	41
50 – 60	55	6	50.1	300.6	90360.36	47
60 – 70	65	3	60.1	180.1	32436.01	50
Total		50		1424.8	433848.36	

(ii) Here N = 50 , $\sum_{i=0}^{n} fi|xi - A| = 1424.8$ where A = $\bar{x}$

Mean Deviation = $\frac{1}{N}\sum_{i=0}^{n} fi|xi - A|$

= 1424.8/50 = 28.496 → Answer

(i) Here N = 50 , N/4 = 12.75 ; 3N/4 = 37.25

The c.f just greater than 12.75 is 19. Hence, the corresponding class 20 – 30 contains Q_1.

$Q_1 = 20 + \frac{10}{8}(12.75 - 11)$

$= 22.19$

The c.f just greater than 37.25 is 41. Hence, the corresponding class 40 – 50 contains Q_3

$Q_3 = 40 + \frac{10}{7}(37.25 - 34) = 44.64$

Hence , Q.D = $(Q_3 - Q_1)/2 = (44.64 - 22.19)/2 = 11.23$

Q.D = 11.23 → answer

(ii) We calculate standard deviation -

$\sigma = \sqrt{\frac{1}{N}\sum_{i=0}^{n} fi(xi - \ddot{x})^2}$

$\sigma = \sqrt{(433848.36)/50} = 93.15$

$\sigma = 93.15$ → answer

UNIT-2

Skewness

Literaly , skewness means ‘ lack of symmetry’. We study skewness to have an idea about the shape of the curve which we can draw with the help of the given data. A distribution is said to be skewed if –

(i) Mean , median and mode fall at different points, i.e. mean ≠ median ≠ mode ;

(ii) Quartiles are not equidistant from median ; and

(iii) The curve drawn with the help of the given data is not symmetrical but stretched more to one side than to the other.

Measures of Skewness -

1. $S_k = M - M_d$
2. $S_k = M - M_0$
3. $S_k = (Q_3 - M_d) - (M_d - Q_1)$

Where, M = mean , M_d = Median M_0 = Mode

Q_3 = 3rd quartiles Q_1 = 1st quartiles

Karl Pearson's measure of Skewness :

In that the mean, median and mode are not equal in a skewed distribution. The Karl Pearson's measure of skewness is based upon the *divergence of mean from mob* in a skewed distribution.

Since Mean = Mode in a symmetrical distribution, (Mean - Mode) can be taken as an *absolute measure of skewness.* The absolute measure of skewness for a distribution depends upon the unit of measurement.

A relative measure, independent of the units of measurement, is defined as the Karl Pearson's Coeficient of Skewness Sk, given by

$$Sk = \frac{Mean - Mode}{s.d.}$$

The sign of Sk gives the direction and its magnitude gives the extent of skewness. If Sk > 0, the distribution is positively skewed, and if S, < 0 it is negatively skewed.

So far we have seen that Sk is strategically dependent upon mode. If mode is not defined for a distribution we cannot find Sk . But empirical relation between mean, median and mode states that, for a moderately symmetrical distribution, we have

$$Mean - Mode = 3\,(Mean - Median)$$

Hence Karl Pearson's coefficient of skewness is defined in t m s of median as

$$Sk = \frac{3(Mean - Median)}{s.d.}$$

Example : Compute the skewness and Karl Pearson's coefficient of skewness fiom the

following data.

(iv)

Marks	No. Of Student
0 – 10	6
10 – 20	5
20 – 30	8
30 – 40	15
40 – 50	7
50 – 60	6
60 – 70	3

Solution :

Marks	Mid-value (x)	No. Of students (f)	$\lvert x_i - \bar{x} \rceil$ $\bar{x} = 4.9$	$f_i \lvert x_i - x \rceil$	$f_i (x_i - x)^2$ —	c.f
0 – 10	5	6	0.1	0.6	0.36	6
10 – 20	15	5	10.1	50.5	2550.25	11
20 – 30	25	8	20.1	160.8	25856.64	19
30 – 40	35	15	30.1	451.5	203852.25	34
40 – 50	45	7	40.1	280.7	78792.49	41
50 – 60	55	6	50.1	300.6	90360.36	47
60 – 70	65	3	60.1	180.1	32436.01	50
Total		50		1424.8	433848.36	

(ii) Here N = 50 , $\sum_{i=0}^{n} fi|xi - A| = 1424.8$ where A = $\bar{x}$

Mean Deviation = $\frac{1}{N}\sum_{i=0}^{n} fi|xi - A|$

= 1424.8/50 = 28.496 → Answer

(iii) Here N = 50 , N/4 = 12.75 ; 3N/4 = 37.25

The c.f just greater than 12.75 is 19. Hence, the corresponding class 20 – 30 contains Q_1.

$Q_1 = 20 + \frac{10}{8}(12.75 - 11)$

$= 22.19$

The c.f just greater than 37.25 is 41. Hence, the corresponding class 40 – 50 contains Q_3

$Q_3 = 40 + . \frac{10}{7}(37.25 - 34) = 44.64$

Hence , Q.D = $(Q_3 - Q_1)/2 = (44.64 - 22.19)/2 = 11.23$

Q.D = 11.23 → answer

(iv) We calculate standard deviation -

$\sigma = \sqrt{\frac{1}{N}\sum_{i=0}^{n} fi(xi - \ddot{x})^2}$

$\sigma = \sqrt{(433848.36)/50} = 93.15$

$\sigma = 93.15$ → answer

We have to calculate mode –

Hence, the corresponding class 30 – 40

Here l = 30 , h = 10 , f1 = 15 , f0 = 8 , f2 = 7

$M_0 = l + \frac{h(f1-f0)}{2f1-f0-f2}$

$= 30 + \frac{10(15-8)}{2X15-8-7}$

$= 30 + 14/3$

$= 34.6$ → answer

We have to calculate skewness $S_k = M - M_0$

$S_k = 4.9 - 34.6 = -29.7$

We have to calculate Karl Pearson's measure of Skewness-

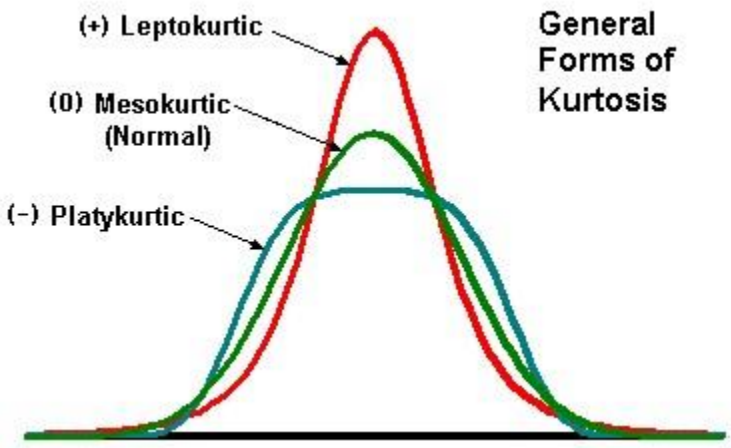
$$Sk = \frac{Mean - Mode}{s.d.}$$

$$Sk = \frac{4.9 - 34.6}{93.15}$$

Kurtosis

Kurtosis is another measure of the shape of a distribution. Whereas skewness measures the lack of symmetry of the frequency curve of a distribution, kurtosis is a measure of the relative peakedness of its fi-equency curve. Various frequency curves can be divided into three categories depending upon the shape of their peak. The three shapes are termed as Leptokurtic, Mesokurtic and Platykurhc as shown in Fig. 6.

Kurtosis is the degree of peakedness of a distribution. A normal distribution is a mesokurtic distribution. A pure leptokurtic distribution has a higher peak than the normal distribution and has heavier tails. A pure platykurtic distribution has a lower peak than a normal distribution and lighter tails.



A measure of kurtosis is given by $\boldsymbol{\beta_2} = \frac{\mu_4}{\mu_2^2}$ a coefficient given by Karl Pearson.

The value of $\beta_2 = 3$ for a Mesokurtic curve. When $\beta_2 > 3$, the curvt: is more peaked than the Mesokurtic curve and is tenned as Leptokurtic. Similarly, when $\beta_2 < 3$, the curve is less peaked than the mesokurtic curve and is called as Platykurtic curve.

Example : The first four central moments of a distribution are 0, 2.5, 0.7 and 18.75. Examine the skewness and kurtosis of the distribution.

To examine skewness , we computeβ_1.

$$\beta_1 = \frac{\mu_3^2}{\mu_2^{32}} = 0.031$$

Since $\mu_3 > 0$ and β_1, is small, the distribution is moderately positively Type equation here. skewed.
Kurtosis is given by the coefficient $\beta_2 = \frac{\mu_4}{\mu_2^2}$ = -18.75 -- 3.0 ,
Hence the curve is mesokurtic.

Probability :-

Since we operate in a world full of uncertainty , mathematician have always been interested to quantify uncertainties with an event so that one may take a better decision when a situation arises. “Probability is a mathematical measures of uncertainty.”

Definitions and Notation :-

Before discussing the rules of probability, we state the following definitions:

- Two events are mutually exclusive or disjoint if they cannot occur at the same time.

All statistical experiments have three things in common:

- The experiment can have more than one possible outcome.
- Each possible outcome can be specified in advance.
- The outcome of the experiment depends on chance.

A coin toss has all the attributes of a statistical experiment. There is more than one possible outcome. We can specify each possible outcome (i.e., heads or tails) in advance. And there is an element of chance, since the outcome is uncertain.

The Sample Space

- A sample space is a set of elements that represents all possible outcomes of a statistical experiment.
- A sample point is an element of a sample space.

- An event is a subset of a sample space - one or more sample points.

Types of events

- Two events are mutually exclusive if they have no sample points in common.
- Two events are independent when the occurrence of one does not affect the probability of the occurrence of the other.

Sample Problems

1. Suppose I roll a die. Is that a statistical experiment?

   Yes. Like a coin toss, rolling dice is a statistical experiment. There is more than one possible outcome. We can specify each possible outcome in advance. And there is an element of chance.

2. When you roll a single die, what is the sample space.

   The sample space is all of the possible outcomes - an integer between 1 and 6.

3. Which of the following are sample points when you roll a die - 3, 6, and 9?

   The numbers 3 and 6 are sample points, because they are in the sample space. The number 9 is not a sample point, since it is outside the sample space; with one die, the largest number that you can roll is 6.

4. Which of the following sets represent an event when you roll a die?

   A. {1}
   B. {2, 4,}
   C. {2, 4, 6}
   D. All of the above

   The correct answer is D. Remember that an event is a subset of a sample

space. The sample space is any integer from 1 to 6. Each of the sets shown above is a subset of the sample space, so each represents an event.

5. Consider the events listed below. Which are mutually exclusive?

   A. {1}
   B. {2, 4,}
   C. {2, 4, 6}

   Two events are mutually exclusive, if they have no sample points in common. Events A and B are mutually exclusive, and Events A and C are mutually exclusive; since they have no points in common. Events B and C have common sample points, so they are not mutually exclusive.

6. Suppose you roll a die two times. Is each roll of the die an independent event?

   Yes. Two events are independent when the occurrence of one has no effect on the probability of the occurrence of the other. Neither roll of the die affects the outcome of the other roll; so each roll of the die is independent.

- Total no. of all possible outcomes in any trial is called Exhaustive events.
- Classical Definition of Probability –

  $P(E) = n(E) / n(S)$

  $$P(E) = \frac{number\ of\ outcomes\ in\ favour\ of\ the\ events}{total\ number\ of\ outcomes}$$

- The probability that Event A occurs, given that Event B has occurred, is called a conditional probability. The conditional probability of Event A, given Event B, is denoted by the symbol P(A|B).
- The complement of an event is the event not occuring. The probability that Event A will not occur is denoted by P(A').
- The probability that Events A and B *both* occur is the probability of the intersection of A and B. The probability of the intersection of Events A and

B is denoted by P(A ∩ B). If Events A and B are mutually exclusive, P(A ∩ B) = 0.

- The probability that Events A or B occur is the probability of the union of A and B. The probability of the union of Events A and B is denoted by P(A ∪ B) .
- If the occurence of Event A changes the probability of Event B, then Events A and B are dependent. On the other hand, if the occurence of Event A does not change the probability of Event B, then Events A and B are independent.

How to Compute Probability:-

The probability of a sample point is a measure of the likelihood that the sample point will occur.

Probability of a Sample Point:-

By convention, statisticians have agreed on the following rules.

- The probability of any sample point can range from 0 to 1.
- The sum of probabilities of all sample points in a sample space is equal to 1.

Example 1

Suppose we conduct a simple statistical experiment. We flip a coin one time. The coin flip can have one of two outcomes - heads or tails. Together, these outcomes represent the sample space of our experiment. Individually, each outcome represents a sample point in the sample space. What is the probability of each sample point?

Solution: The sum of probabilities of all the sample points must equal 1. And the probability of getting a head is equal to the probability of getting a tail. Therefore, the probability of each sample point (heads or tails) must be equal to 1/2.

Example 2

Let's repeat the experiment of Example 1, with a die instead of a coin. If we toss a fair die, what is the probability of each sample point?

Solution: For this experiment, the sample space consists of six sample points: {1, 2, 3, 4, 5, 6}. Each sample point has equal probability. And the sum of

probabilities of all the sample points must equal 1. Therefore, the probability of each sample point must be equal to 1/6.

Probability of an Event:-

The probability of an event is a measure of the likelihood that the event will occur. By convention, statisticians have agreed on the following rules.

- The probability of any event can range from 0 to 1.
- The probability of event A is the sum of the probabilities of all the sample points in event A.
- The probability of event A is denoted by P(A).

Thus, if event A were very unlikely to occur, then P(A) would be close to 0. And if event A were very likely to occur, then P(A) would be close to 1.

Example 1

Suppose we draw a card from a deck of playing cards. What is the probability that we draw a spade?

Solution: The sample space of this experiment consists of 52 cards, and the probability of each sample point is 1/52. Since there are 13 spades in the deck, the probability of drawing a spade is

$$P(\text{Spade}) = (13)(1/52) = 1/4$$

Example 2

Suppose a coin is flipped 3 times. What is the probability of getting two tails and one head?

Solution: For this experiment, the sample space consists of 8 sample points.

$$S = \{TTT, TTH, THT, THH, HTT, HTH, HHT, HHH\}$$

Each sample point is equally likely to occur, so the probability of getting any particular sample point is 1/8. The event "getting two tails and one head" consists of the following subset of the sample space.

$$A = \{TTH, THT, HTT\}$$

The probability of Event A is the sum of the probabilities of the sample points in A. Therefore,

$$P(A) = 1/8 + 1/8 + 1/8 = 3/8$$

What is Probability?

The probability of an event refers to the likelihood that the event will occur.

How to Interpret Probability?

Mathematically, the probability that an event will occur is expressed as a number between 0 and 1. Notationally, the probability of event A is represented by P(A).

- If P(A) equals zero, event A will almost definitely not occur.
- If P(A) is close to zero, there is only a small chance that event A will occur.
- If P(A) equals 0.5, there is a 50-50 chance that event A will occur.
- If P(A) is close to one, there is a strong chance that event A will occur.
- If P(A) equals one, event A will almost definitely occur.

In a statistical experiment, the sum of probabilities for all possible outcomes is equal to one. This means, for example, that if an experiment can have three possible outcomes (A, B, and C), then P(A) + P(B) + P(C) = 1.

How to Compute Probability: Equally Likely Outcomes

Sometimes, a statistical experiment can have *n* possible outcomes, each of which is equally likely. Suppose a subset of *r* outcomes are classified as "successful" outcomes.

The probability that the experiment results in a successful outcome (S) is:

$$P(S) = (\text{Number of successful outcomes}) / (\text{Total number of equally likely outcomes}) = r / n$$

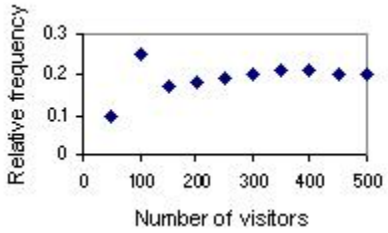
Consider the following experiment. An urn has 10 marbles. Two marbles are red, three are green, and five are blue. If an experimenter randomly selects 1 marble from the urn, what is the probability that it will be green?

In this experiment, there are 10 equally likely outcomes, three of which are green marbles. Therefore, the probability of choosing a green marble is 3/10 or 0.30.

How to Compute Probability: Law of Large Numbers

One can also think about the probability of an event in terms of its *long-run* relative frequency. The relative frequency of an event is the number of times an event occurs, divided by the total number of trials.

$$P(A) = (\text{Frequency of Event A}) / (\text{Number of Trials})$$



For example, a merchant notices one day that 5 out of 50 visitors to her store make a purchase. The next day, 20 out of 50 visitors make a purchase. The two relative frequencies (5/50 or 0.10 and 20/50 or 0.40) differ. However, summing results over many visitors, she might find that the probability that a visitor makes a purchase gets closer and closer 0.20.

The scatterplot (above right) shows the relative frequency as the number of trials (in this case, the number of visitors) increases. Over many trials, the relative frequency converges toward a stable value (0.20), which can be interpreted as the probability that a visitor to the store will make a purchase.

The idea that the relative frequency of an event will converge on the probability of the event, as the number of trials increases, is called the law of large numbers.

Problem:-

A coin is tossed three times. What is the probability that it lands on heads *exactly* one time?

(A) 0.125
(B) 0.250
(C) 0.333

(D) 0.375
(E) 0.500

Solution

The correct answer is (D). If you toss a coin three times, there are a total of eight possible outcomes. They are: HHH, HHT, HTH, THH, HTT, THT, TTH, and TTT. Of the eight possible outcomes, three have exactly one head. They are: HTT, THT, and TTH. Therefore, the probability that three flips of a coin will produce *exactly* one head is 3/8 or 0.375.

Rule of Subtraction-

- The probability of an event ranges from 0 to 1.
- The sum of probabilities of all possible events equals 1.

The rule of subtraction follows directly from these properties.

Rule of Subtraction The probability that event A will occur is equal to 1 minus the probability that event A will not occur.

$$P(A) = 1 - P(A')$$

Suppose, for example, the probability that Bill will graduate from college is 0.80. What is the probability that Bill will not graduate from college? Based on the rule of subtraction, the probability that Bill will not graduate is 1.00 - 0.80 or 0.20.

Rule of Multiplication

The rule of multiplication applies to the situation when we want to know the probability of the intersection of two events; that is, we want to know the probability that two events (Event A and Event B) both occur.

Rule of Multiplication The probability that Events A and B both occur is equal to the probability that Event A occurs times the probability that Event B occurs, given that A has occurred.

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

Example

An urn contains 6 red marbles and 4 black marbles. Two marbles are drawn *without replacement* from the urn. What is the probability that both of the marbles are black?

Solution: Let A = the event that the first marble is black; and let B = the event that the second marble is black. We know the following:

- In the beginning, there are 10 marbles in the urn, 4 of which are black. Therefore, P(A) = 4/10.
- After the first selection, there are 9 marbles in the urn, 3 of which are black. Therefore, P(B|A) = 3/9.

Therefore, based on the rule of multiplication:

$$P(A \cap B) = P(A)\, P(B|A)$$
$$P(A \cap B) = (4/10) * (3/9) = 12/90 = 2/15$$

Rule of Addition

The rule of addition applies to the following situation. We have two events, and we want to know the probability that either event occurs.

Rule of Addition :-The probability that Event A or Event B occurs is equal to the probability that Event A occurs plus the probability that Event B occurs minus the probability that both Events A and B occur.

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

Note: Invoking the fact that $P(A \cap B) = P(A)P(B \mid A)$, the Addition Rule can also be expressed as

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

Example

A student goes to the library. The probability that she checks out (a) a work of fiction is 0.40, (b) a work of non-fiction is 0.30, and (c) both fiction and non-fiction is 0.20. What is the probability that the student checks out a work of fiction, non-fiction, or both?

Solution: Let F = the event that the student checks out fiction; and let N = the event that the student checks out non-fiction. Then, based on the rule of addition:

$$P(F \cup N) = P(F) + P(N) - P(F \cap N)$$
$$P(F \cup N) = 0.40 + 0.30 - 0.20 = 0.50$$

Problem 1

An urn contains 6 red marbles and 4 black marbles. Two marbles are drawn *with replacement* from the urn. What is the probability that both of the marbles are black?

(A) 0.16
(B) 0.32
(C) 0.36
(D) 0.40
(E) 0.60

Solution

The correct answer is A. Let A = the event that the first marble is black; and let B = the event that the second marble is black. We know the following:

- In the beginning, there are 10 marbles in the urn, 4 of which are black. Therefore, P(A) = 4/10.
- After the first selection, we replace the selected marble; so there are still 10 marbles in the urn, 4 of which are black. Therefore, P(B|A) = 4/10.

Therefore, based on the rule of multiplication:

$$P(A \cap B) = P(A)\ P(B|A)$$
$$P(A \cap B) = (4/10)*(4/10) = 16/100 = 0.16$$

Problem 2

A card is drawn randomly from a deck of ordinary playing cards. You win $10 if the card is a spade or an ace. What is the probability that you will win the game?

(A) 1/13
(B) 13/52
(C) 4/13
(D) 17/52
(E) None of the above.

Solution

The correct answer is C. Let S = the event that the card is a spade; and let A = the event that the card is an ace. We know the following:

- There are 52 cards in the deck.
- There are 13 spades, so P(S) = 13/52.
- There are 4 aces, so P(A) = 4/52.
- There is 1 ace that is also a spade, so P(S ∩ A) = 1/52.

Therefore, based on the rule of addition:

$$P(S \cup A) = P(S) + P(A) - P(S \cap A)$$
$$P(S \cup A) = 13/52 + 4/52 - 1/52 = 16/52 = 4/13$$

What is a Random Variable?

When the numerical value of a variable is determined by a chance event, that variable is called a random variable.

Discrete vs. Continuous Random Variables :-

Random variables can be discrete or continuous.

- Discrete. Within a range of numbers, discrete random variables can take on only certain values. Suppose, for example, that we flip a coin and count the number of heads. The number of heads results from a random process - flipping a coin. And the number of heads is represented by an *integer* value - a number between 0 and plus infinity. Therefore, the number of heads is a discrete random variable.
- Continuous. Continuous random variables, in contrast, can take on any value within a range of values. For example, suppose we flip a coin many times and compute the *average* number of heads per flip. The average number of heads per flip results from a random process - flipping a coin. And the average number of heads per flip can take on any value between 0 and 1,

even a non-integer value. Therefore, the average number of heads per flip is a continuous random variable.

Discrete Variables: Finite vs. Infinite

Some references state that continuous variables can take on an infinite number of values, but discrete variables cannot. This is incorrect.

- In some cases, discrete variables can take on only a finite number of values. For example, the number of aces dealt in a poker hand can take on only five values: 0, 1, 2, 3, or 4.
- In other cases, however, discrete variables can take on an infinite number of values. For example, the number of coin flips that result in heads could be infinitely large.

When comparing discrete and continuous variables, it is more correct to say that continuous variables can always take on an infinite number of values; whereas some discrete variables can take on an infinite number of values, but others cannot.

What is a Probability Distribution?

A probability distribution is a table or an equation that links each possible value that a random variable can assume with its probability of occurence.

Discrete Probability Distributions:-

The probability distribution of a discrete random variable can always be represented by a table. For example, suppose you flip a coin two times. This simple exercise can have four possible outcomes: HH, HT, TH, and TT. Now, let the variable X represent the number of heads that result from the coin flips. The variable X can take on the values 0, 1, or 2; and X is a discrete random variable.

The table below shows the probabilities associated with each possible value of X. The probability of getting 0 heads is 0.25; 1 head, 0.50; and 2 heads, 0.25. Thus, the table is an example of a probability distribution for a discrete random variable.

Number of heads, x Probability, P(x)

0	0.25
1	0.50
2	0.25

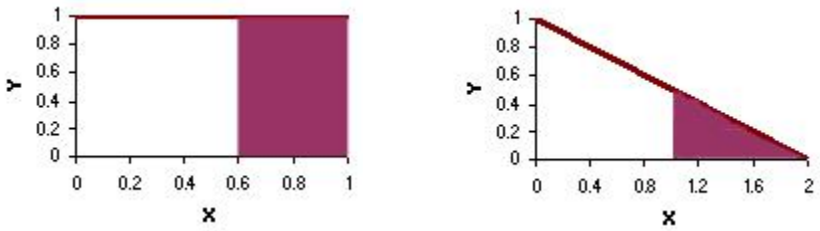
Note: Given a probability distribution, you can find cumulative probabilities. For example, the probability of getting 1 or fewer heads [$P(X \leq 1)$] is $P(X = 0) + P(X = 1)$, which is equal to 0.25 + 0.50 or 0.75.

Continuous Probability Distributions

The probability distribution of a continuous random variable is represented by an equation, called the probability density function (pdf). All probability density functions satisfy the following conditions:

- The random variable Y is a function of X; that is, y = f(x).
- The value of y is greater than or equal to zero for all values of x.
- The total area under the curve of the function is equal to one.

The charts below show two continuous probability distributions. The chart on the left shows a probability density function described by the equation y = 1 over the range of 0 to 1 and y = 0 elsewhere. The chart on the right shows a probability density function described by the equation y = 1 - 0.5x over the range of 0 to 2 and y = 0 elsewhere. The area under the curve is equal to 1 for both charts.



y = 1 y = 1 - 0.5x

The probability that a continuous random variable falls in the interval between *a* and *b* is equal to the area under the pdf curve between *a* and *b*. For example, in the

first chart above, the shaded area shows the probability that the random variable X will fall between 0.6 and 1.0. That probability is 0.40. And in the second chart, the shaded area shows the probability of falling between 1.0 and 2.0. That probability is 0.25.

Note: With a continuous distribution, there are an infinite number of values between any two data points. As a result, the probability that a continuous random variable will assume a particular value is always zero. For example, in both of the above charts, the probability that variable X will equal *exactly* 0.4 is zero.

Problem 1

The number of adults living in homes on a randomly selected city block is described by the following probability distribution.

Number of adults, x	1	2	3	4 or more
Probability, P(x)	0.25	0.50	0.15	???

What is the probability that 4 or more adults reside at a randomly selected home?

(A) 0.10
(B) 0.15
(C) 0.25
(D) 0.50
(E) 0.90

Solution

The correct answer is A. The sum of all the probabilities is equal to 1. Therefore, the probability that four or more adults reside in a home is equal to 1 - (0.25 + 0.50 + 0.15) or 0.10.

Attributes of Random Variables:-

Just like variables from a data set, random variables are described by measures of central tendency (i.e., mean and median) and measures of variability (i.e., standard deviation and variance). This lesson shows how to compute these measures for discrete random variables.

Mean of Mathematical Expectation:-

The mean of the discrete random variable X is also called the expected value of X. Notationally, the expected value of X is denoted by E(X). Use the following formula to compute the mean of a discrete random variable.

$$E(X) = \mu_x = \Sigma [x_i * P(x_i)]$$

where x_i is the value of the random variable for outcome i, μ_x is the mean of random variable X, and $P(x_i)$ is the probability that the random variable will be outcome i.

Example 1

In a recent little league softball game, each player went to bat 4 times. The number of hits made by each player is described by the following probability distribution.

Number of hits, x	0	1	2	3	4
Probability, P(x)	0.10	0.20	0.30	0.25	0.15

What is the mean of the probability distribution?

(A) 1.00
(B) 1.75
(C) 2.00
(D) 2.25
(E) None of the above.

Solution

The correct answer is E. The mean of the probability distribution is 2.15, as defined by the following equation.

$$E(X) = \Sigma [x_i * P(x_i)]$$
$$E(X) = 0*0.10 + 1*0.20 + 2*0.30 + 3*0.25 + 4*0.15 = 2.15$$

Median of a Discrete Random Variable:-

The median of a discrete random variable is the "middle" value. It is the value of X for which $P(X \leq x)$ is greater than or equal to 0.5 *and* $P(X \geq x)$ is greater than or equal to 0.5.

Consider the problem presented above in Example 1. In Example 1, the median is 2; because $P(X \leq 2)$ is equal to 0.60, and $P(X \geq 2)$ is equal to 0.70. The computations are shown below.

$$P(X \leq 2) = P(x=0) + P(x=1) + P(x=2) = 0.10 + 0.20 + 0.30 = 0.60$$

$$P(X \geq 2) = P(x=2) + P(x=3) + P(x=4) = 0.30 + 0.25 + 0.15 = 0.70$$

Variance of Mathematical Expectation :-

The standard deviation of a discrete random variable (σ) is equal to the square root of the variance of a discrete random variable (σ^2). The equation for computing the variance of a discrete random variable is shown below.

$$\sigma^2 = \Sigma \{ [x_i - E(x)]^2 * P(x_i) \}$$

where x_i is the value of the random variable for outcome i, $P(x_i)$ is the probability that the random variable will be outcome i, E(x) is the expected value of the discrete random variable x.

Example 2

The number of adults living in homes on a randomly selected city block is described by the following probability distribution.

Number of adults, x	1	2	3	4
Probability, P(x)	0.25	0.50	0.15	0.10

What is the standard deviation of the probability distribution?

(A) 0.50
(B) 0.62
(C) 0.79
(D) 0.89
(E) 2.10

Solution

The correct answer is D. The solution has three parts. First, find the expected value; then, find the variance; then, find the standard deviation. Computations are shown below, beginning with the expected value.

$$E(X) = \Sigma [x_i * P(x_i)]$$
$$E(X) = 1*0.25 + 2*0.50 + 3*0.15 + 4*0.10 = 2.10$$

Now that we know the expected value, we find the variance.

$$\sigma^2 = \Sigma \{ [x_i - E(x)]^2 * P(x_i) \}$$
$$\sigma^2 = (1 - 2.1)^2 * 0.25 + (2 - 2.1)^2 * 0.50 + (3 - 2.1)^2 * 0.15 + (4 - 2.1)^2 * 0.10$$
$$\sigma^2 = (1.21 * 0.25) + (0.01 * 0.50) + (0.81) * 0.15) + (3.61 * 0.10) = 0.3025 + 0.0050 + 0.1215 + 0.3610 = 0.79$$

And finally, the standard deviation is equal to the square root of the variance; so the standard deviation is sqrt(0.79) or 0.889.

UNIT-3

THEORETICAL DISTRIBUTION

What is a Probability Distribution?

A probability distribution is a table or an equation that links each outcome of a statistical experiment with its probability of occurence.

Probability Distribution Prerequisites

To understand probability distributions, it is important to understand variables. random variables, and some notation.

- A variable is a symbol (*A*, *B*, *x*, *y*, etc.) that can take on any of a specified set of values.
- When the value of a variable is the outcome of a statistical experiment, that variable is a random variable.

Generally, statisticians use a capital letter to represent a random variable and a lower-case letter, to represent one of its values. For example,

- X represents the random variable X.
- P(X) represents the probability of X.
- P(X = x) refers to the probability that the random variable X is equal to a particular value, denoted by x. As an example, P(X = 1) refers to the probability that the random variable X is equal to 1.

Probability Distributions

An example will make clear the relationship between random variables and probability distributions. Suppose you flip a coin two times. This simple statistical experiment can have four possible outcomes: HH, HT, TH, and TT. Now, let the variable X represent the number of Heads that result from this experiment. The variable X can take on the values 0, 1, or 2. In this example, X is a random variable; because its value is determined by the outcome of a statistical experiment.

A probability distribution is a table or an equation that links each outcome of a statistical experiment with its probability of occurence. Consider the coin flip experiment described above. The table below, which associates each outcome with its probability, is an example of a probability distribution.

Number of heads	Probability
0	0.25
1	0.50
2	0.25

The above table represents the probability distribution of the random variable X.

Cumulative Probability Distributions:-

A cumulative probability refers to the probability that the value of a random variable falls within a specified range.

Let us return to the coin flip experiment. If we flip a coin two times, we might ask: What is the probability that the coin flips would result in one or fewer heads? The answer would be a cumulative probability. It would be the probability that the coin flip experiment results in zero heads plus the probability that the experiment results in one head.

$$P(X < 1) = P(X = 0) + P(X = 1) = 0.25 + 0.50 = 0.75$$

Like a probability distribution, a cumulative probability distribution can be represented by a table or an equation. In the table below, the cumulative probability refers to the probability than the random variable X is less than or equal to x.

Number of heads: x	Probability: P(X = x)	Cumulative Probability: P(X < x)
0	0.25	0.25
1	0.50	0.75
2	0.25	1.00

Uniform Probability Distribution:-

The simplest probability distribution occurs when all of the values of a random variable occur with equal probability. This probability distribution is called the uniform distribution.

Uniform Distribution. Suppose the random variable X can assume k different values. Suppose also that the $P(X = x_k)$ is constant. Then,

$$P(X = x_k) = 1/k$$

Example 1

Suppose a die is tossed. What is the probability that the die will land on 6 ?

Solution: When a die is tossed, there are 6 possible outcomes represented by:

S = { 1, 2, 3, 4, 5, 6 }. Each possible outcome is a random variable (X), and each outcome is equally likely to occur. Thus, we have a uniform distribution. Therefore, the P(X = 6) = 1/6.

Example 2

Suppose we repeat the dice tossing experiment described in Example 1. This time, we ask what is the probability that the die will land on a number that is smaller than 5 ?

Solution: When a die is tossed, there are 6 possible outcomes represented by:

S = { 1, 2, 3, 4, 5, 6 }. Each possible outcome is equally likely to occur. Thus, we have a uniform distribution.

This problem involves a cumulative probability. The probability that the die will land on a number smaller than 5 is equal to:

$$P(X < 5) = P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) = 1/6 + 1/6 + 1/6 + 1/6 = 2/3$$

Probability Distributions: Discrete vs. Continuous

All probability distributions can be classified as discrete probability distributions or as continuous probability distributions, depending on whether they define probabilities associated with discrete variables or continuous variables.

Discrete vs. Continuous Variables:-

If a variable can take on any value between two specified values, it is called a continuous variable; otherwise, it is called a discrete variable.

Some examples will clarify the difference between discrete and continuous variables.

- Suppose the fire department mandates that all fire fighters must weigh between 150 and 250 pounds. The weight of a fire fighter would be an example of a continuous variable; since a fire fighter's weight could take on any value between 150 and 250 pounds.
- Suppose we flip a coin and count the number of heads. The number of heads could be any integer value between 0 and plus infinity. However, it could not be any number between 0 and plus infinity. We could not, for example, get 2.5 heads. Therefore, the number of heads must be a discrete variable.

Just like variables, probability distributions can be classified as discrete or continuous.

Discrete Probability Distributions:-

If a random variable is a discrete variable, its probability distribution is called a discrete probability distribution.

An example will make this clear. Suppose you flip a coin two times. This simple statistical experiment can have four possible outcomes: HH, HT, TH, and TT. Now, let the random variable X represent the number of Heads that result from this experiment. The random variable X can only take on the values 0, 1, or 2, so it is a discrete random variable.

The probability distribution for this statistical experiment appears below.

Number of heads	Probability
0	0.25
1	0.50
2	0.25

The above table represents a *discrete* probability distribution because it relates each value of a discrete random variable with its probability of occurrence. In subsequent lessons, we will cover the following discrete probability distributions.

- Binomial probability distribution
- Hypergeometric probability distribution
- Multinomial probability distribution
- Negative binomial distribution
- Poisson probability distribution

Note: With a discrete probability distribution, each possible value of the discrete random variable can be associated with a non-zero probability. Thus, a discrete probability distribution can always be presented in tabular form.

Continuous Probability Distributions:-

If a random variable is a continuous variable, its probability distribution is called a continuous probability distribution.

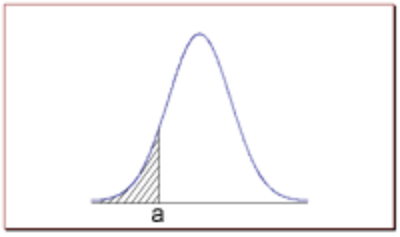
A continuous probability distribution differs from a discrete probability distribution in several ways.

- The probability that a continuous random variable will assume a particular value is zero.
- As a result, a continuous probability distribution cannot be expressed in tabular form.
- Instead, an equation or formula is used to describe a continuous probability distribution.

Most often, the equation used to describe a continuous probability distribution is called a probability density function. Sometimes, it is referred to as a density function, a PDF, or a pdf. For a continuous probability distribution, the density function has the following properties:

- Since the continuous random variable is defined over a continuous range of values (called the domain of the variable), the graph of the density function will also be continuous over that range.
- The area bounded by the curve of the density function and the x-axis is equal to 1, when computed over the domain of the variable.
- The probability that a random variable assumes a value between *a* and *b* is equal to the area under the density function bounded by *a* and *b*.

For example, consider the probability density function shown in the graph below. Suppose we wanted to know the probability that the random variable *X* was less than or equal to *a*. The probability that *X* is less than or equal to *a* is equal to the area under the curve bounded by *a* and minus infinity - as indicated by the shaded area.



Note: The shaded area in the graph represents the probability that the random variable *X* is less than or equal to *a*. This is a cumulative probability. However, the probability that *X* is *exactly* equal to *a* would be zero. A continuous random variable can take on an infinite number of values. The probability that it will equal a specific value (such as *a*) is always zero.

In subsequent lessons, we will cover the following continuous probability distributions.

- Normal probability distribution
- Student's t distribution
- Chi-square distribution
- F distribution

Binomial Probability Distribution:

To understand binomial distributions and binomial probability, it helps to understand binomial experiments and some associated notation; so we cover those topics first.

Binomial Experiment

A binomial experiment (also known as a Bernoulli trial) is a statistical experiment that has the following properties:

- The experiment consists of *n* repeated trials.
- Each trial can result in just two possible outcomes. We call one of these outcomes a success and the other, a failure.
- The probability of success, denoted by *P*, is the same on every trial.
- The trials are independent; that is, the outcome on one trial does not affect the outcome on other trials.

Consider the following statistical experiment. You flip a coin 2 times and count the number of times the coin lands on heads. This is a binomial experiment because:

- The experiment consists of repeated trials. We flip a coin 2 times.
- Each trial can result in just two possible outcomes - heads or tails.
- The probability of success is constant - 0.5 on every trial.
- The trials are independent; that is, getting heads on one trial does not affect whether we get heads on other trials.

Notation:-

The following notation is helpful, when we talk about binomial probability.

- x: The number of successes that result from the binomial experiment.
- n: The number of trials in the binomial experiment.
- P: The probability of success on an individual trial.
- Q: The probability of failure on an individual trial. (This is equal to 1 - P.)
- b(x; n, P): Binomial probability - the probability that an n-trial binomial experiment results in exactly x successes, when the probability of success on an individual trial is P.
- ${}_nC_r$: The number of combinations of n things, taken r at a time.

Binomial Distribution:-

A binomial random variable is the number of successes x in n repeated trials of a binomial experiment. The probability distribution of a binomial random variable is called a binomial distribution (also known as a Bernoulli distribution).

Suppose we flip a coin two times and count the number of heads (successes). The binomial random variable is the number of heads, which can take on values of 0, 1, or 2. The binomial distribution is presented below.

Number of heads	Probability
0	0.25
1	0.50
2	0.25

The binomial distribution has the following properties:

- The mean of the distribution (μ_x) is equal to $n * P$.

- The variance (σ^2_x) is $n * P * (1 - P)$.
- The standard deviation (σ_x) is sqrt[$n * P * (1 - P)$].

Binomial Probability:-

Binomial distribution was discovered by James Bernoulli(1654-1705) in the year 1700 and was first published posthumously in the 1713, eight years after his death.

Let a random experiment be performed repeatedly, each repetition being called a trial and let the occurrence of an event in a trial be called a success and it's non – occurrence a failure. Consider a set of n independent Bernoulli an trials (n being finite) in which the probability 'p' of success in any trial is constant for each trial, then q = 1-p, is the probability of failure in any trial.

The probability of x successes and consequently (n – x) failures in n independent trials, in a specified order (say) SSFSFFFS….FSF (where S represent success and F represents failure) is given by compound probability theorem by the expression :

$$P(SSFSFFFS….FSF) = P(S)P(S)P(F)P(S)…..P(F)P(S)P(F)$$
$$= p.p.q.p……q.p.q$$
$$= p.p.p….p \ . \ q.q.q….q$$
$$\{ x \text{ factor } \} . \{(n – x) \text{ factor}\}$$
$$= p^x .q^{n-x}$$

But x successes in n trials can occur in ${}^nc_x$ ways and the probability of x successes in n trials in any order is given by the addition theorem of probability by the expression ${}^nc_x \, p^x .q^{n-x}$

He probability distribution of the number of successes, so obtained is called the Binomial probability distribution, for the obvious reason that the probabilities of 0,1,2…..,n successes , viz, q^n, ${}^nc_1q^{n-1}p$, ${}^nc_2q^{n-2}p^2$,…..,p^n, are the successive terms of the binomial expansion $(q+p)^n$.

DEFINITION: A random variable X is said to follow binomial distribution if it assumes only non-negative values and its probability mass function is given by :

$$P(X = x) = p(x) = \{ {}^nc_x \, p^x .q^{n-x} \ ; x = 0, 1,2,…..,n; q = 1 – p$$
$$\{ 0 , \text{otherwise}$$

The 2 independent constants n and p in the distribution are known as the parameters of the distribution . 'n' is also sometimes , known as the degree of the binomial distribution.

Binomial distribution is a discrete distribution as X can take only the integral values , viz 0,1,2,….,n. Any random variable which follows binomial distribution is known as binomial variate.

We shall use the notation X ~ B(n,p) to denote that the random variable X follows binomial distribution with parameters n and p.

Remark - $1 = (x + a)^n = \sum_{k=0}^{n} \binom{n}{k} x^k a^{n-k}$

Example 1

Suppose a die is tossed 5 times. What is the probability of getting exactly 2 fours?

Solution: This is a binomial experiment in which the number of trials is equal to 5, the number of successes is equal to 2, and the probability of success on a single trial is 1/6 or about 0.167. Therefore, the binomial probability is:

$$b(2; 5, 0.167) = {}_5C_2 * (0.167)^2 * (0.833)^3$$
$$b(2; 5, 0.167) = 0.161$$

Example 1

Bob is a high school basketball player. He is a 70% free throw shooter. That means his probability of making a free throw is 0.70. During the season, what is the probability that Bob makes his third free throw on his fifth shot?

Solution: This is an example of a negative binomial experiment. The probability of success (*P*) is 0.70, the number of trials (*x*) is 5, and the number of successes (*r*) is 3.

To solve this problem, we enter these values into the negative binomial formula.

$$b^*(x; r, P) = {}_{x-1}C_{r-1} * P^r * Q^{x-r}$$
$$b^*(5; 3, 0.7) = {}_4C_2 * 0.7^3 * 0.3^2$$
$$b^*(5; 3, 0.7) = 6 * 0.343 * 0.09 = 0.18522$$

Thus, the probability that Bob will make his third successful free throw on his fifth shot is 0.18522.

Example 2

Let's reconsider the above problem from Example 1. This time, we'll ask a slightly different question: What is the probability that Bob makes his first free throw on his fifth shot?

Solution: This is an example of a geometric distribution, which is a special case of a negative binomial distribution. Therefore, this problem can be solved using the negative binomial formula or the geometric formula. We demonstrate each approach below, beginning with the negative binomial formula.

The probability of success (*P*) is 0.70, the number of trials (*x*) is 5, and the number of successes (*r*) is 1. We enter these values into the negative binomial formula.

$$b^*(x; r, P) = {}_{x-1}C_{r-1} * P^r * Q^{x-r}$$
$$b^*(5; 1, 0.7) = {}_4C_0 * 0.7^1 * 0.3^4$$
$$b^*(5; 3, 0.7) = 0.00567$$

Now, we demonstate a solution based on the geometric formula.

$$g(x; P) = P * Q^{x-1}$$
$$g(5; 0.7) = 0.7 * 0.3^4 = 0.00567$$

Notice that each approach yields the same answer.

Poisson Distribution:-

A Poisson distribution is the probability distribution that results from a Poisson experiment.

Attributes of a Poisson Experiment:-

A Poisson experiment is a statistical experiment that has the following properties:

- The experiment results in outcomes that can be classified as successes or failures.
- The average number of successes (μ) that occurs in a specified region is known.
- The probability that a success will occur is proportional to the size of the region.
- The probability that a success will occur in an extremely small region is virtually zero.

Note that the specified region could take many forms. For instance, it could be a length, an area, a volume, a period of time, etc.

Notation:-

The following notation is helpful, when we talk about the Poisson distribution.

- *e*: A constant equal to approximately 2.71828. (Actually, *e* is the base of the natural logarithm system.)
- μ: The mean number of successes that occur in a specified region.
- *x*: The actual number of successes that occur in a specified region.
- P(*x*; μ): The Poisson probability that exactly *x* successes occur in a Poisson experiment, when the mean number of successes is μ.

Poisson Distribution:-

A Poisson random variable is the number of successes that result from a Poisson experiment. The probability distribution of a Poisson random variable is called a Poisson distribution.

P.D was discovered by the French mathematician and physicist Simeon Denis Poisson (1781-1840) who published it in 1837. P.D is a limiting case of the binomial distribution under the following conditions:

(i) n, the number of trials is indefinitely large, i.e. $n \to \infty$

(ii) p, the constant probability of success for each trial is indefinitely small, i.e. , $p \to 0$.

(i) $np = \lambda$

Definition : A random variable X is said to follow a Poisson distribution if it assumes only non negative values and its probability mass function is given by-

$$P(x, \lambda) = P(X=x) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!} ; x = 0,1,2,\ldots; \lambda>0 \\ 0 \text{ , otherwise} \end{cases}$$

Remark - it should be noted that

$$\sum_{x=0}^{\infty} P(X = x) = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda} e^{\lambda} = 1$$

Example 1

The average number of homes sold by the Acme Realty company is 2 homes per day. What is the probability that exactly 3 homes will be sold tomorrow?

Solution: This is a Poisson experiment in which we know the following:

- μ = 2; since 2 homes are sold per day, on average.
- x = 3; since we want to find the likelihood that 3 homes will be sold tomorrow.
- e = 2.71828; since *e* is a constant equal to approximately 2.71828.

We plug these values into the Poisson formula as follows:

$P(x; \mu) = (e^{-\mu}) (\mu^x) / x!$
$P(3; 2) = (2.71828^{-2}) (2^3) / 3!$
$P(3; 2) = (0.13534) (8) / 6$
$P(3; 2) = 0.180$

Thus, the probability of selling 3 homes tomorrow is 0.180 .

Normal Distribution -

The normal distribution refers to a family of continuous probability distributions described by the normal equation.

The Normal Equation:-

The normal distribution is defined by the following equation:

Normal equation. The value of the random variable *Y* is:

$$Y = \{ 1/[\sigma * \text{sqrt}(2\pi)] \} * e^{-(x - \mu)2/2\sigma2}$$

where *X* is a normal random variable, μ is the mean, σ is the standard deviation, π is approximately 3.14159, and *e* is approximately 2.71828.

The random variable *X* in the normal equation is called the normal random variable. The normal equation is the probability density function for the normal distribution.

The Normal Curve:-

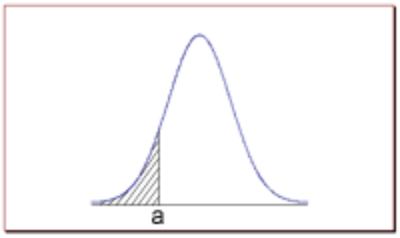
The graph of the normal distribution depends on two factors - the mean and the standard deviation. The mean of the distribution determines the location of the center of the graph, and the standard deviation determines the height and width of the graph. When the standard deviation is large, the curve is short and wide; when the standard deviation is small, the curve is tall and narrow. All normal distributions look like a symmetric, bell-shaped curve, as shown below.

The curve on the left is shorter and wider than the curve on the right, because the curve on the left has a bigger standard deviation.

Probability and the Normal Curve:-

The normal distribution is a continuous probability distribution. This has several implications for probability.

- The total area under the normal curve is equal to 1.
- The probability that a normal random variable *X* equals any particular value is 0.
- The probability that *X* is greater than *a* equals the area under the normal curve bounded by *a* and plus infinity (as indicated by the *non-shaded* area in the figure below).
- The probability that *X* is less than *a* equals the area under the normal curve bounded by *a* and minus infinity (as indicated by the *shaded* area in the figure below).



Additionally, every normal curve (regardless of its mean or standard deviation) conforms to the following "rule".

- About 68% of the area under the curve falls within 1 standard deviation of the mean.
- About 95% of the area under the curve falls within 2 standard deviations of the mean.
- About 99.7% of the area under the curve falls within 3 standard deviations of the mean.

Collectively, these points are known as the empirical rule or the 68-95-99.7 rule. Clearly, given a normal distribution, most outcomes will be within 3 standard deviations of the mean.

To find the probability associated with a normal random variable, use a graphing calculator, an online normal distribution calculator, or a normal distribution table. In the examples below, we illustrate the use of Stat Trek's Normal Distribution Calculator, a free tool available on this site. In the next lesson, we demonstrate the use of normal distribution tables.

Example 1

An average light bulb manufactured by the Acme Corporation lasts 300 days with

a standard deviation of 50 days. Assuming that bulb life is normally distributed, what is the probability that an Acme light bulb will last at most 365 days?

Solution: Given a mean score of 300 days and a standard deviation of 50 days, we want to find the cumulative probability that bulb life is less than or equal to 365 days. Thus, we know the following:

- The value of the normal random variable is 365 days.
- The mean is equal to 300 days.
- The standard deviation is equal to 50 days.

We enter these values into the Normal Distribution Calculator and compute the cumulative probability. The answer is: P(X < 365) = 0.90. Hence, there is a 90% chance that a light bulb will burn out within 365 days.

Example 2

Suppose scores on an IQ test are normally distributed. If the test has a mean of 100 and a standard deviation of 10, what is the probability that a person who takes the test will score between 90 and 110?

Solution: Here, we want to know the probability that the test score falls between 90 and 110. The "trick" to solving this problem is to realize the following:

$$P(90 < X < 110) = P(X < 110) - P(X < 90)$$

We use the Normal Distribution Calculator to compute both probabilities on the right side of the above equation.

- To compute P(X < 110), we enter the following inputs into the calculator: The value of the normal random variable is 110, the mean is 100, and the standard deviation is 10. We find that P(X < 110) is 0.84.
- To compute P(X < 90), we enter the following inputs into the calculator: The value of the normal random variable is 90, the mean is 100, and the standard deviation is 10. We find that P(X < 90) is 0.16.

We use these findings to compute our final answer as follows:

$$P(90 < X < 110) = P(X < 110) - P(X < 90)$$
$$P(90 < X < 110) = 0.84 - 0.16$$
$$P(90 < X < 110) = 0.68$$

Thus, about 68% of the test scores will fall between 90 and 110.

Standard Normal Distribution:-

The standard normal distribution is a special case of the normal distribution. It is the distribution that occurs when a normal random variable has a mean of zero and a standard deviation of one.

Standard Score (aka, z Score)

The normal random variable of a standard normal distribution is called a standard score or a z-score. Every normal random variable X can be transformed into a z score via the following equation:

$$z = (X - \mu) / \sigma$$

where X is a normal random variable, μ is the mean mean of X, and σ is the standard deviation of X.

Standard Normal Distribution Table:-

A standard normal distribution table shows a cumulative probability associated with a particular z-score. Table rows show the whole number and tenths place of the z-score. Table columns show the hundredths place. The cumulative probability (often from minus infinity to the z-score) appears in the cell of the table.

For example, a section of the standard normal table is reproduced below. To find the cumulative probability of a z-score equal to -1.31, cross-reference the row of the table containing -1.3 with the column containing 0.01. The table shows that the

probability that a standard normal random variable will be less than -1.31 is 0.0951; that is, P(Z < -1.31) = 0.0951.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
...	...	...	...	...	...	...	...	...	...	...
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0722	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
...	...	...	...	...	...	...	...	...	...	...
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Of course, you may not be interested in the probability that a standard normal random variable falls between minus infinity and a given value. You may want to know the probability that it lies between a given value and plus infinity. Or you may want to know the probability that a standard normal random variable lies between two given values. These probabilities are easy to compute from a normal distribution table. Here's how.

- Find P(Z > a). The probability that a standard normal random variable (z) is greater than a given value (a) is easy to find. The table shows the P(Z < a). The P(Z > a) = 1 - P(Z < a).

  Suppose, for example, that we want to know the probability that a z-score will be greater than 3.00. From the table (see above), we find that P(Z < 3.00) = 0.9987. Therefore, P(Z > 3.00) = 1 - P(Z < 3.00) = 1 - 0.9987 = 0.0013.

- Find P(a < Z < b). The probability that a standard normal random variables lies between two values is also easy to find. The P(a < Z < b) = P(Z < b) - P(Z < a).

  For example, suppose we want to know the probability that a z-score will be greater than -1.40 and less than -1.20. From the table (see above), we find

that P(Z < -1.20) = 0.1151; and P(Z < -1.40) = 0.0808. Therefore, P(-1.40 < Z < -1.20) = P(Z < -1.20) - P(Z < -1.40) = 0.1151 - 0.0808 = 0.0343.

In school or on the Advanced Placement Statistics Exam, you may be called upon to use or interpret standard normal distribution tables. Standard normal tables are commonly found in appendices of most statistics texts.

The Normal Distribution as a Model for Measurements:-

Often, phenomena in the real world follow a normal (or near-normal) distribution. This allows researchers to use the normal distribution as a model for assessing probabilities associated with real-world phenomena. Typically, the analysis involves two steps.

- Transform raw data. Usually, the raw data are not in the form of z-scores. They need to be transformed into z-scores, using the transformation equation presented earlier: $z = (X - \mu) / \sigma$.
- Find probability. Once the data have been transformed into z-scores, you can use standard normal distribution tables, online calculators (e.g., Stat Trek's free normal distribution calculator), or handheld graphing calculators to find probabilities associated with the z-scores.

The problem in the next section demonstrates the use of the normal distribution as a model for measurement.

Problem 1

Molly earned a score of 940 on a national achievement test. The mean test score was 850 with a standard deviation of 100. What proportion of students had a higher score than Molly? (Assume that test scores are normally distributed.)

(A) 0.10
(B) 0.18
(C) 0.50
(D) 0.82
(E) 0.90

Solution

The correct answer is B. As part of the solution to this problem, we assume that test scores are normally distributed. In this way, we use the normal distribution as a model for measurement. Given an assumption of normality, the solution involves three steps.

- First, we transform Molly's test score into a z-score, using the z-score transformation equation.

$$z = (X - \mu) / \sigma = (940 - 850) / 100 = 0.90$$

- Then, using an online calculator (e.g., Stat Trek's free normal distribution calculator), a handheld graphing calculator, or the standard normal distribution table, we find the cumulative probability associated with the z-score. In this case, we find $P(Z < 0.90) = 0.8159$.
- Therefore, the $P(Z > 0.90) = 1 - P(Z < 0.90) = 1 - 0.8159 = 0.1841$.

Thus, we estimate that 18.41 percent of the students tested had a higher score than Molly.

CORRELATION AND REGRESSION

5.1: Introduction

So far we have confined our discussion to the distributions involving only one variable. Sometimes, in practical applications, we might come across certain set of data, where each item of the set may comprise of the values of two or more variables.

Suppose we have a set of 30 students in a class and we want to measure the heights and weights of all the students. We observe that each individual (unit) of the set assumes two values – one relating to the height and the other to the weight. Such a distribution in which each individual or unit of the set is made up of two values is called a bivariate distribution. The following examples will illustrate clearly the meaning of bivariate distribution.

(i) In a class of 60 students the series of marks obtained in two subjects by all of them.

(ii) The series of sales revenue and advertising expenditure of two companies in a particular year.

(iii) The series of ages of husbands and wives in a sample of selected married couples.

Thus in a bivariate distribution, we are given a set of pairs of observations, wherein each pair represents the values of two variables.

In a bivariate distribution, we are interested in finding a relationship (if it exists) between the two variables under study.

The concept of 'correlation' is a statistical tool which studies the relationship between two variables and Correlation Analysis involves various methods and techniques used for studying and measuring the extent of the relationship between the two variables.

"Two variables are said to be in correlation if the change in one of the variables results in a change in the other variable".

5.2: Types of Correlation

There are two important types of correlation. They are (1) Positive and Negative correlation and (2) Linear and Non – Linear correlation.

5.2.1: Positive and Negative Correlation

If the values of the two variables deviate in the same direction i.e. if an increase (or decrease) in the values of one variable results, on an average, in a corresponding increase (or decrease) in the values of the other variable the correlation is said to be positive.

Some examples of series of positive correlation are:

(i) Heights and weights;
(ii) Household income and expenditure;
(iii) Price and supply of commodities;
(iv) Amount of rainfall and yield of crops.

Correlation between two variables is said to be negative or inverse if the variables deviate in opposite direction. That is, if the increase in the variables deviate in opposite direction. That is, if increase (or decrease) in the values of one variable results on an average, in corresponding decrease (or increase) in the values of other variable.

Some examples of series of negative correlation are:

(i) Volume and pressure of perfect gas;

(ii) Current and resistance [keeping the voltage constant] $(R=\frac{V}{I})$;

(iii) Price and demand of goods.

Graphs of Positive and Negative correlation:

Suppose we are given sets of data relating to heights and weights of students in a class. They can be plotted on the coordinate plane using x – axis to represent heights and y – axis to represent weights. The different graphs shown below illustrate the different types of correlations.

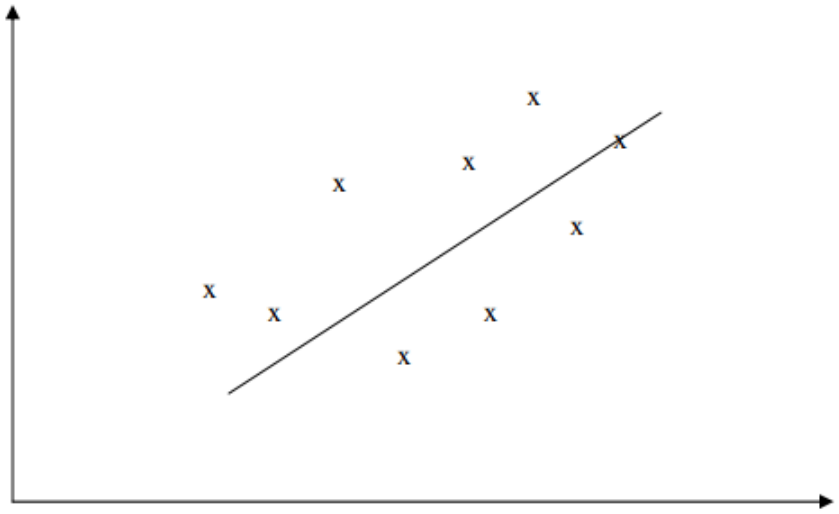


Figure for positive correlation

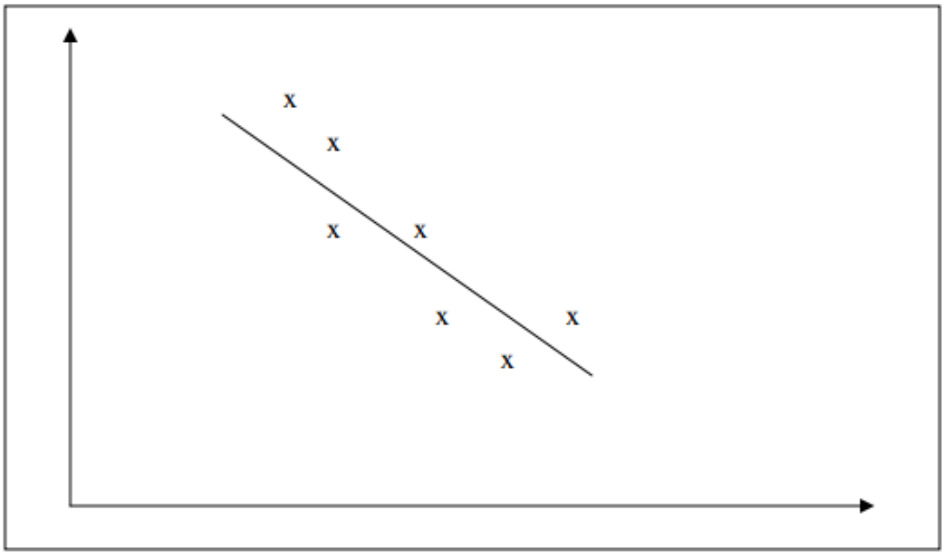


Figure for negative correlation

Note:

(i) If the points are very close to each other, a fairly good amount of correlation can be expected between the two variables. On the other hand if they are widely scattered a poor correlation can be expected between them.

(ii) If the points are scattered and they reveal no upward or downward trend as in the case of (d) then we say the variables are uncorrelated.

(iii) If there is an upward trend rising from the lower left hand corner and going upward to the upper right hand corner, the correlation obtained from the graph is said to be positive. Also, if there is a downward trend from the upper left hand corner the correlation obtained is said to be negative.

(iv) The graphs shown above are generally termed as **scatter diagrams**.

Example:1: The following are the heights and weights of 15 students of a class. Draw a graph to indicate whether the correlation is negative or positive.

Heights (cms)	Weights (kgs)
170	65
172	66
181	69
157	55
150	51
168	63
166	61
175	75
177	72
165	64
163	61
152	52
161	60
173	70
175	72

Since the points are dense (close to each other) we can expect a high degree of correlation between the series of heights and weights. Further, since the points reveal an upward trend, the correlation is positive. Arrange the data in increasing order of height and check that , as height increases, the weight also increases, except for some (stray) cases..

EXERCISES

(1) A Company has just brought out an annual report in which the capital investment and profits were given for the past few years. Find the type of correlation (if it exists).

Capital Investment (crores)	10 16 18 24 36 48 57
Profits (lakhs)	12 14 13 18 26 38 62

(2) Try to construct more examples on the positive and negative correlations.

(3) Construct the scattered diagram of the data given below and indicate the type of correlation.

(Average Value in Lakhs of Rs.)

Years	1965	1970	1975	1980	1985	1990
Raw cotton import	42	60	112	98	118	132
Cotton manufacture exports	68	79	88	86	106	114

5.3: Linear and Non – Linear Correlation

The correlation between two variables is said to be **linear** if the change of one unit in one variable result in the corresponding change in the other variable over the entire range of values.

For example consider the following data.

X	2	4	6	8	10
Y	7	13	19	25	31

Thus, for a unit change in the value of x, there is a constant change in the corresponding values of y and the above data can be expressed by the relation

$$y = 3x + 1$$

In general two variables x and y are said to be **linearly related**, if there exists a relationship of the form

$$y = a + bx$$

where 'a' and 'b' are real numbers. This is nothing but a straight line when plotted on a graph sheet with different values of x and y and for constant values of a and b. Such relations generally occur in physical sciences but are rarely encountered in economic and social sciences.

The relationship between two variables is said to be **non – linear** if corresponding to a unit change in one variable, the other variable does not change at a constant rate but changes at a fluctuating rate. In such cases, if the data is plotted on a graph sheet we will not get a straight line curve. For example, one may have a relation of the form

$$y = a + bx + cx^2$$

or more general polynomial.

5.4: The Coefficient of Correlation

One of the most widely used statistics is the **coefficient of correlation** 'r' which measures the degree of association between the two values of related variables given in the data set. It takes values from + 1 to – 1. If two sets or data have r = +1, they are said to be perfectly **correlated positively** if r = -1 they are said to be perfectly **correlated negatively**; and if r = 0 they are **uncorrelated.**

The coefficient of correlation 'r' is given by the formula

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

The following example illustrates this idea.

Example:2: A study was conducted to find whether there is any relationship between the weight and blood pressure of an individual. The following set of data was arrived at from a clinical study. Let us determine the coefficient of correlation for this set of data. The first column represents the serial number and the second and third columns represent the weight and blood pressure of each patient.

S. No.	Weight	Blood Pressure
1.	78	140
2.	86	160
3.	72	134
4.	82	144
5.	80	180
6.	86	176
7.	84	174
8.	89	178
9.	68	128
10.	71	132

Solution:

x	y	x^2	y^2	xy
78	140	6084	19600	10920
86	160	7396	25600	13760
72	134	5184	17956	9648
82	144	6724	20736	11808
80	180	6400	32400	14400
86	176	7396	30976	15136
84	174	7056	30276	14616
89	178	7921	31684	15842
68	128	4624	16384	8704
71	132	5041	17424	9372
796	1546	63,776	243036	1242069

Then

$$r=\frac{10(124206)-(796)(1546)}{\sqrt{\left[(10)63776-(796)^2\right]\left[(10)(243036)-(1546)^2\right]}}$$

$$=\frac{11444}{\sqrt{(1144)(40244)}}$$

$$=0.5966$$

5.4: Rank Correlation

Data which are arranged in numerical order, usually from largest to smallest and numbered 1,2,3 ---- are said to be in **ranks** or **ranked data.**. These ranks prove useful at certain times when two or more values of one variable are the same. The coefficient of correlation for such type of data is given by **Spearman rank difference correlation coefficient** and is denoted by R.

In order to calculate R, we arrange data in ranks computing the difference in rank 'd' for each pair. The following example will explain the usefulness of R. R is given by the formula

$$R=1-6\frac{(\sum d^2)}{n(n^2-1)}$$

Example:3: The data given below are obtained from student records. Calculate the rank correlation coefficient 'R' for the data.

Subject	Grade Point Average (x)	Graduate Record exam score (y)
1.	8.3	2300
2.	8.6	2250
3.	9.2	2380
4.	9.8	2400
5.	8.0	2000
6.	7.8	2100
7.	9.4	2360
8.	9.0	2350
9.	7.2	2000
10.	8.6	2260

Note that in the G. P. A. column we have two students having a grade point average of 8.6 also in G. R. E. score there is a tie for 2000.

Now we first arrange the data in descending order and then rank 1,2,3,---- 10 accordingly. In case of a tie, the rank of each tied value is the mean of all positions they occupy. In x, for instance, 8.6 occupy ranks 5 and 6. So each has a rank $\frac{5+6}{2}=5.5$;

Similarly in 'y' 2000 occupies ranks 9 and 10, so each has rank $\frac{9+10}{2}=9.5$.

Now we come back to our formula $R=1-\frac{6\sum d^2}{n(n^2-1)}$

We compute 'd' , square it and substitute its value in the formula.

Subject	x	y	Rank of x	Rank of y	d	d^2
1.	8.3	2300	7	5	2	4
2.	8.6	2250	5.5	7	-1.5	2.25
3.	9.2	2380	3	2	1	1
4.	9.8	2400	1	1	0	0
5.	8.0	2000	8	9.5	-1.5	2.25
6.	7.8	2100	9	8	1	1
7.	9.4	2360	2	3	-1	1
8.	9.0	2350	4	4	0	0
9.	7.2	2000	10	9.5	0.5	0.25
10.	8.6	2260	5.5	6	-0.5	0.25

So here, n = 10, sum of $d^2 = 12$. So

$$R=1-\frac{6(12)}{10\,(100-1)}$$
$$=1-0.0727 \quad =0.9273$$

Note: If we are provided with only ranks without giving the values of x and y we can still find Spearman rank difference correlation R by taking the difference of the ranks and proceeding in the above shown manner.

EXERCISES

1. A horse owner is investigating the relationship between weight carried and the finish position of several horses in his stable. Calculate r and R for the data given

Weight Carried	Position Finished
110	2
113	6
120	3
115	4
110	6
115	5
117	4
123	2
106	1
108	4
110	1
110	3

2. The top and bottom number which may appear on a die are as follows

Top	1	2	3	4	5	6
bottom	5	6	4	3	1	2

Calculate r and R for these values. Are the results surprising?

3. The ranks of two sets of variables (Heights and Weights) are given below. Calculate the Spearman rank difference correlation coefficient R.

	1	2	3	4	5	6	7	8	9	10
Heights	2	6	8	4	7	4	9.5	4	1	9.5
Weights	9	1	9	4	5	9	2	7	6	3

5.5: Regression

If two variables are significantly correlated, and if there is some theoretical basis for doing so, it is possible to predict values of one variable from the other. This observation leads to a very important concept known as 'Regression Analysis'.

Regression analysis, in general sense, means the estimation or prediction of the unknown value of one variable from the known value of the other variable. It is one of the most important statistical tools which is extensively used in almost all sciences – Natural, Social and Physical. It is specially used in business and economics to study the relationship between two or more variables that are related causally and for the estimation of demand and supply graphs, cost functions, production and consumption functions and so on.

Prediction or estimation is one of the major problems in almost all the spheres of human activity. The estimation or prediction of future production, consumption, prices, investments, sales, profits, income etc. are of very great importance to business professionals. Similarly, population estimates and population projections, GNP, Revenue and Expenditure etc. are indispensable for economists and efficient planning of an economy.

Regression analysis was explained by M. M. Blair as follows: "Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data."

5.5.1: Regression Equation

Suppose we have a sample of size 'n' and it has two sets of measures, denoted by x and y. We can predict the values of 'y' given the values of 'x' by using the equation, called the REGRESSION EQUATION.

$$y^* = a + bx$$

where the coefficients a and b are given by

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

$$a = \frac{\sum y - b\sum x}{n}$$

The symbol y* refers to the predicted value of y from a given value of x from the regression equation.

Example: 4 : Scores made by students in a statistics class in the mid - term and final examination are given here. Develop a regression equation which may be used to predict final examination scores from the mid – term score.

STUDENT	MID – TERM	FINAL
1.	98	90
2.	66	74
3.	100	98
4.	96	88
5.	88	80
6.	45	62
7.	76	78
8.	60	74
9.	74	86
10.	82	80

Solution:

We want to predict the final exam scores from the mid term scores. So let us designate ‘y’ for the final exam scores and ‘x’ for the mid – term exam scores. We open the following table for the calculations.

Stud	x	y	X^2	xy
1	98	90	9604	8820
2	66	74	4356	4884
3	100	98	10,000	9800
4	96	88	9216	8448
5	88	80	7744	7040
6	45	62	2025	2790
7	76	78	5776	5928
8	60	74	3600	4440
9	74	86	5476	6364
10	82	80	6724	6560
Total	785	810	64,521	65,071

Numerator of b = 10 * 65,071 – 785 * 810 = 6,50,710 – 6,35,850 = 14,860
Denominator of b = 10 * 64, 521 – $(785)^2$ = 6,45,210 – 6,16,225 = 28,985

Therefore, b = 14,860 / 28,985 = 0.5127

Numerator of a = 810 – 785 * 0.5127 = 810 – 402.4695 = 407.5305
Denominator of a = 10

Therefore a = 40.7531

Thus , the regression equation is given by

$$y^* = 40.7531 + (0.5127)\, x$$

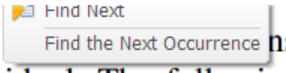
We can use this to find the projected or estimated final scores of the students.
For example, for the midterm score of 50 the projected final score is
$y^* = 40.7531 + (0.5127)\, 50 = 40.7531 + 25.635 = 66.3881$
which is a quite a good estimation.
To give another example, consider the midterm score of 70. Then the projected final score is
$y^* = 40.7531 + (0.5127)\, 70 = 40.7531 + 35.889 = 76.6421,$
which is again a very good estimation.

This brings us to the end of this chapter. We close with some problems for you.

EXERCISES

1. The data given below are obtained from student records. Calculate the regression equation and compute the estimated GRE scores for GPA = 7.5, 8.5..

Subject	Grade Point Average (x)	Graduate Record exam score (y)
11.	8.3	2300
12.	8.6	2250
13.	9.2	2380
14.	9.8	2400
15.	8.0	2000
16.	7.8	2100
17.	9.4	2360
18.	9.0	2350
19.	7.2	2000
20.	8.6	2260

2. A study was conducted to find whether 

nship between the weight and blood pressure of an individual. The following set of data was arrived at from a clinical study.

S. No.	Weight	Blood Pressure
1.	78	140
2.	86	160
3.	72	134
4.	82	144
5.	80	180
6.	86	176
7.	84	174
8.	89	178
9.	68	128
10.	71	132

3. A horse was subject to the test of how many minutes it takes to reach a point from the starting point. The horse was made to carry luggage of various weights on 10 trials.. The data collected are presented below in the table.

Trial No.	Weight (in Kgs)	Time taken (in mins)
1	11	13
2	23	22
3	16	16
4	32	47
5	12	13
6	28	39
7	29	43
8	19	21
9	25	32
10	20	22

Find the regression equation between the load and the time taken to reach the goal. Estimate the time taken for the loads of 35 Kgs , 23 Kgs, and 9 Kgs. Are the answers in agrrement with your intuitive feelings? Justify.

The least squares regression line:-

The least squares regression line is the line which produces the smallest value of the sum of the squares of the residuals. A residual is the vertical distance from a point on a scatter diagram to the line of best fit. Therefore the least squares regression line can be seen as the best line of best fit.

The equation of the least squares regression line of y on x is:

$$y-\overline{y}=b\left(x-\overline{x}\right)$$

Where b is:

$$b=\frac{S_{xy}}{S_{xx}}=\frac{\sum\left(x_i-\overline{x}\right)\left(y_i-\overline{y}\right)}{\sum\left(x_i-\overline{x}\right)^2}=\frac{\frac{\sum x_i y_i}{n}-\overline{x}\,\overline{y}}{\frac{\sum x_i^2}{n}-\overline{x}^2}$$

As you can see there are 3 different ways to calculate the value for b, based on what information you are given in the question

Example

Calculate the least squared regression line of y on x from the following data:

x	20	30	40	50	60	70
y	2.49	2.41	2.38	2.14	1.97	2.03

Firstly draw up a table of the values you need and fill it out. In this example we'll use the third equation for b so we need all the values of x^2 and $x_i y_i$:

	x	y	x^2	$x_i y_i$
	20	2.49	400	49.9
	30	2.41	900	72.3
	40	2.38	1600	95.2
	50	2.14	2500	107
	60	1.97	3600	118.2
	70	2.03	4900	142.1
Sum:	270	13.42	13900	584.7

Next step is to calculate the means of the x and y values:

$$\overline{x} = \frac{270}{6} = 45$$

$$\overline{y} = \frac{13.42}{6} = 2.24$$

Now the value of b can be calculated:

$$b = \frac{\frac{584.7}{6} - 45 \text{ x } 2.24}{\frac{13900}{6} - 45^2} = -0.01$$

Hence the equation is therefore:

$$y - 2.24 = -0.01(x - 45)$$

Cleaning it up:

$$y = -0.01x + 2.69$$

Properties of the Regression Line -

- When the regression parameters (b0 and b1) are defined as described above, the regression line has the following properties.
- The line minimizes the sum of squared differences between observed values (the y values) and predicted values (the y values computed from the regression equation).
- The regression line passes through the mean of the X values (x) and through the mean of the Y values (y).
- The regression constant (b0) is equal to the y intercept of the regression line.
- The regression coefficient (b1) is the average change in the dependent variable (Y) for a 1-unit change in the independent variable (X). It is the slope of the regression line.

- The least squares regression line is the only straight line that has all of these properties.

Unit – 4

Curve Fitting and The Method of Least Squares

Relationship between variables:-

Very often in practice a relationship is found to exist between two (or more) variables. For example: The weights of adults males depend to some degree on their heights; circumstances and areas of circles depend on their radii; and the pressure of a given mass of gas depends on its temperature and volume.

The last two mentioned type of relation can be proven purely mathematically/physically. But what about the lack of Ozon-stratum and occurences of skin cancer. Or women's inclination to get abortions related to the where the women live, in cities and towns or in the country.

It is frequently desiable to express the relations in mathematical form by determining an equation connecting the variables.

Curve Fitting:-

To aid in determining an equation connecting variables, a first step is the collection of data showing corresponding values of the variables under consideration.

For example, suppose X and Y denote respectively the height and the weight of adult males. Then a sample N individuals would reveal the heights X1, X2,…..XN and the corresponding weight Y1, Y2,…..YN

A next step is to plot the points (X1,Y1), (X2,Y2)…..(XN,YN) on a rectangular coordinate system. The resulting set of point is sometimes called a scatter diagram.

From the scatter diagram it is often possible to visualize a smooth curve approximating the data. Such a curve is called an approximating curve. In the next diagram, for example, the data appears to be aproximated well by a straight line and perhaps a linear relationship exists between the two variables. In the diagram to the right some other relation between the two variables seems to exist, perhaps some curve.

The general problem of finding equations of approximating curves which fit given sets of data is called curve fitting.

Equations of Approximating Curves

For purposes of reference we have listed below several common type of approximating curves and equations. All letters other than X and Y represent constants. The variables X and Y are often referred to as independent and dependent variables respectively, although these roles can be interchanging.

1	Y = a0 + a1X	Straight line
2	Y = a0 + a1X + a2X^2	Parabola curve
3	Y = a0 + a1X + a2X^2 + a3X^3	Cubic curve

4	Y = a0 + a1X + a2X^2 + a3X^3 + a4X^4	Quartic curve
5	Y = a0 + a1X + a2X^2 +…+anX^n	*n*th degree curve
6	Y= 1/(a0 +a1X) or 1/Y = a0 + a1X	Hyperbola
7	Y = ab^X or logY = loga + Xlogb = a0 + a1X	Exponential curve
8	Y = ab^X or log Y = log a + blogX	Geometric curve
9	Y = ab^X + g	Modified exponential curve
10	Y = aX^b + g	Modified geometric curve
11	Y = pq^b^X or logY = logp + b^Xlogq = ab^X + g	Gompertz curve
12	Y = pq^b^X + h	Modified Gombertz curve
13	Y = 1/(ab^X +g) or 1/Y = ab^X + g	Logistic curve
14	Y = a0 + a1(logX) + a2(logX)^2	Logistic curve

The right sides of the above equations (1-5) are called polynomials of the first, second, third, fourth and *n*'th degree respectively. The functions defined by the first four of these equations are sometimes called linear, squared, cubic and quartic functions respectively.

To decide which curve should be used, it is helpful to obtain scatter diagrams of transformed variables. For example, if a scatter diagram of Log Y vs X shows a linear relationship the equation has the form (7), while if Log Y vs Log X shows a linear relationship the equation has the form (8). Logarithmic or double-logarithmic paper to show the functions when one or both scales are calibrated logarithmically.

Freehand Method of Curve Fitting

Individual judgment can often be used to draw an approximating curve to fit a set of data. This is called a freehand method of curve fitting. If the type of equation of this curve is known, it is possible to obtain the constants in the equation by choosing as many points on the curve as there are constants in the equation. For example, if the curve is a straight line, two points a necessary, if it is a parabola, three points are necessary. This method has the disadvantage that different observers will obtain different curves and equation. The method is *unequivocable*.

The straight line

The simplest type of approximating curve is a straight line, whose equation can be written:

$Y = a0 + a1X$ (15)

Given the two points (X1,Y1) and (X2,Y2) on the line, the constants a0 and a1 can be determined. The resulting equation can be writtten:

$Y - Y1 = (X – X1)*(Y2 -Y1)/(X2 –X1)$ or $Y - Y1 = m(X - X1)$, where $m = (Y2 – Y1)/(X2 –X1)$ is called the slope of the line and represents the change in Y divided by the corresponding change in X.

When the equation is written in the form (15) the constant a1 is the slope *m*. The constant a0, which is the value of Y when X = 0, is called the U intercept.

The Method of Least Squares

To avoid invidual judgment in constructing lines, parabolas or other approximating curves to fit sets of data, it is necessary to agree on a definition of a "best fitting line", "best fitting parabola", etc.

To motivate a possible definition, consider the next figure in which the data points are given by (X1,Y1), (X2, X2)…(Xn,Yn). For a given value of X, say X1, there will be a difference between the value Y1 and the corresponding value determined from the curve. As indicated in the figure we denote this difference by D1, D2, D3, D4, *D5*…*Dn* , which is sometimes referred to as a derivation, error or residual and may be positive, negative, or zero. Similarly, corresponding to the values X1,…Xn we obtain D1 …Dn.

A measure of the "goodness of fit" of the curve to the given data is provided by the quantity $D1^2 + D2^2 + D3^2 +….Dn^2$. Somebody might propose D1 + D2 +D3 +….Dn to calculate the total divergence.

But it does not work as it sums to zero, if the curve has drawn accurately, half of the Ds are as positive, as the rest are negative all together.

Best fitting curve: □ D^2 is a minimum, where □ sums all the D^2 from 1 to n. A curve having this property is said to fit the data in the least square sense and it is called the Least Square Curve.

Thus a line having this property is called the least square line, a parabola with this property is called a least square parabola, etc.

It is customary to employ the above definition when X is a independent variable and Y is the dependent variable. If X is the dependent variable the definition is modified by considering the horizontal instead of the vertical deviation, which amounts to an interchange of the X and Y axes. These two definitions in general lead to different least square curves. Unless otherwise specified we shall consider Y as the dependent and X as the independent variable.

It is possible to define another least square curve by considering perpendicular distances from each of the data points to the curve instead of either vertical or horizontal distances. However, this is not often used.

The Least Square Line
The least square line approximating the set of points (X1,Y1), (X2,Y2)…(Xn,Yn) has an equation:

Y = a0 + a1X (18)

where the constants a0 and a1 are determined by solving simultaneously the equations:

□ Y = a0N + a1□ X

□ XY = a0□ X + a1□ X^2 (19)

which are called the normal equations for the least square line (18)

a0 = [(□ Y)(□ X^2) - (□ X)(□ XY)]/[N□ X^2 - (□ X^2)]

a1 = [N□ XY – (□ X)(□ Y)]/[N□ X^2 – (□ X)^2] (20)

The normal equation (19) are easily remembered by observing that the first equation can be obtained formally by summing on both sides of (18), i.e. □ Y = □ (a0 + a1X) = a0N + a1□□X, while the second equation is obtained by first multiplying both sides of (18) by X and then summing, □ XY = □ X(a0 + a1X) = a0□ X + a1□ X^2 . Note that this is not a derivation of the normal equation but simply a mean for remembering them. For a derivation using the calculus. The sum of squared derivations is minimized by derivating this sum S and equalizing it to zero (that is a simple mathematical rule not to be proved here).

S = (a0 + a1X1 - Y1)^2 +….+(a0 + a1Xn +Yn)^2

The labor involved in finding a least square line can sometimes be shortened by transforming the data so that x = X - Xg and y = Y - Yg, where the notation g means average. The equation of the least square line can then be written :

y = (□ xy/□ x^2)x or x = (□ xy/□ y^2)y (21)

This is easily proved but not included here.

If particular X is such that □ X = 0, i.e. Xg =0, this becomes:

Y = Yg + (□ XY/□ X^2)X (22)

From these equations it is at once evident that the least square line passes through the point (Xg,Yg), called the centroid or the center of gravity of data.

If the variable X is taken as dependent instead of independent variable, we write (18) X = b0 + b1Y. Then the above results hold if X and Y are interchanged and a0 and a1 are replaced by b0 and b1 respectively. The resulting least square line, however, is in general not the same as that obtained above.

Non-linear Relationships
Non-linear relationships can sometimes be reduced to linear relationships by appropriate transformation of variables. The logarithm-function can sometimes be used. But it is often the problem to find out if the original data fit a known mathematical function

The least square Parabola
The least square parabola approximating the set of points (X1,Y1)…(Xn,Yn) has the equation:

Y = a0 + a1X + a2X^2

where the constants a0, a1 and a2 are determined by solving simultaneously the equations:

□ Y = a0N + a1□ X +a2□ X^2

□ XY = a0□ X + a1□ X^2 + a2□ X^3

□ X^2Y = a0□ X^2 + a1□ X^3 + a2□ X^4 (24)

called the normal equation for the least square parabola (23)

A least square cubic or quartic curve can easily be determined by extending the multiplying technic mentioned in the section following (20).

Regression
Often, on the basis of sample data, we wish to estimate the value of a variable Y corresponding to a given value of a variable X. This can be accomplished by

estimating the value of Y from a least square curve which fits the sampled data. The resulting curve is called a regression curve of Y on X, since Y is estimated from X.

If we desired to estimate the value of X from a given value of Y we would use a regression curve of X on Y, which amounts to interchanging the variables in the scatter diagram so that X is the dependent variable and Y is the independent variable. This is equivalent to replacing vertical deviations in the definition of least square curve by horizontal deviations. The last is based on simple mathematic not included here. More about regression and correlation in another link.

Applications to Time Series
If the independent variable X is time, the data shows the values of Y at various times. Data arranged according to time are called time series. The regression line or curve of Y on X in this case is often called a trend line or trend curve and is often used for purposes of estimation, prediction or forecasting.

Problems Involving More Than Two Variables
Problems involving more than two variables can be treated in a manner analogious to that for two variables. For example, there may be a relationship between the three variables X, Y and Z which can be described by the equation:

Z = a0 + a1X + a2Y (25)

which is called a linear equation in the variables X, Y and Z.

In the three dimensional rectangular coordinate system this equation represents a plan and the actual sample points (X1,Y1,Z1),…..(Xn,Yn,Zn,) might "scatter" not too far from this plane which we can call an approximating plane.

By extension of the method of least squares, we can speak of a least square plane approximating data. If we are estimating Z from given values of X and Y, this would be called a regression plan of Z on X and Y. The normal equations corresponding to the least square plane (25) are given by:

□ Z = a0N + a1□ X + a2□Y

□ XZ = a0□ X + a1□X^2 + a2□ XY

□ YZ = a0 □□Y + a1□ XY + a2□ Y^2 (26)

and can be remembered as obtained from (25) by multiplying by 1, X and Y succesively and then summing.

More complicated equations than (25) can also be considered. These represent regression surfaces. If the number of variables exceeds three, geometric intuition is lost since we then require four, five … dimentional spaces.

Problems involving estimation of a variable from two or more variables are called problems of multi-regression and will be dealt with in more detail under another link.

Problems

1.

Corresponding data of the variables X and Y, height and weight of males

Height (inches)	70	63	72	60	66	70	74	65	62	67	65	68
Weight (pounds)	155	150	180	135	156	168	178	160	132	145	139	152

Obtain a scatter diagram from the data:

The scatter diagram is obtained by plotting the points (70,155), 63,150)….(68,152). By using a ruler you find several straight line which apparently suites the relation in question. Choosing any two point on the line just drawn you can account the slope of a fitting line.

$Y - Y1 = (X - X1)(Y2 - Y1)/(X2 - X1)$

$(170 - 156)/(68 - 66) = 7$

$Y - 156 = 7 (X - 66)$

Y = 7X – 306

If X=63 then Y= 7*63 - 306 = 135 provided that the line expresses the relation between height and weight among females in right way. We chose the best fitting line in the diagram, we hope. As we shall see below this method is certainly not exactly. Instead of the point (66,156) we could have chosen (65,139) and got the result: Y = 10.33X – 316. Train a more exact method below.

2.

Find the least square line to the following data: (1,1), (3,2), (4,4), (6,4), (8,5), (9,7), (11,8), (14,9)

The equation of the line is Y = a0 + a1X. The normal equations are:

□□Y = a0N + a1□ X

□ XY = a0□ X + a1□ X^2

The work involved in computing the sums can be arranged as in the following table. Although the last column is not needed for this part of the problem, it has been added to the table for use with X as an dependent variable which gives quite another result. The last called regression of X on Y.

X	Y	X^2	XY	Y^2
1	1	1	1	1
3	2	9	6	4
4	4	16	16	16
6	4	36	24	16
8	5	64	40	25
9	7	81	63	49
11	8	121	88	64
14	9	196	126	81
□ X = 56	□ Y = 40	□ X^2 = 524	□ XY = 364	□ Y^2 = 256

Since there are 8 pairs of values of X and Y, N =8 and the normal equations become:

8a0 + 56a1 = 40

56a0 + 524a1 = 364

Solved simultaneously, a0 =
6/11 or 0.545, a1= 7/11 or 0.636

Another method:

a0 = [(□ Y)(□ X^2) - (□ X)(□ XY)]/[N□ X^2 - (□ X)^2] = [(40*524) - (56*364)]/[(8*524) - (56)^2] = 6/11 or 0.545

a1 = [N□ XY - (□ X)(□ Y)]/[N□ X^2 - (□ X)^2] = [(8*364) - (56*40)]/[(8*524) – (56)^2] =

7/11 or 0.636

Perhaps we should try to estimate the regression line from the example with heigts and weights a little more exactly:

Height X	Weight Y	x=X-Xg	y=Y-Yg	xy	x^2	y^2
70	155	3.2	0.8	2.56	10.24	0.64
63	150	-3.8	-4.2	15.96	14.44	17.64
72	180	5.2	25.8	134.16	27.04	665.64
60	135	-6.8	-19.2	130.56	46.24	368.64
66	156	-0.8	1.8	-1,44	0.64	3.24
70	168	3.2	13.8	44.16	10.24	190.44
74	178	7.2	23.8	171.36	51.84	566.44
65	160	-1.8	5.8	-10.44	3.24	33.64
62	132	-4.8	-22.2	106.56	23.04	492.84
67	145	0.2	-9.2	-1.84	0.04	84.64
65	139	-1.8	-15.2	27.36	3.24	231.04

68	152	1.2	-2.2	-2.64	1.44	4.84
□	□			□ xy=	□ x^2=	□ y^2=
X=802	Y=1850					
				616.32	191.68	2659.68

Xg=66.8 Yg=154.2

The required least square line has this equation:

Y - 154.2 = 3.22(X – 66.8) or

Y= 3.22X - 60.9

Sometimes when the raw figures are large it is an advantage to subtract a large figure at least from one of, perhaps from both of the variables before accounting. Then you must remember to add the same figures to the averages you account to get the right result.

Example of least square

The given example explains you that how to find the equation of straight line or least square line by using the method of least square, which is very useful in statistics as well as in mathematics.

Example:

Fit a least square line to the following data. Also find trend values and show that $\Sigma\left(Y-\hat{Y}\right)=0$.

X	1	2	3	4	5
Y	2	5	3	8	7

Solution:

X	Y	XY	X^2	$\hat{Y}=1.1+1.3X$	$Y-\hat{Y}$
1	2	2	1	2.4	-0.4
2	5	10	4	3.7	+1.3
3	3	9	9	5.0	-2

4	8	32	16	6.3	1.7
5	7	35	25	7.6	-0.6
$\sum X = 15$	$\sum Y = 25$	$\sum XY = 88$	$\sum X^2 = 55$	Trend Values	$\sum\left(Y - \hat{Y}\right) = 0$

The equation of least square line $Y = a + bX$

Normal Equation for '*a*' $\sum Y = na + b\sum X$ $25 = 5a + 15b$ ---- (1)

Normal Equation for '*b*' $\sum XY = a\sum X + b\sum X^2$ $88 = 15a + 55b$ ----(2)

Eliminate '*a*' from equation (1) and (2), multiply equation (2) by 3 and subtract form equation (2), we get the values of '*a*' and '*b*'.

Here $a = 1.1$ and $b = 1.3$, the equation of least square line becomes $Y = 1.1 + 1.3X$.

For the trends values, put the values of '*X*' in above equation, see the above table column 4.

Contingency tables and Chi-squared Tests

In this type of analysis we have two characteristics, such as gender and eye colour, which cannot be measured but which can be used to group people by variations within them. These characteristics may, or may not, be associated in some way. How can we decide?

We can take a random sample from the population, note which variation of each characteristic is appropriate for each case and then cross-tabulate the data. It is then analysed in order to see if the proportions of each characteristic in the sub-samples are the same as the overall proportions - easier to do than describe! As an example, if there is no relationship between gender and eye colour we would expect similar proportions of males and females to have blue eyes.

The Variables

If the variables are, as is usually the case, nominal, (described by name only), frequencies may be cross-tabulated by each category within each variable. Ordinal

variables may be used if there are only a limited number of orders so that each one can be classified as a separate category. Continuous variables may be grouped and then tabulated similarly though the results will then vary according to the grouping categories.

Contingency Tables (Cross-Tabs)

You have met this type of table before as a contingency table when calculating probabilities. As a reminder: cases are allotted to categories and their frequencies cross-tabulated; e.g. in the gender / eye colour example there might be blue eyed males, blue eyed females, brown eyed males, brown eyed females, etc. These tables are known as contingency tables. All possible 'contingencies' are included in the 'cells' which are themselves mutually exclusive. The table is completed by calculating the 'row totals', the 'column totals' and the 'grand total'.

Expected values

If the two variables, the characteristics under scrutiny, are completely independent, the proportions within the sub-totals of the contingency table would be expected to be the same as those of the totals for each variable. In practice we work with frequencies rather than proportions, distinguishing between 'observed' and 'expected' frequencies by enclosing the latter within brackets. If gender and eye colour are independent and if a third of the population has blue eyes, we would expect a third of males to be blue eyed and a third of females to be blue eyed.

These proportions are obviously contrived so as to be easy to work with. How can we cope with more awkward numbers? In the on-going example, the proportions are first calculated as fractions which are then multiplied by the total frequency to find the expected individual cell frequencies. This produces a formula which is applicable in all cases:

For any cell the expected frequency is calculated by:

$$\frac{\textbf{Row total} \times \textbf{Column total}}{\textbf{Overall total}},$$

where the relevant row and column are those crossing in that particular cell.

Chi-squared (X^2) Test for Independence

The hypothesis test which is carried out in order to see if there is any association between categorical variables, such as gender and eye colour, is known as the Chi-squared, (χ^2), test,

Example 1

The following table compiled by a personnel manager relates to a random sample of 180 staff taken from the whole workforce of the supermarket chain. We shall, in this example, test for association between a member of staff's gender and his/her type of job, at the 5% level of significance.

	Male	Female	Total
Supervisor	20	15	
Shelf stacker	20	30	
Till operator	10	35	
Cleaner	10	40	
Total			

Completing the row and column totals, as previously with probability, gives the full table.

In this example we randomly selected a sample of 180 supermarket staff and found that two thirds, 120, of them were female and one third, 60, were male. Assuming there is no association between gender and job category and finding that we have 45 till operators, we would expect two thirds, 30, of them to be female and one third, 15, of them to be male.

Note that these figures are a quarter of each gender respectively, which checks since till operators, 45, form a quarter of the total staff, 180.

We now calculate the other **expected frequencies** from the probabilities and put them into the table: (See Section 4.8)

- P (Supervisor) = 35/180
- P (Male) = 60/180
- P (Supervisor and male) = 35/180 x 60/180 assuming independence.

Therefore the **expected number** of Male supervisors = $\frac{35}{180} \times \frac{60}{180} \times 180 = 11.67$

This is the expected frequency for members of staff who are both male and a supervisor.

Note that it is a theoretical number which does not have to be an integer.

This simplifies to: $\frac{\text{Row total} \times \text{Column total}}{\text{Overall total}}$, $\frac{35 \times 60}{180} = 11.67$

Calculating the other expected frequencies and inserting them in the table (in brackets):

	Male	Female	Total
Supervisor	20 (11.67)	15 (23.33)	35
Shelf stacker	20 (16.67)	30 (33.33)	50
Till operator	10 (15.00)	35 (30.00)	45
Cleaner	10 (16.67)	40 (33.33)	50
Total	60	120	180

These are the frequencies which would be expected if there is no association between gender and job category at the supermarket.

If the expected frequencies are observed to actually occur in practice then we can deduce that the two variables are indeed independent.

We would obviously not expect to get exact agreement with the expected frequencies, so some critical amount of difference is allowed and we compare the difference from our observations with that allowed by the use of a standard table. Are the values observed so different to those expected that we must reject the idea of independence? Or are the results just due to sampling errors, with the variables actually being independent?

It is to be hoped that you recognise the need for a hypothesis test!

The chisquared (X^2) Hypothesis test:-

To find the answer, we analyse the data and compare the result to a standard table figure. We carry out a formal hypothesis test at 5% significance: **the chi-squared test**.

1) State Null Hypothesis, H_0, (that of no association) and Alternative Hypothesis, H_1.

2) Record observed frequencies, O, in each cell of the contingency table.

3) Calculate row, column and grand totals.

4) Calculate expected frequency, E, for each cell : $\frac{\text{row total x column total}}{\text{grand total}}$

Note that: No expected frequency should be less than 1 and the number of expected frequencies below 5 should not be over 20% of the total number of cells. Otherwise the test is invalid.

5) Find critical value from chi-square table, as appended, with $(r - 1) \times (c - 1)$ degrees of freedom where r and c are the number of rows and columns respectively.

6) Calculate test statistic: $\sum \frac{(O - E)^2}{E}$

7) Compare the two values and conclude whether the variables are independent or not.

In example 1, we have already carried out steps 2, 3 and 4 of the procedure by calculating the expected values. Whether these are calculated before, or during, the test is up to personal preference. Some statisticians also prefer to calculate the test

statistic, as this procedure is rather lengthy, before starting the test and to then insert the calculated value in the formal hypothesis test.

The test statistic is an overall measure of the difference between the expected and observed frequencies. Each cell difference is squared so that positive and negative differences have the same weighting and proportioned by the size of the expected cell contents. When the contributions from each cell are totalled their sum is compared with a critical value from the chi-squared table – hence the name of this test.

Null Hypothesis (H_0): There is **no association** between gender and job category. (Remember that 'null' means none.)

Alternative Hypothesis (H_1): There is an association between gender and job category.

Critical Value: from the chi-squared table

Number of degrees of freedom (ν) = (r - 1)(c - 1) = (4 - 1)(2 - 1) = 3 x 1 = 3;

X^2 table, as appended, is always one tailed. Level of significance = 5%

$X^2_{5\%,\ \nu=3} = 7.816$

Test statistic

The test statistic is calculated from the contingency table which includes both the observed and the expected values for the frequency of staff. The data may be tabulated, as we shall do in this example, or the contribution of each cell may be

calculated directly as $\frac{(O-E)^2}{E}$ and then the test statistic found as the sum of these contributions:

	Male	Female	Total
Supervisor	20 (11.67)	15 (23.33)	35
Shelf stacker	20 (16.67)	30 (33.33)	50
Till operator	10 (15.00)	35 (30.00)	45
Cleaner	10 (16.67)	40 (33.33)	50
Total	60	120	180

Test statistic $= \sum \frac{(O-E)^2}{E}$

O	E	(O - E)	$(O - E)^2/E$
20	11.67	8.33	5.946
15	23.33	-8.33	2.974

20	16.67	3.33	0.665
30	33.33	-3.33	0.333
10	15.00	-5.00	1.667
35	30.00	5.00	0.833
10	16.67	-6.67	2.669
40	33.33	6.67	1.335
		Total	16.422

Test Statistic: 16.422

Conclusion: Test statistic > Critical value therefore reject H_0 .

Conclude that there is an association between gender and job category in the supermarket chain.

Looking again at the data we can see that far more males than expected were supervisors or shelf stackers and more females were cleaners or till operators.

Example 2 In this example we first have to set up the contingency table from the following information collected from a questionnaire:

In a recent survey within a Supermarket chain, a random sample of 160 employees: stackers, sales staff and administrators, were asked to grade their attitude towards future wage restraint on the scale:

Very favourable; favourable; unfavourable; very unfavourable.

Of the 40 stackers interviewed, 7 gave the response 'favourable', 24 the response 'unfavourable', and 8 the response 'very unfavourable'. There were 56 sales staff and from these, 10 responded 'very unfavourable', 9 responded 'favourable' and 3 responded 'very favourable'. The rest of the sample were administrators. Of these, 16 gave the response 'very favourable' and 2 the response 'very unfavourable'. In the whole survey, exactly half the employees interviewed responded 'unfavourable'.

We first draw up a contingency table showing these results and then test whether attitude towards future wage restraint is dependent on the type of employment.

Setting up the table: in this example there are three types of employee giving four different responses, i. e. we have a 3 x 4 (or a 4 x 3) table.

Adding extra rows and columns for the subtotals and titles we need 5 x 6 cells.

Have a go at compiling the table. As you come to each number in the frequency of response above insert it into the appropriate cell; then find the missing figures by difference. There is sufficient information here to enable you to complete your table. When complete, check with that below before calculating the expected values.

	V.favourable	Favourable	Unfavourable	V.unfavourable	Total
Stackers	1	7	24	8	40
Sales staff	3	9	34	10	56
Administrators	16	24	22	2	64
Total	20	40	80	20	160

The expected values can next be calculated: $\frac{\text{Row total} \times \text{Column total}}{\text{Overall total}}$ and inserted.

	V.favourable	Favourable	Unfavourable	V.unfavourable	Total
Stackers	1 (5)	7 (10)	24 (20)	8 (5)	40
Sales staff	3 (7)	9 (14)	34 (28)	10 (7)	56
Administrators	16 (8)	24 (16)	22 (32)	2 (8)	64
Total	20	40	80	20	160

Hypothesis test

Null Hypothesis (H_0): There is **no association** between job category and attitude towards wage restraint.

Alternative Hypothesis (H_1): There is an association between job category and attitude towards wage restraint.

Level of Significance: 5% Level of significance

Critical value:

Number of degrees of freedom (□) = (r - 1)(c - 1) = (3 - 1)(4 - 1) = 2 x 3 = 6

Level of significance = 5%

X^2 table, Table 5 in Appendix D, 5%, 6 degrees of freedom = 12.59

Test statistic $\sum \frac{(O - E)^2}{E}$

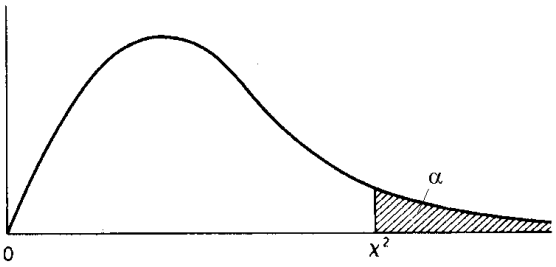
O	E	(O - E)	$(O - E)^2/E$
1	5	-4	3.200
7	10	-3	0.900
24	20	+4	0.800
8	5	+3	1.800
3	7	-4	2.286
9	14	-5	1.786
34	28	+6	1.286
10	7	+3	1.286
16	8	+8	8.000
24	16	+8	4.000
22	32	-10	3.125
2	8	-6	4.500
		Total	32.969

Test static: 32.969

Conclusion: Test statistic > Critical value therefore reject H_0

Conclude that there is an association between job category and attitude towards future wage restraint. The administrators were for it but the others against it.

Table 5 PERCENTAGE POINTS OF THE □^{2}-D1STRIB



ν	10%	5%	2 5%	1%	0.l%
1	2.706	3.841	5.024	6.635	10.83
2	4.605	5.991	7.378	9.210	13.82
3	6.252	7.816	9.351	11.35	16.27
4	7.780	9.488	11.14	13.28	18.47
5	9.236	11.07	12.83	15.08	20.51
6	10.64	12.59	14.45	16.81	22.46
7	12.02	14.07	16.02	18.49	24.36
8	13.36	15.51	17.53	20.09	26.13
9	14.68	16.92	19.02	21.67	27.89

10	15.99	18.31	20.48	23.21	29.59
11	17.28	19.68	21.92	24.72	31.26
12	18.55	21.03	23.34	26.22	32.91
13	19.81	22.36	24.74	27.69	34.51
14	21.06	23.68	26.12	29.14	36.12
15	22.31	25.00	27.49	30.58	37.70
16	23.54	26.30	28.85	32.00	39.25
17	24.77	27.59	30.19	33.41	40.79
18	25.99	28.87	31.53	34.81	42.31
19	27.20	30.14	32.85	36.19	43.82
20	28.41	31.41	34.17	37.57	45.32
21	29.62	32.67	35.48	38.93	46.80
22	30.81	33.92	36.78	40.29	48.27
23	32.01	35.17	38.08	41.64	49.73
24	33.20	36.42	39.36	42.98	51.18
25	34.38	37.65	40.65	44.31	52.62
26	35.56	38.89	41.92	45.64	54.05
27	36.74	40.11	43.19	46.96	55.48
28	37.92	41.34	44.46	48.28	56.89
29	39.09	42.56	45.72	49.59	58.30
30	40.26	43.77	46.98	50.89	59.70
40	51.81	55.76	59.34	63.69	73.42
50	63.17	67.50	71.42	76.15	86.66

60	74.40	79.08	83.30	88.38	99.61
70	85.53	90.53	95.02	100.4	112.3
80	96.58	101.9	106.6	112.3	124.8
90	107.6	113.1	118.1	124.1	137.2
100	118.5	124.3	129.6	135.8	149.5

Chi-squared Test of Independence

Two random variables x and y are called independent if the probability distribution of one variable is not affected by the presence of another.

Assume f_{ij} is the observed frequency count of events belonging to both i-th category of x and j-th category of y. Also assume e_{ij} to be the corresponding expected count if x and y are independent. The null hypothesis of the independence assumption is to be rejected if the p-value of the following Chi-squared test statistics is less than a given significance level α.

$$\chi^2 = \sum_{i,\ j} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

Tests of Significance

Once sample data has been gathered through an observational study or experiment, statistical inference allows analysts to assess evidence in favor or some claim about the population from which the sample has been drawn. The methods of inference used to support or reject claims based on sample data are known as ***tests of significance***.

Every test of significance begins with a ***null hypothesis H_0***. H_0 represents a theory that has been put forward, either because it is believed to be true or because it is to be used as a basis for argument, but has not been proved. For example, in a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on average, than the current drug. We would write H_0: there is no difference between the two drugs on average.

The ***alternative hypothesis***, H_a, is a statement of what a statistical hypothesis test is set up to establish. For example, in a clinical trial of a new drug, the alternative hypothesis might be that the new drug has a different effect, on average, compared to that of the current drug. We would write H_a: the two drugs have different effects, on average. The alternative hypothesis might also be that the new drug is better, on average, than the current drug. In this case we would write H_a: the new drug is better than the current drug, on average.

The final conclusion once the test has been carried out is always given in terms of the null hypothesis. We either "reject H_0 in favor of H_a" or "do not reject H_0"; we never conclude "reject H_a", or even "accept H_a".

If we conclude "do not reject H_0", this does not necessarily mean that the null hypothesis is true, it only suggests that there is not sufficient evidence against H_0 in favor of H_a; rejecting the null hypothesis then, suggests that the alternative hypothesis may be true.

Hypotheses are always stated in terms of population parameter, such as the mean μ. An alternative hypothesis may be ***one-sided*** or ***two-sided***. A one-sided hypothesis claims that a parameter is either larger *or* smaller than the value given by the null hypothesis. A two-sided hypothesis claims that a parameter is simply *not equal* to the value given by the null hypothesis -- the direction does not matter.

Hypotheses for a one-sided test for a population mean take the following form:

H_0: $\mu = k$

H_a: $\mu > k$

or

H_0: $\mu = k$

H_a: $\mu < k$.

Hypotheses for a two-sided test for a population mean take the following form:

H_0: $\mu = k$

H_a: $\mu \neq k$.

A ***confidence interval*** gives an estimated range of values which is likely to include an unknown population parameter, the estimated range being calculated from a given set of sample data.)

Example

Suppose a test has been given to all high school students in a certain state. The mean test score for the entire state is 70, with standard deviation equal to 10. Members of the school board suspect that female students have a higher mean score on the test than male students, because the mean score $\overline{x}$ from a random sample of 64 female students is equal to 73. Does this provide strong evidence that the overall mean for female students is higher?

The null hypothesis H_0 claims that there is no difference between the mean score for female students and the mean for the entire population, so that μ = 70. The alternative hypothesis claims that the mean for female students is higher than the entire student population mean, so that μ > 70.

Significance Tests for Unknown Mean and Known Standard Deviation

Once null and alternative hypotheses have been formulated for a particular claim, the next step is to compute a ***test statistic***. **For claims about a population mean from a population with a normal distribution or for any sample with large sample size *n* (for which the sample mean will follow a normal distribution by the Central Limit Theorem), if the standard deviation σ is known, the appropriate significance test is known as the *z-test*, where the test statistic is defined as z =** $\dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$ **.**

The test statistic follows the standard normal distribution (with mean = 0 and standard deviation = 1). The test statistic *z* is used to compute the ***P-value*** for the standard normal distribution, the probability that a value at least as extreme as the test statistic would be observed under the null hypothesis. Given the null hypothesis that the population mean μ is equal to a given value μ_0, the *P-values* for testing H_0 against each of the possible alternative hypotheses are:

$P(Z \geq z)$ for H_a: $\mu > \mu_0$

$P(Z \leq z)$ for H_a: $\mu < \mu_0$
$2P(Z \geq |z|)$ for H_a: $\mu \neq \mu_0$.

The probability is doubled for the two-sided test, since the two-sided alternative hypothesis considers the possibility of observing extreme values on *either* tail of the normal distribution.

Example

In the test score example above, where the sample mean equals 73 and the population standard deviation is equal to 10, the test statistic is computed as follows:
z = (73 - 70)/(10/sqrt(64)) = 3/1.25 = 2.4. Since this is a one-sided test, the *P-value* is equal to the probability that of observing a value greater than 2.4 in the standard normal distribution, or $P(Z > 2.4) = 1 - P(Z \leq 2.4) = 1 - 0.9918 = 0.0082$. The *P-value* is less than 0.01, indicating that it is highly unlikely that these results would be observed under the null hypothesis. The school board can confidently reject H_0 given this result, although they cannot conclude any additional information about the mean of the distribution.

Significance Levels

The ***significance level*** α for a given hypothesis test is a value for which a *P-value* less than or equal to α is considered statistically significant. Typical values for α are 0.1, 0.05, and 0.01. These values correspond to the probability of observing such an extreme value by chance. In the test score example above, the *P-value* is 0.0082, so the probability of observing such a value by chance is less that 0.01, and the result is significant at the 0.01 level.

In a one-sided test, α corresponds to the critical value z^* such that $P(Z \geq z^*) = \alpha$. For example, if the desired significance level for a result is 0.05, the corresponding value for *z* must be greater than or equal to $z^* = 1.645$ (or less than or equal to -1.645 for a one-sided alternative claiming that the mean is less than the null hypothesis). For a two-sided test, we are interested in the probability that $2P(Z \geq z^*) = \alpha$, so the critical value z^* corresponds to the $\alpha/2$ significance level. To achieve a significance level of 0.05 for a two-sided test, the absolute value of the test statistic (|*z*|) must be greater than or equal to the critical value 1.96 (which corresponds to the level 0.025 for a one-sided test).

Another interpretation of the significance level α, based in ***decision theory***, is that α corresponds to the value for which one chooses to reject or accept the null hypothesis H_0. In the above example, the value 0.0082 would result in rejection of the null hypothesis at the 0.01 level. The probability that this is a mistake -- that, in fact, the null hypothesis is true given the z-statistic -- is less than 0.01. In decision theory, this is known as a ***Type I error***. The probability of a Type I error is equal to the significance level α, and the probability of rejecting the null hypothesis when it is in fact false (a correct decision) is equal to 1 - α. To minimize the probability of Type I error, the significance level is generally chosen to be small.

Example

Of all of the individuals who develop a certain rash, suppose the mean recovery time for individuals who do not use any form of treatment is 30 days with standard deviation equal to 8. A pharmaceutical company manufacturing a certain cream wishes to determine whether the cream shortens, extends, or has no effect on the recovery time. The company chooses a random sample of 100 individuals who have used the cream, and determines that the mean recovery time for these individuals was 28.5 days. Does the cream have any effect?

Since the pharmaceutical company is interested in *any* difference from the mean recovery time for all individuals, the alternative hypothesis H_a is two-sided: $\mu \neq$ 30. The test statistic is calculated to be z = (28.5 - 30)/(8/sqrt(100)) = -1.5/0.8 = -1.875. The *P-value* for this statistic is $2P(Z \geq 1.875)$ = 2(1 - P((Z < 1.875) = 2(1-0.9693) = 2(0.0307) = 0.0614. This is not significant at the 0.05 level, although it is significant at the 0.1 level.

Example

In the test score example, for a fixed significance level of 0.10, suppose the school board wishes to be able to reject the null hypothesis (that the mean = 70) if the mean for female students is in fact 72. To determine the power of the test against this alternative, first note that the critical value for rejecting the null hypothesis is z^* = 1.282. The calculated value for z will be greater than 1.282 whenever ($\overline{x}$ - 70)/(1.25) > 1.282, or $\overline{x}$ > 71.6. The probability of rejecting the null hypothesis (mean = 70) given that the alternative hypotheses (mean = 72) is true is calculated by:

$P((\overline{x} \geq 71.6 \mid \mu = 72)$
$= P((\overline{x} - 72)/(1.25) \geq (71.6 - 72)/1.25)$
$= P(Z \geq -0.32) = 1 - P(Z \leq -0.32) = 1 - 0.3745 = 0.6255$. The power is about 0.60, indicating that although the test is more likely than not to reject the null hypothesis for this value, the probability of a Type II error is high.

Significance Tests for Unknown Mean and Unknown Standard Deviation

In most practical research, the standard deviation for the population of interest is not known. In this case, the standard deviation σ is replaced by the estimated standard deviation *s*, also known as the ***standard error***. Since the standard error is an estimate for the true value of the standard deviation, the distribution of the sample mean $\overline{x}$ is no longer normal with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. Instead, the sample mean follows the ***t distribution*** with mean μ and standard deviation $\frac{s}{\sqrt{n}}$. The *t* distribution is also described by its ***degrees of freedom*. For a sample of size *n*, the *t* distribution will have *n-1* degrees of freedom.** The notation for a *t* distribution with *k* degrees of freedom is *t(k)*. As the sample size *n* increases, the *t* distribution becomes closer to the normal distribution, since the standard error approaches the true standard deviation σ for large *n*.

For claims about a population mean from a population with a normal distribution or for any sample with large sample size *n* (for which the sample mean will follow a normal distribution by the Central Limit Theorem) with unknown standard deviation, the appropriate significance test is known as the *t-test*, where the test statistic is defined as t = $\frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}}$ **.**

The test statistic follows the *t* distribution with *n-1* degrees of freedom. The test statistic *z* is used to compute the ***P-value*** for the *t* distribution, the probability that a value at least as extreme as the test statistic would be observed under the null hypothesis.

Coefficients for Measuring Association

The following are a few of the many measures of association used with chi-square and other contingency table analyses. When using the chi-square statistic, these coefficients can be helpful in interpreting the relationship between two variables once statistical significance has been established. The logic for using measures of association is as follows:

> Even though a chi-square test may show statistical significance between two variables, the relationship between those variables may not be substantively important. These and many other measures of association are available to help evaluate the relative strength of a statistically significant relationship. In most cases, they are not used in interpreting the data unless the chi-square statistic first shows there is statistical significance (i.e., it doesn't make sense to say there is a strong relationship between two variables when your statistical test shows this relationship is not statistically significant).

Nominal and Ordinal Variables

Phi

> Only used on 2x2 contingency tables. Interpreted as a measure of the relative (strength) of an association between two variables ranging from 0 to 1.

$$Phi = \sqrt{\frac{X^2}{n}}$$

Pearson's Contingency Coefficient (C)

It is interpreted as a measure of the relative (strength) of an association between two variables. The coefficient will always be less than 1 and varies according to the number of rows and columns.

$$C = \sqrt{\frac{X^2}{n + X^2}}$$

Cramer's V Coefficient (V)

Useful for comparing multiple X^2 test statistics and is generalizable across contingency tables of varying sizes. It is not affected by sample size and therefore is very useful in situations where you suspect a statistically significant chi-square was the result of large sample size instead of any substantive relationship between the variables. It is interpreted as a measure of the relative (strength) of an association between two variables. The coefficient ranges from 0 to 1 (perfect association). In practice, you may find that a Cramer's V of .10 provides a good minimum threshold for suggesting there is a substantive relationship between two variables.

$$V = \sqrt{\frac{X^2}{n(q-1)}}$$ where q = smaller # of rows or columns

Describing Strength of Association

Characterizations

>.5	high association
.3 to .5	moderate association
.1 to .3	low association
0 to .1	little if any association

Proportional Reduction of Error (PRE)

Lambda

This is a proportional reduction in error (PRE) measure that ranges from 0 to 1. Lambda indicates the extent to which the independent variable reduces the error associated with predicting the value of a dependent variable. Multiplied by 100, it represents the percent reduction in error.

Ordinal Variables Only

Gamma

Another PRE measure ranging from -1 to 1 that estimates the extent errors are reduced in predicting the order of paired cases. Gamma ignores ties.

Kendall's Tau b

Similar to Gamma but includes ties. Ranges from -1 to 1 but since standardization is different from Gamma, it provides no clear explanation of PRE.

Inter-rater Agreement

Cohen's Kappa

Measures agreement beyond chance. Although a negative value is possible, it commonly ranges from 0 to 1 (perfect agreement). This measure requires a balanced table where the number of rows is the same as the number of columns. The diagonal cells represent agreement.

The Yule's Q contingency coefficient

The Yule's Q contingency coefficient (Yule (1900)), is a measure of correlation, which can be calculated for 2 × 2 contingency tables.

$$Q = \frac{O_{11}O_{22} - O_{12}O_{21}}{O_{11}O_{22} + O_{12}O_{21}},$$

where O_{11}, O_{12}, O_{21}, O_{22} - observed frequencies in a contingency table.

The Q coefficient value is included in a range of <-1, 1>. The closer to 0 the value of the Q is, the weaker dependence joins the analysed features, and the closer to –1 or +1, the stronger dependence joins the analysed features. There is one disadvantage of this coefficient. It is not much resistant to small observed frequencies (if one of them is 0, the coefficient might wrongly indicate the total dependence of features).

The statistic significance of the Yule's Q coefficient is defined by the Z test. Hypotheses:

- $\mathcal{H}_0$: $Q = 0$,
- $\mathcal{H}_1$: $Q \neq 0$.

The φ contingency coefficient

The φ contingency coefficient is a measure of correlation, which can be calculated for 2 × 2 contingency tables.

$$\phi = \sqrt{\frac{\chi^2}{n}},$$

The - coefficient value is included in a range of < 0; 1 >. The closer to 0 the value of φ is, the weaker dependence joins the analysed features, and the closer to 1, the stronger dependence joins the analysed features.

The φ contingency coefficient is considered as statistically significant, if the *p* value calculated on the basis of the χ^2 test (designated for this table) is equal to or less than the significance level α.

The Cramer's V contingency coefficient

The Cramer's V contingency coefficient (Cramer (1946)), is an extension of the φ coefficient on r × c contingency tables.

$$V = \sqrt{\frac{\chi^2}{n(w-1)}},$$

where χ^2 - value of the χ^2 test statistic,
n - total frequency in a contingency table,

w – jthe smaller the value out of r and c.

The V coefficient value is included in a range of < 0; 1 >.The closer to 0 the value of V is, the weaker dependence joins the analysed features, and the closer to 1, the stronger dependence joins the analysed features. The V coefficient value depends also on the table size, so you should not use this coefficient to compare different sizes of contingency tables.

The *V* contingency coefficient is considered as statistically significant, if the *p* value contingency coefficient is considered as statistically significant, if the χ^2 test (designated for this table) is equal to or less than the significance leveli α.

The Pearson's C contingency coefficient

The Pearson's C contingency coefficient is a measure of correlation, which can be calculated for r × c contingency tables

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}},$$

gdzie χ^2 - value of the χ^2 test statistic,
n - total frequency in a contingency table.

The C coefficient value is included in a range of < 0; 1). The closer to 0 the value of C is, the weaker dependence joins the analysed features, and the farther from 0, the stronger dependence joins the analysed features. The C coefficient value depends also on the table size (the bigger table, the closer to 1 C value can be), that is why it should be calculated the top limit, which the C coefficient may gain for the particular table size.

The *C* contingency coefficient is considered as statistically significant, if the *p* value calculated on the basis of the χ^2 test (designated for this table) is equal to or less than significance level α.

Example (EN_sex-exam.pqs file)

There is a sample of 170 persons (n = 170), who have 2 features analysed (X=sex, Y =passing the exam). Each of these features occurs in 2 categories (X1=f, X2=m, Y1=yes, Y2=no). Basing on the sample, we would like to get to know, if there is any dependence between sex and passing the exam in an analysed population. The data distribution is presented in a contingency table:

The chi-square test statistic value is 16.33 and the p value calculated for it: p = 0.00005. The result indicates that there is a statistically significant dependence between sex and passing the exam in the analysed population. Coefficient values, which are based on the chi-square test, so the strength of the correlation between analysed features are:
C_{adj}-Pearson = 0.42,
V-Cramer = φ = 0.31,
Q-Yule = 0.58, and the p value of the *Z* test (similarly to the chi-square test) indicates the statistically significant dependence between the analysed features.

Basics of Sampling Theory

P = { x_1, x_2, ……, x_N}

where P = population

x_1, x_2, ……, x_N are real numbers

Assuming x is a random variable;

Mean/Average of x,

$$\overline{X} = \sum_{i=0}^{n} xi$$

Standard Deviation,

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=0}^{n} fi(xi - \ddot{x})^2}$$

Variance,

$$\sigma^2 = \frac{1}{N}\sum_{i=0}^{n} fi(xi - \ddot{x})^2$$

Basics of Sampling Theory---
Theorem About Mean

picking random numbers x, mean = x

picking random numbers y, mean = y

$x = y$

Picking another number z,

mean $z = x = y$

$z = c_1x + c_2y$; c_1, c_2 are constants

$\overline{z} = \overline{x} + \overline{y}$

Independence

two events are independent if the occurrence of one of the events gives no information about whether or not the other event will occur; that is, the events have no influence on each other

for example a, b and c are independent if:

- a and b are independent; a and c are independent; and b and c are independent

Theorem About Variances/Sampling Theorem

$z = (x + y)/2$; $\sigma_z^2 = ?$ $\sigma_z^2 < \sigma_x^2$

Taking, $z = (x + y)/2$

$\sigma_z^2 = (\sigma_x^2 + \sigma_y^2)/4$

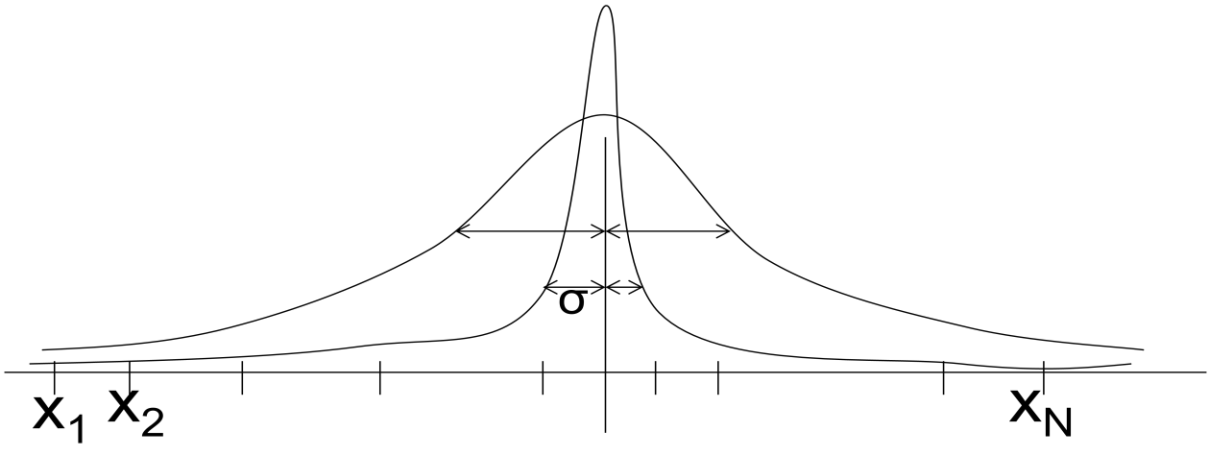
Taking k sample, $z = (x + x' + x'' + \ldots + x''^{\ldots k})/k$

$\sigma_z^2 = (k\sigma_x^2)/k^2$

$\sigma_z^2 = \sigma_x^2)/k$

Basics of Sampling Theory

Normal Distribution curve



- K = number of samples
- $\bar{z}$ = sample mean
- as k increases, $\bar{z}$ comes closer to $\bar{x}$

Assuming numbers are sorted

Variance:-

In probability theory and statistics, the **variance** is a measure of how far a set of numbers is spread out. It is one of several descriptors of a probability distribution, describing how far the numbers lie from the mean (expected value). In particular, the variance is one of the moments of a distribution. In that context, it forms part of a systematic approach to distinguishing between probability distributions. While other such approaches have been developed, those based on moments are advantageous in terms of mathematical and computational simplicity.

The variance is a parameter describing in part either the actual probability distribution of an observed population of numbers, or the theoretical probability distribution of a sample (a not-fully-observed population) of numbers. In the latter

case a sample of data from such a distribution can be used to construct an estimate of its variance: in the simplest cases this estimate can be the **sample variance**, defined below.

Examples

The **variance** of a random variable or distribution is the expectation, or mean, of the squared deviation of that variable from its expected value or mean. Thus the variance is a measure of the amount of variation of the values of that variable, taking account of all possible values and their probabilities or weightings (not just the extremes which give the range).

For example, a perfect six-sided die, when thrown, has expected value of

$$\frac{1}{6}(1+2+3+4+5+6) = 3.5.$$

Its expected absolute deviation—the mean of the equally likely absolute deviations from the mean—is

$$\frac{1}{6}(|1-3.5|+|2-3.5|+|3-3.5|+|4-3.5|+|5-3.5|+|6-3.5|) = \frac{1}{6}(2.5+1.5+0.5+0.5-$$

But its expected *squared* deviation—its variance (the mean of the equally likely squared deviations)—is

$$\frac{1}{6}(2.5^2 + 1.5^2 + 0.5^2 + 0.5^2 + 1.5^2 + 2.5^2) = 17.5/6 \approx 2.9.$$

As another example, if a coin is tossed twice, the number of heads is: 0 with probability 0.25, 1 with probability 0.5 and 2 with probability 0.25. Thus the expected value of the number of heads is:

$$0.25 \times 0 + 0.5 \times 1 + 0.25 \times 2 = 1,$$

and the variance is:

$$0.25 \times (0-1)^2 + 0.5 \times (1-1)^2 + 0.25 \times (2-1)^2 = 0.25 + 0 + 0.25 = 0.5.$$

[edit] Units of measurement

Unlike expected absolute deviation, the variance of a variable has units that are the square of the units of the variable itself. For example, a variable measured in inches will have a variance measured in square inches. For this reason, describing data sets via their standard deviation or root mean square deviation is often preferred over using the variance. In the dice example the standard deviation is $\sqrt{2.9} \approx 1.7$, slightly larger than the expected absolute deviation of 1.5.

The standard deviation and the expected absolute deviation can both be used as an indicator of the "spread" of a distribution. The standard deviation is more amenable to algebraic manipulation than the expected absolute deviation, and, together with variance and its generalization covariance, is used frequently in theoretical statistics; however the expected absolute deviation tends to be more robust as it is less sensitive to outliers arising from measurement anomalies or an unduly heavy-tailed distribution.

[edit] Estimating the variance

Real-world distributions such as the distribution of yesterday's rain throughout the day are typically not fully known, unlike the behavior of perfect dice or an ideal distribution such as the normal distribution, because it is impractical to account for every raindrop. Instead one estimates the mean and variance of the whole distribution as the computed mean and variance of a sample of n observations drawn suitably randomly from the whole sample space, in this example the set of all measurements of yesterday's rainfall in all available rain gauges.

This method of estimation is close to optimal, with the caveat that it underestimates the variance by a factor of $(n - 1) / n$. (For example, when $n = 1$ the variance of a single observation is obviously zero regardless of the true variance). This gives a bias which should be corrected for when n is small by multiplying by $n / (n - 1)$. If the mean is determined in some other way than from the same samples used to estimate the variance then this bias does not arise and the variance can safely be estimated as that of the samples.

To illustrate the relation between the population variance and the sample variance, suppose that in the (not entirely observed) population of numerical values, the value 1 occurs 1/3 of the time, the value 2 occurs 1/3 of the time, and the value 4 occurs 1/3 of the time. The population mean is $(1/3)[1 + 2 + 4] = 7/3$. The equally likely deviations from the population mean are $1 - 7/3$, $2 - 7/3$, and $4 - 7/3$. The

population variance — the expected squared deviation from the mean 7/3 — is $(1/3)[(-4/3)^2 + (-1/3)^2 + (5/3)^2] = 14/9$. Now suppose for the sake of a simple example that we take a very small sample of $n = 2$ observations, and consider the nine equally likely possibilities for the set of numbers within that sample: (1, 1), (1, 2), (1,4), (2, 1), (2,2), (2, 4), (4,1), (4, 2), and (4, 4). For these nine possible samples, the sample variance of the two numbers is respectively 0, 1/4, 9/4, 1/4, 0, 4/4, 9/4, 4/4, and 0. With our plan to observe two values, we could end up computing any of these sample variances (and indeed if we hypothetically could observe a pair of numbers many times, we would compute each of these sample variances 1/9 of the time). So the expected value, over all possible samples that might be drawn from the population, of the computed sample variance is $(1/9)[0 + 1/4 + 9/4 + 1/4 + 0 + 4/4 + 9/4 + 4/4 + 0] = 7/9$. This value of 7/9 for the expected value of our sample variance computation is a substantial underestimate of the true population variance, which we computed as 14/9, because our sample size of just two observations was so small. But if we adjust for this downward bias by multiplying our computed sample variance, whichever it may be, by $n/(n - 1) = 2/(2 - 1) = 2$, then our estimate of the population variance would be any one of 0, 1/2, 9/2, 1/2, 0, 4/2, 9/2, 4/2, and 0. The average of these is indeed the correct population variance of 14/9, so on average over all possible samples we would have the correct estimate of the population variance.

The variance of a real-valued random variable is its second central moment, and it also happens to be its second cumulant. Just as some distributions do not have a mean, some do not have a variance. The mean exists whenever the variance exists, but the converse is not necessarily true.

[edit] Definition

If a random variable X has the expected value (mean) $\mu = \mathrm{E}[X]$, then the variance of X is given by:

$$\operatorname{Var}(X) = \mathrm{E}\left[(X - \mu)^2\right].$$

That is, the variance is the expected value of the squared difference between the variable's realization and the variable's mean. This definition encompasses random variables that are discrete, continuous, or neither (or mixed). It can be expanded as follows:

$$\begin{aligned}\operatorname{Var}(X) &= \mathrm{E}\left[(X-\mu)^2\right] \\ &= \mathrm{E}\left[X^2 - 2\mu X + \mu^2\right] \\ &= \mathrm{E}\left[X^2\right] - 2\mu\ \mathrm{E}[X] + \mu^2 \\ &= \mathrm{E}\left[X^2\right] - 2\mu^2 + \mu^2 \\ &= \mathrm{E}\left[X^2\right] - \mu^2 \\ &= \mathrm{E}\left[X^2\right] - (\mathrm{E}[X])^2.\end{aligned}$$

A mnemonic for the above expression is "mean of square minus square of mean". The variance of random variable *X* is typically designated as Var(*X*), σ_X^2, or simply σ^2 (pronounced "sigma squared").

Continuous random variable

If the random variable *X* is continuous with probability density function *f*(*x*), then the variance equals the second central moment, given by

$$\operatorname{Var}(X) = \int (x-\mu)^2 f(x)\,dx\,,$$

where μ is the expected value,

$$\mu = \int x\, f(x)\,dx\,,$$

and where the integrals are definite integrals taken for *x* ranging over the range of *X*.

If a continuous distribution does not have an expected value, as is the case for the Cauchy distribution, it does not have a variance either. Many other distributions for which the expected value does exist also do not have a finite variance because the integral in the variance definition diverges. An example is a Pareto distribution whose index *k* satisfies $1 < k \leq 2$.

[edit] Discrete random variable

If the random variable *X* is discrete with probability mass function $x_1 \mapsto p_1$, ..., $x_n \mapsto p_n$, then

$$\operatorname{Var}(X) = \sum_{i=1}^{n} p_i \cdot (x_i - \mu)^2$$

where $\mu$is the expected value, i.e.

$$\mu = \sum_{i=1}^{n} p_i \cdot x_i \,.$$

(When such a discrete weighted variance is specified by weights whose sum is not 1, then one divides by the sum of the weights.) That is, it is the expected value of the square of the deviation of *X* from its own mean. In plain language, it can be expressed as "The mean of the squares of the deviations of the data points from the average". It is thus the *mean squared deviation*.

Examples

Exponential distribution

The exponential distribution with parameter λ is a continuous distribution whose support is the semi-infinite interval [0,∞). Its probability density function is given by:

$$f(x) = \lambda e^{-\lambda x},$$

and it has expected value $\mu = \lambda^{-1}$. Therefore the variance is equal to:

$$\int_0^\infty f(x)(x-\mu)^2\,dx = \int_0^\infty \lambda e^{-\lambda x}(x-\lambda^{-1})^2\,dx = \lambda^{-2}.$$

So for an exponentially distributed random variable $\sigma^2 = \mu^2$.

[edit] Fair dice

A six-sided fair die can be modelled with a discrete random variable with outcomes 1 through 6, each with equal probability $\frac{1}{6}$. The expected value is (1 + 2 + 3 + 4 + 5 + 6)/6 = 3.5. Therefore the variance can be computed to be:

$$\sum_{i=1}^{6} \tfrac{1}{6}(i-3.5)^2 = \tfrac{1}{6}\sum_{i=1}^{6}(i-3.5)^2 = \tfrac{1}{6}\left((-2.5)^2+(-1.5)^2+(-0.5)^2+0.5^2+1.5^2+2.5\right.$$
$$= \tfrac{1}{6}\cdot 17.50 = \tfrac{35}{12} \approx 2.92.$$

The general formula for the variance of the outcome *X* of a die of n sides is:

$$\sigma^2 = E(X^2) - (E(X))^2 = \frac{1}{n}\sum_{i=1}^{n} i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} i\right)^2$$
$$= \tfrac{1}{6}(n+1)(2n+1) - \tfrac{1}{4}(n+1)^2$$
$$= \frac{n^2-1}{12}.$$

Properties

Variance is non-negative because the squares are positive or zero.

$$\operatorname{Var}(X) \geq 0.$$

The variance of a constant random variable is zero, and if the variance of a variable in a data set is 0, then all the entries have the same value.

$$P(X = a) = 1 \Leftrightarrow \operatorname{Var}(X) = 0.$$

Variance is invariant with respect to changes in a location parameter. That is, if a constant is added to all values of the variable, the variance is unchanged.

$$\operatorname{Var}(X + a) = \operatorname{Var}(X).$$

If all values are scaled by a constant, the variance is scaled by the square of that constant.

$$\operatorname{Var}(aX) = a^2 \operatorname{Var}(X).$$

The variance of a sum of two random variables is given by:

$$\operatorname{Var}(aX + bY) = a^2 \operatorname{Var}(X) + b^2 \operatorname{Var}(Y) + 2ab \operatorname{Cov}(X, Y),$$
$$\operatorname{Var}(X - Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) - 2 \operatorname{Cov}(X, Y),$$

In general we have for the sum of N random variables:

$$\operatorname{Var}\left(\sum_{i=1}^{N} X_i\right) = \sum_{i,j=1}^{N} \operatorname{Cov}(X_i, X_j) = \sum_{i=1}^{N} \operatorname{Var}(X_i) + \sum_{i \neq j} \operatorname{Cov}(X_i, X_j).$$

The variance of a finite sum of *uncorrelated* random variables is equal to the sum of their variances. This stems from the above identity and the fact that for uncorrelated variables the covariance is zero.

$$\operatorname{Cov}(X_i, X_j) = 0 \ (i \neq j) \Rightarrow \operatorname{Var}\left(\sum_{i=1}^{N} X_i\right) = \sum_{i=1}^{N} \operatorname{Var}(X_i).$$

These results lead to the variance of a linear combination as:

$$\begin{aligned}
\operatorname{Var}\left(\sum_{i=1}^{N} a_i X_i\right) &= \sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j \operatorname{Cov}(X_i, X_j) \\
&= \sum_{i=1}^{N} a_i^2 \operatorname{Var}(X_i) + \sum_{i \neq j} a_i a_j \operatorname{Cov}(X_i, X_j) \\
&= \sum_{i=1}^{N} a_i^2 \operatorname{Var}(X_i) + 2 \sum_{i<j} a_i a_j \operatorname{Cov}(X_i, X_j).
\end{aligned}$$

Suppose that the observations can be partitioned into equal-sized **subgroups** according to some second variable. Then the variance of the total group is equal to the mean of the variances of the subgroups plus the variance of the means of the subgroups. This property is known as variance decomposition or the law of total variance and plays an important role in the analysis of variance. For example, suppose that a group consists of a subgroup of men and an equally large subgroup of women. Suppose that the men have a mean height of 180 and that the variance of their heights is 100. Suppose that the women have a mean height of 160 and that the variance of their heights is 50. Then the mean of the variances is (100 + 50) / 2 = 75; the variance of the means is the variance of 180, 160 which is 100. Then, for the total group of men and women combined, the variance of the height will be 75 + 100 = 175. Note that this uses N for the denominator instead of N − 1.

In a more general case, if the subgroups have unequal sizes, then they must be weighted proportionally to their size in the computations of the means and

variances. The formula is also valid with more than two groups, and even if the grouping variable is continuous.

This formula implies that the variance of the total group cannot be smaller than the mean of the variances of the subgroups. Note, however, that the total variance is not necessarily larger than the variances of the subgroups. In the above example, when the subgroups are analyzed separately, the variance is influenced only by the man-man differences and the woman-woman differences. If the two groups are combined, however, then the men-women differences enter into the variance also.

Many computational formulas for the variance are based on this equality: **The variance is equal to the mean of the square minus the square of the mean:**

$$\operatorname{Var}(X) = \mathrm{E}[X^2] - \mathrm{E}[X]^2.$$

For example, if we consider the numbers 1, 2, 3, 4 then the mean of the squares is (1 × 1 + 2 × 2 + 3 × 3 + 4 × 4) / 4 = 7.5. The regular mean of all four numbers is 2.5, so the square of the mean is 6.25. Therefore the variance is 7.5 − 6.25 = 1.25, which is indeed the same result obtained earlier with the definition formulas. Many pocket calculators use an algorithm that is based on this formula and that allows them to compute the variance while the data are entered, without storing all values in memory. The algorithm is to adjust only three variables when a new data value is entered: The number of data entered so far (*n*), the sum of the values so far (*S*), and the sum of the squared values so far (*SS*). For example, if the data are 1, 2, 3, 4, then after entering the first value, the algorithm would have *n* = 1, *S* = 1 and *SS* = 1. After entering the second value (2), it would have *n* = 2, *S* = 3 and *SS* = 5. When all data are entered, it would have *n* = 4, *S* = 10 and *SS* = 30. Next, the mean is computed as *M* = *S* / *n*, and finally the variance is computed as *SS* / *n* − *M* × *M*. In this example the outcome would be 30 / 4 − 2.5 × 2.5 = 7.5 − 6.25 = 1.25. If the unbiased sample estimate is to be computed, the outcome will be multiplied by *n* / (*n* − 1), which yields 1.667 in this example.

Properties, formal

Sum of uncorrelated variables (Bienaymé formula)

See also: Sum of normally distributed random variables

One reason for the use of the variance in preference to other measures of dispersion is that the variance of the sum (or the difference) of uncorrelated random variables is the sum of their variances:

$$\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \operatorname{Var}(X_i).$$

This statement is called the Bienaymé formula.[1] and was discovered in 1853. It is often made with the stronger condition that the variables are independent, but uncorrelatedness suffices. So if all the variables have the same variance σ^2, then, since division by *n* is a linear transformation, this formula immediately implies that the variance of their mean is

$$\operatorname{Var}\left(\overline{X}\right) = \operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \operatorname{Var}\left(X_i\right) = \frac{\sigma^2}{n}.$$

That is, the variance of the mean decreases when *n* increases. This formula for the variance of the mean is used in the definition of the standard error of the sample mean, which is used in the central limit theorem.

[edit] Product of independent variables

If two variables X and Y are independent, the variance of their product is given by[2][3]

$$\operatorname{Var}(XY) = [E(X)]^2 \operatorname{Var}(Y) + [E(Y)]^2 \operatorname{Var}(X) + \operatorname{Var}(X) \operatorname{Var}(Y).$$

[edit] Sum of correlated variables

In general, if the variables are correlated, then the variance of their sum is the sum of their covariances:

$$\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \sum_{j=1}^{n} \operatorname{Cov}(X_i, X_j).$$

(Note: This by definition includes the variance of each variable, since $\operatorname{Cov}(X_i,X_i) = \operatorname{Var}(X_i)$.)

Here Cov is the covariance, which is zero for independent random variables (if it exists). The formula states that the variance of a sum is equal to the sum of all elements in the covariance matrix of the components. This formula is used in the theory of Cronbach's alpha in classical test theory.

So if the variables have equal variance σ^2 and the average correlation of distinct variables is ρ, then the variance of their mean is

$$\operatorname{Var}(\overline{X}) = \frac{\sigma^2}{n} + \frac{n-1}{n}\rho\sigma^2.$$

This implies that the variance of the mean increases with the average of the correlations. Moreover, if the variables have unit variance, for example if they are standardized, then this simplifies to

$$\operatorname{Var}(\overline{X}) = \frac{1}{n} + \frac{n-1}{n}\rho.$$

This formula is used in the Spearman–Brown prediction formula of classical test theory. This converges to ρ if n goes to infinity, provided that the average correlation remains constant or converges too. So for the variance of the mean of standardized variables with equal correlations or converging average correlation we have

$$\lim_{n\to\infty} \operatorname{Var}(\overline{X}) = \rho.$$

Therefore, the variance of the mean of a large number of standardized variables is approximately equal to their average correlation. This makes clear that the sample mean of correlated variables does generally not converge to the population mean, even though the Law of large numbers states that the sample mean will converge for independent variables.

[edit] Weighted sum of variables

The scaling property and the Bienaymé formula, along with this property from the covariance page: $\operatorname{Cov}(aX, bY) = ab \operatorname{Cov}(X, Y)$ jointly imply that

$$\operatorname{Var}(aX + bY) = a^2 \operatorname{Var}(X) + b^2 \operatorname{Var}(Y) + 2ab \operatorname{Cov}(X, Y).$$

This implies that in a weighted sum of variables, the variable with the largest weight will have a disproportionally large weight in the variance of the total. For example, if X and Y are uncorrelated and the weight of X is two times the weight of Y, then the weight of the variance of X will be four times the weight of the variance of Y.

The expression above can be extended to a weighted sum of multiple variables:

$$\operatorname{Var}\left(\sum_i a_i X_i\right) = \sum_i a_i^2 \operatorname{Var}(X_i) + 2\sum_i \sum_{j>i} a_i a_j \operatorname{Cov}(X_i, X_j)$$

[edit] Decomposition

The general formula for variance decomposition or the law of total variance is: If X and Y are two random variables and the variance of X exists, then

$$\operatorname{Var}(X) = \operatorname{Var}(\operatorname{E}(X|Y)) + \operatorname{E}(\operatorname{Var}(X|Y)).$$

Here, $\operatorname{E}(X|Y)$ is the conditional expectation of X given Y, and $\operatorname{Var}(X|Y)$ is the conditional variance of X given Y. (A more intuitive explanation is that given a particular value of Y, then X follows a distribution with mean $\operatorname{E}(X|Y)$ and variance $\operatorname{Var}(X|Y)$. The above formula tells how to find $\operatorname{Var}(X)$ based on the distributions of these two quantities when Y is allowed to vary.) This formula is often applied in analysis of variance, where the corresponding formula is

$$MS_{\text{Total}} = MS_{\text{Between}} + MS_{\text{Within}};$$

here MS refers to the Mean of the Squares. It is also used in linear regression analysis, where the corresponding formula is

$$MS_{\text{Total}} = MS_{\text{Regression}} + MS_{\text{Residual}}.$$

This can also be derived from the additivity of variances, since the total (observed) score is the sum of the predicted score and the error score, where the latter two are uncorrelated.

Similar decompositions are possible for the sum of squared deviations (sum of squares, SS):

$$SS_{\text{Total}} = SS_{\text{Between}} + SS_{\text{Within}},$$

$$SS_{\text{Total}} = SS_{\text{Regression}} + SS_{\text{Residual}}.$$

Computational formula

Main article: computational formula for the variance

See also: algorithms for calculating variance

The **computational formula for the variance** follows in a straightforward manner from the linearity of expected values and the above definition:

$$\begin{aligned}\operatorname{Var}(X) &= \operatorname{E}(X^2 - 2\,X\,\operatorname{E}(X) + (\operatorname{E}(X))^2) \\ &= \operatorname{E}(X^2) - 2(\operatorname{E}(X))^2 + (\operatorname{E}(X))^2 \\ &= \operatorname{E}(X^2) - (\operatorname{E}(X))^2.\end{aligned}$$

This is often used to calculate the variance in practice, although it suffers from catastrophic cancellation if the two components of the equation are similar in magnitude.

Characteristic property

The second moment of a random variable attains the minimum value when taken around the first moment (i.e., mean) of the random variable, i.e. $\operatorname{argmin}_m \operatorname{E}((X-m)^2) = \operatorname{E}(X)$. Conversely, if a continuous function φ satisfies $\operatorname{argmin}_m \operatorname{E}(\varphi(X-m)) = \operatorname{E}(X)$ for all random variables *X*, then it is necessarily of the form $\varphi(x) = ax^2 + b$, where *a* > 0. This also holds in the multidimensional case.[4]

Calculation from the CDF

The population variance for a non-negative random variable can be expressed in terms of the cumulative distribution function *F* using

$$2\int_0^\infty uH(u)\,du - \left(\int_0^\infty H(u)\,du\right)^2.$$

where $H(u) = 1 - F(u)$ is the right tail function. This expression can be used to calculate the variance in situations where the CDF, but not the density, can be conveniently expressed.

[edit] Matrix notation for the variance of a linear combination

Let's define X as a column vector of n random variables $X_1, \ldots, X_n$, and c as a column vector of N scalars $c_1, \ldots, c_n$. Therefore $c^T X$ is a linear combination of

these random variables, where c^T denotes the transpose of vector c. Let also be Σ the variance-covariance matrix of the vector X. The variance of $c^T X$ is given by[5]:

$$\operatorname{Var}(c^T X) = c^T \Sigma c.$$

Approximating the variance of a function

The delta method uses second-order Taylor expansions to approximate the variance of a function of one or more random variables: see Taylor expansions for the moments of functions of random variables. For example, the approximate variance of a function of one variable is given by

$$\operatorname{Var}\left[f(X)\right] \approx \left(f'(\operatorname{E}[X])\right)^2 \operatorname{Var}[X]$$

provided that *f* is twice differentiable and that the mean and variance of *X* are finite.

Population variance and sample variance

In general, the ***population variance*** of a *finite* population of size *N* is given by

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

where

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

is the population mean.

In many practical situations, the true variance of a population is not known *a priori* and must be computed somehow. When dealing with extremely large populations, it is not possible to count every object in the population.

A common task is to estimate the variance of a population from a sample.[6] We take a sample with replacement of *n* values y_1, ..., y_n from the population, where $n < N$, and estimate the variance on the basis of this sample. There are several good estimators. Two of them are well known:[7]

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \left(\frac{1}{n}\sum_{i=1}^{n} y_i^2\right) - \overline{y}^2,$$ and

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

The first estimator, also known as the second central moment, is called the ***biased sample variance***. The second estimator is called the ***unbiased sample variance***. Either estimator may be simply referred to as the ***sample variance*** when the version can be determined by context. Here, $\overline{y}$ denotes the sample mean:

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

The two estimators only differ slightly as can be seen, and for larger values of the sample size n the difference is negligible. While the first one may be seen as the variance of the sample considered as a population, the second one is the unbiased estimator of the population variance, meaning that its expected value $\mathrm{E}[s^2]$ is equal to the true variance of the sampled random variable; the use of the term $n - 1$ is called Bessel's correction. In particular,

$$\mathrm{E}[s^2] = \sigma^2,$$

while, in contrast,

$$\mathrm{E}[s_n^2] = \frac{n-1}{n}\sigma^2.$$

The unbiased sample variance is a U-statistic for the function $f(x_1, x_2) = (x_1 - x_2)^2/2$, meaning that it is obtained by averaging a 2-sample statistic over 2-element subsets of the population.

Distribution of the sample variance

Being a function of random variables, the sample variance is itself a random variable, and it is natural to study its distribution. In the case that y_i are independent observations from a normal distribution, Cochran's theorem shows that s^2 follows a scaled chi-squared distribution:[citation needed]

$$(n-1)\frac{s^2}{\sigma^2} \sim \chi^2_{n-1}.$$

As a direct consequence, it follows that $\mathrm{E}(s^2) = \sigma^2$.

If the y_i are independent and identically distributed, but not necessarily normally distributed, then[citation needed]

$$\mathrm{E}[s^2] = \sigma^2, \quad \mathrm{Var}[s^2] = \sigma^4 \left(\frac{2}{n-1} + \frac{\kappa}{n} \right),$$

where κ is the excess kurtosis of the distribution. If the conditions of the law of large numbers hold, s^2 is a consistent estimator of σ^2.[citation needed]

[edit] Samuelson's inequality

Samuelson's inequality is a result that states, given that the sample mean and variance have been calculated from a particular sample, bounds on the values that individual values in the sample can take.[8] Values must lie within the limits $m \pm s\,(n-1)^{1/2}$.

[edit] Relations with the harmonic and arithmetic means

It has been shown[9] that for a sample of real numbers that

$$\mathrm{Var} \leq 2M(A-H)$$

where M is the maximum of the sample, A is the arithmetic mean, H is the harmonic mean of the sample and Var is the variance of the sample.

This bound has been improved on and it is known that variance is bounded by

$$\text{Var} \leq \frac{M(A-H)(M-A)}{M-H}$$

$$\text{Var} \geq \frac{m(A-H)(A-m)}{H-m}$$

where *m* is the minimum of the sample.[10]

Generalizations

If X is a vector-valued random variable, with values in $\mathbb{R}^n$, and thought of as a column vector, then the natural generalization of variance is $\mathrm{E}((X-\mu)(X-\mu)^{\mathrm{T}})$, where $\mu = \mathrm{E}(X)$ and X^{T} is the transpose of X, and so is a row vector. This variance is a positive semi-definite square matrix, commonly referred to as the covariance matrix.

If X is a complex-valued random variable, with values in $\mathbb{C}$, then its variance is $\mathrm{E}((X-\mu)(X-\mu)^{\dagger})$, where $X^{\dagger}$ is the conjugate transpose of X. This variance is also a positive semi-definite square matrix.

Introduction to Hypothesis Testing

Significance testing is used to help make a judgment about a claim by addressing the question, Can the observed difference be attributed to chance? We break up significance testing into three (or four) steps:

Step A: Null and alternative hypotheses

The first step of hypothesis testing is to convert the research question into null and alterative hypotheses. We start with the **null hypothesis (H_0)**. The null hypothesis is a claim of "no difference." The opposing hypothesis is the **alternative hypothesis (H_1)**. The alternative hypothesis is a claim of "a difference in the population," and is the hypothesis the researcher often hopes to bolster. It is important to keep in mind that the null and alternative hypotheses reference population values, and *not* observed statistics.

Step B: Test statistic

We calculate a **test statistic** from the data. There are different types of test statistics. This chapter introduces the one-sample *z*-statistics. The z statistic will compare the observed sample mean to an expected population mean μ_0. Large test

statistics indicate data are far from expected, providing evidence against the null hypothesis and in favor of the alternative hypothesis.

Step C: *p* Value and conclusion

The test statistic is converted to a conditional probability called a *P*-value. The *P*-value answers the question "If the null hypothesis were true, what is the probability of observing the current data or data that is more extreme?"
Small *p* values provide evidence against the null hypothesis because they say the observed data are unlikely when the null hypothesis is true. We apply the following **conventions**:

- When *p* value > .10 → the observed difference is "not significant"
- When *p* value ≤ .10 → the observed difference is "marginally significant"
- When *p* value ≤ .05 → the observed difference is "significant"
- When *p* value ≤ .01 → the observed difference is "highly significant"

Use of "significant" in this context means "the observed difference is not likely due to chance." It does *not* mean of "important" or "meaningful."

Step D: Decision (optional)

Alpha (α) is a probability threshold for a decision. If $P \leq \alpha$, we will reject the null hypothesis. Otherwise it will be retained for want of evidence.
Page

One-Sample *z* Test

The one-sample z test is used to compare a mean from a single sample to an expected "norm." The norm for the test comes from a hypothetical value or observations in prior studies, and does not come from the current data. In addition, this test is used only when the population standard deviation σ is known from a prior source. Finally, data represent a SRS, and measurements that comprise the data are assumed to be accurate and meaningful.
Example ("Lake Wobegon"). Garrison Keller claims the children of Lake Wobegon are above average. You take a simple random sample of 9 children from Lake Wobegon and measure their intelligence with a Wechsler test and find the following scores: {116, 128, 125, 119, 89, 99, 105, 116, and 118}. The mean of this sample (x) is 112.8. We know Wechsler scores are scaled to be Normally distributed with a mean of 100 and standard deviation of 15. Is this sample mean sufficiently different from a population mean μ of 100 to reject the null hypothesis of "no difference?"

The null and alternative hypotheses

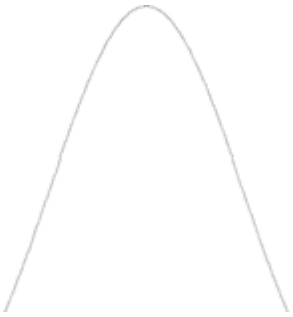
The claim being made in the illustrative example is that the population has higher than average intelligence. The null hypothesis is the population has average intelligence. Since an average intelligence score is 100, H_0: $\mu = 100$.

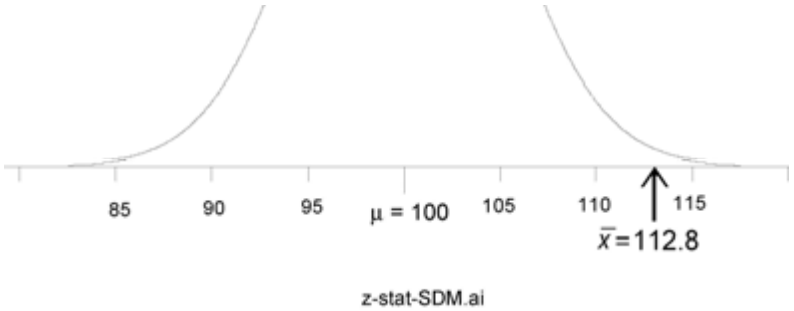
The alternative hypothesis claims the population has a higher than average intelligence. Therefore, H_1: $\mu > 100$. (The alternative hypothesis resembles the claim the investigator wishes to bolster.) It would be **incorrect** to state H_0: $x = 100$.

Inferential statements address the population, not the sample. The alternative hypothesis is **one-sided.** It is interested only in whether the population has a *higher* average core. It is not interested a *lower* than average score.

Test statistic

We use our knowledge of sampling distributions of the means (SDM) to help make judgments. Assuming the null hypothesis is true, the sampling distribution of $x$based on $n = 9$ would be Normal with a mean of 100 and standard error of $\sigma / \sqrt{n} = 15/\sqrt{9} = 5$. Therefore, under H_0, $x \sim N(100, 5)$.





z-stat-SDM.ai

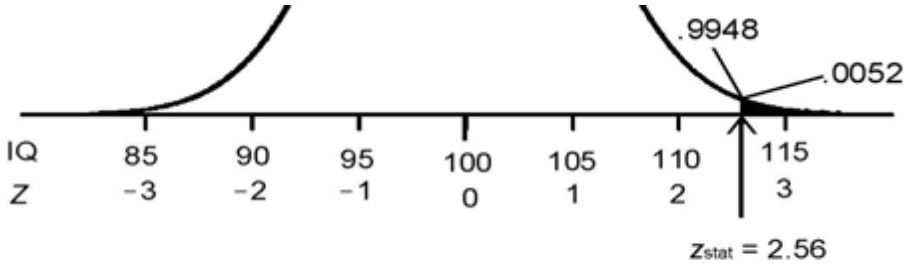
The observed *x* of 112.8 is out in the right-tail of the SDM.
The **test statistic** compares *x* to the hypothesized value (μ_0) as follows:

$$SEMxzstat0\mu-=$$

where *SEM* represents the standard error of the mean and is equal to $\sigma / \sqrt{n}$. The population standard deviation σ must be known in order to use this test statistic. (The z_{stat} quantifies how far *x* is from μ_0 in standard deviation units.)
For the illustrative example, $51008.112-=statz= 2.56$.

P-value and conclusion

To convert a z_{stat} to a *P*-value, find the area under the curve beyond the z_{stat} on a Standard Normal distribution. Use the Z table or a statistical package (e.g., *StaTable*) for this purpose. For the current problem we have:



z-stat-one-sided.ai

Therefore, $P = 0.0052$. This provides good evidence against H_0. The jargon is to say the difference is "significant." You can reject H_0.

Two-Sided Alternative

The two-sided *z* test makes no presupposition about the direction of the difference. The null hypothesis is the same as in the one-sided test: H_0: $\mu = \mu_0$. The alternative hypothesis is H_1: $\mu \neq \mu_0$. The test statistic is the same as the one-sample *z* test. Because we must consider sample means that might be above or below μ_0, the *p* value is "two-tailed."

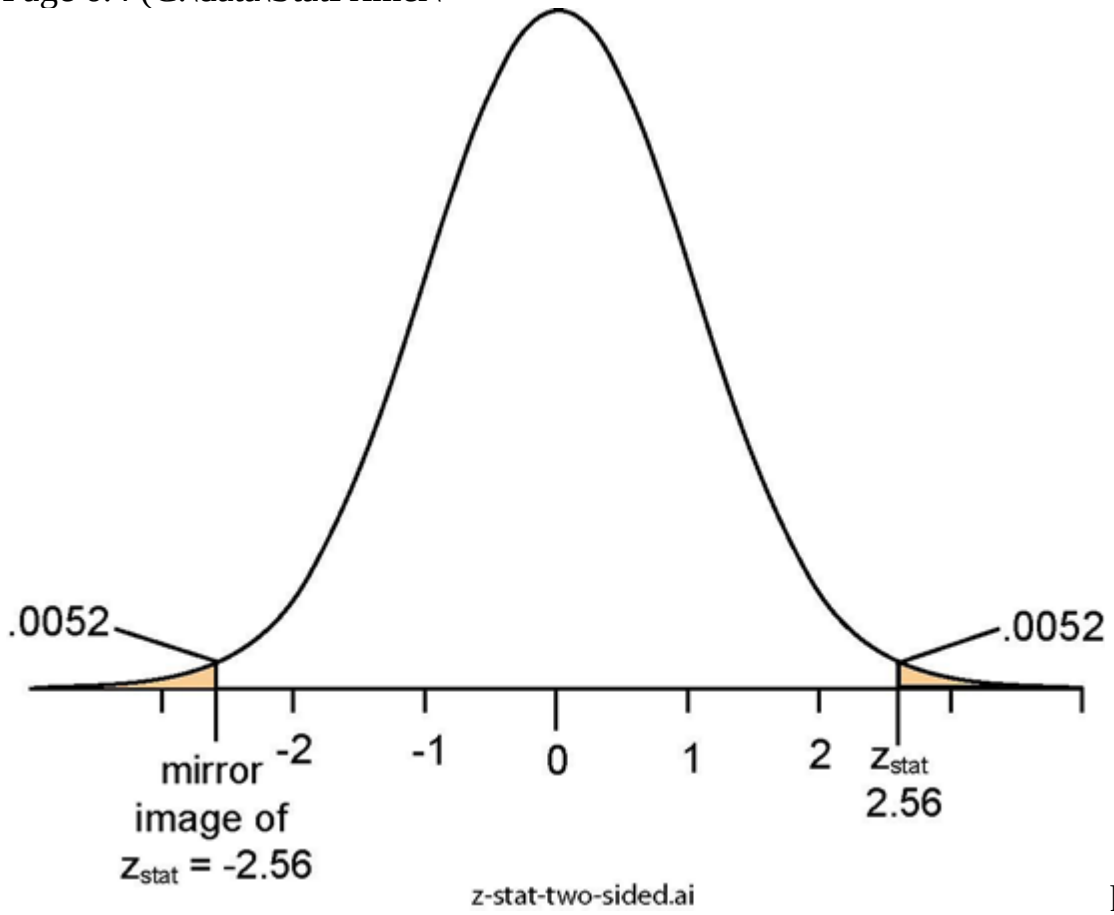
Illustrative example (Two-sided alternative). We return to the Lake Wobegon illustrative example, where children are assumed to be "above average." The two-sided alternative allows for unanticipated findings that are either "up" or "down" from expected.

Let μ_0 represent the expected value of the population mean *under* the null hypothesis. In the Lake Wobegon example, $\mu_0 = 100$. Therefore, we test H_0: $\mu = 100$ versus H_1: $\mu \neq 100$.

The test statistic is56.251008.1120=−=−=*SEMxzstat*μ (same as before).

With two-sided alternatives, we reject the null hypothesis in favor of the alternative if the sample mean is either significantly greater or less than μ_0. Rejection regions for the test statistic lie in both tails of SDM. The effect is to double the size of the one-sided *p* value. The one-sided *p* value for the Lake Wobegon illustration was .0052. Therefore, the two-sided *p* value = 2 × .0052 = .0104.

The *P*-value of 0.0104 provides good evidence against the null hypothesis.


.0052
.0052
mirror image of z_{stat} = -2.56
-2
-1
0
1
2
z_{stat} 2.56
z-stat-two-sided.ai

Fallacies of Statistical Hypothesis Testing

The results of statistical tests are frequently misunderstood. Therefore, I'm going to list some of the fallacies of hypothesis testing here. It will be helpful to refer back to this list as you grapple with the interpretation of hypothesis tests results.

1. Failure to reject the null hypothesis leads to its acceptance. (WRONG! Failure to reject the null hypothesis implies insufficient evidence for its rejection.)
2. The *p* value is the probability that the null hypothesis is incorrect. (WRONG! The *p* value is the probability of the current data or data that is more extreme assuming H_0 is true.)
3. $\alpha = .05$ is a standard with an objective basis. (WRONG! $\alpha = .05$ is merely a convention that has taken on unwise mechanical use. There is no sharp distinction between "significant" and "insignificant" results, only increasingly strong evidence as the *p* value gets smaller. Surely god loves $p = .06$ nearly as much as $p = .05$)
4. Small *p* values indicate large effects. (WRONG! *p* values tell you next to nothing about the size of an effect.)
5. Data show a theory to be true or false. (WRONG! Data can at best serve to bolster or refute a theory or claim.)
6. Statistical significance implies importance. (WRONG! WRONG! WRONG! Statistical significance says very little about the importance of a relation.)

Difference between Small and Large Samples:-

Though it is difficult t draw a clear-cut line of demarcation between large and small samples it is normally agreed amongst statisticians that a sample is to be recorded as large only if its size exceeds 30. The tests of significance used for dealing with problems samples for the reason that the assumptions that we make in case of large samples do not hold good for small samples.

The assumption made while dealing with problems relating to large samples are:-

(i) The random sampling distribution of a statistic is approximately normal. and

(ii) Values given by the samples are sufficiently close to the population value and can be used in its place for calculating the standard error of the estimate.

(Large Sample) Testing the significance of the difference between the means of two samples.)

To compare the means of two population we must understand the theory concerning the distribution of differences of sample means. Statisticians have determined that the distribution distribution difference between mean d (d Mean's) is approximately normal for large samples of n1 and n2. That is the distribution of differences of sample means is normal as long as neither n1 nor n2 Is less than 30. We can therefore use the probabilities associated with the normal distribution to construct confidence intervals and to perform hypothesis tests associated with this distribution.

PROCUEDURS:-

1. To compare the ($\mu 1$) mean of population 1 with the mean ($\mu 2$), of population 2 two independent random random samples of sizes n1 and n2 are to be selected from population 1 and population 2 respectively.

By independent we mean that the sample drawn from population 1, in no way affects the sample drawn from population 2 fro example drawing two samples from men population and women population

2. Compute (Mean1) and (Mean 2) i.e., mean of the sample 1 and 2

3. Computer the difference in the two samples means, d (mean) i.e,. d(Mean) = (Mean1 -Mean2).

Thus for each pair of sample means of (Mean1) and (Mean2). a value of d(Mean) is obtained. The result is therefore a distribution of d(Mean)s.

4. If $\mu 1$ and $\sigma 1$ are the parameters of population 1. and $\mu 2$ and $\sigma 2$ are the parameters of population 2, then for the distribution of d(Mean)s the menu $\mu d(Mean)s$ is given by the equation

$\mu d(Mean)s = \mu 1 - \mu 2$ the mean of the difference of the distribution of mean is the difference of the means of the two populations being compared.

5. The standard deviation (or standard error) of the distribution of d(mean)s (written as $\sigma d(Mean)s$) is given by the equation

(Large Sample) Testing the significance of the difference between the means of two samples.)

1. Point Estimation:- According to Central Limit Theorem for large samples the means of sampling distribution are normally distributed. The procedure that is frequently used to obtain a point estimate for the m of some population involves the following steps:

(a) Select a representation (random) sample of the population.

(b) Determine the mean (Mean) of the sample data
(c) Assert that the value of M is the corresponding value of (Mean) i.e., = μ.

2. Interval Estimation:-

An extension of the above method of obtaining an estimate for μ is with the confidence interval, i.e., an interval estimate for μ.

The advantages of interval estimate are:

1. Interval estimate is more likely to be correct than the point estimate.
2. We can calculate the probability that a given interval contains the mean of a population. We therefore speak of a specific interval as having "90' per cent probability of containing μ.
3. We can choose the value of the probability we want for a given interval before we actually construct it.

Recall that the central limit theorem asserts that for large sample sizes, the means are normally distributed. Furthermore, we know that any given mean (Mean) value can be standardized with the equation.

Where μ = Mean of the population
μ.(Mean) = Mean of the sampling distribution of means.
σ (Mean) = standard error or sampling distribution
Since μ.(Mean) μ we can write the following equation
Now, with a given pair of Z values associated with some percentage of the Z distribution and equation, we can determine an upper and lower boundary for the same percentage of (Mean) values in the given mean distribution.

Small And Large Sampling:-

Then instead of

$$x = \mu \pm z.\ se(popn\ mean)$$

we can write

$$\mu = x \pm z.\ se(sample\ mean)$$

where

$$se(sample\ mean) = s / \sqrt{n}$$

T- Test :

- Hypothesis testing involves making a decision concerning some hypothesis or statement about a population parameter such as the population mean, using the sample mean, to decide whether this statement about the value of is valid or not.
- **The steps of the hypothesis testing :**

1- The first step is to formulate a null hypothesis written . The statement for is usually expressed as an equation or inequality as follows:

$$H_{0:}\mu = \text{given value}$$

$$H_{0:}\mu \leq \text{given value}$$

$$H_{0:}\mu \geq \text{given value}$$

Also in this step it is stated an alternative hypothesis, written , a statement that indicates the opinion of the conductor of the test as to the actual value of . is expressed as follows:

$$H_{a:}\mu \neq \text{given value}$$

$$H_{a:}\mu < \text{given value}$$

$$H_{a:}\mu > \text{given value}$$

We conduct a hypothesis test on a given value to find out if actual observation would lead us to reject the stated value.

The alternative hypothesis suggests the direction of the actual value of the parameter relative to the stated value. The statement of in the form of an inequality that indicates that the investigator has no opinion as to whether the actual value of is more than or less than the stated value but the feeling is that the stated value is incorrect. In this case the test is two-tail test. Statements in the form

of strictly greater than or strictly less than relationship indicate that the investigator has an opinion as to the direction of the value of the parameter relative to the stated value. In this case it is called one-tail test.

- Case 1: The variable has a normal distribution and is known. In this case the test statistic is

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

which has a standard normal distribution if .

- Case 2: The variable has a normal distribution and is unknown. The test statistic is

$$Z = \frac{\bar{x} - \mu_0}{S / \sqrt{n}}$$

which has a t_{n-1} distribution if H_0 is true.

- Case 3: The variable is not normal but n is large (which n>30), may be known or unknown.
- The test statistic is

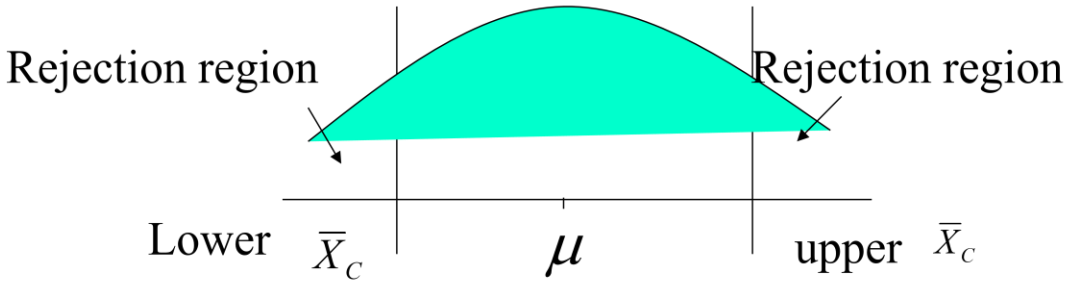
$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \text{ if } \sigma^2 \text{ is known}$$

$$\text{or } Z = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} \text{ if } \sigma^2 \text{ is unknown}$$

- By central limit theorem it has approximately standard normal distribution (0,1) if is true.

  4- Determine the boundary (or boundaries) for the area of rejection regions using either or values. A critical value is the boundary or limit value that requires as to reject the statement of the null hypothesis.

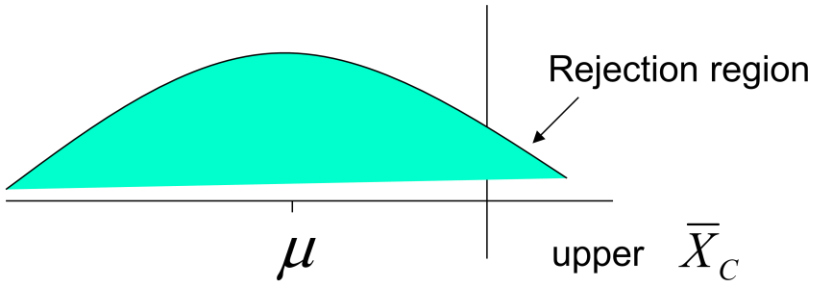
T-Test



In directional test there are two critical values when:

$$H_a : \mu \neq \mu_o$$

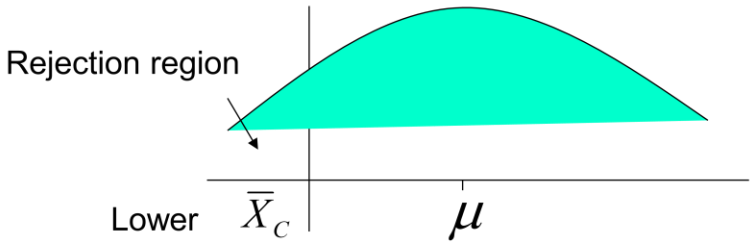
T-Test



In directional test there is one critical value (upper boundary) when:

$$H_a : \mu > \mu_o$$

T-Test



In directional test there is one critical value (lower boundary) when:

$$H_a : \mu < \mu_o$$

- The critical value is simply the maximum or minimum value that we are willing to accept as being consistent with the stated parameter . The mean of the distribution is given by:

$$\mu_{\bar{x}} = \mu$$

- The standard deviation of the distribution is given by:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

- 5- Formulate a decision rule on the basis of the boundary values obtained in step 4. When we conduct an hypothesis test, we are required to make one of two decisions:
- a- Reject H_o or
- B- Accept H_o
- It is possible to make two errors in decision . One error is called a type I error or .We make a type I error whenever we reject the statement of ,when is in fact true. The probability of making a type I

error is the level of significance of the test. The second error we can make in an hypothesis test is called a type II error, or B-error. We commit a type II error if we fail to reject the statement of ,when is in fact false. The four combinations of truth values of and the resulting decisions are summarizing below:

	True H_0	False H_0
Reject H_0	Type I Error	Correct Decision
Accept H_0	Correct Decision	Type II error

When we lower the level of significance of an hypothesis test we always increase the possibility of committing a B-error.

6- State a conclusion for the hypothesis test based on the sample data obtained and the decision rule stated in steps.

P-value of a test:

- The p- value is the probability of getting a value more extreme than one observed value of the test statistic, it is denoted by Z_{obs} When H_a is as follows: $H_a \neq$
- P-value= 2p (Z >| Z_{obs} |)
- When H_a is :>

- p-value= p (Z > z_{obs})
- When H_a is :<
- P-value = p (Z < z_{obs})

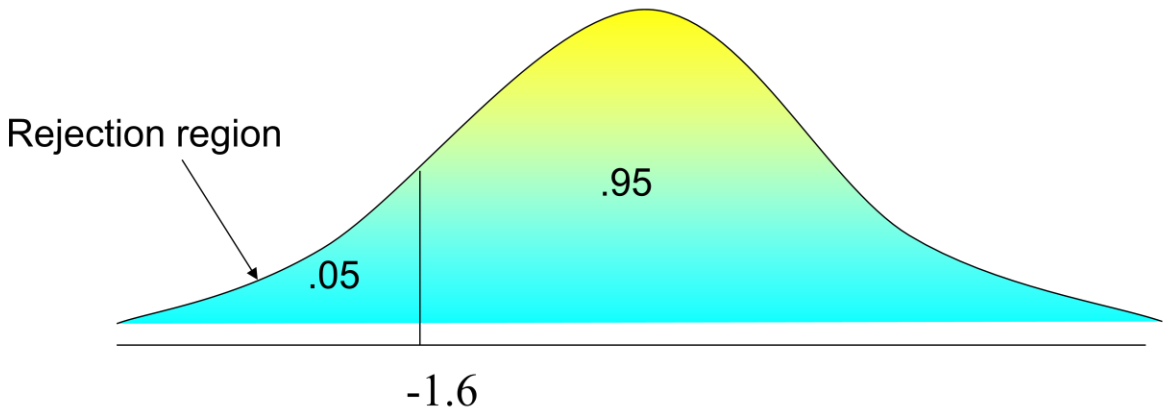
- If we have a T statistic with a t_{n-1} distribution and observe value t_{obs} , these p-values becomes:
- $\neq$ alternative :p-value = 2p (t_{n-1} >| t_{obs} |)

- > alternative :p-value = p (t_{n-1} > t_{obs})
- < alternative :p-value = p(t_{obs} < t_{obs})
- Thus is rejected if p-value < . When data is collected from a normally distributed population and the sample size is small, the t values of the student t distribution must be used in the hypothesis test not the Z values of the normal distribution. This is due to the fact that her central limit theorem does not apply when n < 30.
- Ex:
- Suppose we measure the sulfur content (as a percent) of 15 samples of crude oil from a particular Middle Eastern area obtaining:
- 1.9,2.3,2.9,2.5,2.1,2.7,2.8,2.6,2.6,2.5,2.7,2.2,2.8,2.7,3.
- Assume that sulfur content are normally distributed . Can we conclude that the average sulfur content in this area is less than 2.6? Use a level of significance of .05.

$n = 15 \quad \bar{X} = 2.533 \quad S = .3091 \quad \alpha = .05$

$H_0 : \mu = 2.6$

$H_a : \mu < 2.6$



One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
X	15	2.5533	.3091	7.980E-02

One-Sample Test

	Test Value = 2.6					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
X	-.585	14	.568	-4.667E-02	-.2178	.1245

- Testing for the Difference in Two Population means:

- Often we have two populations for which we would like to compare the means. Independent random samples of sizes and are selected from the two populations with no relationship between the elements we drawn from the two populations. The statistical hypothesis are given by:

$$H_0 : \mu_1 = \mu_2 \text{ vs } \quad H_a : \mu_1 \neq \mu_2$$

$$\text{or} \quad H_a : \mu_1 > \mu_2$$

$$\text{or} \quad H_a : \mu_1 < \mu_2$$

- The e population variances.

- Case1: $\sigma_1^2 \quad and \quad \sigma_2^2$

- Population variances are known for normal populations (or non normal populations with both n1 and n2 large). In this case the test statistic is to be :

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

- Case2:

- Populations are unknown but are to be equal $\sigma_1^2 = \sigma_2^2 = \sigma^2$

- in normal populations. In this case, we pool our estimates to get the pooled two- sample variance

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

- And the test statistic is to be

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_p^2(\frac{1}{n_1} + \frac{1}{n_2})}}$$

- Which has a $t_{n_1+n_2-2}$ distribution if H_O is true.

Case 3:

- σ_1^2 and σ_2^2 are unknown and unequal normal populations . In this case the test statistic is given by:

$$T' = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

which does not have a known distribution.

Ex:

The amount of solar ultraviolet light of wavelength from 290 to 320 nm which reached the earths surface in the Riyadh area was measured for independent samples of days in cooler months (October to March) and in warmer months (April to September):

- Cooler:5.31,4.36,3.71,3.74,4.51,4.58,4.64,3.83,3.16,3.67,4.34,2.95,3.62,3.29 ,2.45.
- Warmer:4.07,3.83,4.75,4.84,5.03,5.48,4.11,4.15,3.9,4.39,4.55,4.91,4.11,3.16 ,2.99,3.01,3.5,3.77.
- Assuming normal distributions with equal variances , test whether there is a difference in the average ultraviolet light reaching Riyadh in the cooler and warmer months . Use a level of significance of .05.

$n_1 = 15$ $n_2 = 18$

$\overline{X}_1 = 3.877$ $\overline{X}_2 = 4.142$

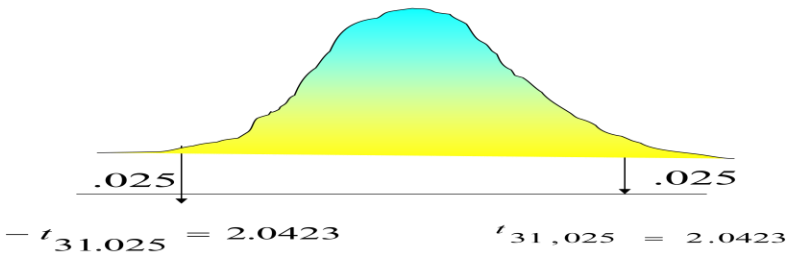
$S_1 = .751$ $S_2 = .709$

- The pooled two sample variance is

$$S_p^2 = \frac{(n_1 - 1)S_2^1 + (n_2 - 1)S_2^2}{n_1 + n_2 - 1} = .531$$

- And the test statistic is to be

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = -1.033$$



Group Statistics

	VAR00002	N	Mean	Std. Deviation	Std. Error Mean
VAR00001	1.00	15	3.8773	.7507	.1938
	2.00	18	4.1417	.7088	.1671

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
VAR00001	Equal variances assumed	.091	.764	-1.039	31	.307	-.2643	.2545	-.7834	.2548
	Equal variances not assumed			-1.033	29.238	.310	-.2643	.2559	-.7875	.2588

- Since the value of the test statistic is in the acceptance region , then H_o is accepted at $\alpha = 05$
- It means that there is no difference in the average ultraviolet light reaching Riyadh in the cooler and warmer months .

What is a Monte Carlo simulation:-

- In a Monte Carlo simulation we attempt to follow the `time dependence' of a model for which change, or growth, does not proceed in some rigorously predefined fashion (e.g. according to Newton's equations of motion) but

rather in a stochastic manner which depends on a sequence of random numbers which is generated during the simulation.

Random Walk:

- Markov chain is a sequence of events with the condition that the probability of each succeeding event is uninfluenced by prior events
- Choosing from Probability Distribution: Any random variable has a probability distribution for its occurrence. We need to choose a random variable which mimics that probability distribution
- Best way to relate random number to a random variable is to use cumulative probability distribution and equating it to the random number

Random Numbers:

- Uniformly distributed numbers in [0,1]
- Most useful method for obtaining random numbers for computer use is a pseudo random number generator
- How random are these pseudo random numbers?

Application to Microscale Heat Transfer :-

- Boltzmann Transport Equation (BTE) for phonons best describes the heat flow in solid nonmetallic thin films
- difficult to solve analytically or even numerically using deterministic approaches
- alternative is to solve the BTE using stochastic or Monte Carlo techniques

Boltzmann Transport Equation for Particle Transport

Distribution Function of Particles: $f = f(\mathbf{r},\mathbf{p},t)$

--probability of particle occupation of momentum **p** at location **r** and time t

Equilibrium Distribution:

f_0, i.e. Fermi-Dirac for electrons, Bose-Einstein for phonons, Plank for photons, etc.

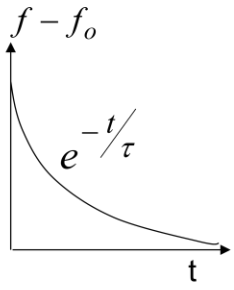
Non-equilibrium, e.g. in a high electric field or temperature gradient:

$$\frac{\partial f}{\partial t} + \mathbf{v}\cdot\nabla_{\mathbf{r}} f + \mathbf{F}\cdot\nabla_{p} f = \left(\frac{\partial f}{\partial t}\right)_{scat}$$

Relaxation Time Approximation

$$\left(\frac{\partial f}{\partial t}\right)_{scat} = \sum_{\mathbf{p}'}\left[\underbrace{W(\mathbf{p},\mathbf{p}')f(\mathbf{p}')}_{\mathbf{p}'\to\mathbf{p}} - \underbrace{W(\mathbf{p}',\mathbf{p})f(\mathbf{p})}_{\mathbf{p}\to\mathbf{p}'}\right] \approx \frac{f_o - f}{\tau(\mathbf{r},\mathbf{p})}$$

Relaxation time



Monte Carlo Solution Technique:-

- Phonons are drawn from the six individual stochastic spaces, including three wave-vector components and the three position vector components
- Phonons are then allowed to drift (or unrestrained motion) and scatter in time, and their statistics is collected at various points in time and space, and processed to extract the necessary information

Initial Conditions:-

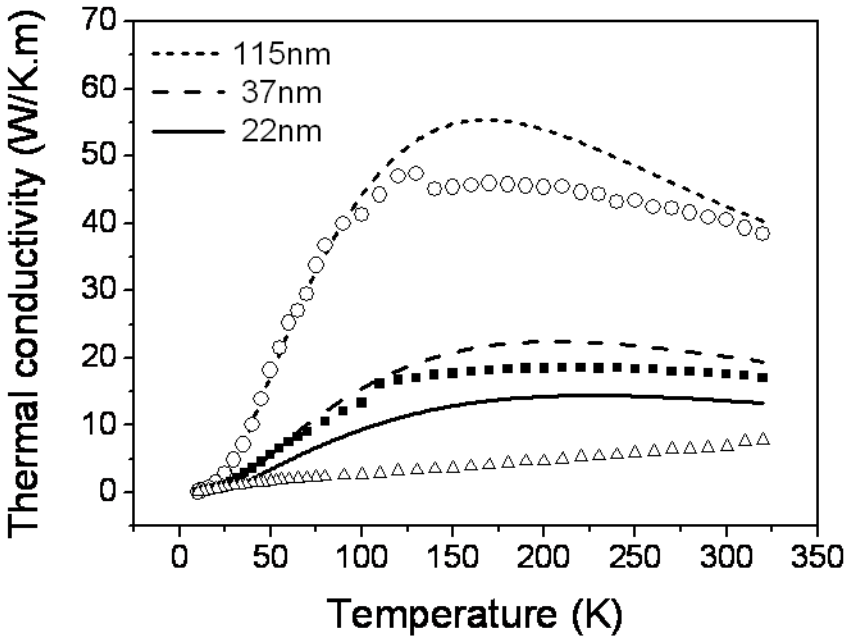
- number of phonons per unit volume and polarization (p) is usually an extremely large number
- a scaling factor is used to simulate only a fraction of the phonons

- A series of random numbers properly distributed to match the equilibrium distribution are drawn to initialize the positions, frequencies, polarizations, and wavevectors of the ensemble of phonons chosen for the simulation
- Mazumdar and Majumdar developed a numerical scheme to obtain the number of phonons within the *i*th frequency interval D*w* as:

$$N_i = \langle n(\omega_{0,i}, LA)\rangle D(\omega_{0,i}, LA)\Delta\omega_i + 2\langle n(\omega_{0,i}, TA)\rangle D(\omega_{0,i}, TA)\Delta\omega_i$$

Monte Carlo Simulation of Silicon Nanowire Thermal Conductivity:-

- Boundary scattering play an important role in thermal resistance as the structure size decreases to nanoscale



-