Draft under revision

Regional Workshop on use of Sampling for Agricultural Census and Surveys

14-18 May, 2012, Bangkok, Thailand

Material* for Review
Bridge course for the Workshop

13 May 2012

Food and Agriculture Organization of the United Nations
Statistics Division

* Prepared by M.K Srivastava, Team Leader, Agricultural Census and Survey and Adriana Neciu, Consultant

Acknowledgement

The authors would like to acknowledge the contributions of numerous websites which provide open access teaching material. The material used in this document has been selectively picked-up purely as an aid for teaching without any profit motive. The picked-up material has been edited and supplemented to suit the requirements of specific set of target audience.

The authors alone are responsible for any mistakes in the compilation and not the organization for which they work or the websites or books from where the material has been borrowed.

Content

Chapter 1 Reading and Understanding Formula

1.1 Order of mathematical operations

Mathematical convention must be followed for all computations. The conventional sequence of operations is:

B- Brackets
O- Of
D- Division
M-Multiplication
A-Addition
S-Subtraction

When solving equations or constructing formulas in a spreadsheet like Microsoft Excel, you need to ensure that the equation or formula is solved in the correct order of operations. If the formula or equation is not solved in the correct order, you may find the result you get, is not the result you were looking for. Please note that the brackets always occur in pairs.

1. Parts of an algebraic expression or numbers enclosed in 'Brackets' must be solved first. There are no exceptions to this rule. For Example; (5 + 3) * 7 = 56, whereas 5 + 3 * 7 = 26.
2. There exist a hierarchy of brackets and they should be opened from smallest to highest in order:

- Line brackets, smallest e.g. $a - \overrightarrow{b - c} = a - b + c$
- Parenthesis $2 * a - \overrightarrow{b - c}$
- Curly brackets $\left\{4 + 3(2 * a - \overrightarrow{b - c})\right\}^2$
- Square brackets $10 + \left[9 * \left\{4 + 3(2 * a - \overrightarrow{b - c})\right\}^2\right]$

3. Please note that absence of Sign of Operation means multiplication. For example, in the above expressions 3 is to be multiplied with the result of solving of quantities within the parenthesis.
4. 'Of' or 'Exponent' (e.g. Square, cube) must be solved next. There are no exceptions to this rule. Most computer programs like Microsoft Excel do not have a mathematical 'Of', so you can ignore the 'Of'. This operation is a verbal expression of operation of Multiplication and division.
5. Next, the parts of the equation that contain 'Division' and 'Multiplication' are calculated. For example; 3 + 9 - 2 * 4 + 5 = 3 + 9 - 8 + 5
6. Where division and multiplication follow one another, then regardless of their order that part of the equation is solved left to right. For example; 3 + 9 - 2 * 6 / 4 + 5 = 3 + 9 - 3 + 5. The correct order is 2 * 6 = 12 then 12 / 4 = 3 even though 6 / 4 = 1.5 and 1.5 * 2 = 3. You will find it very important to get this order correct when doing addition and subtraction.
7. Last but not least, the parts of the equation that contain 'Addition' and 'Subtraction' should be calculated.

As with division and multiplication, where addition and subtraction follow one and other, then regardless of their order that part of the equation is solved left to right. For example; 3 + 9 - 8 + 5 = 12 - 8 + 5, not 12 - 13.

Note the difference as: 12 - 8 = 4, and 4 + 5 = 9, but 12 - 13 = -1
A totally different answer.

Example

2*6+3-4/2-5+20/5*3+50

Solution

12+3-4/2-5+20/5*3+50
12+3-2-5+20/5*3+50
12+3-2-5+4*3+50
12+3-2-5+12+50
15-2-5+12+50
13-5+12+50
8+12+50
20+50=70

Exercise 1

Do these practice exercises by hand

a) (3+3-5)*(15-5)*10-99

(6-5)*(15-5)*10-99
1*(15-5)*10-99
1*10*10-99
10*10-99
100-99 =**1**

b) 2*6+3-4/2-5+20/5*3+50

12+3-4/2-5+20/5*3+50
12+3-2-5+20/5*3+50
12+3-2-5+4*3+50
12+3-2-5+12+50
15-2-5+12+50
13-5+12+50
8+12+50
20+50=**70**

c) 50/5-7*2+11+3*10/2-2+6*5

10-7*2+11+3*10/2-2+6*5
10-14+11+3*10/2-2+6*5
10-14+11+30/2-2+6*5
10-14+11+15-2+6*5
10-14+11+15-2+30
-4+11+15-2+30
7+15-2+30
22-2+30
20+30 =**50**

Exercise 2

Do further exercises available on the website below using ms excel
http://www.abacustraining.biz/bodmasExercises.htm

Exercise 3

The solution for the quadratic equation: $ax^2 + bx + c = 0$

is given by: $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$

Use the above formula to find the value of x from the following equations:

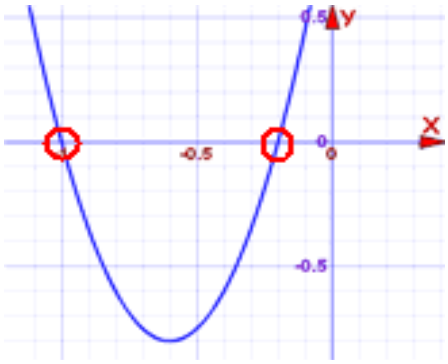
(i) $5x^2 + 6x + 1 = 0$

(ii) For at least 10 arbitrary values of x calculate the value of $Y = (5x^2 + 6x + 1)$ using MS Excel

(iii) Plot the values of x and corresponding Y using Excel.

(iv) Check where the curve cuts X axis.

(v) Compare these values with the solution (i)



1.2 Use of $\sum$ (sigma) notation

☞Variables are represented by capitals English alphabets, e.g. X, Y, Z, ...

☞Values of the variables are represented by corresponding small alphabets, e.g. ''x'' represent a value of the variables ''X''.

☞ Different values of the variables (or observation) are represented by using subscripts e.g. x_1, $x_2 \ldots x_i \ldots x_n$ are the n observation (or values) on the variable X. x_i is the i^{th} observation on X.

☞ $\sum_{i=1}^{n} x_i$ means $(x_1 + x_2 + x_3 + \cdots + x_i \ldots + x_n)$

☞ $\sum_{i=1}^{n} x_i^2$ means $(x_1^2 + x_2^2 + x_3^2 + \cdots + x_i^2 \ldots + x_n^2)$

☞ $\sum_{i=1}^{n}(x_i - \mu)^2$ means $[(x_1 - \mu)^2 + (x_2 - \mu)^2 + (x_3 - \mu)^2 + \cdots (x_i - \mu)^2 + \cdots (x_n - \mu)^2]$

☞ $\vec{x} = \frac{1}{N}\sum_{i=1}^{n} x_i =$ mean of the observation

☞ $\sigma_x^2 =$ variance (X)

$$= \frac{1}{N}\sum_{i=1}^{n}(x_i - \overrightarrow{x})^2$$

$$= \frac{1}{N}\left[(x_1 - \overrightarrow{x})^2 + (x_2 - \overrightarrow{x})^2 + (x_3 - \overrightarrow{x})^2 + \cdots + (x_i - \overrightarrow{x})^2 + \cdots (x_n - \overrightarrow{x})^2\right]$$

☞ $\sigma_x -$ standard deviation (X)

$$= \sqrt{\sigma_X^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(x_i - \overrightarrow{x})^2}$$

Exercise 4

A seed merchant took delivery of a sack of beans. He took a sample of 15 beans from the sack and measured their lengths. His observations measures in centimeters were as follows:

2.4, 2.7, 3.0, 2.5, 3.2, 2.8, 2.8, 2.3, 3.0, 2.8, 2.8, 2.4, 2.9, 2.8, 3.1

Calculate the following measures of variability using the formulae below:

(i) The sample variance

$\sigma^2 = \frac{1}{n-1}\left[\sum x^2 - \frac{(\sum x)^2}{n}\right]$ Where n is number of observations; Answer $= 0.070952$

(ii) The sample standard deviation

$\sigma = \sqrt{\frac{1}{n-1}\left[\sum x^2 - \frac{(\sum x)^2}{n}\right]}$; Answer $= 0.26637$ or 0.27

(ii) Coefficient of variation = Standard deviation /mean (usually expressed as a percentage)

$CV = \frac{\sigma}{\mu}$; Answer = 9.6%

Exercise 5

A book contains 90 Sudoku puzzles at four levels of difficulty: Easy, Mild, Difficult, and Fiendish. I have recorded the times, in minutes, it took me to solve all the puzzles and I have done some analysis, as follows.

Easy	Number of puzzles 4 Times taken 11,9,8,8
Mild	Number of puzzles 16 Times taken 14, 13, 14,12,11,12,13,12,10,11,12,10,11,12,10,11
Difficult	Number of puzzles 45 Mean time taken 18.4 Standard deviation 2.3 minutes
Fiendish	Number of puzzles 25 Mean time taken 25.3 minutes Standard deviation 3.4 minutes

a. Calculate the means and standard deviation of the times taken to solve the easy puzzles and mid puzzles using MS Excel function.

b. Calculate the coefficient of variation of the times taken in each of the four categories using

$CV = \frac{\text{mean}}{\text{standard deviation}}$

c. Calculate the pooled mean (combined for all types of puzzles) and standard deviation using the formulae:

Combined mean $\vec{X} = \frac{n_1\overrightarrow{x_1}+n_2\overrightarrow{x_2}+n_3\overrightarrow{x_3}+n_4\overrightarrow{x_4}}{n_1+n_2+n_3+n_4}$

$$\sigma^2 = \frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2) + n_3(\sigma_3^2 + d_3^2) + n_4(\sigma_4^2 + d_4^2)}{n_1 + n_2 + n_3 + n_4}$$

Where: $n_i -$ number of observation in i^{th}group, $i = 1, 2, 3, 4$
$\overrightarrow{x_i}$ – mean of i^{th} group
d_i- $(\vec{X}_I - \vec{X})$, where $\vec{X}$ is pooled mean
$\sigma_i^2 -$ variance of i^{th}group

Easy puzzle $CV = \frac{\sigma}{\mu}$; Answer = 13%
Mild puzzles $CV = \frac{\sigma}{\mu}$; Answer = 11%
Difficult puzzles $CV = \frac{\sigma}{\mu}$ Answer = 12.5%
Fiendish $CV = \frac{\sigma}{\mu}$; Answer = 13%

Exercise 6
An estimate is required of the mean value of a variable defined on a large population. The population contains two strata (sub-population) in proportions p : (1 – p), where 0 < p < 1.

A random sample of size *n* is to be drawn, comprising n_1 observations from stratum 1 and n_2 observations from stratum 2, where $n_1 + n_2 = n$. For each stratum, the sample means, $\overrightarrow{x_1}$ and $\overrightarrow{x_2}$, and the sample standard deviations, $S_1$and S_2, were computed. The formula for the estimate of the population mean is . $\overrightarrow{x_1} = p\overrightarrow{x_1} + (1 - p)\overrightarrow{x_2}$. An estimate for the standard deviation of (the standard error of the estimate of mean) is:

$$se = \sqrt{\frac{p^2 s_1^2}{n_1} + \frac{(1-p)^2 * S_2^2}{n_2}}$$

(i) Supposing that n_1 = 20, n_2 = 10, p =2/3 , $\overrightarrow{x_1}$ = 5, $\overrightarrow{x_2} = 8$, S_1 = 1 and S_2 = 3, compute $\overrightarrow{x}$and se.
(ii) Supposing that n_1 = 12, n_2 = 18, p =2/3 , $\overrightarrow{x_1}$ =5, $\overrightarrow{x_2} = 8$, S_1 = 1 and S_2 = 3, compute $\overrightarrow{x}$and se.

Exercise 7
Particulars regarding the income of two villages are given below:

	Village X	**Village Y**
Number of people	600	500
Average income	175	186
Variance of income	100	81

Calculate the coefficient of variation for both villages using the formula:

$\text{CV} = \frac{\sigma}{X} * 100;$ where

$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(\text{x}_\text{i} - \vec{\text{x}})^2}$ or

$\sigma = \sqrt{\sum \frac{d^2}{N}}$, where d^2 = is a score's deviation from the mean squared

Exercise 8

Find the missing information from the following:

	Group 1	**Group 2**	**Group 3**	**Combined**
Number	50	?	90	200
Standard deviation	9	7	?	7746
Mean	113	?	115	116

(i) Find the number of observation in the second group given that $\text{N}_1+\text{N}_2+\text{N}_3= 200$. Substitute the known value to find the N_2 .

(ii) Find the mean of the second group: $\overrightarrow{\text{X}_2}$, by using the formula:

Combined mean of the three groups: $\vec{\text{X}}_{123} = \frac{\text{N}_1\overrightarrow{\text{X}_1}+\text{N}_2\overrightarrow{\text{X}_2}+\text{N}_3\overrightarrow{\text{X}_3}}{\text{N}_1+\text{N}_2+\text{N}_3}$,

and solving the equation.

(iii) Find the standard deviation of the third group using the formula,

$$\sigma = \sqrt{\frac{\text{N}_1\sigma_1^2 + \text{N}_2\sigma_2^2 + \text{N}_3\sigma_3^2 + \text{N}_1\text{d}_1^2 + \text{N}_2\text{d}_2^2 + \text{N}_3\text{d}_3^2}{\text{N}_1 + \text{N}_2 + \text{N}_3}}$$

Where: $\text{d}_1 = \overrightarrow{\text{X}_1} - \vec{\text{X}}_{123}$

$\text{d}_2 = \overrightarrow{\text{X}_2} - \vec{\text{X}}_{123}$

$\text{d}_3 = \overrightarrow{\text{X}_3} - \vec{\text{X}}_{123}$

Chapter 2 Graphical Presentation of Data

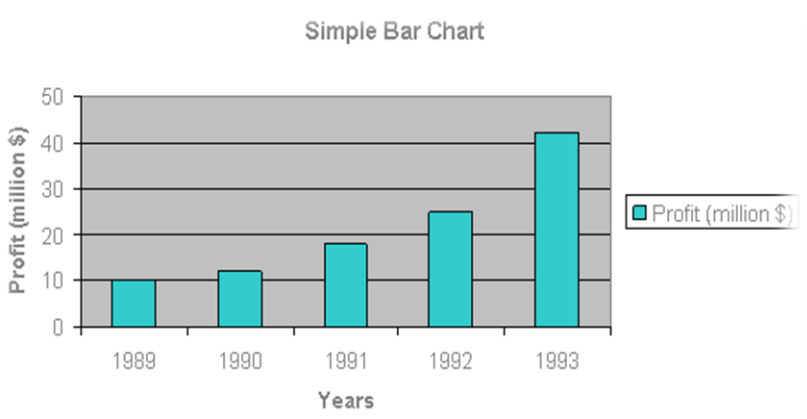
Types of diagrams/charts:

1. Simple Bar Chart
2. Multiple Bar Chart or Cluster Chart
3. Component Bar Chart
   - Simple Component Bar Chart
   - Pie Chart
4. Histograms

Simple Bar Charts

A simple bar chart is used to represent data involving only one variable classified on spatial, quantitative or temporal basis. In simple bar chart, we make bars of equal width but variable length, i.e. the magnitude of a quantity is represented by the height or length of the bars.

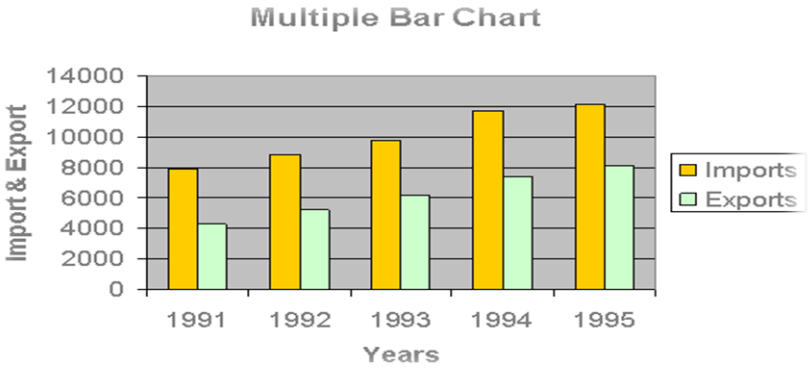
Simple bar chart showing the profit of a bank for 5 years.



Multiple Bar Chart

By multiple bars diagram two or more sets of inter-related data are represented (multiple bar diagram facilities comparison between more than one phenomena). The technique of simple bar chart is used to draw this diagram but the difference is that we use different shades, colors, or dots to distinguish between different phenomena. *We draw multiple bar charts if the total of different phenomena is meaningless.*

Simple bar chart showing the import and export of Canada from 1991 – 1995.

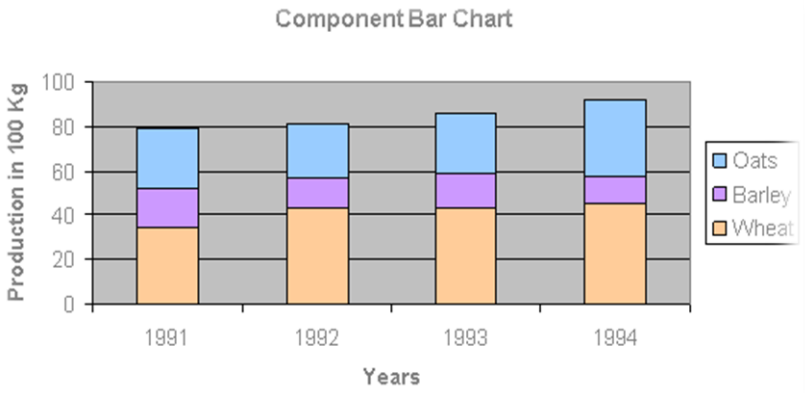


Component Bar Chart

Sub-divided or component bar chart is used to represent data in which the total magnitude is divided into different components.

In this diagram, first we make simple bars for each class taking total magnitude in that class and then divide these simple bars into parts in the ratio of various components. This type of diagram shows the variation in different components within each class as well as between different classes. Sub-divided bar diagram is also known as component bar chart or staked chart.

An example of this diagram is given below:

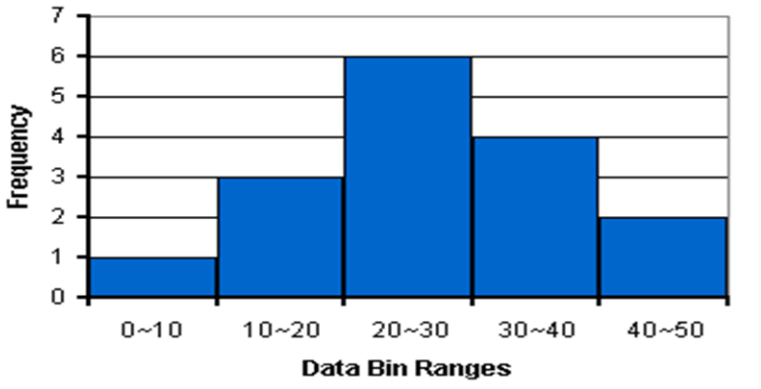


Histograms

A histogram is made up of columns plotted on a graph. Usually, there is no space between adjacent columns. Here is how to read a histogram.

- The columns are positioned over a label that represents a quantitative variable.
- The column label can be a single value or a range of values.
- The height of the column indicates the size of the group defined by the column label.

The histogram below shows frequency for five groups.



The Difference Between Bar Charts and Histograms

Which bar charts, each column represents a group defined by a **categorical variable**; and with histograms, each column represents a group defined by a **quantitative variable**.

Pie Chart

Pie chart can used to compare the relation between the whole and its components. Pie chart is a circular diagram and the area of the sector of a circle is used in pie chart. Circles are drawn with radii proportional to the square root of the quantities because the area of a circle is $2\pi r^2$. To construct a pie chart (sector diagram), we draw a circle with radius (square root of the total). The total angle of the circle is 360^0. The angles of each component are calculated by the formula.

$$\text{Angle of sector} = \frac{\text{Component Part}}{\text{Total}} * 360^0$$

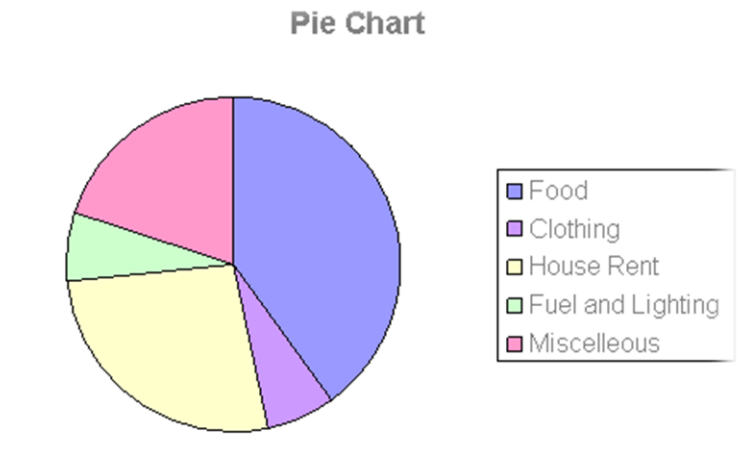
These angles are made in the circle by mean of a protractor to show different components. The arrangement of the sectors is usually anti-clock wise.

The following table gives the details of monthly budget of a family. Represent these figures by a suitable diagram.

Item of Expenditure	Family Budget
Food	$ 600
Clothing	$ 100
House Rent	$ 400
Fuel and Lighting	$ 100
Miscellaneous	$ 300
Total	**$ 1500**

$$\text{Angle of sector} = \frac{\text{Component Part}}{\text{Total}} * 360^0$$

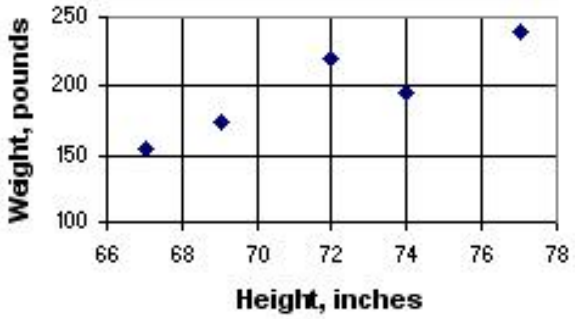
Items	Family Budget		
	Expenditure $	Angle of Sectors	Cumulative Angle
Food	600	144^0	144^0
Clothing	100	24^0	168^0
House Rent	400	96^0	264^0
Fuel and Lighting	100	24^0	288^0
Miscellaneous	300	72^0	360^0
Total	**1500**	**360^0**	



Scatter plots

A scattered plot consist of an X axis (the horizontal axis), a Y axis(the vertical axis) and a series of dots. Each dot on the scatter plot represent one observation from a data set. The position of the dot on the scatter plot represents its X and Y values. This plot is useful for observing covariance tendency of two variables observed on the same set of units.

Height, inches	Weight, pounds
67	155
72	220
77	240
74	195
69	175



Chapter 3. Permutations and Combinations

3.1 What's the difference?

In English we use the word "combination" loosely, without thinking if the **order of things** is important or not. Let us see the difference between two statements give below:

> In saying ***"My fruit salad is a combination of apples, grapes and bananas"*** we don't care what order the fruits are in, they could also be "bananas, grapes and apples" or "grapes, apples and bananas", its the same fruit salad.
>
> In stating ***"The combination to the safe was 472"***. Now we **do** care about the order. "724" would not work, nor would "247". It has to be exactly **4-7-2**.

But, in Mathematics we use more *precise* language.

Combination: .If the order **does not matter**

Permutation: If the order **does matter**

Is this a “Combination Lock” or "Permutation Lock"?

3.2 Permutations

By the permutation of the letters *abc* we mean all of their possible arrangements:
abc, acb, bac, bca, cab, cba.

There are 6 permutations of three different things. As the number of things (letters) increases, their permutations grow astronomically. For example, if twelve different things are permuted, then the number of their permutations is 479,001,600. Now, this enormous number was not found by counting them. It is derived theoretically from the Fundamental Principle of Counting.

If something can be chosen, or can happen, or be done, in *m* different ways, and, *after that has happened*, something else can be chosen in *n* different ways, then the number of ways of choosing both of them is $m \cdot n$.

For example, imagine putting the letters *a, b, c, d* into a hat, and then drawing two of them in succession. We can draw the first in 4 different ways: either *a* or *b* or *c* or *d*. After that has happened, there are 3 ways to choose the second. That is, to *each* of those 4 ways there correspond 3 different ways. Therefore, there are $4 \cdot 3$ or 12 possible ways to choose two letters from four.

All possible draws can be described as:

ab *ba* *ca* *da*
ac *bc* *cb* *db*
ad *bd* *cd* *dc*

ab means that *a* was chosen first and *b* second; *ba* means that *b* was chosen first and *a* second; and so on.

Let us now consider the total number of permutations of all four letters. There are 4 ways to choose the first. 3 ways remain to choose the second, 2 ways to choose the third, and 1 way to choose the last. Therefore the number of permutations of 4 different things is: 4· 3· 2· 1 = 24.

Thus the number of permutations of 4 different things taken 4 at a time is 4!. In general,
The number of permutations of n different things taken n at a time is ***n! = n.(n-1).(n-2)….3.2.1***

Example 1
Five different books are on a shelf. In how many different ways could you arrange them?
Solution

5! = 1· 2· 3· 4· 5 = 120

Example 2
There are 6! Permutations of the 6 letters of the word *square*.
a) In how many of them is *r* the second letter? _ *r* _ _ _ _
b) In how many of them are *q* and *e* next to each other?
Solution
a) Let *r* be the second letter. Then there are 5 ways to fill the first spot. After that has happened, there are 4 ways to fill the third, 3 to fill the fourth, and so on. There are 5! such permutations.
b) Let *q* and *e* be next to each other as *qe*. Then we will be permuting the 5 units *qe, s, u a, r.*. They have 5! permutations. But *q* and *e* could be together as *eq*. Therefore, the total number of ways they can be next to each other is 2· 5! = 240.

Permutations of less than all
We have seen that the number of ways of choosing 2 letters from 4 is 4· 3 = 12. We call this
"The number of permutations of 4 different things taken 2 at a time."
We will symbolize this as ${}^4P_2 = 4\cdot 3$

In general,
"The number of permutations of n different things taken r at a time are:

${}^nP_r$= n.(n-1).(n-2).(n-3)…….r terms = n!/(n-r)!

3.3 Combinations

In permutations, the order is all important i.e. we count *abc* as different from *bca.* But in combinations we are concerned only that a, b, and c have been selected. *abc* and *bca* are the same combination.

Here are all the combinations of abcd taken three at a time: *abc, abd, acd* and *bcd.*
There are four such combinations. We call this as "the number of combinations of 4 distinct things taken 3 at a time". We will denote this number as ${}^4C_3$.

In general,
${}^nC_k$ = The number of combinations of n distinct things taken k at a time.

Now, how are the number of combinations ${}^nC_k$ related to the number of permutations, ${}^nP_k$? To be specific, how are the combinations ${}^4C_3$ related to the permutations ${}^4P_3$?

Since the order does not matter in combinations, there are clearly fewer combinations than permutations. The combinations are contained among the permutations i.e., they are a "subset" of the permutations. Each of those four combinations, in fact, will give rise to 3! permutations:

abc	*abd*	*acd*	*bcd*
acb	*adb*	*adc*	*bdc*
bac	*bad*	*cad*	*cbd*
bca	*bda*	*cda*	*cdb*
cab	*dab*	*dac*	*dbc*
cba	*dba*	*dca*	*dcb*

Each column is the 3! permutations of that combination. But they are all *one* combination -- because the order does not matter. Hence there are 3! times as many permutations as combinations. ${}^4C_3$, therefore, will be ${}^4P_3$ divided by 3! -- the number of *permutations* that each combination generates.

$${}^4C_3 = \frac{{}^4P_3}{3!} = \frac{4 \cdot 3 \cdot 2}{1 \cdot 2 \cdot 3}$$

Notice: The numerator and denominator have the same number of factors, 3, which is indicated by the lower index. The numerator has 3 factors starting with the *upper* index and going down, while the denominator is 3!.

In general,

$$C_k^n = \frac{P_k^n}{k!} = \frac{n(n-1)(n-2)\ldots \text{to k factors}}{k!}$$

Example 1
How many combinations are there of 5 distinct things taken 4 at a time?

Solution: $C_4^5 = \frac{5*4*3*2}{1*2*3*4} = 5$

Again, both the numerator and denominator have the *number of factors* indicated by the lower index, which in this case is 4. The numerator has four factors beginning with the upper index 5 and going backwards. The denominator is 4!.

Example 2

$$C_6^8 = \frac{8 * 7 * 6 * 5 * 4 * 3}{1 * 2 * 3 * 4 * 5 * 6} = 28$$

Notice that both the numerator and denominator have 6 factors. The entire denominator cancels into the numerator. *This will always be the case.*

Example 3

$$C_2^8 = \frac{8*7}{1*2} = 28$$

We see that ${}^8C_2$, the number of ways of taking 2 things from 8, is equal to ${}^8C_6$ (Example 2), the number of ways of taking 8 *minus* 2, or 6. For, the number of ways of taking 2, is the same as the number of ways of leaving 6 behind.

Always, $\mathbf{{}^nC_k = {}^nC_{n-k}}$

The bottom indices, k on the left and $n - k$ on the right, together add up to n.

Example 4

$$C_3^n = \frac{n(n-1)(n-2)}{1*2*3}$$

Factorial representation

In terms of factorials, the number of selections (combinations) of *n* distinct things taken *k* at a time, can be represented as follows:

$\mathbf{{}^nC_k = \frac{n!}{(n-k)!\,k!}}$; This is $\mathbf{{}^nP_k}$ divided by **k!.**

3.4 Applications of concepts

Exercise 1

A country has 4 local varieties of maize (A,B,C, D) and has two new imported varieties, J and K.

(i) How many new hybrid varieties of maize are possible if every variety could cross with any other?

(ii) How many new varieties are possible if only local varieties can be cross-bred with imported varieties?

(iii) How many new hybrid varieties are possible if it is known that:

b) no hybrid is possible within the imported varieties

c) variety A is not compatible with variety B, and variety C is not compatible with variety D.

Exercise 2
A village comprises 100 agricultural holdings. In how many ways can you select a sample for 10 farmers? Will it make a difference if you select a sample one by one, ensuring that if a holding has been selected it cannot be selected again, or select all the 10 holdings all together in one go?

Exercise 3
A village includes 20 small, 30 medium and 50 large farms. You are asked to select a sample of 5 farms from each size class of farmers. The selected sample is pooled together. How many 'pooled' samples are possible?

Exercise 4
A district of a country has: 50 small farms; 30 medium farms and 20 large farms. A stratified sample of 10 farms is to be drawn from this population, including 5 from small farms, 3 from medium farms and 2 from large farm. How many samples are possible? If the restriction of specified sample sizes according to size of farms was not there, and you had the liberty to draw a sample of 10 farms freely from the entire population of 100 farms, without any regard to size, how many samples are possible, if no farm can be selected twice? How will the number of samples change if the sample selection was done one-by-one and the farm selected once was given another chance to be included in the sample. Which approach to sampling you will prefer and why?

Chapter 4 Basic Statistical Concepts

4.1 Variables and attributes

An attribute is a characteristic of an object (person, thing, etc.). While an attribute is often intuitive, the **variable** is the operationalized way in which the attribute is represented for further data processing. Attributes are often observed as "Presence or Absence" of a specified characteristic, e.g., Smoker or Non-smoker, whereas the variables are measured and expressed in numbers. Thus the "*attribute is a qualitative concept"* whereas *"variable is quantitative concept".* When the characteristic of interest is measurable, it is expressed as a **variable.**

"Variables could be converted into Attributes" and "Attributes could be converted into Variables" for statistical treatments.

Example

Age is an attribute that can be operationalized in many ways. It can be dichotomized so that only two values - "old" and "young" - are allowed for further data processing. In this case the attribute "age" is operationalized as a **binary variable**. If more then two values are possible and they can be ordered, the attribute is represented by **ordinal variable**, such as "young", "middle age", and "old". Next, it can be expressed as **rational values**, such as 1, 2, 3.... 99.

The "social class" attribute can be operationalized in similar ways as age, including "lower", "middle" and "upper class" and each class could be further differentiated (divided) between upper and lower, transforming thus changing the three attributes into six or it could use different terminology.

Values (observation on different units) of each variable statistically "vary" (or are distributed) across the **domain** of the variable.

Domain is a set of all possible values that a variable is allowed to have. Domains can be bigger or smaller. The smallest possible domains have those variables that can only have two values, also called *binary* (or dichotomous) variables. A Domain may have **finite, countably infinite** and **infinite** elements (possible values of the variable) in it.

Discrete Variable: A variable which can take only isolated set of values in a given range, e.g., number of students in a class; population of city; oranges in a basket.

Continuous Variables: A variable which can take infinite set of values in a given range, e.g., length of a child. Notwithstanding the limitations of measuring instrument this variable can take many possible values in a range. 2.5 Mt, 2.48 Mt, 2.494 Mt, 2.4952 Mt.

Observations on discrete variable could be treated as observations on a continuous variable under certain conditions for obtaining approximate results using formulas developed for continuous variable.

Use of Symbols to represent variable and their values.
While the variable are represented with upper case symbols (e.g. X, Y, U, V etc), their values are represented with corresponding lower case symbols e.g. x, y, u, v. The observed values of a variable on a specific unit of observation is expressed by assigning a subscript to the value of the variable.

For example,
x_1 is the observation on the first (1^{rst}) unit based for the variable X.
x_i is the observation on the i^{th} unit based for the variable X.

If there are n units under consideration. The observations on X will be $\{ X_1 + X_2 + X_3 + \cdots + X_i \ldots + X_n \}$.

One unit of observation may have several characteristics of interests e.g. A household may be observed on the following variables.

X: Number of persons in the household
Y: Monthly Income
Z: Expenditure on consumption
W: Per capita calorie consumption

Thus the observations on these variable could be expressed a ordered set of values (observed measurement of the variables on the unit) e, g. (x,y)., (x,y,z) and (x,y,z,w). These ordered sets of data are useful for studying association between variables or for calculating correlation, regression, partial correlation, multiple regression or establishing models for forecasting.

Population: A set of all possible units under consideration constitute a ***Population or Universe.*** Population reflects the *zone of consideration* for which estimates are to be prepared. E.g. All households in a city could be the population for a household sample survey whose results are to used as representative for the city. The list of all the units in the reference population is called a ***sampling frame***. A sampling frame include all units about which a statement is to be made on the basis of the survey of a sample survey. There is no duplication or omission in the list.

A population could also be considered as being composed of sub-populations (clusters or strata). E.g. The localities in a city could be considered as sub-populations of the city (population). The number of Units in the Population is usually denoted by "N". The total number of units in all the sub-populations is equal to the number of units in the Population.

Sample: is a subset of units in the population.

Representative Sample: A representative sample is drawn from the population according to scientific procedures so that the results of investigation (or estimates) based on the sample are valid for the population.

Purposive Sample: A sample drawn with a specific purpose in mind is called a purposive sample. The validity or representativeness of the sample depends upon how the sample was drawn.

4.2 Concept of distribution of a variable

Frequency: how often something occurs.

Example 1
Sam played football on

- Saturday Morning,
- Saturday Afternoon
- Thursday Afternoon

The frequency was 2 on Saturday, 1 on Thursday and 3 for the whole week. A *frequency distribution* is an arrangement of the values that one or more variables take in a sample. Each entry in the table contains the frequency or count of the occurrences of values within a particular group or interval, and in this way, the table summarizes the distribution of values in the sample.

Frequency Distribution: values and their frequency (how often each value occurs). By counting frequencies we can make a Frequency Distribution table.

Example 2: Number of seeds in pea pods.
In a yield estimation exercise following numbers of pea seeds were observed in the a sample of 14 pods: 2, 3, 1, 2, 1, 3, 2, 3, 4, 5, 4, 2, 2, 3

No. of seeds	Frequency
1	2
2	5
3	4
4	2
5	1
TOTAL	**14**

From the table we can see interesting things such as

- getting 2 seeds happens most frequently
- only one pod had 5 seeds

Example 3: Newspapers sale
The numbers of newspapers sold at a local shop over the last 10 days are: 22, 20, 18, 23, 20, 25, 22, 20, 18, 20. Let us count how many of each number there is:

Papers Sold	Frequency
18	2
19	0
20	4
21	0
22	2
23	1
24	0
25	1

It is also possible to **group** the values. Here they are grouped in 5s:

Papers Sold	Frequency
15-19	2
20-24	7
25-29	1

This information can be used by the shop keeper for taking a decision about optimal daily supply requirement.

A **frequency distribution** shows us a summarized grouping of data divided into *mutually exclusive classes* and the number of occurrences in a class. It is a way of showing unorganized data in an organized manner so that meaningful conclusions could be drawn, e.g., to show results of an election, income of people for a certain region, sales of a product within a certain period, student loan amounts of graduates, etc. Some of the graphs that can be used with frequency distributions are histograms, line graphs, bar charts and pie charts. Frequency distributions are used for both qualitative and quantitative data. The above are examples of **Univariate Frequency Distributions.**

Cumulative frequency
Cumulative frequency (less than type) for a class interval is the number of observations less than or equal to the upper limit of the class interval. It could also be expressed in percentages (relative frequency). Cumulative frequency corresponding to a particular value is the sum of all the *frequencies up to and including that value.*

Years	Frequency	Cumulative frequency
1900-20	12	12
1920-40	15	27
1940-60	9	36
1960-80	15	51
1980-00	21	72

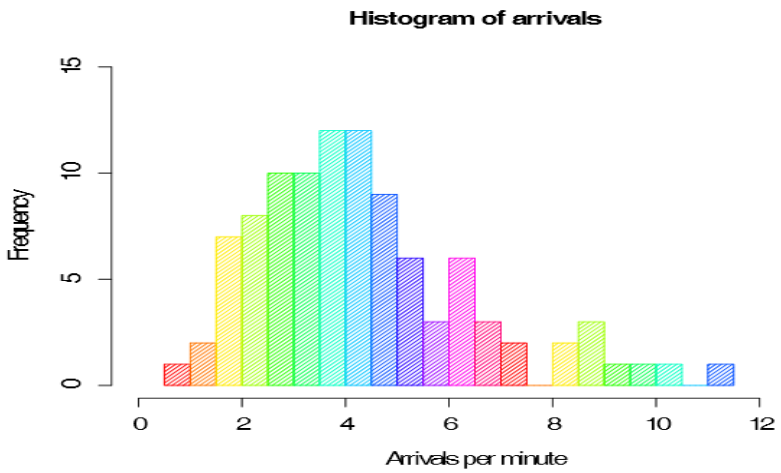
Bi-Variate or Joint Frequency Distributions are often presented as (two-way) contingency table:

Two-way contingency table with marginal frequencies				
	Dance	**Sports**	**TV**	**Total**
Men	2	10	8	20
Women	16	6	8	30
Total	18	16	16	50

The row *total* and column *total* report the *marginal frequencies* or ***marginal distribution***, while the body of the table reports the *joint frequencies or* ***Joint Frequency Distribution*****.**

4.3 Frequency and probability

The **frequency** of an event *i* is the number n_i of times the event occurred in the experiment or the study. These frequencies are often graphically represented in histograms.



We speak of **absolute frequencies**, when the counts n_i themselves are given and of **relative frequencies**, when those are normalized by the total number of events:

$$f_i = \frac{n_i}{N} = \frac{n_i}{\sum_i n_i}$$

Taking the f_i for all *i* and tabulating or plotting them leads to a frequency distribution.

The concept **relative frequency density** is used in case of grouped data or continuous variables. It is often termed as the empirical probability or experimental probability. Relative frequency density help to compare data among different classes of an equal width. It is a procedure to normalize the relative frequency by class width.

For example: If the lower extreme of the class you are measuring the density of is 15 and the upper extreme of the class you are measuring is 30, given a relative frequency of 0.0625, you would calculate the frequency density for this class to be:

Relative frequency / (Upper extreme of class − lower extreme of class) = density
0.0625 / (30 − 15) = 0.0625 / 15 = 0.0041666.. That is: 0.00417 to 5 decimal places.

The **limiting relative frequency** of an event over a long series of trials is the conceptual foundation of the frequency interpretation of probability. In this framework, it is assumed that as the length of the series increases without bound, the fraction of the experiments in which we observe the event will stabilize. This when the number of trials are infinitely large, the relative frequency (density) converges to probability.

Chapter 5 Measure of Location and Dispersion

5.1 Central tendency

The tendency of the data to cluster around a value (called central value, mean, average) is called the Central Tendency of the data. The Measures of Central Tendency are the formulae designed to discover this value about which most data is concentrated.

The Mean is the average value of a set of two or more numbers. The mean for a given set of numbers can be computed in more than one way, including the arithmetic mean method, which is based on the sum of the numbers in the series, and the geometric mean method which is based on multiplication of numbers. However, all of the primary methods for computing a simple average of a normal number series produce the same approximate result most of the time.

The most basic statistical formula:

$$\vec{X} = \frac{\sum X}{N}$$

Where:
$\bar{X}$ (called the X-bar) is the symbol for the arithmetic mean.
Σ (the Greek letter sigma) is the symbol for summation.
X is the symbol for the scores.
N is the symbol for the number of scores.

So this formula simply says you get the mean by summing up all the scores and dividing the total by the number of scores

If you have just 6 numbers (3, 9, 10, 8, 6, and 5), you insert them into the formula for the mean, and do the math:

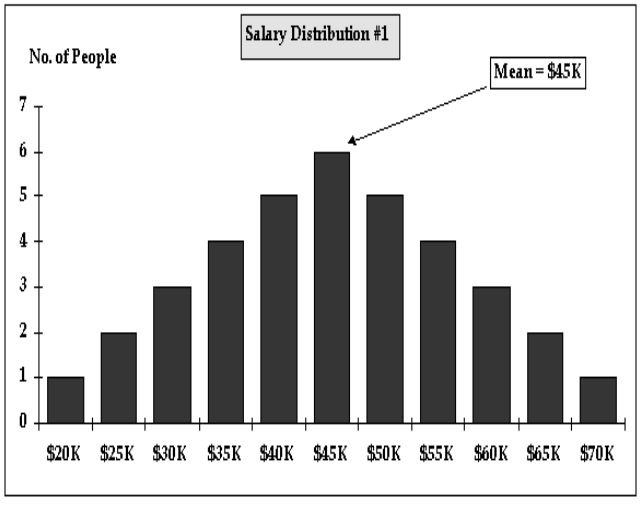
$$\vec{X} = \frac{\sum X}{N} = \frac{(3+9+10+8+5)}{6} = \frac{41}{6} = 6.83$$

Example 1: Mean salary

We're going to compute the mean salary of 36 people. Column A of Table below show the salaries (ranging from $20K to $70K), and column B shows how many people earned each of the salaries.

A	B	C
Salary (X)	**Frequency (f)**	**fX**
$20k	1	20
$25K	2	50
$30K	3	90
$35K	4	140
$40K	5	200
$45K	6	270
$50K	5	250
$55K	4	220
$60K	3	180
$65K	2	130
$70K	1	70
Sum	**36**	**1,620**

Distribution of salaries



- To get the ΣX for our formula, we multiply the number of people in each salary category by the salary for that category (e.g., 1 x 20, 2 x 25, etc.)
- Then total those numbers (the ones in column C).

$$\vec{X} = \frac{\sum X}{N} = \frac{\text{Total salary earned}}{\text{Number of salaries earned}} = \frac{1620}{36} = 45$$

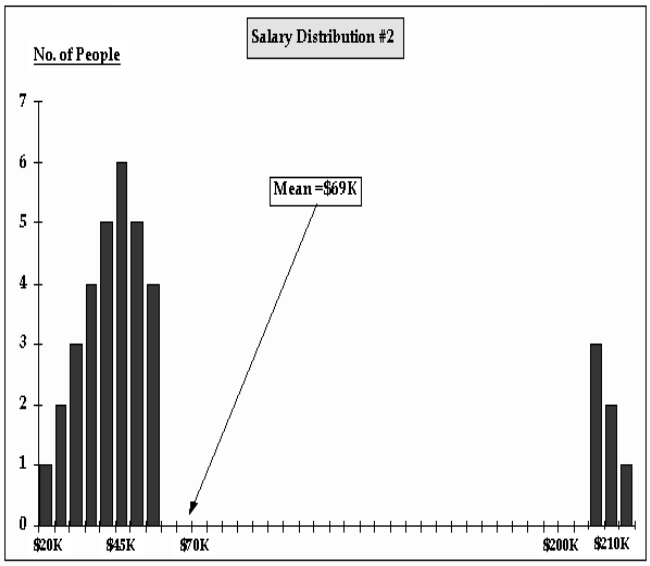
The scores in this distribution are said to be **normally distributed,** i.e., clustered around a central value, with decreasing numbers of cases as you move to the extreme ends of the range. Thus the term **normal curve.**

Example 2

Let's suppose that the three people who made \$60K actually made \$220K, and that the two who made \$65K made \$205K, and the one person who made \$70K made \$210K. The revised salary table is the same except for these changes.

A	B	C
Salary (X)	**Frequency (f)**	**fX**
\$20k	1	20
\$25K	2	50
\$30K	3	90
\$35K	4	140
\$40K	5	200
\$45K	6	270
\$50K	5	250
\$55K	4	220
\$200K	3	600
\$205K	2	410
\$210K	1	210
Sum	**36**	**2,460**

Distribution of salaries



Now, we compute the mean as follows:

$$\vec{X} = \frac{\sum X}{N} = \frac{2460}{36} = 68.3$$

What this shows is that changing the salaries of just six individuals to extreme values greatly affects the mean. In this case, it raised the mean from \$45K to \$68.3K (an increase of 52%), even though all the other scores remained the same. In fact, the mean is a figure that no person in the group has—hardly a figure we would think of as "average" for the group. *(The important lesson here is that the mean is intended to be a measure of central tendency, but it works usefully as such only if the data on which it is based are more or less normally distributed).* The presence of extreme scores distorts the mean, because we note that in this case, mean salary (\$68.3K) is not a very good indication of the "average" salary of this group of 36 individuals. So if we know or suspect that our data may have some extreme scores (called **outliers)** that would distort the mean, the measure that can we use to give us a better measure of central tendency is the ''median'' which is less insensitive to presence of outliers or extreme values.

The Median is the point in the distribution above which and below which 50% of the scores lie. If we list the scores in order from highest to lowest (or lowest to highest) and find the middle-most score, that's the median.

Suppose we have the following scores: 2, 12, 4, 11, 3, 7, 10, 5, 9, 6. The next step is to array them in order from lowest to highest. 2,3,4,5,6,7,9,10,11,12. Since we have 10 scores, and 50% of 10 is 5, we want the point above which and below which there are five scores. If you count up from the bottom, you might think the median is 6. But that's not right because there are 4 scores below 6 and 5 above it.

In statistics, a measurement or a score is regarded not as a point but as an interval ranging from half a unit below to half a unit above the value. So in this case, the actual midpoint or median of this distribution—the point above which and below which 50% of the scores lie—is 6.5

When the observations are arranged in ascending order,

$$\text{Median} = \left(\frac{N+1}{2}\right)^{th}, \text{observation when N is odd}$$

$$\text{Median} = \text{Average of } \left(\frac{N}{2}\right)^{th} \text{ and } \left(\frac{N}{2}+1\right)^{th}, \text{observation when N is even}$$

For grouped data,

$$\text{Med} = L + \frac{\frac{N}{2} - c.f}{f} * h$$

Where: N – the number of scores
L- the lower limit of the median class
c.f.- cumulative frequency of the class preceding the median class
f- simple frequency of the median class
h-the class interval of the median class

The median class is identified by calculating cumulative frequencies.

Example 1

Salary	Range	Frequency	Cumulative Frequency
$20K	$19.5K-20.5K	1	1
$25K	$24.5K-25.5K	2	3
$30K	$29.5K-30.5K	3	6
$35K	$34.5K-35.5K	4	10
$40K	$39.5K-40.5K	5	15
$45K	**$44.5K-45.5K**	6	21
$50K	$49.5K-50.5K	5	26
$55K	$54.5K-55.5K	4	30
$60K	$59.5K-60.5K	3	33
$65K	$64.5K-65.5K	2	35
$70K	$69.5K-70.5K	1	36
Sum		36	

To calculate the median we order the salaries from lowest to highest (there are already). Then determine how many individuals (ratings, scores, or whatever) we have. Those are shown in the frequency column, and the total is 36. So our **N = 36.**

We want to find the salary point above which and below which 50%, or 18, of the individuals fall. If we count up from the bottom through the $40K level, we have 15, and we need three more. But if we include the $45K level (in which there are 6), we have 21, three more than we need. Thus, we need 3, or 50%, of the 6 cases in the $45K category. We add this value (.5) to the lower limit of the interval in which we know the median lies ($44.5K-$45.5K), and this gives us value of $45K.

In this case, **the mean and the median are the same as they always are in normal distributions**. So in situations like this, the mean is the preferred measure.

Exercise
Calculate mean and median from the following distributions and comment on their equality or otherwise based on the normally or otherwise on the distribution.

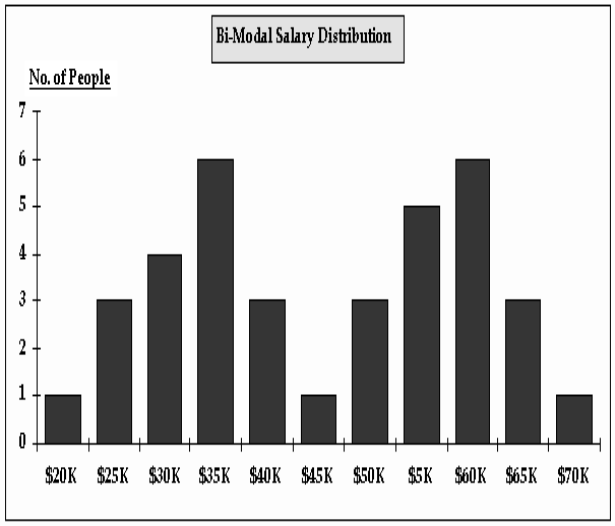
Salary	**Range**	**Frequency**
$20k	$19.5K-20.5K	1
$25K	$24.5K-25.5K	2
$30K	$29.5K-30.5K	3
$35K	$34.5K-35.5K	4
$40K	$39.5K-40.5K	5
$45K	$44.5K-45.5K	6
$50K	$49.5K-50.5K	5
$55K	$54.5K-55.5K	4
$200K	$199.5K-200.5K	3
$205K	$204.5K-205.5K	2
$210K	$209.5K-210.5K	1
Sum		**36**

The Mode is the most frequently occurring score or value. Example 1 above, that value is 45K. But sometimes we may have odd distributions in which there may be two peaks. Even if the peaks are not exactly equal, they're referred to as bi-modal distributions.

Let's assume we have such a bi-modal distribution of salaries as shown in Example below.

Example Bi-Modal Distribution of Salaries

A	B	C
Salary (X)	**Frequency (f)**	**fX**
$20K	1	20
$25K	3	75
$30K	4	120
$35K	6	210
$40K	3	120
$45K	1	45
$50K	3	150
$55K	5	275
$60K	6	360
$65K	3	195
$70K	1	70
Sum	**36**	**1,640**



Before we talk about the mode, using the formula and calculation procedures you've just learned, calculate the mean and median for the salaries in Table (the fx and the Σx data are in Column C).

When you look at this distribution of salaries, as shown graphically in Figure, it's hard to discern any central tendency. The mean is $45K, which only one person earns, and the median is also $45K, which, while it is the middle-most value (50% of the cases are above and below it), certainly does not give us a meaningful indication of the central tendency in this distribution—because it is not there. Therefore, the most informative general statement we can make about this distribution is to say the it is **bi-modal.**

Readers are suggested to follow standard text books on calculation of mode; there are methods to deal with bi-modal or multi-modal distributions.

Quartiles are the three values which divide the data in four parts, when it is arranged in ascending or descending order. Median (Q_2) is the second quartile, (Q_3) has 75% of the total number of observations below it, (Q_1) has 25%.

Percentiles and related characteristics

A percentile is the value of a variable below which a certain percent of observations fall. For example, the 20th percentile is the value (or score) below which 20 percent of the observations may be found.

The 25th percentile is also known as the first quartile (Q_1), the 50th percentile as the median or second quartile (Q_2), and the 75th percentile as the third quartile (Q_3). Percentile describes the relative location of points anywhere along the range of a distribution. A score that is at a certain percentile falls even with or above that percent of scores. The median score of a distribution is at the 50th percentile. It is the score at which 50% of other scores are below (or equal) and 50% are above.

Quintiles or partition values
Quantiles are used to divide the distribution in n parts. When n=10, these partition values are called deciles. Median is the 5^{th} decile.

5.2 Dispersion

The dispersion in a set of data is the variation among the set of data values. It measures whether the values or individual data are all close together, or more scattered. The measures of dispersion give us an idea of degree of scatterness.

5.2.1 Range is the simplest measure of dispersion. It can be thought of in two ways:

- As a quantity; the difference between the highest and lowest score in a distribution, e.g., the range of scores in an examination was 32.
- As an interval; the lowest and highest scores may be reported as the range, e.g.: The range was "62 to 94'' which is written as (62, 94).

The range is easy to calculate but it depends only on the largest and smallest values which may sometimes be extremes or outliers[1]; it does not use the whole pattern including the central values. Range is determined by the farthest outliers at either end of the distribution. Range is of limited use as a measure of dispersion, because it reflects information about extreme values but not necessarily about "typical" values. Only when the range is "narrow" (meaning that there are no outliers) does it tell us about typical values in the data.

5.2.2 Inter quartile range is the differences between the upper quartile (Q_3) and the lower quartile (Q_1). Q_3 has 75% of the total number of observations below it, and Q_1 has 25%. Once the data are arranged in rank order Q_1, and Q_3 are easy to locate, and the quartile value is not affected by the most extreme data points.

5.2.3 Variance and Standard Deviation
The variance is a measure of how far a set of numbers are spread out. Variance is a mathematical expectation (of the average) of squared deviations from the mean. [Its value is seriously influenced by extreme values in a set of data as the deviation are squared]

Standard deviation too is a measure of the dispersion of a set of data from its mean. Standard deviation is calculated as the positive square root of variance. Standard deviation is very useful in theoretical work and in statistical methods and inference.

[1] An outlier is an extreme score, i.e., an infrequently occurring score at either tail of the distribution which has a large impact on calculated measures like mean, variance. Very often procedures are needed to detect and remove the outliers before tabulation of survey results. Outliers may be real values or they might have arisen due to mistakes in data capturing processes.

The formula for the standard deviation is:

$$\sigma = \sqrt{\sum \frac{d^2}{N}}$$

Where:
σ (little sigma) is the standard deviation.
d^2 is squared deviation from the mean.
N is the number of cases.

Variance

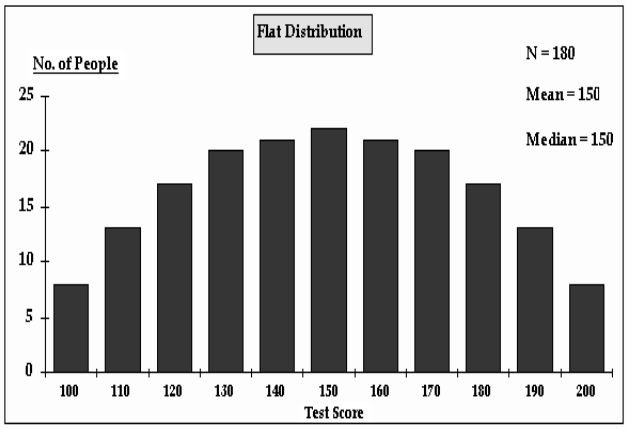
$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2$$

Standard Deviation

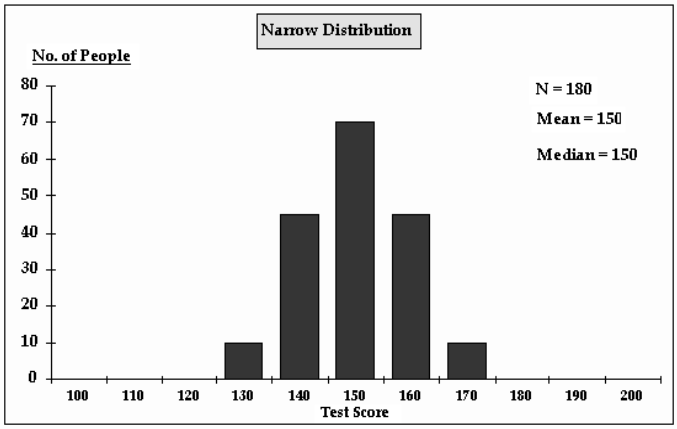
$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2}$$

To understand the concept of variation in a distribution, compare the distribution of test scores in Figures 1 and 2 (below). The first is flat and spread out, while the second is concentrated and bunched up closely around the mean.

1. **Graphic Display of Flat or Spread Out Score Distribution**



2. **Display of a Narrow or Concentrated Distribution**



Note that he mean and median of these two quite different distributions are the same ($\bar{X}$ = 150, Mdn = 150), so simply calculating and reporting those two measures of central tendency would fail to reveal how different the dispersion of scores is between the two groups. But we can do this by calculating the standard deviation

The numbers we need to calculate the standard deviation for Figure 1, the flat distribution, are in the following table.

The Flat Distribution

A	B	C	D	E
Test Score (X)	**Frequency (f)**	**X–Mean (d)**	**fd**	**fd^2**
100	8	50	400	20,000
110	13	40	520	20,800
120	17	30	510	15,300
130	20	20	400	8,000
140	21	10	210	2,100
150	22	0	0	0
160	21	-10	-210	2,100
170	20	-20	-400	8,000
180	17	-30	-510	15,300
190	13	-40	-520	20,800
200	8	-50	-400	20,000
SUM	**180**			**132,400**

Column A displays the test scores (X).
Column B shows how many people got each test score (f).
Column C - the test score minus the mean (X minus the mean or d).
Column D - the sum of the deviations in column C (fd).
Column E contains the squares of all the deviations.

Of course, to get the deviation of each score from the mean (column C), we have to first calculate the mean, and you already know how to do that. We now have what we need to calculate the standard deviation for the flat distribution in Figure 4:

$$\sigma = \sqrt{\sum \frac{d^2}{N}} \text{ or } \sigma = \sqrt{\frac{132{,}400}{180}} = 27$$

You can do the last part of this calculation, the square root of 132,400/180 (which is 736) by using the square-root button on your little hand calculator or raising the number to exponent $^{1/2}$

Now let's compute the standard deviation for the data in Figure 2. The data are in the following table, and you follow the same steps we've just completed.

Narrow or Concentrated Distribution

A	B	C	D	E
Test Score (X)	**Frequency (f)**	**X - Mean (d)**	**fd**	**fd^2**
100	0	50	0	0
110	0	40	0	0
120	0	30	0	0
130	10	20	200	4,000
140	45	10	450	4,500
150	70	0	0	0
160	45	-10	-450	4,500
170	10	-20	-200	4,000
180	0	-30	0	0
190	0	-40	0	0
200	0	-50	0	0
SUM	180			17,000

$\sigma = \sqrt{\sum \frac{d^2}{N}}$ or $\sigma = \sqrt{\frac{17,000}{180}} = 10$

The two standard deviations provide a statistical indication of the how different the distributions are: 27 for the spread-out distribution and 10 for the bunched-up distribution. So, once we know the mean and median, why do we need to know the standard deviation? What use is it?

The standard deviation is important because regardless of the mean, it makes a great deal of difference whether the distribution is spread out over a broad range or bunched up closely around the mean.

Example 1

Suppose you have two classes whose mean reading scores are the same. With only that information, you would be inclined to teach the two classes in the same way. But suppose you discover that the standard deviation of one of the classes is 27 and the other is 10, as in the examples we just finished working with. That means that in the first class (the one where $\sigma = 27$), you have many students throughout the entire range of performance. You'll need to have teaching strategies for both the gifted and the challenged. But in the second class (the one where $\sigma = 10$), you do not have any gifted or challenged students. They are all average, and your teaching strategy will be entirely different.

Example 2

In a number of diverse agro-climatic agricultural trials two varieties Y_1 and Y_2 of wheat have shown the same average yield of 500 Kg per hectare. However, it was noted that the variety Y has an standard deviation which is three time the standard deviation of Y_2. Which variety will you choose for national promotion? And Why? Normally, Y_2, because the its is less risky (measured by standard deviation) in terms of productivity. Obviously, you do not want to take risk with the national production in a country which has a fragile food security situation and divergent agro-climatic condition which are subject to vagaries of weather.

5.2.4 Coefficient of variation

The coefficient of variation (CV) is a normalized measure of dispersion. It is also known as *unitized risk* or the *variation coefficient*. The absolute value of the CV is sometimes known as *relative standard deviation* (RSD), which is expressed in percentages. CV should not be used interchangeably with RSD (i.e. one term should be used consistently).

The coefficient of variation (CV) is defined as the ratio of the standard deviation to the mean μ.

$\mathrm{CV} = \frac{\sigma}{\mu}$, which is the inverse of the signal-to-noise ratio.

The coefficient of variation should be computed only for data measured on a ratio scale, which are measurements that can only take non-negative values. The coefficient of variation may not have any meaning for data on an interval scale.

For example, most temperature scales are interval scales (e.g. Celsius, Fahrenheit etc.), they can take both positive and negative values. The Kelvin scale has an absolute null value, and no negative values can naturally occur. Hence, the Kelvin scale is a ratio scale. While the standard deviation (SD) can be derived on both the Kelvin and the Celsius scale (with both leading to the same SDs), the CV could only be derived for the Kelvin scale.

Often, laboratory values that are measured based on chromatographic methods are log-normally distributed. In this case, the CV would be constant over a large range of measurements, while SDs would vary depending on the actual range that has been measured.

The CV is sometimes expressed as a percent, in which case the CV is multiplied by 100%. The CV is often preferred to standard deviation, as it is unit free.

5.2.5 Comparison of measures of dispersion

When data are described by a measure of central tendency (mean, median or mode), all the scores are summarized by a single value. Reports of central tendency are commonly supplemented and complemented by including a measure of dispersion. The measure of dispersion you have just seen differ in ways that will help determine which one is most useful in a particular situation.

Range. Of all the measure of dispersion, the range is the easiest to determine. It is commonly used as a preliminary indicator of dispersion. However, because it takes into account only scores that lie at two extremes, it is of limited use.

Quartiles Scores are based on more information than the range, and are not affected by outliers. However, they are only infrequently used to describe dispersion because they are not as easy to calculate as the range and they do not have the mathematical properties that make them so useful as standard deviation and variance.

The **standard deviation** and **variance** are more complete measure of dispersion which take into account every score in a distribution. The other measures of dispersion we have discussed are based on considerably less information. However, because variance relies on the squared differences of scores from the mean, a single outlier has greater impact on the size of the variance than does a single score near the mean. Some statisticians view this property as a shortcoming of variance as a measure of dispersion, especially when there is reason to doubt the reliability of some of the extreme scores.

For example, a researcher might believe that a person who reports watching television on an average of 24 hours per day may have misunderstood the question. Just one such a extreme score might result in a appreciably larger standard deviation, especially if the sample is small. Fortunately, since all the scores are used in the calculation of variance, the many non-extreme scores (those closer to the mean) will tend to offset the misleading impact of any extreme scores.

The standard deviation and variance are the most commonly used measure of dispersion in the social science because:

- Both take into account the difference between each score and the mean. Consequently, these measures are based on a maximum amount of information.
- The standard deviation is the baseline for defining the concept of standardized score or ‘’z-score’’.
- Variance in a set of scores on some dependent variable is a baseline for measuring the correlation between two or more variables.(the degree to which they are related).

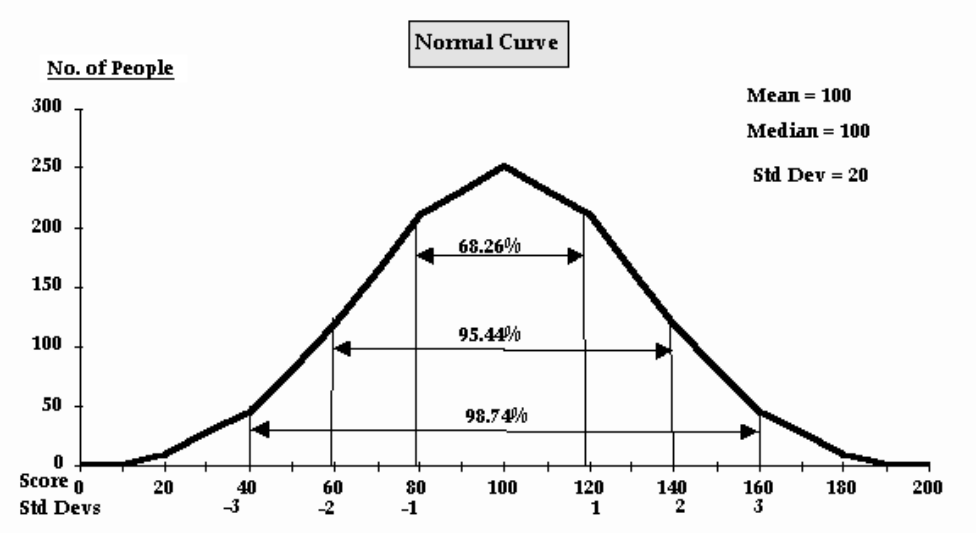
5.3 The normal curve

The reason for it to be called so is that so many things in life are distributed in the shape of this curve: IQ, strength, height, weight, musical ability, resistance to disease, and so on. Not everything is normally distributed, but most things are. Thus the term normal curve.

In Figure below, we have a set of scores which are normally distributed. The range is from 0 to 200, the mean and median are 100, and the standard deviation is 20. In a normal curve, the standard deviation indicates precisely how the scores are distributed. Note that the percentage of scores is marked off by standard deviations on either side of the mean. In the range between 80 and 20 (that’s one standard deviation on either side of the mean), there are 68.26% of the cases. In other words, in a normal distribution, roughly two thirds of the scores lie between one standard deviation on either side of the mean. If we go out to two standard deviations on either side of the mean, we will include 95.44% of the scores; and if we go out three standard deviations, that will encompass 98.74% of the scores; and so on.

Another way to think about this is to realize that in this distribution, if you have a score that’s within one standard deviation of the mean, i.e., between 80 and 120, that’s pretty average—two thirds of the people are concentrated in that range. But if you have a score that’s two or three standard deviations away from the mean, that is clearly a deviant score, i.e., very high or very low. Only a small percent of the cases lie that far out from the mean. This is valuable to understand in its own right, and will become useful when we take up determining the significance of difference between means.

Figure
Normal Curve Showing the Percent of Cases Lying Within 1, 2, and 3 Standard Deviations From the Mean.



5.4 Standardized distribution scores, or ‘’Z- Scores’’

Actual scores from a distribution are commonly known as a ‘’raw scores’’ . These are expressed in terms of empirical units like dollars, years, tons, etc. We might say ‘’The Smith family’s income is $ 29 418. The compare a raw score to the mean we might say something like ‘’The mean household income in the US $ is 2 232 above the Smith family’s income’’. The differences is an absolute deviation of 2 232 empirical units from the mean.

When we are given an absolute deviation from the mean, expressed in terms of empirical units, it is difficult to tell if the difference is large or small compared to other members of the data set. In the above example, are there many families that make less money than the Smith family or only few? We are not given enough information to decide.

We get more information about deviation from the mean when we use the standard deviation measure presented earlier in this tutorial. Raw scores expressed in empirical units can be converted to ‘’’standardized’’ scores, called **z-scores**. The **z-scores** is a measure of how many units of standard deviation the raw score is from the mean. Thus the **z-score** is a relative measure instead of an absolute measure. This is because every individual in the dataset affects value for the standard deviation. Raw scores are converted to standardized z- scores by the following equations.

Population z-score $$Z = \frac{x - \mu}{\sigma}$$

Sample z-score $$z = \frac{x - \bar{X}}{s}$$

where μ is the population mean, $\bar{X}$ is the sample mean, σ is the population standard deviation, ***s*** is the sample standard deviation, and *x* is the raw score being converted.

For example, if the mean of a sample of I.Q. scores is 100 and the standard deviation is 15, then an I.Q. of 128 would correspond to:

$$z = \frac{x - \vec{X}}{s} = \frac{128 - 100}{15} = 1.87$$

For the same distribution, a score of 90 would correspond to:

$$z = \frac{90 - 100}{15} = -0.67$$

Chapter 6 Correlation

The degree of relationship between the variables under consideration is measured through the correlation analysis. The measure of correlation called **correlation coefficient** or correlation index summarizes in one figure the direction and degree of correlation. The correlation analysis refers to the techniques used in measuring the closeness of the relationship between the variables.

Correlation analysis:
- deals with the association between two or more variables
- attempts to determine the *''degree of relationship between variables''*
- is an analysis of the co variation between two or more variables

The problem of analyzing the relationship between different series should be broken into three steps:
1. Determining whether a **relation exists** and if it does, **measuring it.**
2. Testing whether **it is significant**.
3. Establishing the **cause and effect** relation, if any.

The study of correlation is of immense use in practical life because of the following reasons:
- Most of the variables show some kind of relationship.
- Once we know that two variables are closely related we can estimate the value of one variable given the value of another.
- Correlation analysis contributes to the understanding of economic behavior, aids in locating the critically important variables on which other depend.

The explanation of a significant degree of correlation may be one, or combination of the following reasons:
- The correlation may be due to pure chance, especially in a small sample.
- Both the correlated variables may be influenced by one or more other variables
- Both the variables may be mutually influencing each other so that neither can be designated as the cause and the other effect.

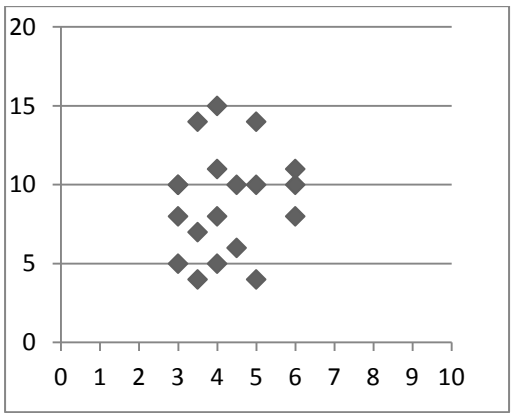
Correlation is described or classified in several different ways. Three of the most important ways of classifying correlation are:
1. Positive or negative
2. Simple, partial or multiple
3. Linear and non linear.

<u>Methods of studying correlation:</u>
- Scatter Diagram Method
- Graphic method
- **Karl Pearson's coefficient of correlation**

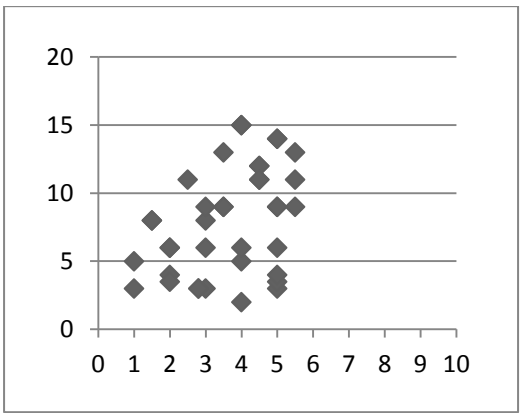
Scatter diagram

The simplest device for ascertain whether two variables are related is to prepare a dot chart called scatter diagram. When this method is used the given data are plotted on a graph paper or simply scatter plot in the form of dots, for each pair of X and Y values we put a dot and thus obtain as many points as the number of observations. By looking to the scatter of the various points we can form an idea as to whether the variables are related or not.

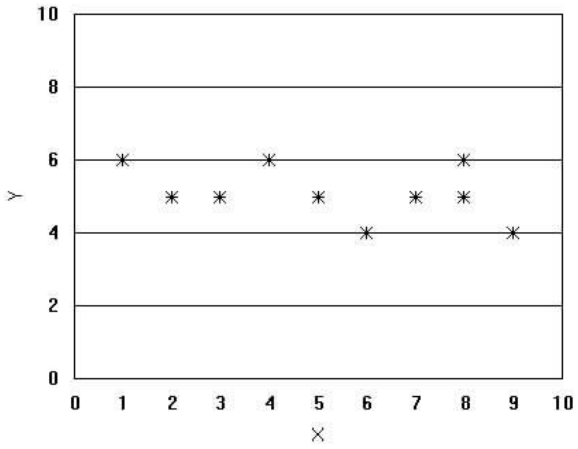
No correlation



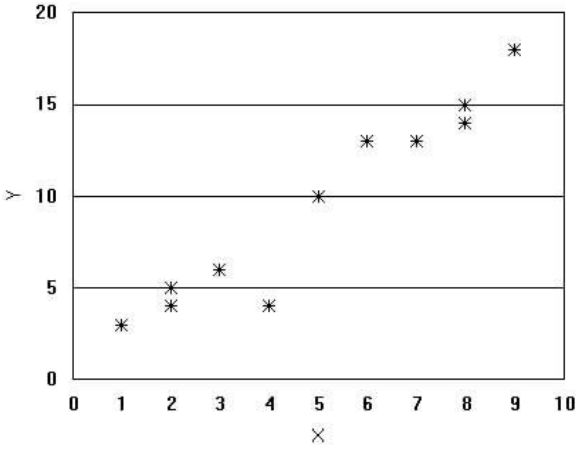
No correlation



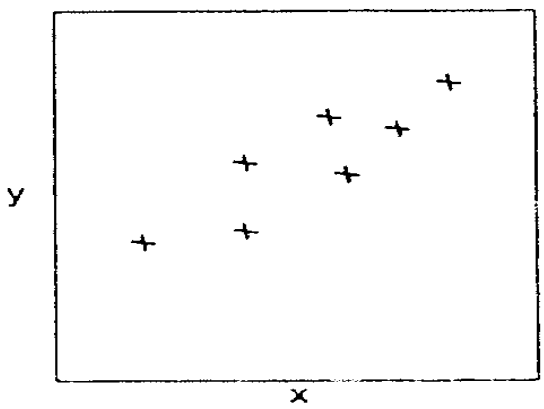
No correlation



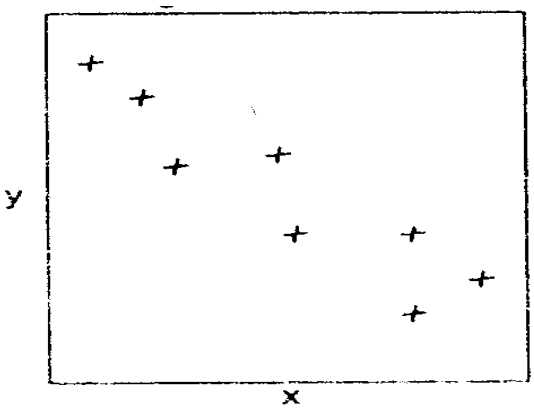
Moderate correlation



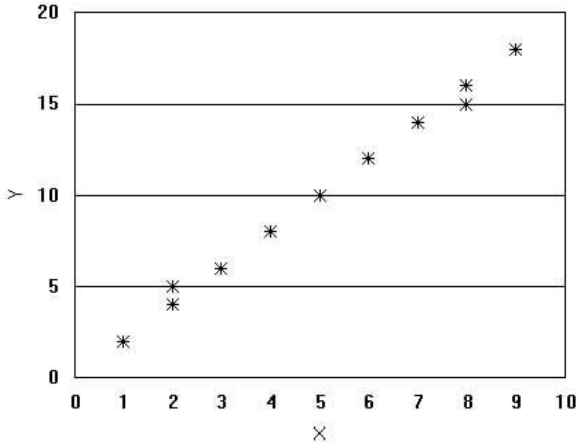
High degree of positive correlation



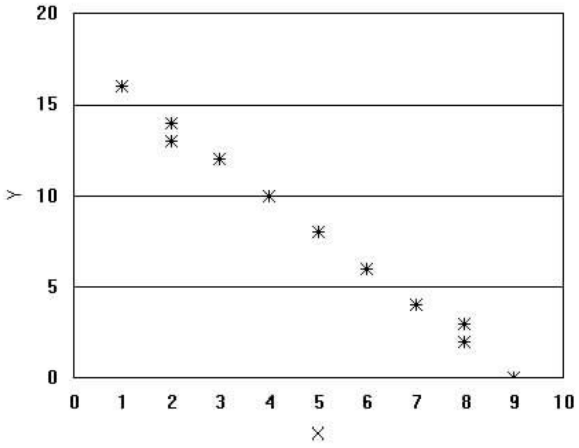
High degree of negative correlation



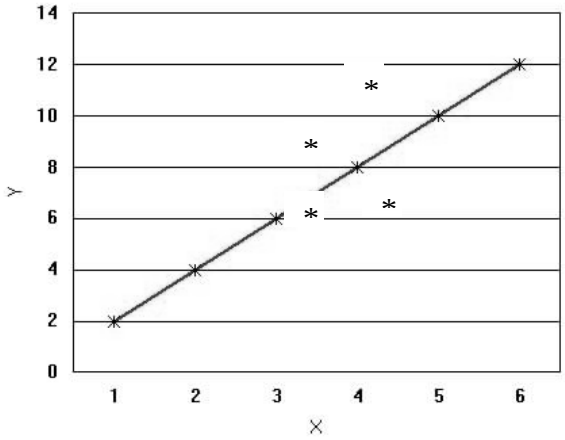
Positive, High Correlation close to +1



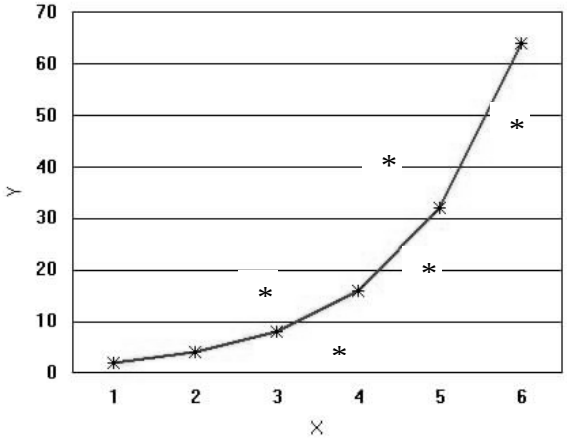
Negative, High Correlation close to -1



Linear Relationship



Non linear Relationship



6.1 Karl Pearson's coefficient of correlation

(product moment correlation coefficient)

Of the several mathematical methods of measuring correlation, the Karl Pearson's method known as Pearson's coefficient of correlation is most widely used in practice. Pearson's correlation coefficient between two variables is defined as the covariance of the two variables divided by the product of their standard deviations:

$$r = \frac{\sum_{i=1}^{n}(X_i - \vec{X})(Y_I - \overleftarrow{Y})}{\sqrt{\sum(X_i - \vec{X})^2} * \sqrt{\sum(Y_I - \overleftarrow{Y})^2}}$$

Covariance gives the idea of degree of co-variation of two variables, viz, the extent to which the two variables move together.

$$Cov_{xy} = \frac{1}{N}\sum_{i=1}^{n}(X_i - \vec{X})(Y_i - \vec{Y})$$

Merits and limitations

The correlation coefficient summarizes in one figure not only the degree of correlation but also the direction whether the correlation is positive or negative. The chief limitation of the method is:

- The correlation coefficient always assumes linear relationship regardless of the fact that wheatear the assumption is correct or not.
- Great care must be exercised in interpreting the value of this coefficient as very often the coefficient is misinterpreted.
- The value of the coefficient is unduly affected by the extreme items.
- As compared with other methods this method takes more time to normally compute the value of correlation coefficient.

Interpreting coefficient of correlation

The following general rules are given which would help in interpreting the value of ''r''':

- When ''r = + 1, it means there is perfect positive relationship between the variables.
- When ''r = -1, it means there is perfect negative relationship between the variables.
- When ''r = 0, it means there is no relationship between the variables, the variables are uncorrelated.

Properties of the coefficient of correlation

- The coefficient of correlation lies between -1 and +1. Symbolically **–1 ≤ r ≤ +1.**
- The coefficient of correlation is independent of change of scale and origin of the variables **X** and **Y** series, i.e. $\text{corr.}(x, y) = \text{corr}(x, y) \text{where } x = X - a;\ y = Y - b.$
- Also, $\text{corr.}(x, y) = \text{corr}(u, v) \text{where } u = \frac{X - a}{h};\ v = \frac{Y - b}{k}.$

6.2 Rank correlation coefficient

The Karl Pearson's method is based on the assumption that the population being studied is normally distributed. When it is known that the population is not normal or when the shape of the distribution is not known there is need to measure of correlation that involves no assumption about the parameter of population.

The method of finding out co variability or the lack of it between two variables was developed by the British Psychologist Charles Edward Spearmen in 1904. Spearman's rank correlation coefficient is defined as:

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

This measure is especially useful when quantitative measures for certain factors cannot be fixed, but the individuals in the group can be arranged in order thereby obtaining for each individual a number indicating his rank in the group.

Where ranks are given

- Take the differences of the two ranks , i.e (R1-R2) and denote these differences by ‘’d’’.
- Square these differences and obtain the totals. Sum(d^2).
- Apply the formula:

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

Where the ranks are not given

- When we are given the actual data and not the ranks, it will be necessary to assign the ranks. Ranks can be assigned by taking either the highest value as 1 or the lowest value as 1.
- But whether we start with the lowest or the highest value we must follow the same method in case of both the variables.
- Apply the formula:

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

Where ranks are equal

- In some cases it may be found necessary to rank two or more individuals or entries as equal. In such a cases it is customary to give each individual an average rank. Thus if two individuals are ranked equal at fifth place they are each given (5+6)/2, so 5.5.

Merits and limitations of the Rank Method

- This method is simpler to understand and easier to apply compared to the Karl Pearson’s method. The answer obtained by this method and the Karl Pearson’s method will be the same provided no value is repeated, i.e. all the items are different.
- Where the data are of a qualitative nature like honesty, efficiency, intelligence, etc, this method can be used with great advantage.

- This method cannot be used for finding out correlation in a grouped frequency distribution.
- Where the number of items exceeds 30 the calculations become quite tedious and require a lot of time. Therefore this method should not be applied where N exceed 30 unless we are given the ranks and not the actual values of the variables.

Chapter 7 Basic Concepts in Probability

The occurrences in the world can be divided with following three categories:

Totally undeterministic *(At random) haphazard*	**Partially deterministic** *(Random)*	**Totally deterministic** *(Certain)*
Nothing is known • When will happen? • What will happen? • How it will happen? • Very vague ideas.	**It is known that** • the outcome of a trial will be one of the several possible outcomes. • there may also be knowledge about past frequencies of occurrences of different possible outcomes. • whether the outcomes are equally likely or not.	**No problem**! (not a subject matter for Statistician)
Example When will moon collapse in earth? When will the oceans evaporate as result of heating of the planet?	Example Outcome after tossing a coin. Outcome after rolling a dice. **Experiment: Roll two dice**. *Event 1:* Sum of number of two dice is even. *Event 2:* (6,,6)	Example Heat the water, it will boil at 100 degree Centigrade. If you cut your finger the blood comes out. If you jump through a window you break bones.

The word ‘’**probability**’’ is commonly used in day to day conversation but generally people have only a vague idea about its meaning. Many common people associate probability and chance with nebulous and mystic ideas. Probability is a basically a *“measure of uncertainty”* about happening of an event in which we are interested. It expressed on 0 to 1 scale (inclusive bounds).

The probability of a given event is an expression of likelihood or chance of occurrence of an event. Probability is a number which ranges from 0 (zero) to 1 (one) – zero for en event which cannot occur (impossible event) and 1 for an event certain (sure) to occur. How the number is assigned would depend on the interpretation of the term ‘’probability’’. The most precise definition of Probability is based on Measure Theory which deals with concepts related to sets and functions.

7.1 Basic concepts on sets

Set – is a collection of well (precisely) defined objects (elements) listed without duplications. Example, set of students of class ’’XB’’ of ’ABC’’ school. The students are the elements of the set, and their collection is a set.

Universal set. Every set is supposed to be a subset of a very big set, which is called the universal set. For example, the set N= {1,2,3,4 …} may be taken as the universal set for every set of counting number. A universal set is generally denoted by ‘’U’’ or ‘’S’’. In sampling theory, it will represent the Population (all the units under consideration). It is basically a reference frame which contains everything (element) under consideration.

Null set or Void set or Empty Set – A set with no elements in it. It is is denoted by ∅.

Sub set – If each element of a set A is a element of set B, the set A is called sub set of B, and is written as $A \subset B$. Every set is a subset of itself and ∅ is a subset of very set.

Equal set – If $A \subset B$ and $B \subset A$, then A and B are said to be equal.

Venn diagrams. A set can be represented by an arbitrary plane figure and its relationship with other set(s) can be explained by the combination of two or more such a figures. All elements of a set can be considered to be the points contained within the boundary of the figure.

7.2 Basic operations on sets

Union of sets. The set of elements which belong to at least one of the sets A and B is called union of A and B and is written as A U B.

Intersection of Sets. The set of all elements, which are common to both the sets A and B is called the intersection of A and B, and is written as $A \cap B$.

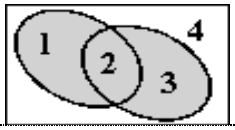
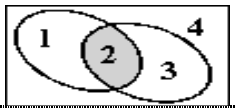
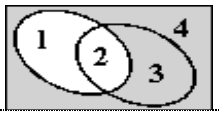
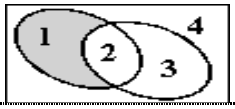
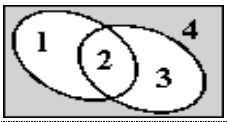
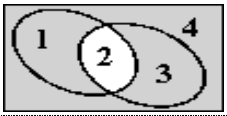
Disjoint sets. Two sets A and B are called disjoint sets if there is no element in common between them, i.e., their intersection is an empty set., i.e., $A \cap B$

Differences of two sets. Let there be two sets A and B. The set of all elements of A which are not in B gives (A-B). Similarly, we can define (B – A).

Complements of a set. Let U denote the universal set. The set which consist of all elements of U which are not element of A is called complements of A and is denoted A^c or $\overline{A}$.

Measure of a set is defined as the number of elements in it. If $A = \{1, 2, 3\}$ then n(A)= measure of A=3. Let's say that our universe contains the numbers 1, 2, 3, and 4. Let *A* be the set containing the numbers 1 and 2; that is, *A* = {1, 2}. (Warning: The curly braces are the customary notation for sets. Avoid using parentheses or square brackets.) Let *B* be the set containing the numbers 2 and 3; that is, *B* = {2, 3}.

Then we have the following relationships, depicted with shaded marking of the solution "regions" in the Venn diagrams:

set notation	pronunciation	meaning	Venn diagram	answer
$A \cup B$	"A union B"	everything that is in either of the sets		{1, 2, 3}
$A \wedge B$ or $A \cap B$	"A intersect B"	only the things that are in both of the sets		{2}
A^c or $\sim A$	"A complement", or "not A"	everything in the universe outside of A		{3, 4}
$A - B$	"A minus B", or "A complement B"	everything in A except for anything in its overlap with B		{1}
$\sim(A \cup B)$	"not (A union B)"	everything outside A and B		{4}
$\sim(A \wedge B)$ or $\sim(A \cap B)$	"not (A intersect B)"	everything outside of the overlap of A and B		{1, 3, 4}

Some Important Results

For any three sets A, B, C, the following results hold:

- $A \cup A = A$
- $A \cap A = A$
- $A \cup \emptyset = A$
- $A \cap \emptyset = \emptyset$
- $A \cup B = B \cup A$
- $A \cap B = B \cap A$
- $A \cup (B \cup C) = (A \cup B) \cup C$
- $A \cap (B \cap C) = (A \cap B) \cap C$
- $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
- $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$
- $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
- $(A \cap B) \cup C = (A \cup B) \cup C$
- $(A - B) = (A \cap B^c)$

7.3 Concepts used in calculation of probability

Random Experiment and Events

The term experiment refers to an act which can be repeated under some given conditions. *Random experiments* are those experiments whose results depend on chance, such as tossing a coin or throwing a dice. The results of random experiments are called *outcomes*. If in an experiment the set of all the possible outcomes is known with certainty but which of these outcomes will occur on a single trial of the experiment is not known, then such an experiment is called a *random experiment* and the outcomes as events or chance events.

Sample Space. The set of all possible outcomes of random experiment is called "'Sample Space" or "Universe."

An **Event** is basically a subsets of sample space. We can define variables based on events. The events are generally denoted by capital letter A, B, C ..etc

Example:
Random Experiment: Roll two dice or roll a die twice.
Sample Space
(1,1) (1,2) (1,3) (1,4) (1,5) (1,6)
(2,1) (2,2) (2,3) (2,4) (2,5) (2,6)
(3,1) (3,2) (3,3) (3,4) (3,5) (3,6)
(4,1) (4,2) (4,3) (4,4) (4,5) (4,6)
(5,1) (5,2) (5,3) (5,4) (5,5) (5,6)
(6,1) (6,2) (6,3) (6,4) (6,5) (6,6)

Variable defined on the sample space and events

X: Sum of numbers on two dice € {2, 3,4,5,6,7,8,9,10,11,12}
Y: Number on first die {1,2,3,4,5,6}

Correspondence between events and variables

X=2 correspond to (1,1)
X=12 correspond to (6,6)
X=7 corresponds to {(1,6) (2,5) (3,4) (4,3) (5,2) (6,1)}

Mutually Exclusive Events

Two events are said to be mutually exclusive or incompatible when both events cannot happen simultaneously in a single trial or, in other words the occurrence of any one of them precludes the occurrence of the other. For example if a single coins is tossed either head can be up or tail can be up; both cannot be up at the same time.

Two events are mutually exclusive if they cannot occur at the same time (i.e., they have no outcomes in common). Symbolically, if A and B are mutually exclusive events, P(A∩B)= 0

Independent and Dependent Events
Two or more events are said to be *independen*t when the outcome of one does not affect and is not affected by the other. For example, if a coin is tossed twice, the result of the second throw would in no way be affected by the results of the first throw. Similarly, the result obtained in a throw of a die are independent of the results obtained from drawal of card from a pack.

Dependent events are those for which the occurrence or non-occurrence of one event in any trial affects the chances of occurrence of other events in subsequent trials.

Equally Likely Events
Events are said to be equally likely when one cannot be expected in preference to other. In practice, over a large number of trials, any one event does not occur more often than the others. For example, if an unbiased coin is thrown, each face may be expected to be observed approximately the same number of times in the long run.

Simple and Compound Events
In case of simple events we consider the probability of the happening or not happening of single events. For example, we might be interested in finding out the probability of drawing a red ball from a bag containing 10 white and 6 red balls. On the other hand, in case of compound events we consider the joint occurrence of two or more events.

Exhaustive Events
Events are said to be exhaustive when their totality includes all the possible outcomes of a random experiment, i.e. the Universal set or sample space. For example, while throwing a dice, the possible outcomes are 1, 2, 3, 4, 5, 6 and hence the exhaustive number of cases is 6. If two dices are thrown once, the all possible outcomes are:

(1,1) (1,2) (1,3) (1,4) (1,5) (1,6)
(2,1) (2,2) (2,3) (2,4) (2,5) (2,6)
(3,1) (3,2) (3,3) (3,4) (3,5) (3,6)
(4,1) (4,2) (4,3) (4,4) (4,5) (4,6)
(5,1) (5,2) (5,3) (5,4) (5,5) (5,6)
(6,1) (6,2) (6,3) (6,4) (6,5) (6,6)
Number of exhaustive cases here are 36.

Complementary Events
Let there be two events A and B. A is called the complementary event of B (and vice versa) if A and B are mutually exclusive and exhaustive. For example, when a dice is thrown, occurrence of an even number (2, 4, 6) and odd number (1, 3, 5) are complementary events.

7.4 Probability

Classical or *a priori* probability
This concept of probability originated with games of chances and deals with calculating probability of a particular outcome out of a given a set of equally likely outcomes. Probability is defined as the ratio of the number of favourable cases to the number of equally likely cases.

Probability of occurrence of event A, denoted by P(A) is defined as:

$$P(A)= \frac{\text{Number of favourable cases}}{\text{Total number of equally likely cases}} = \frac{n(A)}{n(U)}$$

Probability of non-occurrence of an event is denoted by $P(A^c)$

$$P(A^c)= \frac{\text{No.of unfavourable cases}}{\text{Total number of equally likely cases}} = 1- P(A)$$

Example
From a bag containing 10 black and 20 white balls, a ball is drawn at random. What is the probability that is black?

We are interested in A: drawn ball is black
Total number of balls in the bag is 10+20 = 30.
Numbers of black balls is = 10

Probability of getting a black balls,

$$P(A)=\frac{\text{Number of favourable cases for A to happen}}{\text{Total number of equally likely cases}} = \frac{10}{30} = \frac{1}{3}$$

Probability of not getting a black ball ,
$P(A^c) = \frac{20}{30} = \frac{2}{3}$; Thus, $P(A) + P(A^c) = \frac{1}{3} + \frac{2}{3} = 1$

Dependence of Events
Event A is called dependent on event B, if chances of occurrence of A is dependent on prior occurrence or non-occurrence of B.

Example 1
An urn contains 3 white and 2 black balls. The second urn contains 5 white and 4 Black balls. One ball is transferred from the first urn to the second urn, and then a ball is drawn from the second urn. We are interested in the probability of getting a while ball from the draw from second urn. The chances of this will depend upon the color of the ball transferred from the first urn to the second urn.

Example 2
When two students solve a problem sum separately, the chances of each one solving the sum does not depend upon the results of the other one. This an example of independent events.

Conditional Probability
The conditional probability refers to evaluation of probability of occurrence of an event (say A), in the light of information that another event B has already happened. This is represented as

P (A/B), read as "Probability of A given B"
= P (AB)/ P(B), if A depend upon B

P (A/B),
= P(A)* P(B)/ P(B), if A and B are independent: since P(AB) = P(A)*P(B),
= P (A),

Addition Theorem
The addition theorem states that if two or events A and B are mutually exclusive, the probability of the occurrence of either A or B is the sum of their individual probabilities, provided the events cannot occur together.

Symbolically,
$\mathrm{P(A\ or\ B)} = \mathrm{P(A \cup B)} = \mathrm{P(A)} + \mathrm{P(B)}$, for disjoint events

The theorem can be extended to three or more mutually exclusive events, thus:
$\mathrm{P(A\ or\ B\ or\ C)} = \mathrm{P(A \cup B \cup C)} = \mathrm{P(A)} + \mathrm{P(B)} + \mathrm{P(C)}$

When events are not mutually exclusive:
$\mathrm{P(A\ or\ B)} = \mathrm{P(A \cup B)} = \mathrm{P(A)} + \mathrm{P(B)} - \mathrm{P(A \cap B)}$

In case of tree events
$\mathrm{P(A\ or\ B\ or\ C)} = \mathrm{P(A)} + \mathrm{P(B)} + \mathrm{P(C)} - \mathrm{P(A \cap B)} - \mathrm{P(A \cap C)} - \mathrm{P(B \cap C)} + \mathrm{P(A \cap B \cap C)}$

The reader is suggested to think of conceptual proof of these results based on Venn diagram for 3 sets A, B, C**.**

Multiplication Theorem
This theorem states that if two events A and B are independent the probability that they both will occur is equal to the product of their individual probability. Symbolically, if A and B are independent, then:
$P\mathrm{(A\ and\ B)} = \mathrm{P(A \cap B)} = \mathrm{P(A)} * \mathrm{P(B)}$

The theorem can be extended to three or more independent events:
$\mathrm{P(A, B\ and\ C)} = \mathrm{P(A \cap B \cap C)} = \mathrm{P(A)} * \mathrm{P(B)} * \mathrm{P(C)}$

Example 1

Two genetically identically plants of olive are imported in a country. These are planted in two different agro-climatic zones A & B. Past empirical studies estimates the chances of these plants flourishing and yielding fruits in zones A and B as $\frac{1}{2}$; and $\frac{1}{3}$ respectively.

a) What is the probability that the plants will flourish to fructify in at least one zone?
b) What is the probability that the plants will flourish to fructify in zone A but not in zone B?
c) What is the probability that neither plants will flourish to fructify?
d) What is the probability that the plants will flourish to fructify in only one zone?
e) What is the probability that the plants will flourish to fructify in both A and B zones?

Solutions

Let us define the events as:

A = the plants will fructify in zone A

B = the plants will fructify in zone B

We are given, P(A)=1/2 and P(B)=1/3

a) The probability that the plants will flourish to fructify in at least one zone

$= P(A \cup B)$

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

$= P(A) + P(B) - P(A) * P(B)$

$= \frac{1}{2} + \frac{1}{3} - \left(\frac{1}{2} * \frac{1}{3}\right)$

$= \frac{3+2}{6} - \frac{1}{6}$

$= \frac{5}{6} - \frac{1}{6}$

$= \frac{4}{6} = \frac{2}{3}$

b) The probability that the plants will flourish to fructify in zone A but not in zone B

$= P(A \cap \overrightarrow{B})$

$= P(A) * P(\overrightarrow{B})$

$= P(A) * [1 - P(B)]$

$= \frac{1}{2} * \left[1 - \frac{1}{3}\right]$

$= \frac{1}{2} * \left[\frac{3-1}{3}\right]$

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

c) The probability that neither plants will flourish to fructify.

$= P(\vec{A} \cap \overrightarrow{B})$

$= P(\vec{A}) * P(\overrightarrow{B})$

$= [1 - P(A)] * [1 - P(B)]$

$= \left[1 - \frac{1}{2}\right] * \left[1 - \frac{1}{3}\right]$

$= \left[\frac{2-1}{2}\right]\left[\frac{3-1}{3}\right]$

$= \frac{1}{2} * \frac{2}{3} = \frac{2}{6} = \frac{1}{3}$

d) The probability that the plants will flourish to fructify only in one zone

$= P(A \cap \vec{B}) \cup (B \cap \vec{A})$

$= P(A) * P(\vec{B}) + P(B) * P(\vec{A})$

$= P(A)[1 - P(B)] + P(B) * [1 - P(A)]$

$= \frac{1}{2}\left[1 - \frac{1}{3}\right] + \frac{1}{3}\left[1 - \frac{1}{2}\right]$

$= \left(\frac{1}{2} * \frac{1}{3}\right) + \left(\frac{1}{3} * \frac{1}{2}\right)$

$= \frac{1}{3} + \frac{1}{6} = \frac{2}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$

e) The probability that the plants will flourish to fructify in both A and B zones

$= P(A \cap B)$

$= P(A) * P(B)$

$= \frac{1}{2} * \frac{1}{3} = \frac{1}{6}$

Example 2

A bag contain 5 white and 3 black balls. Two balls are drawn at random one after another without replacement. Find the probability that both balls drawn are black.

Solution

Total number of balls is 5+3 = 8.

No of way of drawing 2 balls

$C_2^8 = \frac{8 * 7}{2 * 1} = \frac{56}{2} = 28$ ways

No of ways of drawing 2 black balls out of 3 black balls

$C_2^3 = \frac{3 * 2}{2 * 1} = \frac{6}{2} = 3$ ways

$\text{Probability} = \frac{\text{No of favourable ways}}{\text{Total number of ways}} = \frac{3}{28} = 0.1071428$

Example 3
Find the probabilities of drawing a queen, a king, and a knave in that order from a pack of cards in three consecutive draws, the cards drawn are not being replaced.

Solution
Let us defined the events as:
$A = \text{a card drawn out of } 52 \text{ is a king}$
$B = \text{a card drawn out of } 51 \text{ is a queen}$
$C = \text{a card drawn out of } 50 \text{ is a knave}$
P={drawing a king, queen, knave in that order}

$=P\{A \cap (B/A) \cap [C/(A \cap B)]\}$
$=P(A) * P(B/A) * P(C/A \cap B)$

$$=\frac{4}{52} * \frac{4}{51} * \frac{4}{50}$$

$$= \frac{64}{132600}$$

Exercise 1
A new trace of pathogenic plant virus has been identified. Laboratory trials have indicated following empirical probabilities on results of applications of new pesticide formulation A and B for controlling the virus.

$$P\{\text{plant disease caused by virus is cured after application of formulation A}\} = \frac{3}{4}$$

$$P\{\text{plant disease caused by virus is cured after application of formulation B}\} = \frac{1}{2}$$

A farmer witnessing virus attach on his crop, not willing to take chances, makes application of both pesticides A and B together . Assuming that the two pesticides act independently, find the probability that his crop will be completely cured (virus eliminated) after both applications of pesticides.

Exercise 2
A seed merchant has two 100 Kg of local seeds with 550 Kg seed of high yielding varieties (HYV). It is know that HYV is on an average 10% heavier than the local seed. You take a random sample of from the seed mixture. What is the ratio of number of seeds of two types you will expect in the sample?

References

Text Books

- Robert V.Hogg, Allen T Craig, Sixth Edition, Introduction to Mathematical Statistics
- J.E. Freund, Seventh Edition, Mathematical Statistics with Applications.
- S.P.Gupta, Thirty – seventh Revised Edition, 2008 Statistical Methods

Websites

- Reading and understanding formula

http://www.emathzone.com/tutorials/basic-algebra/algebraic-expression.html
http://www.abacustraining.biz/bodmasExercises.htm

- Graphical Presentation of Data,

http://www.childrensmercy.org/stats/category/Definitions.aspx
http://stattrek.com/statistics/charts/cumulative-plot.aspx?tutorial=ap

- Basic statistical concepts

http://www.wikihow.com/Calculate-Spearman's-Rank-Correlation-Coefficient
http://www.une.edu.au/WebStat/unit_materials/c4_descriptive_statistics/least_squares_regress.html
http://www.investopedia.com/terms/m/mean.asp#ixzz1pYdDla8N
http://simon.cs.vt.edu/SoSci/converted/Dispersion_I/
http://www.investopedia.com/terms/s/standarddeviation.asp#ixzz1pYdnSShX