# Numerical Solution Of Ordinary Differential Equations

Computer Oriented Numerical and Statistical Methods

## Outline

- Introduction
- Euler method
- Runge – Kutta (RK) method

## Introduction

- The subject of ordinary differential equations is not only fascinating part of mathematics but also an essential tool for modeling many physical processes.
- Most scientific laws are expressed in terms of differential equations.
  - Thermodynamics $dT/dt = -0.27(u-60)^{5/4}$
  - Probability $dPr /dt = (r+1)(n-r)P_{r+1} – r(n-r+1)Pr$
  - Mechanics $mdv/dt = mf – kv^2$
  - Economics $dx/dt = Sf(x) – g(x)$.

## Solving A Differential Equations

- Formulation of differential equation is simple but difficult to solve it.
- Use numerical solution i.e. instead of finding an algebraic (analytical) solution we compute (approximately) the numerical values taken by solution.
- Therefore, as a solution of differential equations, instead of finishing up with an expression.
- This is known as the numerical solution of a differential equations.

## Basic Terminology Of Differential Equations

- Differential Equation : A differential equation is an equation containing  an unknown function and its derivatives. $\frac{d^2y}{dx^2}+3\frac{dy}{dx}+ay=0$

  ***y*** is dependent variable and ***x*** is independent variable,

  and this is an ordinary differential equations [i.e. involves only one independent]
- Ordinary Derivative : If y is a function of x i.e. y = f(x) , then dy/dx is called the ordinary derivative. Physically it means the rate of change of the dependent variable with respect to the independent variable.

## Basic Terminology Of Differential Equations

- Partial Derivative : If u is a function of x and y i.e. u = f(x,y) then $\left.\frac{\partial u}{\partial x}\right|_{y}$ is called the partial derivative with respect to x keeping y constant, and $\left.\frac{\partial u}{\partial y}\right|_{x}$ is called the partial derivative with respect to y keeping x constant. Physically it means the rates of change of dependent variable with respect to one of the independent variable keeping others fixed.

## Basic Terminology Of Differential Equations

- Ordinary Differential Equations : An ordinary differential equation is an equation involving only ordinary derivative of one or more function with respect to a single independent variables.

Examples:.

1. $$\frac{dy}{dx} = 2x + 3$$
2. $$\frac{d^2y}{dx^2} + 3\frac{dy}{dx} + ay = 0$$
3. $$\frac{d^3y}{dx^3} + \left(\frac{dy}{dx}\right)^4 + 6y = 3$$

## Basic Terminology Of Differential Equations

- Partial Differential Equations : A partial differential equation is an equation involving a single independent variables .
- Examples:

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \qquad \frac{\partial^4 u}{\partial x^4} + \frac{\partial^4 u}{\partial t^4} = 0 \qquad \frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 u}{\partial t^2} - \frac{\partial u}{\partial t}$$

## Order of  Differential Equation

**The order of the differential equation is order of the highest derivative in the differential equation.**

| **Differential Equation** | **ORDER** |
|---|---|
| $\frac{dy}{dx} = 2x + 3$ | 1 |
| $\frac{d^2y}{dx^2} + 3\frac{dy}{dx} + 9y = 0$ | 2 |
| $\frac{d^3y}{dx^3} + \left(\frac{dy}{dx}\right)^4 + 6y = 3$ | 3 |

## Degree of Differential Equation

**The degree of a differential equation is power of the highest order derivative term in the differential equation.**

| **Differential Equation** | **Degree** |
|---|---|
| $\frac{d^2y}{dx^2} + 3\frac{dy}{dx} + ay = 0$ | 1 |
| $\frac{d^3y}{dx^3} + \left(\frac{dy}{dx}\right)^4 + 6y = 3$ | 1 |
| $\left(\frac{d^2y}{dx^2}\right)^3 + \left(\frac{dy}{dx}\right)^5 + 3 = 0$ | 3 |

## Basic Terminology Of Differential Equations

- Solution of A Differential Equations : Consider the first order ordinary differential equation of type $\frac{dy}{dx} = f(x, y)$

  which can also be written as y’=f(x,y) where the function f(x,y) may be a general non-linear function of x and y or in the form of a table of values.
- The solution of such an ordinary differential equation is a 2-D curve of (x,y) in the xy plane whose slope at every point (x,y) is the specified region is given by the equation $\frac{dy}{dx} = f(x, y)$

## Basic Terminology Of Differential Equations

- Initial value problem : if the parameters of ordinary differential equation is determined based on some given initial values, i.e. initial condition then this system is known as an initial value problem.

# Linear Differential Equation

A differential equation is linear, if

1. dependent variable and its derivatives are of degree one,
2. coefficients of a term does not depend upon dependent variable.

**Example:** 1. $\frac{d^2y}{dx^2}+3\frac{dy}{dx}+9y=0.$

**is linear.**

**Example:** 2. $\frac{d^3y}{dx^3}+\boxed{\left(\frac{dy}{dx}\right)^4}+6y=3$

**is non - linear because in 2nd term is not of degree one.**

# Linear Differential Equation

**Example:** 3. $x^2\frac{d^2y}{dx^2}+\boxed{y\frac{dy}{dx}}=x^3$

**is non - linear because in 2nd term coefficient depends on *y*.**

**Example:** 4. $\frac{dy}{dx}=\boxed{\sin y}$

**is non - linear because** $\sin y = y-\frac{y^3}{3!}+-$ is non – linear

## Numerical Solution Of Differential Equations

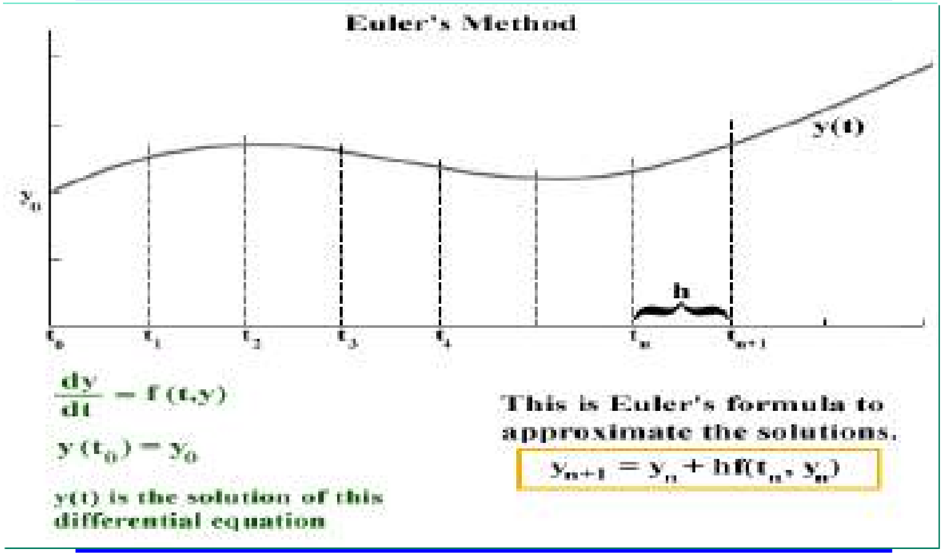
- To describe various numerical methods for the solution of ordinary differential equation we consider the general first order differential equations $\frac{dy}{dx} = f(x,y)$

  with the initial condition $y(x_0) = y_0$
- Numerical solution of differential are classified into two types.
  - A series of y in terms of power of x, from which the value of y can be obtained by direct substitution. These methods are : Taylor series, Picard's Method.
  - A set of tabulated values of x and y. the method are : Euler's method, Runge – Kutta method, Adam- Bashforth method.

## Euler's Method

- The Euler's method is one of the oldest and the simplest method.
- It can be described as a techniques of developing a piecewise linear  approximation to the solution
- In the initial value problem, the starting point of the solution curve and the slope of the curve at the starting point are given.

## Euler’s Method



$$\frac{dy}{dt} = f(t,y)$$

$$y(t_0) = y_0$$

y(t) is the solution of this differential equation

This is Euler's formula to approximate the solutions.

$$\boxed{y_{n+1} = y_n + hf(t_n, y_n)}$$

## Euler’s Method

- Consider again the following first order ordinary differential equation $\frac{dy}{dx} = f(x,y)$ with initial conditions y = $y_0$ for x = $x_0$ and h is a positive increment of x. $x_1$=$x_0$+h
- Divide *l* –$x_0$ into n equal parts. Length of each part is equal to h. So $x_1$=$x_0$+h, $x_2$=$x_1$+h, …..
- The mean value theorem

$$y'(c) = \frac{y(x_1) - y(x_0)}{x_1 - x_0}$$

- If we substitute c = $x_0$ and h = $x_1$ - $x_0$ in the above equation can be written as
  $y(x_1) – y(x_0) = hy’(x_0)$

## Euler's Method

- Now $\frac{dy}{dx} = f(x,y)$
- $\therefore$ y'($x_0$) = f($x_0$,$y_0$)
  y($x_1$) – y($x_0$) = h f($x_0$,$y_0$)
  y($x_1$) = y($x_0$) + h f($x_0$,$y_0$)
  $y_1$ = $y_0$ + h f($x_0$,$y_0$) [because y($x_0$) = $y_0$]
- Using this equation, we can find the second point on the solution curve as ($x_1$,$y_1$)
- Similarly, taking ($x_1$,$y_1$) as the starting point, we get
  $y_2$ = $y_1$ + h f($x_1$,$y_1$)

## Euler's Method

- In general, the (i+1)th point of the solution curve is obtained from the ith point using the following formula.
  **$y_{i+1}$ = $y_i$ + h f($x_i$,$y_i$)** which is Euler's method
- *[The process to find the solution using this method is too slow, and to obtain the reasonable accuracy we must take a very small value of the h.]*

## Runge - Kutta (RK) Method

- The basic objectives of R-K methods are as follows:

1. The method propagate a solution over an interval by combining the information from several Euler-style steps. Here each step is evaluating the function f with different parameters.
2. Using this information obtained to match a Taylor series expansion up to some higher orders.

- Euler's method is less efficient in practical problems since it requires h to be small for obtaining reasonably accuracy.

## Runge - Kutta (RK) Method

- The R-K methods are designed to give greater accuracy and they possess the advantage of requiring only the function values at some selected points on the subintervals.
- There are different Runge – Kutta formulae of various orders and methods:
  - R- K second order method
  - R-K fourth order method

## Runge – Kutta Second Method (R-K 2nd order method)

- The R-K second order methods are actually a family of methods, each of that matches the Taylor series method up to the second terms in h, where h is the step size.
- In these methods the interval $[x_1, x_f]$ is divided into subintervals and a weighted average of derivatives (slopes) at these intervals is used to determine the value of the dependent variable.
- One advantage of these methods is that they, like to evaluate $y_{i+1}$ we need information only at the preceding point $(x_i, y_i)$

## Runge – Kutta Second Method (R-K 2nd order method)

- The second order method can be expressed as follows
- $y_1 = y_0 + h/2(f(x_0,y_0) + f(x_1,y_1))$
- Substitute $y_1 = y_0 + hf(x_0,y_0)$ in the above equation
- $y_1 = y_0 + h/2[f(x_0,y_0) + f(x_1, y_0 + hf(x_0,y_0))]$
- $y_1 = y_0 + h/2[f_0 + f(x_0 + h, y_0 + hf_0)]$ where $f_0 = f(x_0, y_0)$
- We can write $k_1 = hf_0$ and

$$k_2 = hf(x_0 + h, y_0 + k_1)$$

$$y_1 = y_0 + ½(k_1 + k_2)$$

## Runge – Kutta Second Method (R-K 2nd order method)

- In similar we can find $y_2$, $y_3$, …$y_{n+1}$
- $y_{n+1} = y_n + ½(k_1+k_2)$

  where $k_1 = hf(x_n,y_n)$ and

  $k_2 = hf(x_n +h, y_n+k_1)$
- Which is R-K 2nd order formula

## Runge – Kutta Fourth Method (R-K 4th Order Method)

- The fourth order method can be expressed as follows
- $y_1 = y_0 + 1/6(k_1 + 2k_2 + 2k_3 + k_4)$
- Where $k_1 = hf(x_{0,} y_0)$

  $k_2 = hf(x_0 +h/2, y_0+k_1/2)$

  $k_3 = hf(x_0 +h/2, y_0+k_2/2)$

  $k_4 = hf(x_0 +h, y_0+k_3)$
- In similar we can find $y_2$, $y_3$, …$y_{n+1}$

## Runge – Kutta Fourth Method (R-K 4th Order Method)

- $y_{n+1} = y_n + 1/6(k_1+2k_2 +2k_3+k_4)$
- Where $k_1 = hf(x_{n,} y_n)$

  $k_2 = hf(x_n +h/2, y_n+k_1/2)$

  $k_3 = hf(x_n +h/2, y_n+k_2/2)$

  $k_4 = hf(x_n +h, y_n+k_3)$
- Which is R-K 4th order formula

# Chi-Square Tests

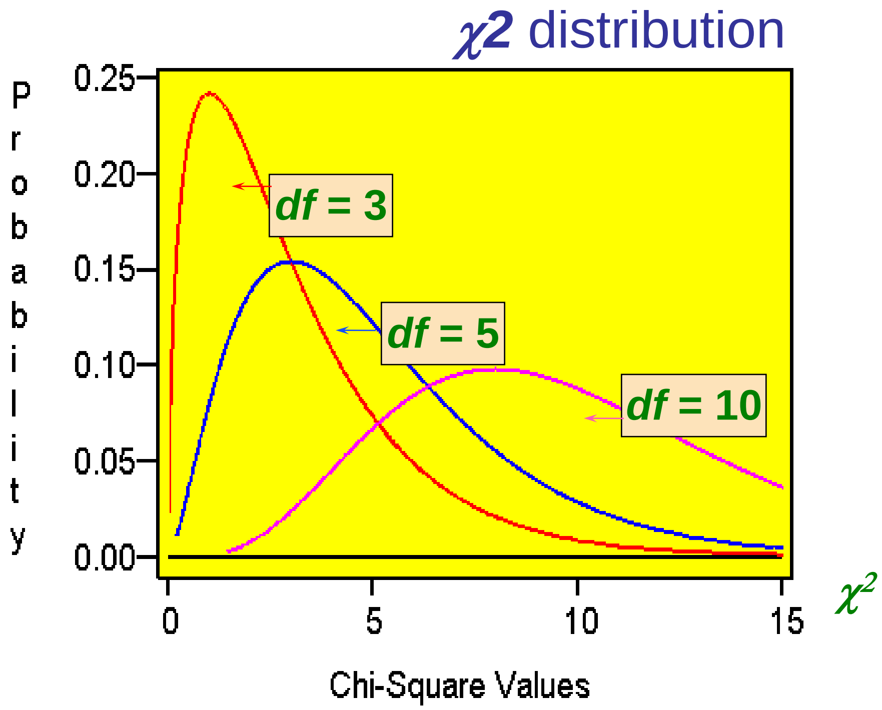
## Chi-Square Test of Independence

- Chi-square test enable us to test whether more than two population proportions can be considered equal.
- Chi-square test allows us to do a lot more than just test for the equality of several proportions. If we classify a population into several categories with respect to two attributes (such as age and job performance), we can then use a chi-square test to determine whether the two attributes are independent of each other.
- The row and columns of a chi-square contingency table must be mutually exclusive categories that exhaust all of the possibilities of the sample.

# Chi-Square Test of Independence

- Hypothesis:
  - $H_0$ : All proportions are equal
  - $H_1$ : At least two proportions are not equal

- The major characteristics of the chi-square distribution are:
  - It is positively skewed
  - It is non-negative
  - There is a family of chi-square distributions



## Procedure of Chi-Square Test

- Describe a contingency table
- Setting up the problem symbolically
- Determining expected frequencies
- Comparing expected and observed frequencies
- Reasoning intuitively about chi-square tests
- Calculating the chi-square statistics
- Interpreting the chi-square statistics

## Contingency Table Example

- Used to classify sample observations according to two or more characteristics
- Also called a cross-classification table.

Left-Handed vs. Gender

Dominant Hand: Left vs. Right

Gender: Male vs. Female

- 2 categories for each variable, so called a 2 x 2 table
- Suppose we examine a sample of size 300

## Contingency Table Example

Sample results organized in a contingency table:

sample size = n = 300:

120 Females, 12 were left handed

180 Males, 24 were left handed

| Gender | Hand Preference | | |
|---|---|---|---|
| | Left | Right | |
| Female | 12 | 108 | 120 |
| Male | 24 | 156 | 180 |
| | 36 | 264 | 300 |

## $\chi^2$ Test for the Difference Between Two Proportions

$H_0$: $\pi_1 = \pi_2$ (Proportion of females who are left handed is equal to the proportion of males who are left handed)

$H_1$: $\pi_1 \neq \pi_2$ (The two proportions are not the same – Hand preference is not independent of gender)

- If $H_0$ is true, then the proportion of left-handed females should be the same as the proportion of left-handed males
- The two proportions above should be the same as the proportion of left-handed people overall

# The Chi-Square Test Statistic

The Chi-square test statistic is:

$$\boxed{\chi^2 = \sum_{\text{all cells}} \frac{(f_o - f_e)^2}{f_e}}$$

- where:

  $f_o$ = observed frequency in a particular cell

  $f_e$ = expected frequency in a particular cell if $H_0$ is true

  $\chi^2$ for the 2 x 2 case has 1 degree of freedom

(Assumed: each cell in the contingency table has expected frequency of at least 5)

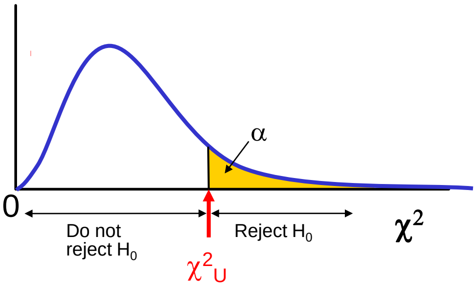
# The Chi-Square Test Statistic

- To use the chi-square test, we must calculate the number of degrees of freedom in the contingency table by applying
  - **Number of degree of freedom = (number of rows -1)*(number of columns -1)**
- A chi-square of zero, on the other hand, indicates that the observed frequencies exactly match the expected frequencies.
- The value of chi-square can never be negative because the differences between the observed and expected frequencies are always squared.

## Decision Rule

The $\chi^2$ test statistic approximately follows a chi-squared distribution with one degree of freedom

Decision Rule:
If $\chi^2 > \chi^2_U$, reject $H_0$, otherwise, do not reject $H_0$



## Example

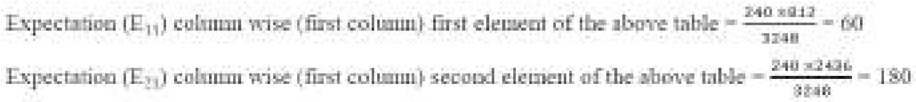
- In an antimalarial campaign in certain area quinine was administered to 812 persons out of total population of 3248. The number of fever cases is shown below:

| Treatment | Fever | No Fever | Total |
|---|---|---|---|
| Quinine | 20 | 792 | 812 |
| No Quinine | 220 | 2216 | 2436 |
| Total | 240 | 3008 | 3248 |

- Discuss the usefulness of quinine in checking malaria.
- (Given: For $\chi^2$ at 0.05 )

# Example

- Let us take the following hypotheses:
- Null Hypothesis $H_0$: Quinine is not effective in checking malaria.
- Alternative Hypothesis $Ha$: Quinine is effective in checking malaria.
- Applying $\chi^2$ test:

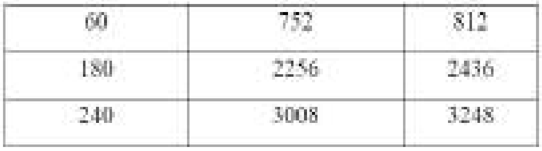


Expectation ($E_{11}$) column wise (first column) first element of the above table = $\frac{240 \times 812}{3248}$ = 60

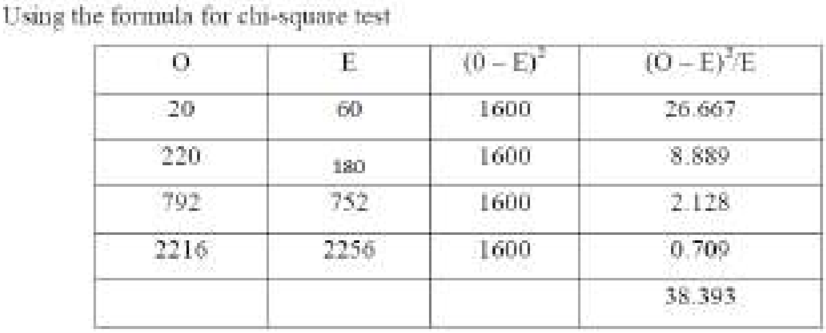
Expectation ($E_{21}$) column wise (first column) second element of the above table = $\frac{240 \times 2436}{3248}$ = 180

# Example

- Expected Frequency is



| | | |
|---|---|---|
| 60 | 752 | 812 |
| 180 | 2256 | 2436 |
| 240 | 3008 | 3248 |



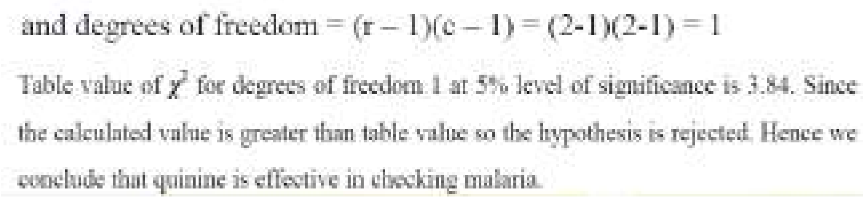
Using the formula for chi-square test

| O | E | $(0 - E)^2$ | $(O - E)^2/E$ |
|---|---|---|---|
| 20 | 60 | 1600 | 26.667 |
| 220 | 180 | 1600 | 8.889 |
| 792 | 752 | 1600 | 2.128 |
| 2216 | 2256 | 1600 | 0.709 |
| | | | 38.393 |

# Example

- The chi square is

$$\chi^2 = \sum \frac{(O-E)^2}{E} \qquad = 38.393$$



and degrees of freedom = (r – 1)(c – 1) = (2-1)(2-1) = 1

Table value of $\chi^2$ for degrees of freedom 1 at 5% level of significance is 3.84. Since the calculated value is greater than table value so the hypothesis is rejected. Hence we conclude that quinine is effective in checking malaria.

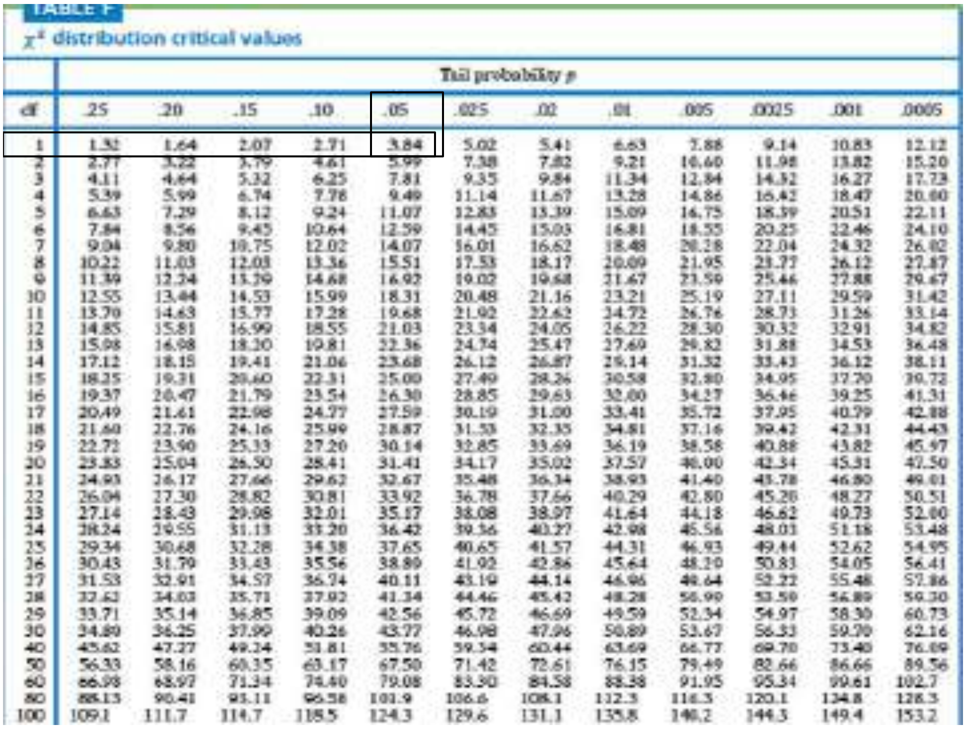


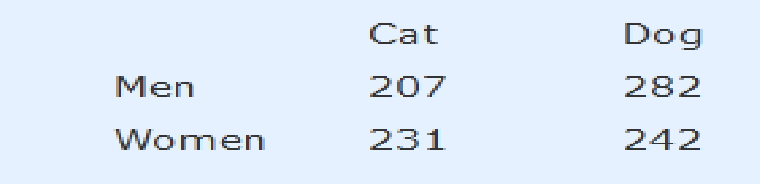
TABLE F

$\chi^2$ distribution critical values

| | Tail probability $p$ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| df | .25 | .20 | .15 | .10 | .05 | .025 | .02 | .01 | .005 | .0025 | .001 | .0005 |
| 1 | 1.32 | 1.64 | 2.07 | 2.71 | 3.84 | 5.02 | 5.41 | 6.63 | 7.88 | 9.14 | 10.83 | 12.12 |
| 2 | 2.77 | 3.22 | 3.79 | 4.61 | 5.99 | 7.38 | 7.82 | 9.21 | 10.60 | 11.98 | 13.82 | 15.20 |
| 3 | 4.11 | 4.64 | 5.32 | 6.25 | 7.81 | 9.35 | 9.84 | 11.34 | 12.84 | 14.32 | 16.27 | 17.73 |
| 4 | 5.39 | 5.99 | 6.74 | 7.78 | 9.49 | 11.14 | 11.67 | 13.28 | 14.86 | 16.42 | 18.47 | 20.00 |
| 5 | 6.63 | 7.29 | 8.12 | 9.24 | 11.07 | 12.83 | 13.39 | 15.09 | 16.75 | 18.39 | 20.51 | 22.11 |
| 6 | 7.84 | 8.56 | 9.45 | 10.64 | 12.59 | 14.45 | 15.03 | 16.81 | 18.55 | 20.25 | 22.46 | 24.10 |
| 7 | 9.04 | 9.80 | 10.75 | 12.02 | 14.07 | 16.01 | 16.62 | 18.48 | 20.28 | 22.04 | 24.32 | 26.02 |
| 8 | 10.22 | 11.03 | 12.03 | 13.36 | 15.51 | 17.53 | 18.17 | 20.09 | 21.95 | 23.77 | 26.12 | 27.87 |
| 9 | 11.39 | 12.24 | 13.29 | 14.68 | 16.92 | 19.02 | 19.68 | 21.67 | 23.59 | 25.46 | 27.88 | 29.67 |
| 10 | 12.55 | 13.44 | 14.53 | 15.99 | 18.31 | 20.48 | 21.16 | 23.21 | 25.19 | 27.11 | 29.59 | 31.42 |
| 11 | 13.70 | 14.63 | 15.77 | 17.28 | 19.68 | 21.92 | 22.62 | 24.72 | 26.76 | 28.73 | 31.26 | 33.14 |
| 12 | 14.85 | 15.81 | 16.99 | 18.55 | 21.03 | 23.34 | 24.05 | 26.22 | 28.30 | 30.32 | 32.91 | 34.82 |
| 13 | 15.98 | 16.98 | 18.20 | 19.81 | 22.36 | 24.74 | 25.47 | 27.69 | 29.82 | 31.88 | 34.53 | 36.48 |
| 14 | 17.12 | 18.15 | 19.41 | 21.06 | 23.68 | 26.12 | 26.87 | 29.14 | 31.32 | 33.43 | 36.12 | 38.11 |
| 15 | 18.25 | 19.31 | 20.60 | 22.31 | 25.00 | 27.49 | 28.26 | 30.58 | 32.80 | 34.95 | 37.70 | 39.72 |
| 16 | 19.37 | 20.47 | 21.79 | 23.54 | 26.30 | 28.85 | 29.63 | 32.00 | 34.27 | 36.46 | 39.25 | 41.31 |
| 17 | 20.49 | 21.61 | 22.98 | 24.77 | 27.59 | 30.19 | 31.00 | 33.41 | 35.72 | 37.95 | 40.79 | 42.88 |
| 18 | 21.60 | 22.76 | 24.16 | 25.99 | 28.87 | 31.53 | 32.35 | 34.81 | 37.16 | 39.42 | 42.31 | 44.43 |
| 19 | 22.72 | 23.90 | 25.33 | 27.20 | 30.14 | 32.85 | 33.69 | 36.19 | 38.58 | 40.88 | 43.82 | 45.97 |
| 20 | 23.83 | 25.04 | 26.50 | 28.41 | 31.41 | 34.17 | 35.02 | 37.57 | 40.00 | 42.34 | 45.31 | 47.50 |
| 21 | 24.93 | 26.17 | 27.66 | 29.62 | 32.67 | 35.48 | 36.34 | 38.93 | 41.40 | 43.78 | 46.80 | 49.01 |
| 22 | 26.04 | 27.30 | 28.82 | 30.81 | 33.92 | 36.78 | 37.66 | 40.29 | 42.80 | 45.20 | 48.27 | 50.51 |
| 23 | 27.14 | 28.43 | 29.98 | 32.01 | 35.17 | 38.08 | 38.97 | 41.64 | 44.18 | 46.62 | 49.73 | 52.00 |
| 24 | 28.24 | 29.55 | 31.13 | 33.20 | 36.42 | 39.36 | 40.27 | 42.98 | 45.56 | 48.03 | 51.18 | 53.48 |
| 25 | 29.34 | 30.68 | 32.28 | 34.38 | 37.65 | 40.65 | 41.57 | 44.31 | 46.93 | 49.44 | 52.62 | 54.95 |
| 26 | 30.43 | 31.79 | 33.43 | 35.56 | 38.89 | 41.92 | 42.86 | 45.64 | 48.29 | 50.83 | 54.05 | 56.41 |
| 27 | 31.53 | 32.91 | 34.57 | 36.74 | 40.11 | 43.19 | 44.14 | 46.96 | 49.64 | 52.22 | 55.48 | 57.86 |
| 28 | 32.62 | 34.03 | 35.71 | 37.92 | 41.34 | 44.46 | 45.42 | 48.28 | 50.99 | 53.59 | 56.89 | 59.30 |
| 29 | 33.71 | 35.14 | 36.85 | 39.09 | 42.56 | 45.72 | 46.69 | 49.59 | 52.34 | 54.97 | 58.30 | 60.73 |
| 30 | 34.80 | 36.25 | 37.99 | 40.26 | 43.77 | 46.98 | 47.96 | 50.89 | 53.67 | 56.33 | 59.70 | 62.16 |
| 40 | 45.62 | 47.27 | 49.24 | 51.81 | 55.76 | 59.34 | 60.44 | 63.69 | 66.77 | 69.70 | 73.40 | 76.09 |
| 50 | 56.33 | 58.16 | 60.35 | 63.17 | 67.50 | 71.42 | 72.61 | 76.15 | 79.49 | 82.66 | 86.66 | 89.56 |
| 60 | 66.98 | 68.97 | 71.34 | 74.40 | 79.08 | 83.30 | 84.58 | 88.38 | 91.95 | 95.34 | 99.61 | 102.7 |
| 80 | 88.13 | 90.41 | 93.11 | 96.58 | 101.9 | 106.6 | 108.1 | 112.3 | 116.3 | 120.1 | 124.8 | 128.3 |
| 100 | 109.1 | 111.7 | 114.7 | 118.5 | 124.3 | 129.6 | 131.1 | 135.8 | 140.2 | 144.3 | 149.4 | 153.2 |

# Example

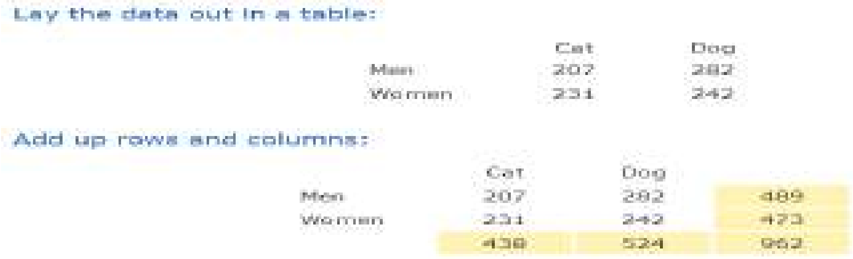
**"Which pet do you prefer?“ The significance at 0.05**



| | Cat | Dog |
|---|---|---|
| Men | 207 | 282 |
| Women | 231 | 242 |

# Example

- The two **hypotheses** are.
- Gender and preference for cats or dogs are **independent**.
- Gender and preference for cats or dogs are **not independent**.

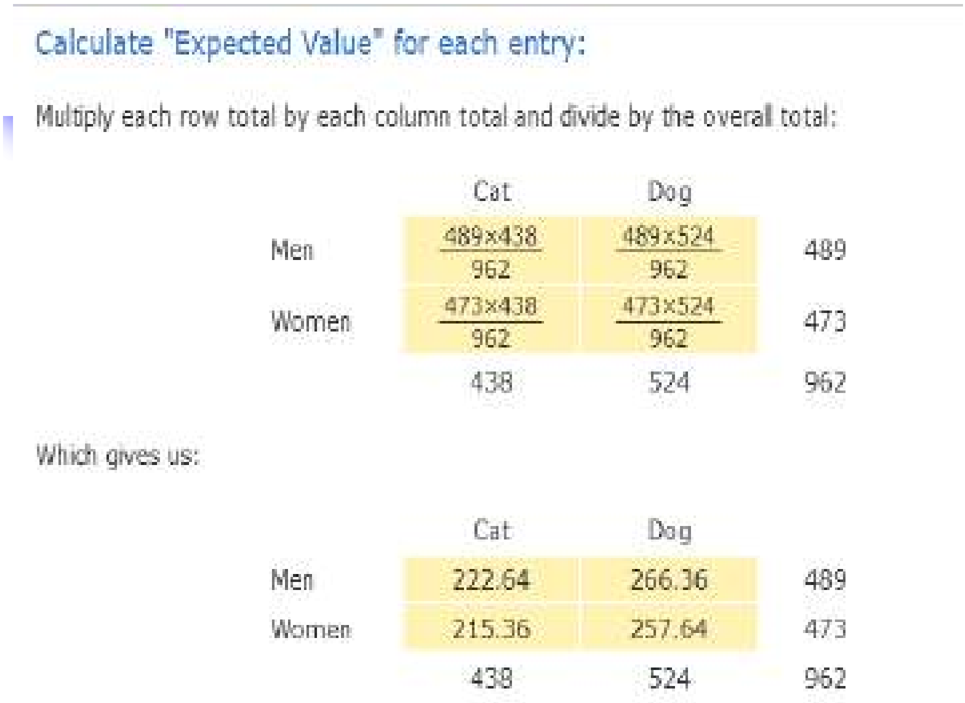


Lay the data out in a table:

| | Cat | Dog |
|---|---|---|
| Men | 207 | 282 |
| Women | 231 | 242 |

Add up rows and columns:

| | Cat | Dog | |
|---|---|---|---|
| Men | 207 | 282 | 489 |
| Women | 231 | 242 | 473 |
| | 438 | 524 | 962 |



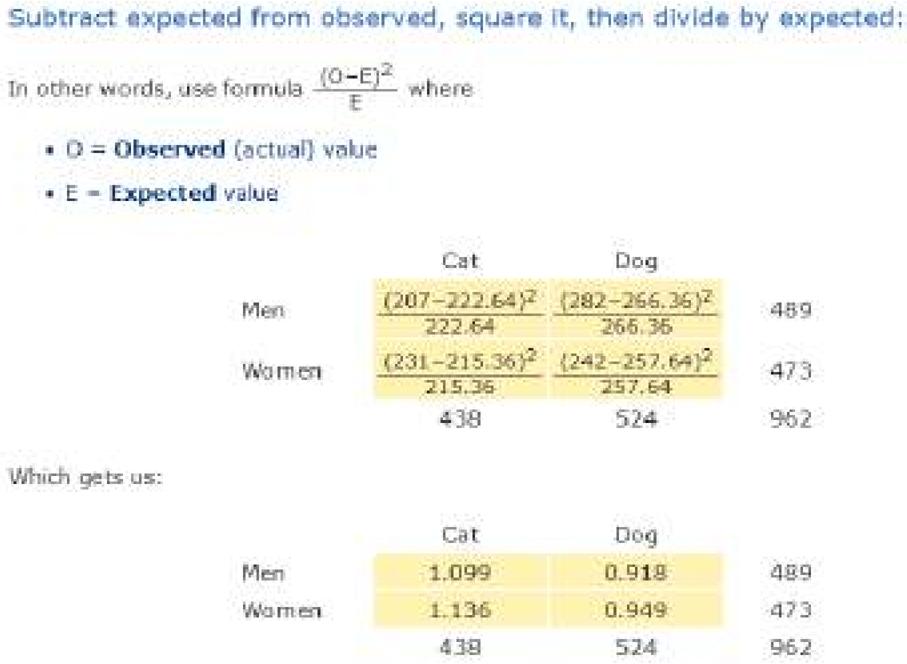
## Calculate "Expected Value" for each entry:

Multiply each row total by each column total and divide by the overall total:

| | Cat | Dog | |
|---|---|---|---|
| Men | $\frac{489\times438}{962}$ | $\frac{489\times524}{962}$ | 489 |
| Women | $\frac{473\times438}{962}$ | $\frac{473\times524}{962}$ | 473 |
| | 438 | 524 | 962 |

Which gives us:

| | Cat | Dog | |
|---|---|---|---|
| Men | 222.64 | 266.36 | 489 |
| Women | 215.36 | 257.64 | 473 |
| | 438 | 524 | 962 |



## Subtract expected from observed, square it, then divide by expected:

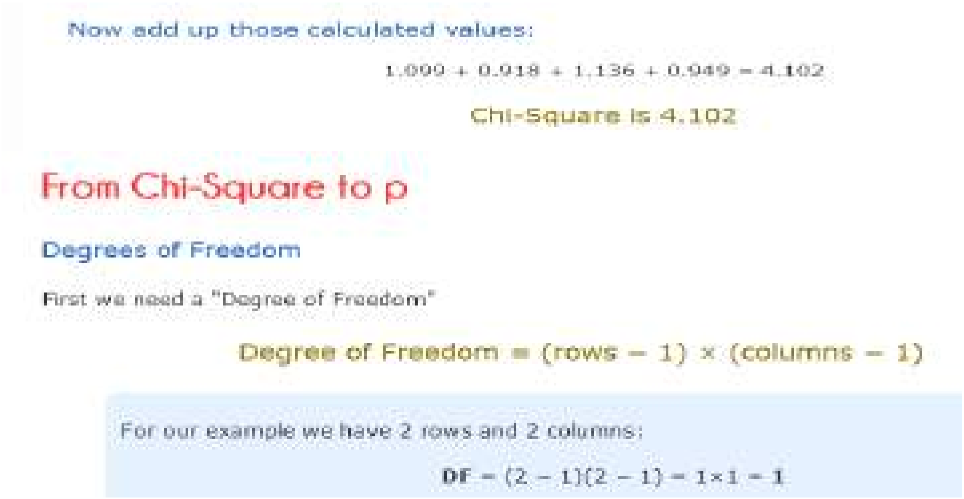
In other words, use formula $\frac{(O-E)^2}{E}$ where

- O = **Observed** (actual) value
- E = **Expected** value

| | Cat | Dog | |
|---|---|---|---|
| Men | $\frac{(207-222.64)^2}{222.64}$ | $\frac{(282-266.36)^2}{266.36}$ | 489 |
| Women | $\frac{(231-215.36)^2}{215.36}$ | $\frac{(242-257.64)^2}{257.64}$ | 473 |
| | 438 | 524 | 962 |

Which gets us:

| | Cat | Dog | |
|---|---|---|---|
| Men | 1.099 | 0.918 | 489 |
| Women | 1.136 | 0.949 | 473 |
| | 438 | 524 | 962 |



Now add up those calculated values:

$$1.099 + 0.918 + 1.136 + 0.949 = 4.102$$

Chi-Square is 4.102

## From Chi-Square to p

### Degrees of Freedom

First we need a "Degree of Freedom"

Degree of Freedom = (rows − 1) × (columns − 1)

For our example we have 2 rows and 2 columns:

**DF** = (2 − 1)(2 − 1) = 1×1 = 1

The rest of the calculation is look it up in a table

The result is:

p = 3.84 (significance level 0.05)

- **Conclusion:**
- $\chi^2 = 4.102 > \chi^2_\alpha = 3.84$ so **reject $H_0$** and conclude that Gender and preference for cats or dogs are **independent**.

# ANOVA

- ANOVA is used for testing the significance of the differences among more than two sample means.
- Assumptions
  - Each sample is randomly drawn from normal population
  - Each of these population have same variance
- Analysis of variance (ANOVA)is based on comparison of two different estimates of the variance $\sigma^2$ ,of overall population.
- Hypothesis:
  - $H_0$ : All means are equal
  - H1 : At least two means are not equal.

# Inferences About a Population Variance

- Sometimes decision makers are interested about the variability in a population.
- Chi-square test can be used to test the variability in a population.
- Assumption:
  - The distribution of data in the underlying population from which the sample is derived is normal.
  - The sample has been randomly selected from the population it represents.

# Inferences About Two Population Variance

- Two population variances can be tested by F-test.
- Assumptions
  - Each sample has been randomly selected from the population it represents.
  - The distribution of data in the underlying population from which each of the samples is derived is normal; and
  - homogeneity of variance assumption, states that the variances of the both populations are equal.
- Hypothesis
  - $H_0$: Both sample have equal variance
  - $H_1$: Both sample have unequal variance

# F Test or The Variance Ratio Test :

- The F test is named in honor of the great statistician R. A. Fisher
- The object of the F test is to find out whether the two independent estimates of population variance differ significantly or whether the two samples may be regarded as drawn from the normal populations having the same variance
- F is defined as
- F = larger estimate of variance / smaller estimate of variance
- $v_1 = n_1 – 1$ and $v_2 = n_2 – 1$

## F Test or The Variance Ratio Test :

- $v_1$ = d. f. for sample having larger variance
- $v_2$ = d. f. for sample having smaller variance
- The calculated value of F is compared with the table value for $v_1$ and $v_2$ at 5% or 1% level of significance
- If calculated value of F is greater than the table value then the F ratio is considered significant and null hypothesis is rejected
- If calculated value of F is smaller than the table value then the F ratio is considered insignificant and null hypothesis is accepted

## F Test or The Variance Ratio Test :

- It is inferred that both samples have come from the population having same variance
- Since the F test is based on the ratio of two variances, it is also known as the Variance Ratio Test

## F Test or The Variance Ratio Test :

- Two random samples were drawn from the two normal populations and their values are :
- A:66 67 75 76 82 84 88 90
  92
- B:64 66 74 78 82 85 87 92
  93 95 97
- Test whether the two populations have the same variance at the 5% level of significance

## F Test or The Variance Ratio Test :

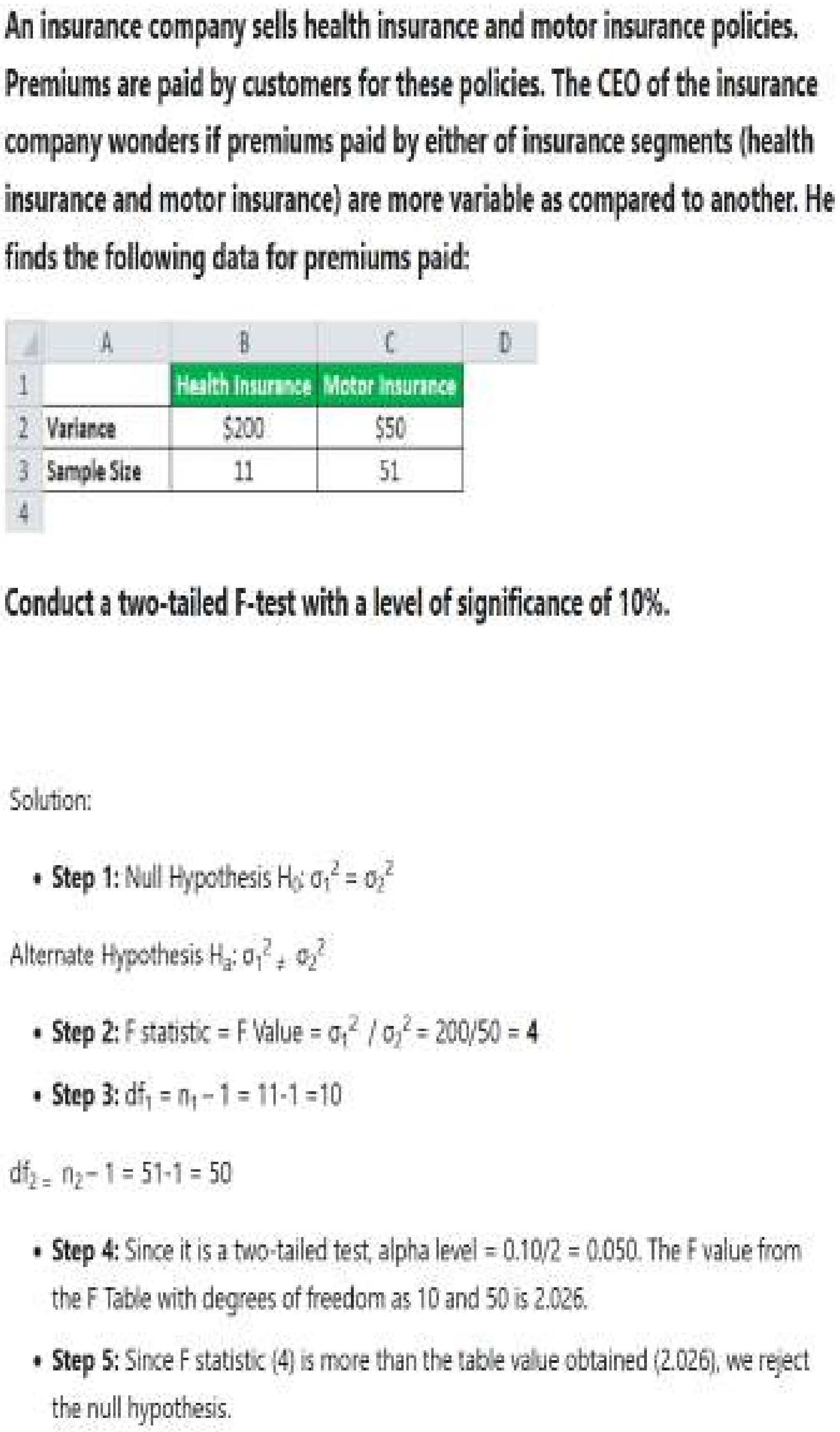
- $H_0$: the two populations have the same variance

$$\overline{x_1} = 80 \ , \ s_1^2 = 91.75$$

$$\overline{x_2} = 83 \ , \ s_2^2 = 129.8$$

$$F = 1.415$$

- $F = (s_1^2)/(s_2^2)$
- For $v_1 = 10$ and $v_2 = 8$
- $F_{0.05} = 3.34$
- $F_{0.01} = 5.82$
- $H_o$ is accepted



**An insurance company sells health insurance and motor insurance policies. Premiums are paid by customers for these policies. The CEO of the insurance company wonders if premiums paid by either of insurance segments (health insurance and motor insurance) are more variable as compared to another. He finds the following data for premiums paid:**

| | A | B | C | D |
|---|---|---|---|---|
| 1 | | Health Insurance | Motor Insurance | |
| 2 | Variance | $200 | $50 | |
| 3 | Sample Size | 11 | 51 | |
| 4 | | | | |

**Conduct a two-tailed F-test with a level of significance of 10%.**

Solution:

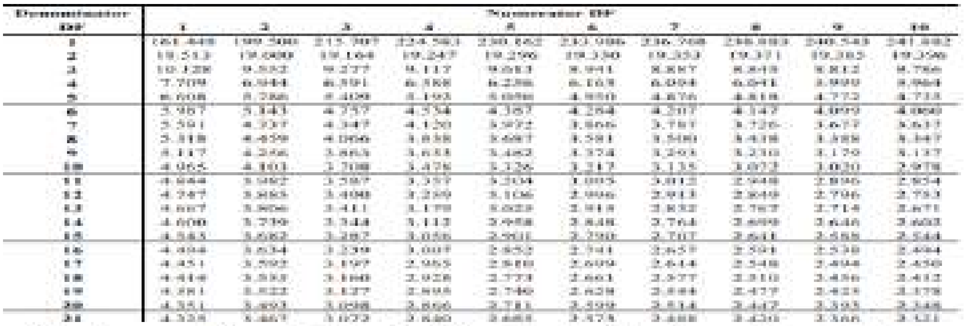
- **Step 1:** Null Hypothesis $H_0$: $\sigma_1^2 = \sigma_2^2$

Alternate Hypothesis $H_a$: $\sigma_1^2 \neq \sigma_2^2$

- **Step 2:** F statistic = F Value = $\sigma_1^2 \ / \sigma_2^2$ = 200/50 = **4**
- **Step 3:** $df_1 = n_1 - 1$ = 11-1 =10

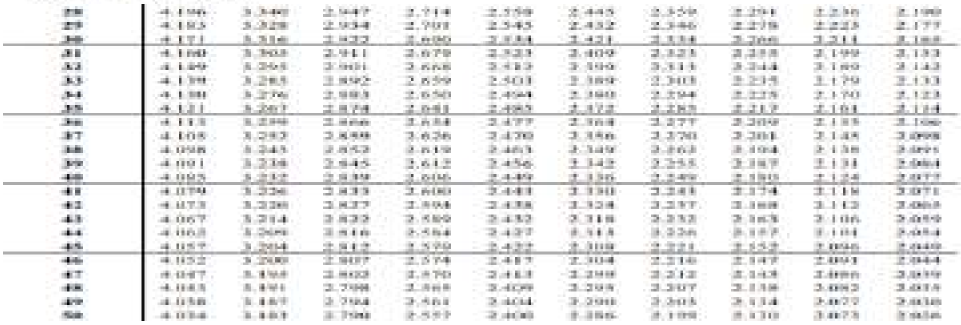
$df_2 = n_2 - 1$ = 51-1 = 50

- **Step 4:** Since it is a two-tailed test, alpha level = 0.10/2 = 0.050. The F value from the F Table with degrees of freedom as 10 and 50 is 2.026.
- **Step 5:** Since F statistic (4) is more than the table value obtained (2.026), we reject the null hypothesis.



| Denominator DF | Numerator DF 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 161.448 | 199.500 | 215.707 | 224.583 | 230.162 | 233.986 | 236.768 | 238.883 | 240.543 | 241.882 |
| 2 | 18.513 | 19.000 | 19.164 | 19.247 | 19.296 | 19.330 | 19.353 | 19.371 | 19.385 | 19.396 |
| 3 | 10.128 | 9.552 | 9.277 | 9.117 | 9.013 | 8.941 | 8.887 | 8.845 | 8.812 | 8.786 |
| 4 | 7.709 | 6.944 | 6.591 | 6.388 | 6.256 | 6.163 | 6.094 | 6.041 | 5.999 | 5.964 |
| 5 | 6.608 | 5.786 | 5.409 | 5.192 | 5.050 | 4.950 | 4.876 | 4.818 | 4.772 | 4.735 |
| 6 | 5.987 | 5.143 | 4.757 | 4.534 | 4.387 | 4.284 | 4.207 | 4.147 | 4.099 | 4.060 |
| 7 | 5.591 | 4.737 | 4.347 | 4.120 | 3.972 | 3.866 | 3.787 | 3.726 | 3.677 | 3.637 |
| 8 | 5.318 | 4.459 | 4.066 | 3.838 | 3.687 | 3.581 | 3.500 | 3.438 | 3.388 | 3.347 |
| 9 | 5.117 | 4.256 | 3.863 | 3.633 | 3.482 | 3.374 | 3.293 | 3.230 | 3.179 | 3.137 |
| 10 | 4.965 | 4.103 | 3.708 | 3.478 | 3.326 | 3.217 | 3.135 | 3.072 | 3.020 | 2.978 |
| 11 | 4.844 | 3.982 | 3.587 | 3.357 | 3.204 | 3.095 | 3.012 | 2.948 | 2.896 | 2.854 |
| 12 | 4.747 | 3.885 | 3.490 | 3.259 | 3.106 | 2.996 | 2.913 | 2.849 | 2.796 | 2.753 |
| 13 | 4.667 | 3.806 | 3.411 | 3.179 | 3.025 | 2.915 | 2.832 | 2.767 | 2.714 | 2.671 |
| 14 | 4.600 | 3.739 | 3.344 | 3.112 | 2.958 | 2.848 | 2.764 | 2.699 | 2.646 | 2.602 |
| 15 | 4.543 | 3.682 | 3.287 | 3.056 | 2.901 | 2.790 | 2.707 | 2.641 | 2.588 | 2.544 |
| 16 | 4.494 | 3.634 | 3.239 | 3.007 | 2.852 | 2.741 | 2.657 | 2.591 | 2.538 | 2.494 |
| 17 | 4.451 | 3.592 | 3.197 | 2.965 | 2.810 | 2.699 | 2.614 | 2.548 | 2.494 | 2.450 |
| 18 | 4.414 | 3.555 | 3.160 | 2.928 | 2.773 | 2.661 | 2.577 | 2.510 | 2.456 | 2.412 |
| 19 | 4.381 | 3.522 | 3.127 | 2.895 | 2.740 | 2.628 | 2.544 | 2.477 | 2.423 | 2.378 |
| 20 | 4.351 | 3.493 | 3.098 | 2.866 | 2.711 | 2.599 | 2.514 | 2.447 | 2.393 | 2.348 |
| 21 | 4.325 | 3.467 | 3.072 | 2.840 | 2.685 | 2.573 | 2.488 | 2.420 | 2.366 | 2.321 |

***Note:*** There are different F Tables for different levels of significance. Above is the F table for alpha = .050.



|  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|
| 28 | 4.196 | 3.340 | 2.947 | 2.714 | 2.558 | 2.445 | 2.359 | 2.291 | 2.236 | 2.190 |
| 29 | 4.183 | 3.328 | 2.934 | 2.701 | 2.545 | 2.432 | 2.346 | 2.278 | 2.223 | 2.177 |
| 30 | 4.171 | 3.316 | 2.922 | 2.690 | 2.534 | 2.421 | 2.334 | 2.266 | 2.211 | 2.165 |
| 31 | 4.160 | 3.305 | 2.911 | 2.679 | 2.523 | 2.409 | 2.323 | 2.255 | 2.199 | 2.153 |
| 32 | 4.149 | 3.295 | 2.901 | 2.668 | 2.512 | 2.399 | 2.313 | 2.244 | 2.189 | 2.142 |
| 33 | 4.139 | 3.285 | 2.892 | 2.659 | 2.503 | 2.389 | 2.303 | 2.235 | 2.179 | 2.133 |
| 34 | 4.130 | 3.276 | 2.883 | 2.650 | 2.494 | 2.380 | 2.294 | 2.225 | 2.170 | 2.123 |
| 35 | 4.121 | 3.267 | 2.874 | 2.641 | 2.485 | 2.372 | 2.285 | 2.217 | 2.161 | 2.114 |
| 36 | 4.113 | 3.259 | 2.866 | 2.634 | 2.477 | 2.364 | 2.277 | 2.209 | 2.153 | 2.106 |
| 37 | 4.105 | 3.252 | 2.859 | 2.626 | 2.470 | 2.356 | 2.270 | 2.201 | 2.145 | 2.098 |
| 38 | 4.098 | 3.245 | 2.852 | 2.619 | 2.463 | 2.349 | 2.262 | 2.194 | 2.138 | 2.091 |
| 39 | 4.091 | 3.238 | 2.845 | 2.612 | 2.456 | 2.342 | 2.255 | 2.187 | 2.131 | 2.084 |
| 40 | 4.085 | 3.232 | 2.839 | 2.606 | 2.449 | 2.336 | 2.249 | 2.180 | 2.124 | 2.077 |
| 41 | 4.079 | 3.226 | 2.833 | 2.600 | 2.443 | 2.330 | 2.243 | 2.174 | 2.118 | 2.071 |
| 42 | 4.073 | 3.220 | 2.827 | 2.594 | 2.438 | 2.324 | 2.237 | 2.168 | 2.112 | 2.065 |
| 43 | 4.067 | 3.214 | 2.822 | 2.589 | 2.432 | 2.318 | 2.232 | 2.163 | 2.106 | 2.059 |
| 44 | 4.062 | 3.209 | 2.816 | 2.584 | 2.427 | 2.313 | 2.226 | 2.157 | 2.101 | 2.054 |
| 45 | 4.057 | 3.204 | 2.812 | 2.579 | 2.422 | 2.308 | 2.221 | 2.152 | 2.096 | 2.049 |
| 46 | 4.052 | 3.200 | 2.807 | 2.574 | 2.417 | 2.304 | 2.216 | 2.147 | 2.091 | 2.044 |
| 47 | 4.047 | 3.195 | 2.802 | 2.570 | 2.413 | 2.299 | 2.212 | 2.143 | 2.086 | 2.039 |
| 48 | 4.043 | 3.191 | 2.798 | 2.565 | 2.409 | 2.295 | 2.207 | 2.138 | 2.082 | 2.035 |
| 49 | 4.038 | 3.187 | 2.794 | 2.561 | 2.404 | 2.290 | 2.203 | 2.134 | 2.077 | 2.030 |
| 50 | 4.034 | 3.183 | 2.790 | 2.557 | 2.400 | 2.286 | 2.199 | 2.130 | 2.073 | 2.026 |

# Interpolation

Computer Oriented Numerical and Statistical Methods

## Outline

- Polynomial Interpolation
- Difference Tables
- Netwon’s Forward and Backward Interpolation Formula
- Lagrange's Formula
- Divided Difference Formula
- Inverse Interpolation

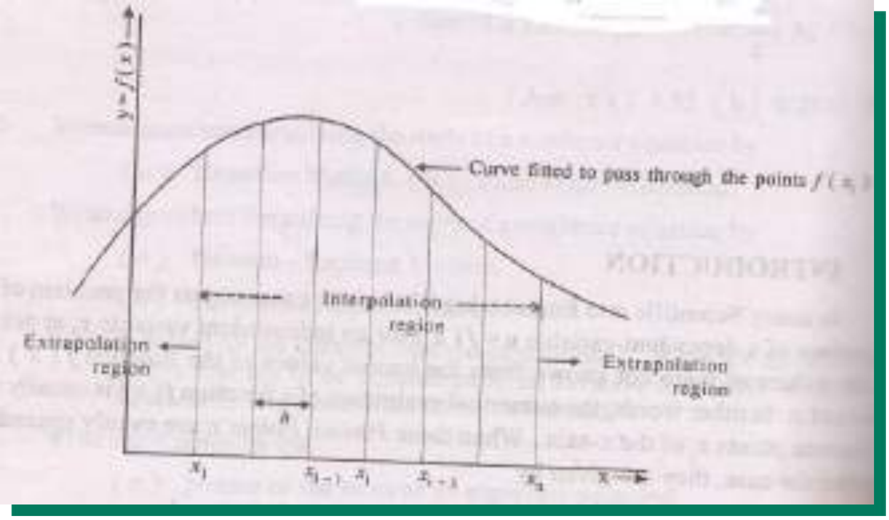
## Introduction

- Suppose x and y are two variables and their relation can be expressed as y = f(x) ; $x_1 \le x \le x_{n.}$ Then we say that x is an independent variable and y is a dependent variable.
- When the form of f(x) is known, then the value of y can be computed directly corresponding to any value of x in the range $x_1 \le x \le x_{n.}$
- However, if the form of f(x) is not known, and only a set of values $(x_1,y_1)$, $(x_2,y_2)$, …., $(x_n,y_n)$ satisfying the relation y = f(x) are known, then the process of estimating the value of independent variable y for a given value of x in the range $x_1 \le x \le x_n$ is known as interpolation.

## Introduction

- However, if we move in opposite direction i.e. estimate the value of dependent variable x for a given value of independent variable y, the process is known as inverse interpolation.
- The process of estimating the value of independent variable y for a given value of x outside the range $x_1 \le x \le x_n$ is known as extrapolation.

## Introduction



## Methods Of Interpolation

- The decision of using a particular method depends in tabulation of the functions.
- The tabulated points $(x_i, y_i)$ i = 1,2,…,n of function y = f(x), can be equally spaced or unequal spaced.
- Methods for equally spaced functions
  - Netwon’s forward interpolation formula
  - Netwon’s backward interpolation formula

## Methods Of Interpolation

- Methods for unequally spaced functions
  - Netwon’s divided difference interpolation formula
  - Lagrangian interpolation (Lagrange’s)
  - These methods also works well for equally spaced function.

## Finite Differences

- The finite differences are either difference between the values of the function or the differences between the past differences.
- There are 3 types of differences
  - Forward Differences
  - Backward Differences
  - Divided Differences

## Forward Differences

- If $y_1, y_2, y_3, \ldots, y_n$ denotes the values of the function of type $y = f(x)$ at $x = x_1, x_2, \ldots, x_n$ then $y_2 - y_1, y_3 - y_2, y_4 - y_3, \ldots, y_n - y_{n-1}$ are called the forward differences of y
- These differences are denoted as $\Delta y_1, \Delta y_2, \Delta y_3, \ldots, \Delta y_{n-1}$

∀ $\therefore\ \Delta y_1 = y_2 - y_1,\ \Delta y_2 = y_3 - y_2, \ldots, \Delta y_{n-1} = y_n - y_{n-1}$ where $\Delta$ is called forward difference operator and $\Delta y_1, \Delta y_2, \Delta y_3, \ldots, \Delta y_{n-1}$ are called first order forward differences.

## Forward Differences

- The differences of the first order forward differences are called second order forward differences and are denoted as $\Delta^2 y_1, \Delta^2 y_2, \Delta^2 y_3, \ldots, \Delta^2 y_{n-1}$

∀ $\therefore\ \Delta^2 y_1 = \Delta y_2 - \Delta y_1 = y_3 - 2y_2 + y_1$

- $\Delta^2 y_2 = \Delta y_3 - \Delta y_2 = y_4 - 2y_3 + y_2$
- In the similar manner, the third order forward differences are

∀ $\Delta^3 y_1 = \Delta^2 y_2 - \Delta^2 y_1 = y_4 - 3y_3 + 3y_2 - y_1$

∀ $\Delta^3 y_2 = \Delta^2 y_3 - \Delta^2 y_2 = y_5 - 3y_4 + 3y_3 - y_2$

- In general, the first order forward differences at the $i^{th}$ point is $\Delta y_i = y_{i+1} - y_i$ and the $j^{th}$ order forward differences at the $i^{th}$ point is $\Delta^j y_i = \Delta^{j-1} y_{i+1} - \Delta^{j-1} y_i$

## Backward Differences

- If $y_1, y_2, y_3, \ldots, y_n$ denotes the values of the function of type $y = f(x)$ at $x = x_1, x_2, \ldots, x_n$ then $y_2 - y_1$, $y_3 - y_2$, $y_4 - y_3$, ….. , $y_n - y_{n-1}$ are called the backward differences of y
- These differences are denoted as $\nabla y_2, \nabla y_3, \ldots, \nabla y_n$

∀ $\therefore\ \nabla y_2 = y_2 - y_1,\ \nabla y_3 = y_3 - y_2, \ldots, \nabla y_n = y_n - y_{n-1}$ where $\nabla$ is called backward difference operator and $\nabla y_2, \nabla y_3, \ldots, \nabla y_n$ are called first order backward differences.

## Backward Differences

- The differences of the first order backward differences are called second order backward differences and are denoted as $\nabla^2 y_3, \nabla^2 y_4$ , …. etc.

∀ $\therefore\ \nabla^2 y_3 = \nabla y_3 - \nabla y_2 = y_3 - 2y_2 + y_1$

- $\nabla^2 y_4 = \nabla y_4 - \nabla y_3 = y_4 - 2y_3 + y_2$
- In the similar manner, the third order backward differences are

∀ $\nabla^3 y_4 = y_4 - 3y_3 + 3y_2 - y_1$

∀ $\nabla^3 y_5 = y_5 - 3y_4 + 3y_3 - y_2$

- In general, the first order backward differences at the ith point is $\nabla y_i = y_i - y_{i-1}$ and the jth order backward differences at the ith point is $\nabla^j y_i = \nabla^{j-1} y_i - \nabla^{j-1} y_{i-1}$

# Divided Differences

- If $y_1, y_2, y_3, \ldots, y_n$ denotes the values of the function of type $y = f(x)$ at $x = x_1, x_2, \ldots, x_n$ then

$$\frac{y_2 - y_1}{x_2 - x_1}\ ,\quad \frac{y_3 - y_2}{x_3 - x_2}\ ,\quad \frac{y_4 - y_3}{x_4 - x_3}\ ,\ldots,\quad \frac{y_n - y_{n-1}}{x_n - x_{n-1}}$$

- are called the divided differences of y and are denoted as $\Delta_d y_1, \Delta_d y_2, \Delta_d y_3, \ldots, \Delta_d y_{n-1}$

∀ $\therefore\ \Delta_d y_1 = (y_2 - y_1) / (x_2 - x_1) = [x_1, x_2]$

∀ $\Delta_d y_2 = (y_3 - y_2) / (x_3 - x_2) = [x_2, x_3]$

∀ $\Delta_d y_{n-1} = (y_n - y_{n-1}) / (x_n - x_{n-1}) = [x_{n-1}, x_n]$

- Where $\Delta_d$ is called divide difference operator and $\Delta_d y_1, \Delta_d y_2, \ldots, \Delta_d y_n$ are called first order divided differences.

# Divided Differences

- The differences of the first order divided differences are called second order divided differences and are denoted as $\Delta^2{}_d y_1, \Delta^2{}_d y_2, \ldots, \Delta^2{}_d y_n$ etc.

∀ $\therefore\ \Delta^2{}_d y_1 = (\Delta_d y_2 - \Delta_d y_1) / (x_3 - x_1)$

∀ $\Delta^2{}_d y_2 = (\Delta_d y_3 - \Delta_d y_2) / (x_4 - x_2)$

- In the similar manner, the third order forward differences are

∀ $\Delta_d{}^3 y_1 = (\Delta_d{}^2 y_2 - \Delta_d{}^2 y_1) / (x_4 - x_1)$

∀ $\Delta_d{}^3 y_2 = (\Delta_d{}^2 y_3 - \Delta_d{}^2 y_2) / (x_5 - x_2)$

- In general, the first order divided differences at the ith point is $\Delta_d y_i = (y_{i+1} - y_i) / (x_{i+1} - x_i)$ and the jth order forward differences at the ith point is $\Delta_d{}^j y_1 = (\Delta_d{}^{j-1} y_{i+1} - \Delta_d{}^{j-1} y_i) / (x_{i+1} - x_i)$

## Differences Tables

- A difference table is a table that lists the differences of the function values and the differences of differences in succession.

## Forward Difference Table

- Let us consider the values $y_1$, $y_2$, $y_3$, $y_4$ of the function type $y = f(x)$ tabulated at equally spaced points $x_1$, $x_2$, $x_3$, $x_4$. The forward difference table along with tabulated points will look like

| Forward Difference Table | | | | | |
|---|---|---|---|---|---|
| i | $x_i$ | $y_i$ | $\Delta y_i$ | $\Delta^2 y_i$ | $\Delta^3 y_i$ |
| 1 | $x_1$ | $y_1$ | $\Delta y_1 = y_2 – y_1$ | $\Delta^2 y_1 = \Delta y_2 – \Delta y_1$ | $\Delta^3 y_1 = \Delta^2 y_2 – \Delta^2 y_1$ |
| 2 | $x_2$ | $y_2$ | $\Delta y_2 = y_3 – y_2$ | $\Delta^2 y_2 = \Delta y_3 – \Delta y_2$ | |
| 3 | $x_3$ | $y_3$ | $\Delta y_3 = y_4 – y_3$ | | |
| 4 | $x_4$ | $y_4$ | | | |

## Forward Difference Table

- The forward difference table for function tabulated at n equally spaced points can be represented by a matrix of size (n-1) * (n-1) where $j^{th}$ order frequency at the $i^{th}$ point ($\Delta^j y_i$) is represented by the element $d_{ij}$ of matrix D.
- Note that only the elements in the column 1 to (n – i), for rows i = 1, 2, …., n-1 are of interest.

## Backward Difference Table

- Let us consider the values $y_1$, $y_2$, $y_3$, $y_4$ of the function type y = f(x) tabulated at equally spaced points $x_1$, $x_2$, $x_3$, $x_4$. The Backward difference table along with tabulated points will look like

| **Backward Difference Table** | | | | | |
|---|---|---|---|---|---|
| i | $x_i$ | $y_i$ | $\nabla y_i$ | $\nabla^2 y_i$ | $\nabla^3 y_i$ |
| 1 | $x_1$ | $y_1$ | | | |
| 2 | $x_2$ | $y_2$ | $\nabla y_2 = y_2 – y_1$ | | |
| 3 | $x_3$ | $y_3$ | $\nabla y_3 = y_3 – y_2$ | $\nabla^2 y_3 = \nabla y_3 – \nabla y_2$ | |
| 4 | $x_4$ | $y_4$ | $\nabla y_4 = y_4 – y_3$ | $\nabla^2 y_4 = \nabla y_4 – \nabla y_3$ | $\nabla^3 y_4 = \nabla^2 y_4 – \nabla^2 y_3$ |

## Divided Difference Table

- Let us consider the values $y_1$, $y_2$, $y_3$, $y_4$ of the function type y = f(x) tabulated at points $x_1$, $x_2$, $x_3$, $x_4$ not necessarily equally spaced . The divide difference table along with tabulated points will look like

## Divided Difference Table

**Divided Difference Table**

| i | $x_i$ | $y_i$ | $\Delta_d y_i$ | $\Delta^2 y_i$ | $\Delta^3 y_i$ |
|---|---|---|---|---|---|
| 1 | $x_1$ | $y_1$ | $\Delta_d y_1 = (y_2 - y_1) / (x_2 - x_1)$ | $\Delta_d{}^2 y_1 = (\Delta_d y_2 - \Delta_d y_1) / (x_3 - x_1)$ | $\Delta_d{}^3 y_1 = (\Delta_d{}^2 y_2 - \Delta_d{}^2 y_1) / (x_4 - x_1)$ |
| 2 | $x_2$ | $y_2$ | $\Delta_d y_2 = (y_3 - y_2) / (x_3 - x_2)$ | $\Delta_d y_2 = (\Delta_d y_3 - \Delta_d y_2) / (x_4 - x_2)$ | |
| 3 | $x_3$ | $y_3$ | $\Delta_d y_3 = (y_4 - y_3) / (x_4 - x_3)$ | | |
| 4 | $x_4$ | $y_4$ | | | |

## Netwon’s Methods Of Interpolation

- It is divided into following methods depending on the type of differences being used.

1. Netwon’s Forward Difference Interpolation Formula
2. Netwon’s Backward Difference Interpolation Formula
3. Netwon’s Divided Difference Interpolation Formula

- If the function is tabulated at equal intervals, then we can use either Netwon’s Forward Difference Interpolation Formula or Netwon’s Backward Difference Interpolation Formula.

## Netwon’s Forward Difference Interpolation Formula

- Let us assume that the function y(x) is tabulated at (n+1) equally spaced (interval size h) points.
- To derive the formula for netwon’s forward difference interpolation assume a polynomial of type
  $y(x) = a_1 + a_2(x-x_1) + a_3(x-x_1)(x-x_2) + \ldots + a_{n+1}(x-x_1)(x-x_2) \ldots (x-x_n)$ ……………………….[a]
- It is the $n^{th}$ order polynomial in x.
- It is an interpolation polynomial for the table values $x_i$ and $y_i$ then the polynomial must pass through all points.

∀ ∴ we can obtain $y_i$ by substituting the corresponding $x_i$ for x

## Netwon’s Forward Difference Interpolation Formula

- At x = $x_1$   $y_1 = a_1$
- x = $x_2$   $y_2 = a_1 + a_2(x_2-x_1)$
- x = $x_3$   $y_3 = a_1 + a_2(x_3-x_1) + a_3(x_3-x_1)(x_3-x_2)$
- And so on.
- Since $x_i$’s are equally spaced therefore we can write $x_{i+1} – x_i = h$ and $x_{i+m} – x_i = mh$ then we can express the function values $y_i$’s in terms of intervals values as

  $y_1 = a_1$

  $y_2 = a_1 + a_2h$

  $y_3 = a_1 + a_2(2h) + a_3(2h)(h)$

  …….

$y_{n+1} = a_1 + a_2(nh) + a_3(nh)((n-1)h)+….+a_{n+1}(nh)((n-1)h)…(h)$

## Netwon’s Forward Difference Interpolation Formula

- Now for $a_1$ , $a_2$, $a_3$, $a_{n+1}$ we get

$$a_1 = y_1$$

$$a_2 = \frac{y_2 - a_1}{h} = \frac{y_2 - y_1}{h}$$

$$a_3 = \frac{y_3 - 2y_2 + y_1}{2!h^2}$$

$$a_{n+1} = \frac{y_{n+1} - ny_n + \ldots. + y_1}{n!h^n}$$

## Netwon’s Forward Difference Interpolation Formula

- Using forward difference table we get

$$a_1 = y_1$$

$$a_2 = \frac{\Delta y_1}{h}$$

$$a_3 = \frac{\Delta^2 y_1}{2!h^2}$$

$$\vdots$$

$$a_{n+1} = \frac{\Delta^n y_1}{n!h^n}$$

## Netwon’s Forward Difference Interpolation Formula

- Substituting these values of $a_1$ , $a_2$, $a_3$, $a_{n+1}$ in equation [a] we get

$$y(x) = y_1 + \frac{\Delta y_1}{h}(x - x_1) + \frac{\Delta^2 y_1}{2!h^2}(x - x_1)(x - x_2) + \cdots\cdots + \frac{\Delta^n y_1}{n!h^n}(x - x_1)(x - x_2)\cdots(x - x_n) \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots b]$$

## Netwon’s Forward Difference Interpolation Formula

- If we use the relation $(x-x_1) / h = u$ then $x - x_1 = hu$
  $x – x_2 = x – (x_1 + h) = h(u-1)$
  $x – x_3 = x – (x_2 + h) = h(u-2)$ …..
  $x – x_n = h(u-(n-1))$
- Substituting these values $(x-x_1)$, $(x-x_2)$, …., $(x – x_n)$ in equation [b] we get

## Netwon’s Forward Difference Interpolation Formula

$$y(x) = y_1 + \frac{\Delta y_1}{h} hu +$$

$$\frac{\Delta^2 y_1}{2!h^2} hu * h(u-1) +$$

$$\ldots + \frac{\Delta^n y_1}{n!h^n} hu * h(u-1) * \ldots * h(u-(n-1))$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots [c]$$

## Netwon’s Forward Difference Interpolation Formula

- Simplifying we get

$$y(x) = y_1 + \Delta y_1 u +$$

$$\frac{\Delta^2 y_1}{2!} u(u-1) +$$

$$\dots + \frac{\Delta^n y_1}{n!} u(u-1) * \dots * (u-(n-1))$$

$$\dots\dots\dots\dots\dots\dots\dots [d]$$

- The above derivation assumes that $x_1 < x < x_2$
- This polynomial is called Netwon’s Forward Difference Interpolation Formula

## Netwon’s Backward Difference Interpolation Formula

- Let us assume that the function y(x) is tabulated at (n+1) equally spaced (interval size h) points.
- To derive the formula for netwon’s backward difference interpolation assume a polynomial of type $y(x) = a_1 + a_2(x-x_n) + a_3(x-x_n)(x-x_{n-1}) + \dots + a_{n+1}(x-x_n)(x-x_{n-1}) \dots (x-x_1)$ ……………………..[a]
- It is the $n^{th}$ order polynomial in x.
- It is an interpolation polynomial for the table values $x_i$ and $y_i$ then the polynomial must pass through all points.

∀ ∴ we can obtain $y_i$ by substituting the corresponding $x_i$ for x

## Netwon’s Backward Difference Interpolation Formula

- At $x = x_n$ $y_n = a_1$
- $x = x_{n-1}$ $y_{n-1} = a_1 + a_2(x_{n-1} - x_n)$
- $x = x_{n-2}$ $y_{n-2} = a_1 + a_2(x_{n-2} - x_n) + a_3(x_{n-2}-x_n)(x_{n-2}-x_{n-1})$
- And so on.
- Since $x_i$’s are equally spaced therefore we can write $x_i – x_{i-1} = h$ and $x_{i-1} – x_i = -h$ and $x_{i-m} – x_i = -mh$ then we can express the function values $y_i$’s in terms of intervals values as

  $y_n = a_1$
  $y_{n-1} = a_1 + a_2(-h)$
  $y_{n-2} = a_1 + a_2(-2h) + a_3(-2h)(-h)$
  …….

$y_1 = a_1 + a_2(-nh) + a_3(-nh)(-(n-1)h)+….+a_{n+1}(-nh)(-(n-1)h)…(-h)$

## Netwon’s Backward Difference Interpolation Formula

- Now for $a_{1\,,}\ a_2,\ a_3,\ a_{n+1}$ we get

$$a_1 = y_n$$

$$a_2 = \frac{y_{n-1} - a_1}{h} = \frac{y_n - y_{n-1}}{h}$$

$$a_3 = \frac{y_n - 2y_{n-1} + yn-2}{2!h^2}$$

$$a_{n+1} = \frac{y1 - ny_2 + \ldots. + y_n}{n!h^n}$$

## Netwon’s Backward Difference Interpolation Formula

- Using backward difference table we get

$$a_1 = y_n$$

$$a_2 = \frac{\nabla y_n}{h}$$

$$a_3 = \frac{\nabla^2 y_n}{2!h^2}$$

$$\vdots$$

$$a_{n+1} = \frac{\nabla^n y_n}{n!h^n}$$

## Netwon’s Backward Difference Interpolation Formula

- Substituting these values of $a_1$, $a_2$, $a_3$, $a_{n+1}$ in equation [a] we get

$$y(x) = y_n + \frac{\nabla y_n}{h}(x - x_n) +$$

$$\frac{\nabla^2 y_n}{2!h^2}(x - x_n)(x - x_{n-1}) +$$

$$\ldots\ldots + \frac{\nabla^n y_n}{n!h^n}(x - x_n)(x - x_{n-1})\ldots(x - x_1)$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots[b]$$

## Netwon’s Backward Difference Interpolation Formula

- If we use the relation $(x-x_n) / h = u$ then $x - x_n = hu$
  $x – x_{n-1} = x – (x_n - h) = h(u+1)$
  $x – x_{n-2} = x – (x_{n-1} - h) = h(u+2)$ …..
  $x – x_1 = h(u+(n-1))$
- Substituting these values $(x-x_n)$, $(x-x_{n-1})$, …., $(x – x_1)$ in equation [b] we get

## Netwon’s Backward Difference Interpolation Formula

$$y(x) = y_n + \frac{\nabla y_n}{h} hu +$$

$$\frac{\nabla^2 y_n}{2!h^2} hu * h(u+1) +$$

$$.... + \frac{\nabla^n y_n}{n!h^n} hu * h(u+1) * ... * h(u+(n-1))$$

$$.................... [c]$$

## Netwon's Backward Difference Interpolation Formula

- Simplifying we get

$$y(x) = y_n + \nabla y_n u + \frac{\nabla^2 y_n}{2!} u(u+1) + \ldots + \frac{\nabla^n y_n}{n!} u(u+1) * \ldots * (u+(n-1)) \qquad \ldots\ldots\ldots\ldots\ldots [d]$$

- The above derivation assumes that $x_{n-1} < x < x_n$
- This polynomial is called Netwon's Backward Difference Interpolation Formula

## Netwon's Divided Difference Interpolation Formula

- Let us assume that the function y(x) is tabulated at (n+1) equally spaced (interval size h) points.
- To derive the formula for netwon's divided difference interpolation assume a polynomial of type $y(x) = a_1 + a_2(x-x_1) + a_3(x-x_1)(x-x_2) + \ldots + a_{n+1}(x-x_1)(x-x_2) \ldots (x-x_n)$ ……………………..[a]
- It is the nth order polynomial in x.
- It is an interpolation polynomial for the table values $x_i$ and $y_i$ then the polynomial must pass through all points.

∀ ∴ we can obtain $y_i$ by substituting the corresponding $x_i$ for x

## Netwon’s Divided Difference Interpolation Formula

- At x = $x_1$ $y_1 = a_1$
- x = $x_2$ $y_2 = a_1 + a_2(x_2-x_1)$
- x = $x_3$ $y_3 = a_1 + a_2(x_3-x_1) + a_3(x_3-x_1)(x_3-x_2)$
- And so on.
- Solving for $a_{1\,,}$ $a_2$, $a_3$, $a_{n+1}$ we get

## Netwon’s Divided Difference Interpolation Formula

$$a_1 = y_1$$

$$a_2 = \frac{y_2 - a_1}{x_2 - x_1} = \frac{y_2 - y_1}{x_2 - x_1}$$

$$a_3 = \frac{1}{(x_3 - x_2)} * [\frac{(y_3 - y_1)}{(x_3 - x_1)} - \frac{(y_2 - y_1)}{(x_2 - x_1)}]$$

$$and so on$$

## Netwon’s Divided Difference Interpolation Formula

- Using divided difference table we get

$$a_1 = y_1$$

$$a_2 = \Delta_d y_1$$

$$a_3 = \Delta_d^2 y_1$$

$$\cdot$$
$$\cdot$$
$$\cdot$$

$$a_{n+1} = \Delta_d^n y_1$$

## Netwon’s Divided Difference Interpolation Formula

- Substituting these values of $a_1$, $a_2$, $a_3$, $a_{n+1}$ in equation [a] we get

$$y(x) = y_1 + \Delta_d y_1 (x - x_1) + \Delta_d^2 y_1 (x - x_1)(x - x_2) + \cdots + \Delta_d^n y_1 (x - x_1)(x - x_2)\ldots(x - x_n) \quad \ldots\ldots\ldots [b]$$

- This polynomial is called Netwon’s Divided Difference Interpolation Formula

## Lagrangian / Lagranges Interpolation Formula

- In order to derive a general formula for lagrangian interpolation, we consider a second order polynomial of type.
  $y(x) = a_1(x-x_2)(x-x_3) + a_2(x-x_1)(x-x_3) + a_3(x-x_1)(x-x_2)$
  ……………………….[a]
  passing through the points $(x_1,y_1)$, $(x_2,y_2)$, and $(x_3,y_3)$ where $a_1$, $a_2$, and $a_3$ are unknown constants whose values are determined as follows.
- At $x = x_1$ $y(x_1) = a_1(x_1 - x_2)(x_1-x_3)$

## Lagrangian / Lagranges Interpolation Formula

$$At\ x = x_1 \quad y(x_1) = a_1(x_1 - x_2)(x_1 - x_3)$$

$$\Rightarrow a_1 = \frac{y_1}{(x_1 - x_2)(x_1 - x_3)}$$

$$At\ x = x_2 \quad y(x_2) = a_2(x_2 - x_1)(x_2 - x_3)$$

$$\Rightarrow a_2 = \frac{y_2}{(x_2 - x_1)(x_2 - x_3)}$$

$$At\ x = x_3 \quad y(x_3) = a_3(x_3 - x_1)(x_3 - x_2)$$

$$\Rightarrow a_3 = \frac{y_3}{(x_3 - x_1)(x_3 - x_2)}$$

## Lagrangian / Lagranges Interpolation Formula

- Substituting these values of a$_1$ , a$_2$, a$_3$ in equation [a] we get

$$y(x) = y_1 * \frac{(x - x_2)(x - x_3)}{(x_1 - x_2)(x_1 - x_3)} +$$

$$y_2 * \frac{(x - x_1)(x - x_3)}{(x_2 - x_1)(x_2 - x_3)} +$$

$$y_3 * \frac{(x - x_1)(x - x_2)}{(x_3 - x_1)(x_3 - x_2)}$$

## Lagrangian / Lagranges Interpolation Formula

$$f(x) = f(x_0) * \frac{(x - x_1)(x - x_2) \ldots (x - x_n)}{(x_0 - x_1)(x_0 - x_1) \ldots (x_0 - x_n)} +$$

$$f(x_1) * \frac{(x - x_0)(x - x_2) \ldots (x - x_n)}{(x_1 - x_0)(x_1 - x_2) \ldots (x_1 - x_n)} + \ldots +$$

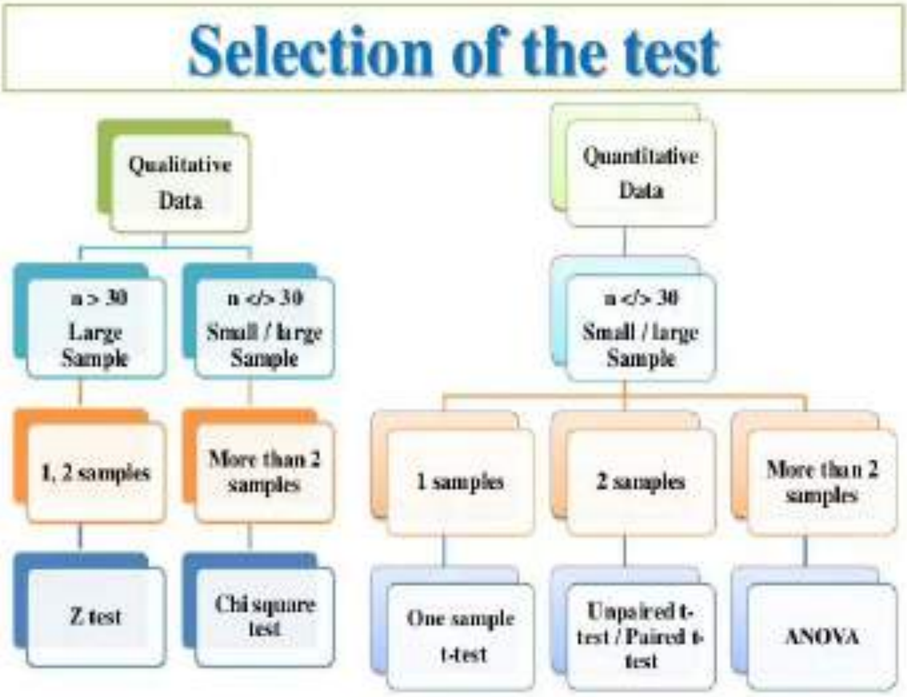
$$f(x_n) * \frac{(x - x_0)(x - x_1) \ldots (x - x_{n-1})}{(x_n - x_0)(x_n - x_1) \ldots (x_n - x_{n-1})} +$$

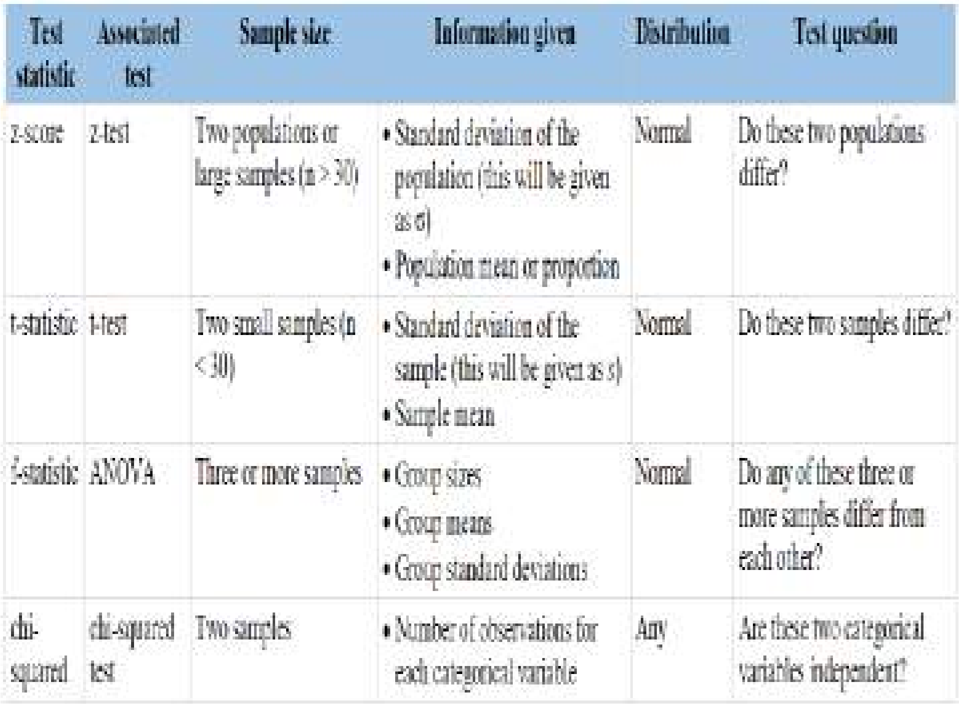
- This polynomial is known as the Lagrange’s polynomial

Minal Shah

# HYPOTESIS EXAMPLES

Computer Oriented Numerical and Statistical Methods





| Test statistic | Associated test | Sample size | Information given | Distribution | Test question |
|---|---|---|---|---|---|
| z-score | z-test | Two populations or large samples (n > 30) | • Standard deviation of the population (this will be given as σ)<br>• Population mean or proportion | Normal | Do these two populations differ? |
| t-statistic | t-test | Two small samples (n < 30) | • Standard deviation of the sample (this will be given as s)<br>• Sample mean | Normal | Do these two samples differ? |
| f-statistic | ANOVA | Three or more samples | • Group sizes<br>• Group means<br>• Group standard deviations | Normal | Do any of these three or more samples differ from each other? |
| chi-squared | chi-squared test | Two samples | • Number of observations for each categorical variable | Any | Are these two categorical variables independent? |

- A manufacturer supplies the rear axles for U.S. postal services mail trucks. These axles must be able to withstand 80,000 pounds per square inch in stress test, but an excessively strong axle raises production cost significantly. Long experience indicates that the S.D. of the strength of its axles is 4000 pounds per square inch. The manufacture selects a sample of 100 axles from production, tests them, and finds that the mean stress capacity of the sample is 79600 pounds per square inch. If the axle manufacturer uses a significance level (α) of 0.05 in testing, will the axles meet his stress requirements?

- $H_0 : \mu = 80,000$ → Null hypothesis
- $H_1 : \mu \neq 80,000$ → Alternative hypothesis
- $\alpha = 0.05$ → level of significance for testing this hypothesis.
- Calculate the standard error

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$$= \frac{4000}{\sqrt{100}}$$

$$= 400\, pounds\ per\, square\, inch$$

- Determining the limits of acceptance region :
  - 0.95 (1 – 0.05) acceptance region contains two equal area of 0.475 (0.95 / 2) each.
- From the normal distribution table we can see that the appropriate z value for 0.475 of the area under the curve is ±1.96.
- Now we can determine the limits of the acceptance region

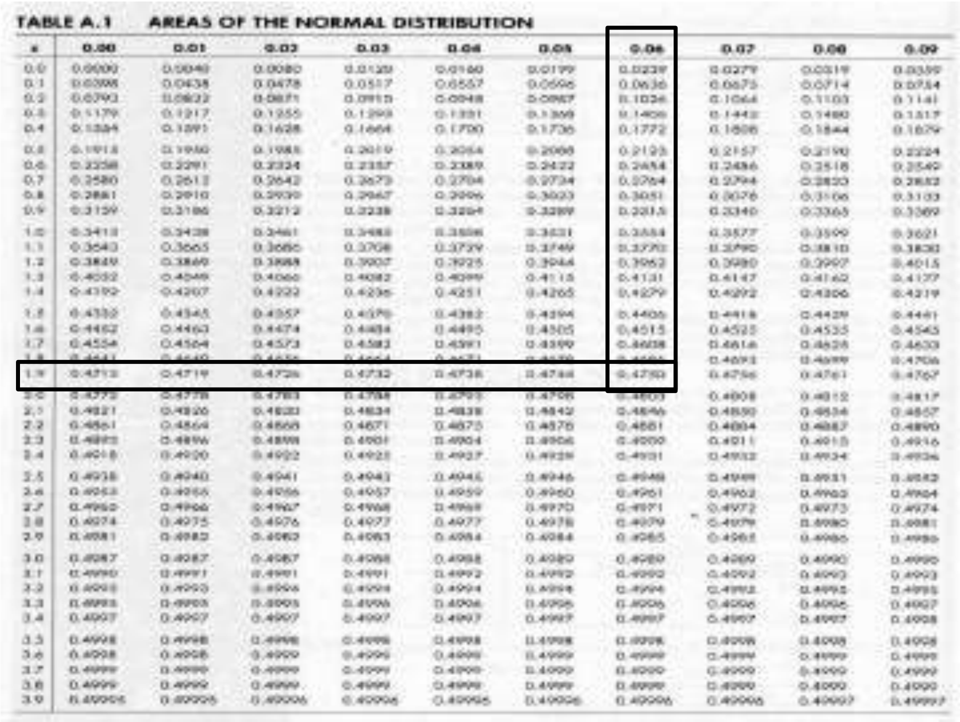
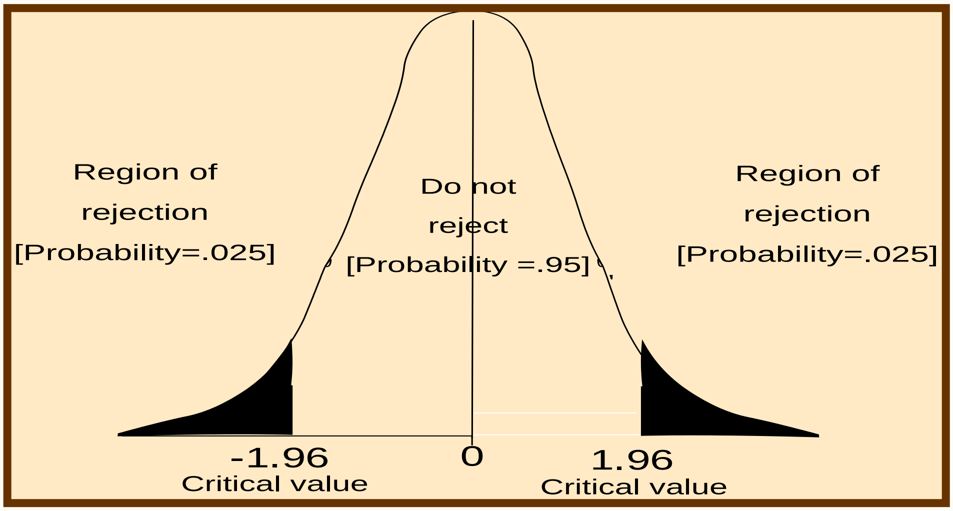
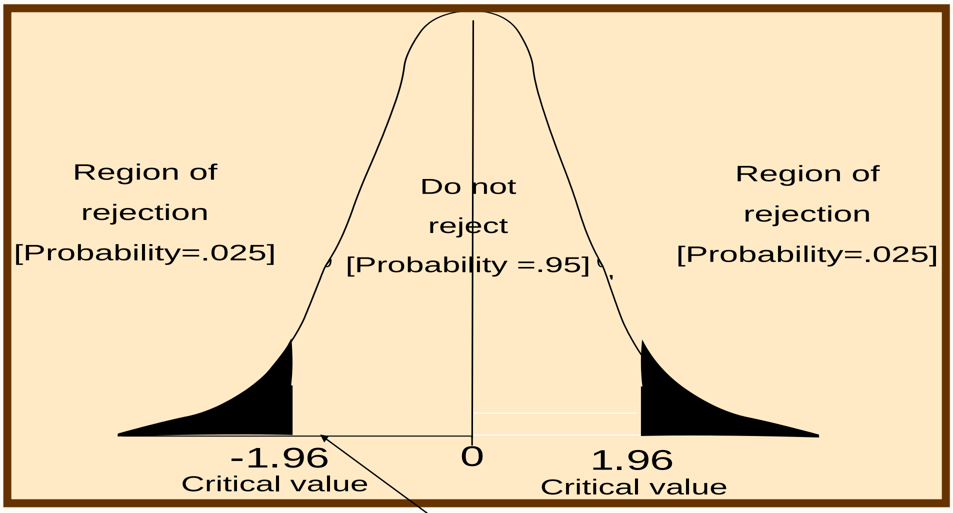


**TABLE A.1 AREAS OF THE NORMAL DISTRIBUTION**

| z | 0.00 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.0 | 0.0000 | 0.0040 | 0.0080 | 0.0120 | 0.0160 | 0.0199 | 0.0239 | 0.0279 | 0.0319 | 0.0359 |
| 0.1 | 0.0398 | 0.0438 | 0.0478 | 0.0517 | 0.0557 | 0.0596 | 0.0636 | 0.0675 | 0.0714 | 0.0754 |
| 0.2 | 0.0793 | 0.0832 | 0.0871 | 0.0910 | 0.0948 | 0.0987 | 0.1026 | 0.1064 | 0.1103 | 0.1141 |
| 0.3 | 0.1179 | 0.1217 | 0.1255 | 0.1293 | 0.1331 | 0.1368 | 0.1406 | 0.1443 | 0.1480 | 0.1517 |
| 0.4 | 0.1554 | 0.1591 | 0.1628 | 0.1664 | 0.1700 | 0.1736 | 0.1772 | 0.1808 | 0.1844 | 0.1879 |
| 0.5 | 0.1915 | 0.1950 | 0.1985 | 0.2019 | 0.2054 | 0.2088 | 0.2123 | 0.2157 | 0.2190 | 0.2224 |
| 0.6 | 0.2258 | 0.2291 | 0.2324 | 0.2357 | 0.2389 | 0.2422 | 0.2454 | 0.2486 | 0.2518 | 0.2549 |
| 0.7 | 0.2580 | 0.2612 | 0.2642 | 0.2673 | 0.2704 | 0.2734 | 0.2764 | 0.2794 | 0.2823 | 0.2852 |
| 0.8 | 0.2881 | 0.2910 | 0.2939 | 0.2967 | 0.2996 | 0.3023 | 0.3051 | 0.3078 | 0.3106 | 0.3133 |
| 0.9 | 0.3159 | 0.3186 | 0.3212 | 0.3238 | 0.3264 | 0.3289 | 0.3315 | 0.3340 | 0.3365 | 0.3389 |
| 1.0 | 0.3413 | 0.3438 | 0.3461 | 0.3485 | 0.3508 | 0.3531 | 0.3554 | 0.3577 | 0.3599 | 0.3621 |
| 1.1 | 0.3643 | 0.3665 | 0.3686 | 0.3708 | 0.3729 | 0.3749 | 0.3770 | 0.3790 | 0.3810 | 0.3830 |
| 1.2 | 0.3849 | 0.3869 | 0.3888 | 0.3907 | 0.3925 | 0.3944 | 0.3962 | 0.3980 | 0.3997 | 0.4015 |
| 1.3 | 0.4032 | 0.4049 | 0.4066 | 0.4082 | 0.4099 | 0.4115 | 0.4131 | 0.4147 | 0.4162 | 0.4177 |
| 1.4 | 0.4192 | 0.4207 | 0.4222 | 0.4236 | 0.4251 | 0.4265 | 0.4279 | 0.4292 | 0.4306 | 0.4319 |
| 1.5 | 0.4332 | 0.4345 | 0.4357 | 0.4370 | 0.4382 | 0.4394 | 0.4406 | 0.4418 | 0.4429 | 0.4441 |
| 1.6 | 0.4452 | 0.4463 | 0.4474 | 0.4484 | 0.4495 | 0.4505 | 0.4515 | 0.4525 | 0.4535 | 0.4545 |
| 1.7 | 0.4554 | 0.4564 | 0.4573 | 0.4582 | 0.4591 | 0.4599 | 0.4608 | 0.4616 | 0.4625 | 0.4633 |
| 1.8 | 0.4641 | 0.4649 | 0.4656 | 0.4664 | 0.4671 | 0.4678 | 0.4686 | 0.4693 | 0.4699 | 0.4706 |
| 1.9 | 0.4713 | 0.4719 | 0.4726 | 0.4732 | 0.4738 | 0.4744 | 0.4750 | 0.4756 | 0.4761 | 0.4767 |
| 2.0 | 0.4772 | 0.4778 | 0.4783 | 0.4788 | 0.4793 | 0.4798 | 0.4803 | 0.4808 | 0.4812 | 0.4817 |
| 2.1 | 0.4821 | 0.4826 | 0.4830 | 0.4834 | 0.4838 | 0.4842 | 0.4846 | 0.4850 | 0.4854 | 0.4857 |
| 2.2 | 0.4861 | 0.4864 | 0.4868 | 0.4871 | 0.4875 | 0.4878 | 0.4881 | 0.4884 | 0.4887 | 0.4890 |
| 2.3 | 0.4893 | 0.4896 | 0.4898 | 0.4901 | 0.4904 | 0.4906 | 0.4909 | 0.4911 | 0.4913 | 0.4916 |
| 2.4 | 0.4918 | 0.4920 | 0.4922 | 0.4925 | 0.4927 | 0.4929 | 0.4931 | 0.4932 | 0.4934 | 0.4936 |
| 2.5 | 0.4938 | 0.4940 | 0.4941 | 0.4943 | 0.4945 | 0.4946 | 0.4948 | 0.4949 | 0.4951 | 0.4952 |
| 2.6 | 0.4953 | 0.4955 | 0.4956 | 0.4957 | 0.4959 | 0.4960 | 0.4961 | 0.4962 | 0.4963 | 0.4964 |
| 2.7 | 0.4965 | 0.4966 | 0.4967 | 0.4968 | 0.4969 | 0.4970 | 0.4971 | 0.4972 | 0.4973 | 0.4974 |
| 2.8 | 0.4974 | 0.4975 | 0.4976 | 0.4977 | 0.4977 | 0.4978 | 0.4979 | 0.4979 | 0.4980 | 0.4981 |
| 2.9 | 0.4981 | 0.4982 | 0.4982 | 0.4983 | 0.4984 | 0.4984 | 0.4985 | 0.4985 | 0.4986 | 0.4986 |
| 3.0 | 0.4987 | 0.4987 | 0.4987 | 0.4988 | 0.4988 | 0.4989 | 0.4989 | 0.4989 | 0.4990 | 0.4990 |
| 3.1 | 0.4990 | 0.4991 | 0.4991 | 0.4991 | 0.4992 | 0.4992 | 0.4992 | 0.4992 | 0.4993 | 0.4993 |
| 3.2 | 0.4993 | 0.4993 | 0.4994 | 0.4994 | 0.4994 | 0.4994 | 0.4994 | 0.4995 | 0.4995 | 0.4995 |
| 3.3 | 0.4995 | 0.4995 | 0.4995 | 0.4996 | 0.4996 | 0.4996 | 0.4996 | 0.4996 | 0.4996 | 0.4997 |
| 3.4 | 0.4997 | 0.4997 | 0.4997 | 0.4997 | 0.4997 | 0.4997 | 0.4997 | 0.4997 | 0.4997 | 0.4998 |
| 3.5 | 0.4998 | 0.4998 | 0.4998 | 0.4998 | 0.4998 | 0.4998 | 0.4998 | 0.4998 | 0.4998 | 0.4998 |
| 3.6 | 0.4998 | 0.4998 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 |
| 3.7 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 |
| 3.8 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 |
| 3.9 | 0.49995 | 0.49995 | 0.49996 | 0.49996 | 0.49996 | 0.49996 | 0.49996 | 0.49996 | 0.49997 | 0.49997 |


Region of rejection [Probability=.025]
Do not reject [Probability =.95]
Region of rejection [Probability=.025]
-1.96
Critical value
0
1.96
Critical value

$$Z = \frac{\bar{x} - \mu}{\sigma\bar{x}} = \frac{79600-80000}{400} = -1.00\ (\sigma\bar{x} = \sqrt{n})$$



$$Z = \frac{\bar{x} - \mu}{\sigma\bar{x}} = \frac{79600-80000}{400} = -1.00 \ (\sigma\bar{x} = \sqrt{n})$$

Manufacturer should accept the production run as meeting stress requirement.

- A Hospital uses large quantities of packaged doses of a particular drug. The individual dose of this drug is 100 cubic cm (100 cc). The action of the drug is such that the body will harmlessly pass off excessive doses on the other hand, insufficient doses so not produce the desired medical effect, and they interfere with patient treatment. The hospital has purchased this drug from the same manufacturer for a number of years and knows that the population S.D. is 2cc. The hospital inspects 50 doses of this drug at random from a very shipment and finds the mean of these doses to be 99.75cc. If the hospital sets a 0.10 significance level and asks us whether the dosages in this shipment are too small, how can we find the answer.

- $H_0 : \mu \geq 100$ → Null hypothesis
- $H_1 : \mu < 100$ → Alternative hypothesis
- $\alpha = 0.10$ → level of significance for testing this hypothesis.
- Choose the appropriate distribution and find the critical value.
  - Here population S.D. is known and n is larger than 30 we use the normal distribution.
  - From the table we can determine that the value of z for 40% (50% - 10% significance) of the area under the curve is 1.28
  - So the critical value for lower tailed test is -1.28 i.e. left tailed test
- Compute the standard error and standardize the sample statistics.

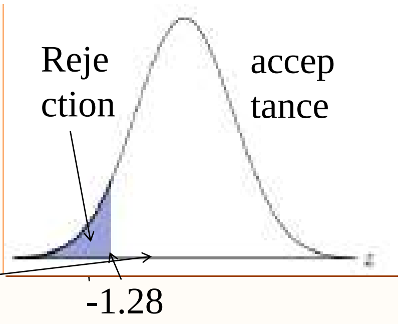
$$\sigma\bar{x} = \frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{50}} = 0.2829cc$$

- Now we use the equation .

$$z = \frac{\bar{x} - \mu H_0}{\sigma \bar{x}}$$

$$= \frac{99.75 - 100}{0.2829}$$

$$= -0.88$$



- Sketch the distribution and mark the sample values and the critical value.
- Placing the standardized value on the z scale shows that this sample mean falls well within the acceptance region.
- Interpret the results : Hospital should accept the null hypothesis.

- A personnel specialist of a major corporation is recruiting a large number of employees for an overseas assignment. During the testing process, management asks how things are going and she replies. “Fine, I think the average score on the aptitude test will be around 90”. When management reviews 20 of the test results compiled, it finds that the mean score is 84 and the S.D. of this score is 11. If management wants to test her hypothesis at the 0.1 level of significance, what is the procedure?

- State your hypothesis type of test, and significance level.
  - $H_0 : \mu = 90 \rightarrow$ Null hypothesis
  - $H_1 : \mu \neq 90 \rightarrow$ Alternative hypothesis
  - $\alpha = 0.10 \rightarrow$ level of significance for testing this hypothesis.
- Choose the appropriate distribution and find the critical value
  - Because the management is interested in knowing whether the true mean score, a two-tailed test is appropriate one to use.
  - The significance level is 0.1, so two area, each containing 0.05 of the area under the t distribution.

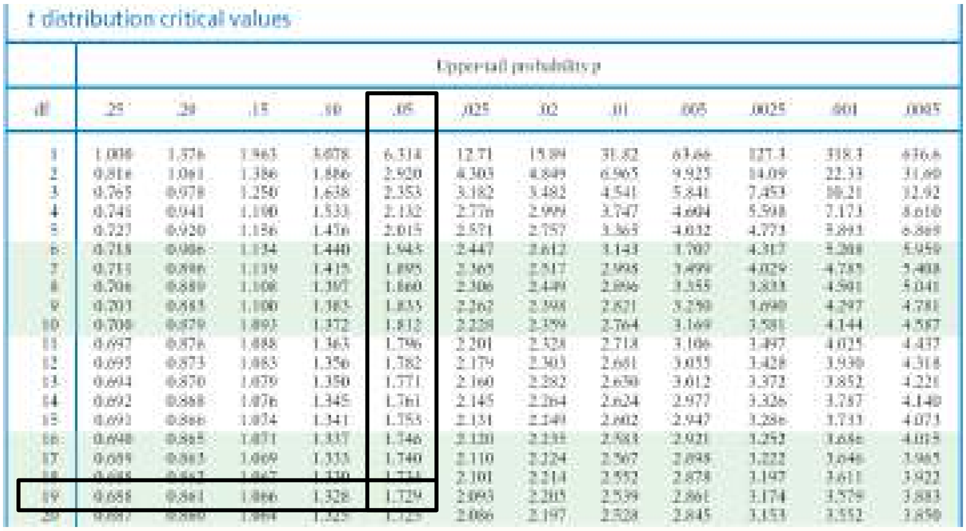
- Because the sample size is 20 i.e. n = 20 the appropriate number of degrees of freedom i.e. n-1 = 19, that is 20 -1.
- From the t table the critical value is 1.729
- Here population S.D. is not known so estimate it using the sample S.D. and equation

$$\hat{\sigma} = S = 11$$

- Compute the standard error and standardize the sample statistics.
- $\hat{\sigma}_{\bar{x}} = \dfrac{\hat{\sigma}}{\sqrt{n}}$

$$= \frac{11}{\sqrt{20}}$$

$$= 2.46$$



t distribution critical values

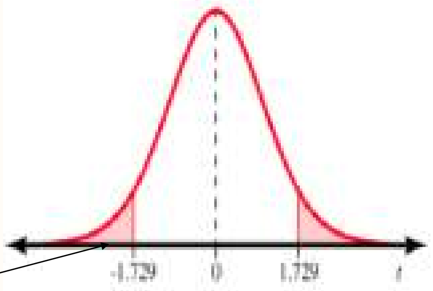
| df | Upper-tail probability p |  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  | .25 | .20 | .15 | .10 | .05 | .025 | .02 | .01 | .005 | .0025 | .001 | .0005 |
| 1 | 1.000 | 1.376 | 1.963 | 3.078 | 6.314 | 12.71 | 15.89 | 31.82 | 63.66 | 127.3 | 318.3 | 636.6 |
| 2 | 0.816 | 1.061 | 1.386 | 1.886 | 2.920 | 4.303 | 4.849 | 6.965 | 9.925 | 14.09 | 22.33 | 31.60 |
| 3 | 0.765 | 0.978 | 1.250 | 1.638 | 2.353 | 3.182 | 3.482 | 4.541 | 5.841 | 7.453 | 10.21 | 12.92 |
| 4 | 0.741 | 0.941 | 1.190 | 1.533 | 2.132 | 2.776 | 2.999 | 3.747 | 4.604 | 5.598 | 7.173 | 8.610 |
| 5 | 0.727 | 0.920 | 1.156 | 1.476 | 2.015 | 2.571 | 2.757 | 3.365 | 4.032 | 4.773 | 5.893 | 6.869 |
| 6 | 0.718 | 0.906 | 1.134 | 1.440 | 1.943 | 2.447 | 2.612 | 3.143 | 3.707 | 4.317 | 5.208 | 5.959 |
| 7 | 0.711 | 0.896 | 1.119 | 1.415 | 1.895 | 2.365 | 2.517 | 2.998 | 3.499 | 4.029 | 4.785 | 5.408 |
| 8 | 0.706 | 0.889 | 1.108 | 1.397 | 1.860 | 2.306 | 2.449 | 2.896 | 3.355 | 3.833 | 4.501 | 5.041 |
| 9 | 0.703 | 0.883 | 1.100 | 1.383 | 1.833 | 2.262 | 2.398 | 2.821 | 3.250 | 3.690 | 4.297 | 4.781 |
| 10 | 0.700 | 0.879 | 1.093 | 1.372 | 1.812 | 2.228 | 2.359 | 2.764 | 3.169 | 3.581 | 4.144 | 4.587 |
| 11 | 0.697 | 0.876 | 1.088 | 1.363 | 1.796 | 2.201 | 2.328 | 2.718 | 3.106 | 3.497 | 4.025 | 4.437 |
| 12 | 0.695 | 0.873 | 1.083 | 1.356 | 1.782 | 2.179 | 2.303 | 2.681 | 3.055 | 3.428 | 3.930 | 4.318 |
| 13 | 0.694 | 0.870 | 1.079 | 1.350 | 1.771 | 2.160 | 2.282 | 2.650 | 3.012 | 3.372 | 3.852 | 4.221 |
| 14 | 0.692 | 0.868 | 1.076 | 1.345 | 1.761 | 2.145 | 2.264 | 2.624 | 2.977 | 3.326 | 3.787 | 4.140 |
| 15 | 0.691 | 0.866 | 1.074 | 1.341 | 1.753 | 2.131 | 2.249 | 2.602 | 2.947 | 3.286 | 3.733 | 4.073 |
| 16 | 0.690 | 0.865 | 1.071 | 1.337 | 1.746 | 2.120 | 2.235 | 2.583 | 2.921 | 3.252 | 3.686 | 4.015 |
| 17 | 0.689 | 0.863 | 1.069 | 1.333 | 1.740 | 2.110 | 2.224 | 2.567 | 2.898 | 3.222 | 3.646 | 3.965 |
| 18 | 0.688 | 0.862 | 1.067 | 1.330 | 1.734 | 2.101 | 2.214 | 2.552 | 2.878 | 3.197 | 3.611 | 3.922 |
| 19 | 0.688 | 0.861 | 1.066 | 1.328 | 1.729 | 2.093 | 2.205 | 2.539 | 2.861 | 3.174 | 3.579 | 3.883 |
| 20 | 0.687 | 0.860 | 1.064 | 1.325 | 1.725 | 2.086 | 2.197 | 2.528 | 2.845 | 3.153 | 3.552 | 3.850 |

- Standardize statistics

$$t = \frac{\overline{x} - \mu H_0}{\hat{\sigma}\overline{x}}$$

$$= \frac{84 - 90}{2.46}$$

$$= -2.44$$



- Sketch the distribution and mark the sample value and the critical value.
- Drawing the results on a sketch of the sampling distribution, we see that the sample mean falls outside the acceptance region.
- Interpret the result : Management should reject the null hypothesis.

THANKS!
Thanks!
Thanks
THANKS
Thanks
THANKS
THANKS
THANKS

# Numerical Integration

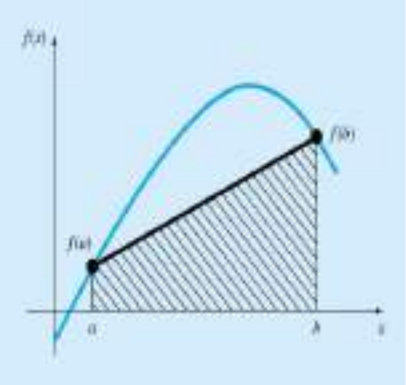
Computer Oriented Numerical and Statistical Methods

## Outline

- Introduction
- Trapezoidal Rule
- Simpson’s Rule

## Introduction

- The concept of definite integral is essential for this, the calculation regarding area of regions bounded by curve, volume of solid figures, centre of gravity, length of arc, work, velocity, movement of interia etc. are the problems of definite integrals.
- Suppose f is a positive functio and increasing in the interval
- Hence for a ≤ x ≤ b we have f the graph of y = f(x) in the inte continuous curve.



## Introduction

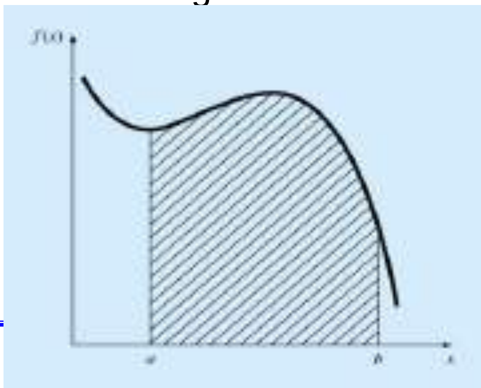
- The definite integral of function f on the interval [a,b] and represented symbolically as $\int_{a}^{b} f(x)dx$

  where a is called the lower limit and b is called the upper limit of given integral.
- Numerical integration is the process of computing the value of a definite integral from a set of numerical values of the function referred to as integrand.
- It is important because there are many integrals whose value cannot be obtained in closed form.
- The integrals have to be computed numerically.

# Introduction

- When we applied to the integration of a function of a single variable, the process is sometimes called numerical integration or numerical quadrature.
- Graphically integration is simply to find the area under a certain curve between the 2 integration limits.

$$I = \int_a^b f(x).dx = A$$



# Numerical Integration

- In practice, given set of values for a function f(x) the following table of data.

| **X** | a | $x_1$ | $x_2$ | … | b |
|---|---|---|---|---|---|
| **f(x)** | f(a) | $f(x_1)$ | $f(x_2)$ | … | f(b) |

is used to compute the value of the integral $\int_a^b f(x)dx$

- By dividing the interval (a,b) into a finite number of equal intervals, finding the areas of those subintervals and summing all such area of the required integration is possible

## Quadrature Formula

$$\int_a^b f(x)dx = h[ny_0 + \frac{n^2}{2}\Delta y_0 + (\frac{n^3}{3} - \frac{n^2}{2})\frac{\Delta^2 y_0}{2} +$$

$$(\frac{n^4}{4} - n^3 + n^2)\frac{\Delta^3 y_0}{6} \;......]$$

where $x_1 - x_0 = x_2 - x_1 = ...... \; x_n - x_{n-1} = h$
$\Delta y_0 = y_1 - y_0$
$\Delta y_1 = y_2 - y_1$

## Trapezoidal Rule

- Take n = 1 in the quadrature formula we get trapezoidal rule.
- As n = 1 the values of $\Delta^2 y_0$ , $\Delta^3 y_0$ , $\Delta^4 y_0$ etc is zero

$$\int_a^b f(x)dx = h[y_0 + \frac{1}{2}\Delta y_0]$$

$$= \frac{h}{2}[2 * y_0 + \Delta y_0]$$

$$= \frac{h}{2}[2 * y_0 + y_1 - y_0]$$

$$= \frac{h}{2}[y_0 + y_1]$$

$$\int_{x_0}^{x_0 + h} f(x)dx = = \frac{h}{2}[y_0 + y_1] \quad ...........(1)$$

## Trapezoidal Rule

- Similarly

$$\int_{x_0+h}^{x_0+2h} f(x)dx = \frac{h}{2}[y_1 + y_2] \quad ...........(2)$$

$$\int_{x_0+2h}^{x_0+3h} f(x)dx = \frac{h}{2}[y_2 + y_3] \quad ...........(3)$$

$$\vdots$$

$$\int_{x_0+(n-1)h}^{x_0+nh} f(x)dx = \frac{h}{2}[y_{n-1} + y_n] \quad ...........(n)$$

## Trapezoidal Rule

- Add (1), (2), (3) , ……, (n)

$$\int_{x_0}^{x_0+h} f(x)dx + \int_{x_0+h}^{x_0+2h} f(x)dx + \int_{x_0+2h}^{x_0+3h} f(x)dx + ..... + \int_{x_0+(n-1)h}^{x_0+nh} f(x)dx$$

$$= \frac{h}{2}[y_0 + y_1 + y_1 + y_2 + ...... + y_{n-1} + y_n]$$

# Trapezoidal Rule

$$\int_{x_0}^{x_0+nh} f(x)dx = \frac{h}{2}[y_0 + 2*(y_1 + y_2 + ...... + y_{n-1}) + y_n]$$

- This formula is called trapezoidal rule.

# Trapezoidal Rule

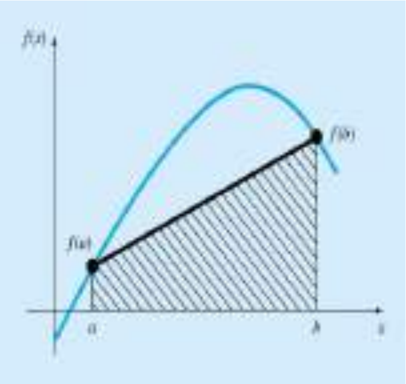
- The trapezoidal rule uses a polynomial of the first degree to replace the function to be integrated.

$$I = \int_a^b f(x)\,dx \cong \int_a^b f_1(x)\,dx$$

$$f_1(x) = f(a) + \frac{f(b)-f(a)}{b-a}(x-a)$$

$$I = \int_a^b \left[ f(a) + \frac{f(b)-f(a)}{b-a}(x-a) \right] dx$$

$$\boxed{I = (b-a)\frac{f(a)+f(b)}{2}}$$



## Trapezoidal Rule (Example)

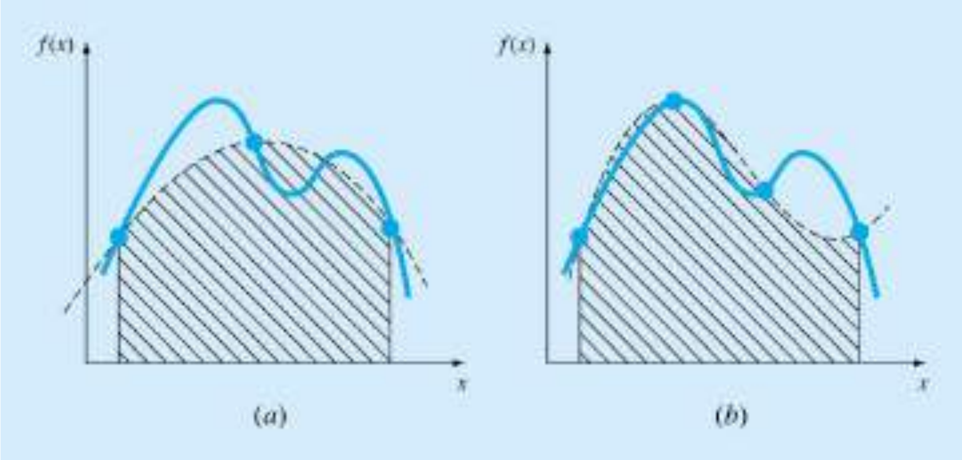
- Evaluate $$\int_0^1 \frac{dx}{1+x^2}$$
- By the trapezoidal rule where the interval (0,1) is sub-divided into 6 equal parts.
- The general trapezoidal formula is

$$=\frac{h}{2}[y_0+2y_1+2y_2+2y_3+.......+2y_{n-1}+y_n]$$

## Simpson's Rule

- One drawback of the trapezoidal rule is that the error is related to the second derivative of the function.
- More accurate estimate of an integral is obtained if a high-order polynomial is used to connect the points.
- The formulas that result from taking the integrals under such polynomials are called ***Simpson's Rules***.
- **Simpson's 1/3 Rule :** Results when a second-order interpolating polynomial is used.
- **Simpson's 3/8 Rule :** Results when a third-order (cubic) interpolating polynomial is used.

## Simpson’s Rules

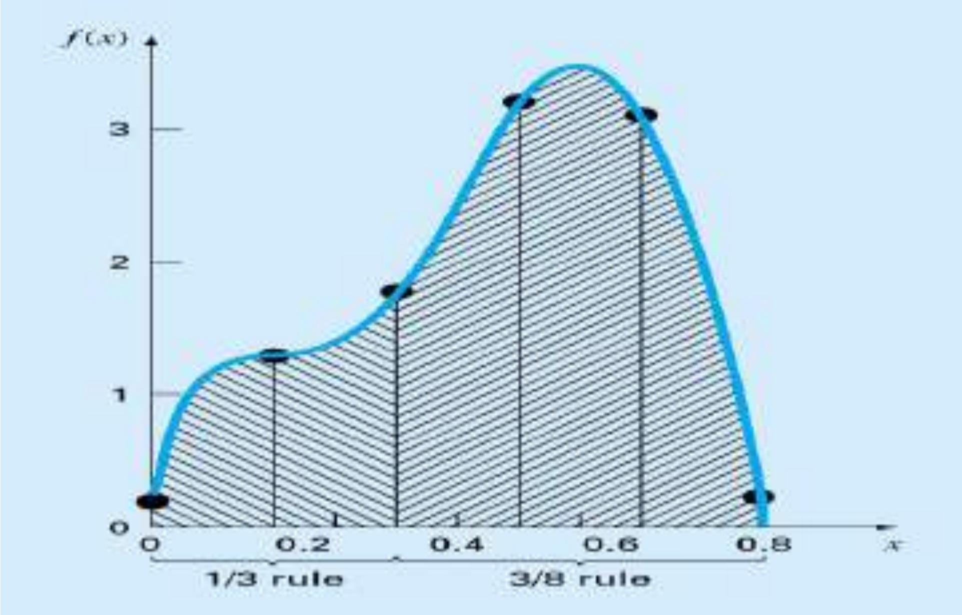


Simpson’s 1/3 Rule

Simpson’s 3/8 Rule

Minal Shah

## Simpson’s Rules



## Simpson's 1/3 Rule

- By putting n = 2 in the quadrature formula we get Simpson's 1/3 rule.
- If n = 2 then the value of $\Delta^3 y_0$ , $\Delta^4 y_{0,\ldots}$ etc is zero

$$\int_a^b f(x)dx = h[2y_0 + 2\Delta y_0 + (\tfrac{8}{3} - 2)\frac{\Delta^2 y_0}{2}]$$

$$= h\,[2*y_0 + 2(y_1 - y_0) + \tfrac{2}{3}\frac{\Delta^2 y_0}{2}]$$

$$= h\,[2*y_0 + 2*y_1 - 2*y_0 + \tfrac{1}{3}\Delta^2 y_0]$$

$$= h\,[2*y_0 + 2*y_1 - 2*y_0) + \tfrac{1}{3}(y_2 - 2*y_1 + y_0)]$$

## Simpson's 1/3 Rule

$$\int_a^b f(x)dx = h[2y_0 + 2\Delta y_0 + (\tfrac{8}{3} - 2)\frac{\Delta^2 y_0}{2}]$$

$$= h\,[2*y_0 + 2(y_1 - y_0) + \tfrac{2}{3}\frac{\Delta^2 y_0}{2}]$$

$$= h\,[2*y_0 + 2*y_1 - 2*y_0 + \tfrac{1}{3}\Delta^2 y_0]$$

$$= h\,[(2*y_0 + 2*y_1 - 2*y_0) + \tfrac{1}{3}(y_2 - 2*y_1 + y_0)]$$

$$= \frac{h}{3}\,[6*y_1 + y_2 - 2*y_1 + y_0]$$

$$= \frac{h}{3}\,[y_2 + 4*y_1 + y_0]$$

## Simpson's 1/3 Rule

$$\int_{x_0}^{x_0+2h} f(x)dx = \frac{h}{3}[y_0 + 4y_1 + y_2]........(1)$$

$$\int_{x_0+2h}^{x_0+4h} f(x)dx = \frac{h}{3}[y_2 + 4y_3 + y_4]........(2)$$

$$\int_{x_0+4h}^{x_0+6h} f(x)dx = \frac{h}{3}[y_4 + 4y_5 + y_6]........(3)$$

$$\vdots$$

$$\int_{x_0+(n-2)h}^{x_0+nh} f(x)dx = \frac{h}{3}[y_{n-2} + 4y_{n-1} + y_n]........(n)$$

## Take sum of integration

$$\int_{x_0}^{x_0+2h} f(x)dx + \int_{x_0+2h}^{x_0+4h} f(x)dx + \int_{x_0+4h}^{x_0+6h} f(x)dx + ....+$$

$$\int_{x_0+(n-2)h}^{x_0+nh} f(x)dx = \frac{h}{3}[(y_0 + 4y_1 + y_2) + (y_2 + 4y_3 + y_4) +$$

$$(y_4 + 4y_5 + y_6) + ....+(y_{n-2} + 4y_{n-1} + y_n)]$$

## Simpson's  1/3 Rule

$$\int_{x_0}^{x_0+nh} f(x)dx = \frac{h}{3}[y_0 + 4(y_1 + y_3 + y_5 + .... + y_{n-1})$$
$$+ 2(y_2 + y_4 + ... + y_{n-2}) + y_n]$$

- Note : It should be noted that this rule requires the divisions of the whole range into an even number of subintervals of width h.

## Simpson's  3/8 Rule

- By putting n = 3 in the quadrature formula we get Simpson's  3/8 rule.
- If n = 3 then the value of  $\Delta^4 y_0$ , $\Delta^5 y_{0,....}$  etc is zero

$$\int_a^b f(x)dx = h[ny_0 + \frac{n^2}{2}\Delta y_0 + (\frac{n^3}{3} - \frac{n^2}{2})\frac{\Delta^2 y_0}{2} + (\frac{n^4}{4} - n^3 + n^2)\frac{\Delta^3 y_0}{6}]$$

$$\int_a^b f(x)dx = h[3y_0 + \frac{9}{2}\Delta y_0 + (9 - \frac{9}{2})\frac{\Delta^2 y_0}{2} + (\frac{81}{4} - 27 + 9)\frac{\Delta^3 y_0}{6}]$$

## Simpson’s 3/8 Rule

$$\int_a^b f(x)dx = h[3y_0 + \frac{9}{2}(y_1 - y_0) + (9 - \frac{9}{2})(\frac{y_2 - 2y_1 + y_0}{2})$$

$$+ (\frac{81}{4} - 27 + 9)(\frac{y_3 - 3y_2 + 3y_1 - y_0}{6})]$$

$$\int_a^b f(x)dx = \frac{3h}{8}[8y_0 + 12(y_1 - y_0) + 6( y_2 - 2y_1 + y_0)$$

$$+ (y_3 - 3y_2 + 3y_1 - y_0)]$$

## Simpson’s 3/8 Rule

$$\int_{x_0}^{x_0+3h} f(x)dx = \frac{3h}{8}[y_0 + 3(y_1 + y_2) + y_3]........(1)$$

$$\int_{x_0+3h}^{x_0+6h} f(x)dx = \frac{3h}{8}[y_3 + 3(y_4 + y_5) + y_6]........(2)$$

$$\int_{x_0+6h}^{x_0+9h} f(x)dx = \frac{3h}{8}[y_6 + 3(y_7 + y_8) + y_9]........(3)$$

## Simpson’s 3/8 Rule

- 
- 
-

$$\int_{x_0+(n-3)h}^{x_0+nh} f(x)dx = \frac{3h}{8}[y_{n-3} + 3(y_{n-2}+y_{n-1}) + y_n]........(n)$$

Minal Shah

## Add all integration

$$\int_{x_0}^{x_0+3h} f(x)dx + \int_{x_0+3h}^{x_0+6h} f(x)dx + ..... +$$

$$\int_{x_0+(n-3)h}^{x_0+nh} f(x)dx = \frac{3h}{8}[y_0 + 3(y_1+y_2) + y_3 +$$

$$y_3 + 3(y_4+y_5) + y_6 +$$

$$..... + y_{n-3} + 3(y_{n-2}+y_{n-1}) + y_n]$$

Minal Shah

## Simpson’s 3/8 Rule

$$\int_{x_0}^{x_0+nh} f(x)dx = \frac{3h}{8}[y_0 + 3(y_1 + y_2 + y_4 + y_5 + \ldots. + y_{n-2} + y_{n-1}) + 2(y_3 + y_6 + \ldots + y_{n-3}) + y_n]$$

- Note : It should be noted that the entire range must be divided into an even number of subintervals of width h Or Total number of points must be odd.

# Skewness

Computer Oriented Numerical and Statistical Methods

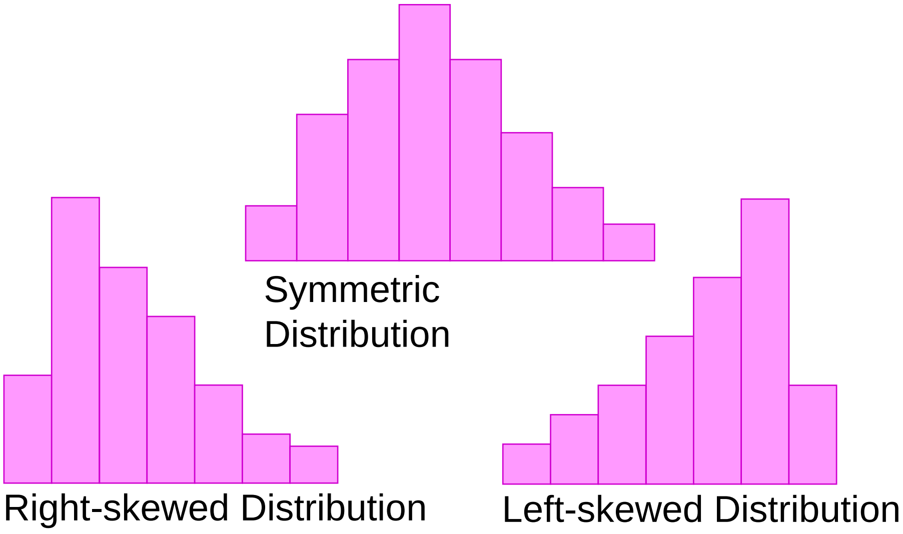
## Outline

- Introduction
- Types of Skewness
- Measure of skewness.
- Karl Pearson’s Measure
- Bowley’s Measure

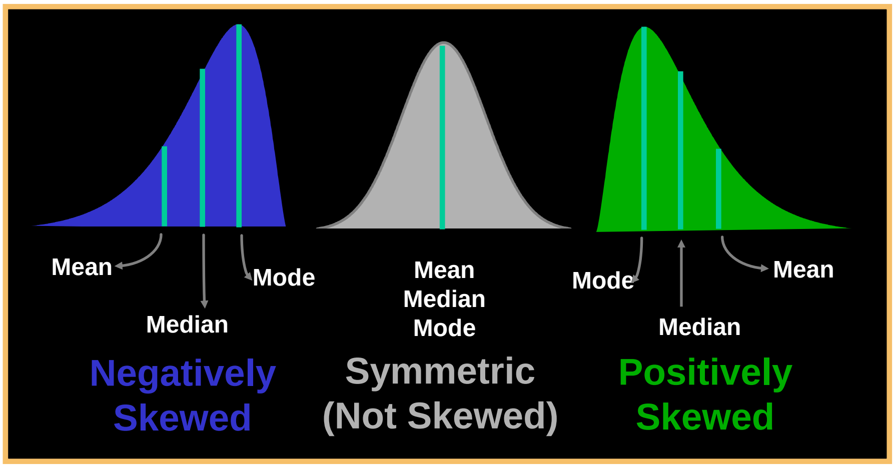
## Introduction

- Measure of central tendency gives us an estimate of the representative value of a series, the measure of dispersion gives an indication of the extent to which the items cluster around or scatter away from the central value and the skewness is a measure that refers to the extent of symmetry or asymmetry in a distribution.
- It describes the shape of a distribution,

## Skewness In The Form Of Histogram

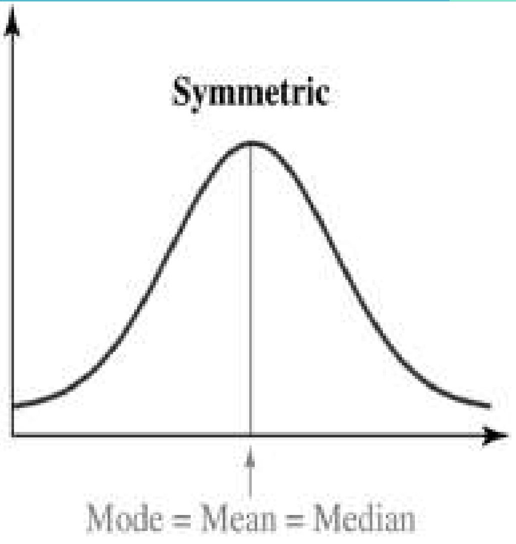


## Skewness



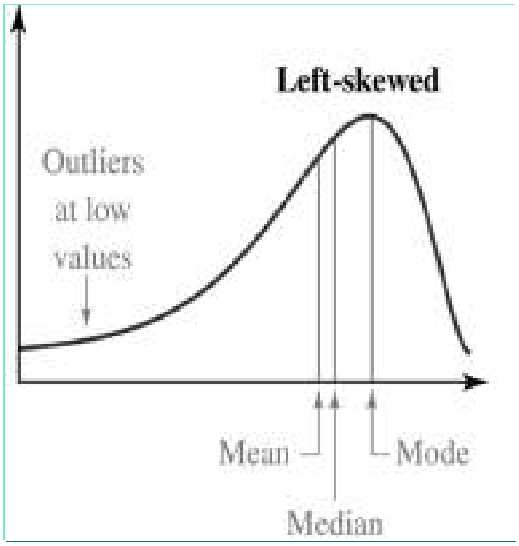
## What is the relationship between mean, median and mode of a skewed distribution?

- Find the mean median and mode of: 1, 2, 2, 3, 3, 3, 4, 4, 4, 4, 4, 4, 5, 5, 5, 6, 6, 7
- Mean is 4.
- Median is 4.
- Mode is 4.



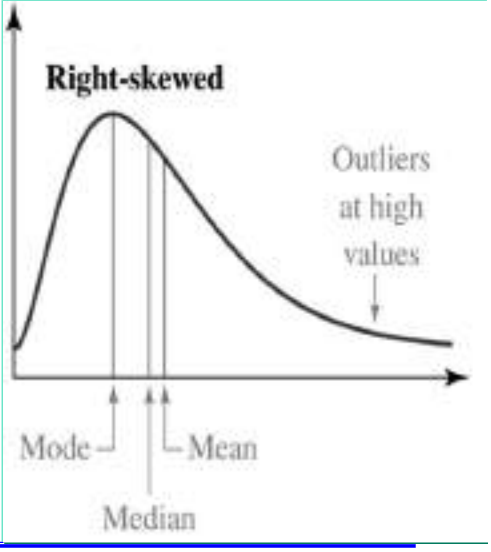
## What is the relationship between mean, median and mode of a left-skewed distribution?

- Find the mean, median and mode of:
  0, 5, 10, 20, 40, 45, 45, 50, 50, 50, 60, 60, 60, 60, 60, 60, 70, 70, 70, 70, 70, 70, 70, 70
- The mean is 51.5.
- The median is 60.
- The mode is 70.



## What is the relationship between mean, median and mode of a right-skewed distribution?

- Find the mean, median, and mode of:
  20, 20, 20, 20, 20, 20, 20, 20, 30, 30, 30, 30, 30, 30, 45, 45, 45, 50, 50, 60, 70, 90
- The mean is 36.1.
- The median is 30.
- The mode is 20.



## Various Distribution And The Position Of Average

| Size | Frequency (c) | Frequency (b) | Frequency (a) |
|---|---|---|---|
| 0 – 5 | 10 | 10 | 10 |
| 5 – 10 | 90 | 30 | 20 |
| 10 – 15 | 50 | 50 | 30 |
| 15 – 20 | 40 | 70 | 40 |
| 20 – 25 | 30 | 50 | 50 |
| 25 – 30 | 20 | 30 | 90 |
| 30 – 35 | 10 | 10 | 10 |
| skewness | positive | symmetry | Negative |
| average | $\bar{x} > M_d > M_0$ | $\bar{x} = M_d = M_0$ | $\bar{x} < M_d < M_0$ |
| Quartiles | $Q_3 – M_d > M_d – Q_1$ | $Q_3 – M_d = M_d – Q_1$ | $Q_3 – M_d < M_d – Q_1$ |
| curve | Skewed to the right | Normal | Skewed to the left |

## Objectives of Skewness

- It helps in finding out the nature and the degree of concentration whether it is in higher or the lower values.
- The empirical relations of mean and median and mode are based on a moderately skewed distribution. The measure of skewness will reveal to what extent such empirical relationship holds good.
- It helps in knowing if the distribution is normal. Many statistical measures, such as the error of the mean are based on the assumption of a normal distribution.

## Measures of Skewness

- To find out the direction and the extent of asymmetry in a series statistical measures of skewness are employed.
- This measure can be absolute or relative.
- Absolute measure of skewness tell us the extent of asymmetry and whether it is positive or negative.
- The absolute skewness is based on the difference between mean and mode. Symbolically,

  absolute $S_k$ = Mean – Mode.
- If the value of mean is greater than the mode, skewness will be positive.

## Measures of Skewness

- If the value of mean is less than the mode, skewness will be negative.
- Absolute measure of skewness is not adequate because it cannot be used for comparison of skewness in two distributions if they are in different units, since difference between mean and mode will be in terms of units of distribution.
- For comparison purpose we use the relative measure of skewness known as coefficient of skewness.

## Measures of Skewness

- There are four types of relative measure of skewness :
  1. The Karl Pearson's Coefficient of Skewness.
  2. The Bowley's Coefficient of Skewness.
  3. The Kelly's Coefficient of Skewness.
  4. Measure of Skewness based on Moments and Kurtosis.

## Karl Pearson's Coefficient of Skewness

- Karl Pearson's Coefficient of Skewness or Pearsonian Coefficient of skewness is given by the formula:

$$S_k = \frac{\text{Mean} - \text{Mode}}{\text{StandardDeviation}}$$

- If in a particular frequency distribution, it is difficult to determine precisely the mode, or the mode is ill-defined, the coefficient of skewness can be determined by the following formula:

$$S_k = \frac{3(\text{Mean} - \text{Median})}{\text{StandardDeviation}}$$

## Karl Pearson’s Coefficient of Skewness

- Theoretically, skewness lies between the limits ± 3, but these limits are rarely attained in practice.

## Bowley’s Coefficient of Skewness

- Bowley’s coefficient of skewness also known as Quartile coefficient of skewness and is especially useful :
  - When the mode is ill-defined and extreme observations are  present in the data.
  - When the distribution has open-end classes or unequal class-interval.
- The quartile measure depends upon the fact that normally $Q_3$ and $Q_1$ are equidistance from the median, i.e. for symmetrical distribution $Q_3 – M_d = M_d – Q_1$.
- If a distribution is asymmetrical, then one quartile will be farther from the median than the other.

## Bowley’s Coefficient of Skewness

- In such a case skewness can be measured by the following formula given by Bowley :

  Skewness = $(Q_3 – M_d) – (M_d – Q_1)$

  Skewness = $Q_3 + Q_1 – 2M_d$
- If the first part is more than the second part, the skewness is positive and in the reverse situation it is negative.

## Bowley’s Coefficient of Skewness

- To make the measure a readily comparable, the coefficient of skewness is obtained by dividing it by quartile range viz $Q_3 – Q_1$

$$S_k = \frac{(Q_3 - M_d) - (M_d - Q_1)}{(Q_3 - M_d) + (M_d - Q_1)}$$

$$S_k = \frac{Q_3 + Q_1 - 2M_d}{Q_3 - Q_1}$$

## Bowley’s Coefficient of Skewness

- The range of variation under this method is ± 1.
- The main drawback of this measure is that it is based on the central 50% of the data and ignores the remaining 50% of the data towards the extremes.

# Theoretical Distribution

Computer Oriented Numerical and Statistical Methods

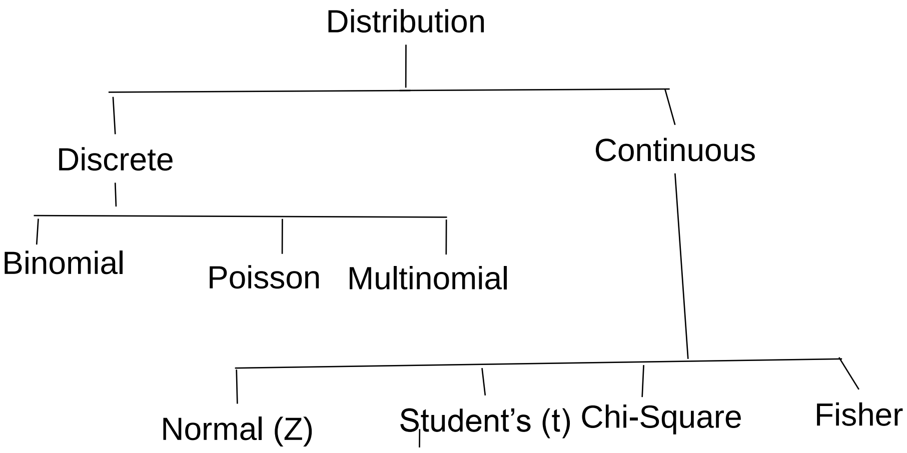
## Outline

- Introduction
- Types of Theoretical Distribution

## Introduction

- Theoretical Distribution refers to mathematical models of relative frequencies of a finite number of observations of a variable.
- It is a systematic arrangement of probabilities associated with mutually exclusive and collectively exhaustive elementary elements in an experiment.
- Where the relative frequency distributions are based on actual observations, the above distribution are based on mathematical functions.
- Important features of these distributions is that with some known parameters like mean and S.D. or the number of trials and the chances of success, the probabilities of various values of a variate can be found in the form of a complete distribution.

## Types of Theoretical Distribution



## Discrete Distribution

- The Discrete probability distributions are known as point functions defined over a sample space.
- The random variables in these takes only a finite integer value.
- These are normally represented by line graphs when not grouped and by histograms when grouped, each bar is raised on the mid-value of the class.
- The cumulative probabilities in this case are represented by a staircase type of histogram.

## Binomial Distribution

- Binomial is also known as the “Bernoulli Distribution” after the Swiss mathematician James Bernoulli (1654-1705)
- The distribution can be used under the following conditions:
  - The random experiment is performed repeatedly a finite and fixed number of times.
  - The result of any trial can be classified only under two mutually exclusive categories called success (the occurrence of event) and failure (the non-occurrence of event).

# Binomial Distribution

- – The proportion of outcomes falling in the "success" category are denoted generally by p and the proportion of item falling in the category of "failure" by q = 1 - p
- – The probability of success in each trial remains constant and does not change from one trail to another.
- – The trails are independent, so that the result of any trial in ineffective by the result of previous trials.

# Probability Function Of Binomial Distribution

- If a trial of an experiment can result in success with the probability p and failure with probability q = 1 – p, the probability of exactly x success in n trails is given by P(X=x) = P(x) = $^nC_x p^x q^{n-x}$ ; x = 0, 1, 2,…n
- The quantities n and p are called parameters of the binomial distribution and the notation b(x:n,p) reads "the binomial probability of x given n and p."
- The entire probabilities distribution of x = 0, 1, 2, …, n success with two types of expression, where p stands for success and q for non-success can be written as follows :

## Probability Function Of Binomial Distribution

| BINOMIAL PROBABILITY DISTRIBUTION | | | |
|---|---|---|---|
| No. of<br>success<br>x | $(p+q)^n$<br>Prob. p(x)<br>p(x) | No. of<br>success<br>x | $(p+q)^n$<br>Prob. p(x)<br>p(x) |
| n | $^nC_np^nq^0$ | 0 | $^nC_0p^0q^n$ |
| n-1 | $^nC_{n-1}p^{n-1}q^1$ | 1 | $^nC_1p^1q^{n-1}$ |
| n-2 | $^nC_{n-2}p^{n-2}q^2$ | 2 | $^nC_2p^2q^{n-2}$ |
| ….. | …… | ….. | …… |
| 0 | $^nC_0p^0q^n$ | n | $^nC_np^nq^0$ |
| Total | 1 | | 1 |

## Obtaining Coefficient Of The Binomial

- To find the term of the expression of $(q+p)^n$
- The first term is $q^n$
- The second term is $^nC_1q^{n-1}p$
- In each succeeding term the power of q is reduced by 1 and the power of p is increased by 1.
- The coefficient of any term is found by Pascal's triangle.
- The mean of binomial distribution is n*p
- The S.D. of binomial distribution is $\sqrt{npq}$

# Using Binomial Tables

| n = 10 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| x | … | p=.20 | p=.25 | p=.30 | **p=.35** | p=.40 | p=.45 | p=.50 | |
| 0 | … | 0.1074 | 0.0563 | 0.0282 | 0.0135 | 0.0060 | 0.0025 | 0.0010 | 10 |
| 1 | … | 0.2684 | 0.1877 | 0.1211 | 0.0725 | 0.0403 | 0.0207 | 0.0098 | 9 |
| 2 | … | 0.3020 | 0.2816 | 0.2335 | 0.1757 | 0.1209 | 0.0763 | 0.0439 | 8 |
| **3** | … | 0.2013 | 0.2503 | 0.2668 | **0.2522** | 0.2150 | 0.1665 | 0.1172 | 7 |
| 4 | … | 0.0881 | 0.1460 | 0.2001 | 0.2377 | 0.2508 | 0.2384 | 0.2051 | 6 |
| 5 | … | 0.0264 | 0.0584 | 0.1029 | 0.1536 | 0.2007 | 0.2340 | 0.2461 | 5 |
| 6 | … | 0.0055 | 0.0162 | 0.0368 | 0.0689 | 0.1115 | 0.1596 | 0.2051 | 4 |
| 7 | … | 0.0008 | 0.0031 | 0.0090 | 0.0212 | 0.0425 | 0.0746 | 0.1172 | 3 |
| 8 | … | 0.0001 | **0.0004** | 0.0014 | 0.0043 | 0.0106 | 0.0229 | 0.0439 | **2** |
| 9 | … | 0.0000 | 0.0000 | 0.0001 | 0.0005 | 0.0016 | 0.0042 | 0.0098 | 1 |
| 10 | … | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0001 | 0.0003 | 0.0010 | 0 |
| | … | p=.80 | **p=.75** | p=.70 | p=.65 | p=.60 | p=.55 | p=.50 | **x** |

Examples:

n = 10, p = 0.35, x = 3: P(x = 3|n =10, p = 0.35) = 0.2522

n = 10, p = 0.75, x = 2: P(x = 2|n =10, p = 0.75) = 0.0004

# Poisson Distribution

- In Binomial distribution it was found that there is a sample of a definite size so that it is possible to count the number of times an event is observed; in other words, n is precisely known.
- There are certain situations where this may not be possible.
- The basic reason for is that the events is the rare and causal.
- Successful events in the total events space are few e.g. the events like accidents on a road, defects in a product, goals scared at a football match, etc.

## Poisson Distribution

- In these, we known the number of times an event occurs but not how many times it does not occur.
- The total number of trials in regard to a given experiment are not precisely known.
- The Poisson distribution is very suitable in case of such rare events.
- Only the average chance of occurrence based on past experience or a small sample extracted for the purpose will enable us to construct the whole distribution.

## Poisson Distribution

- The binomial distribution requires only two parameters (p and n) this distribution requires only the value of m which is the mean of the occurrence of an event (np) based on existing knowledge on the matter.
- Poisson distribution was derived in 1837 by a French-Mathematician Simeon D. Poisson.
- Poisson distribution may be obtained as a limiting case of Binomial probability distribution under the following conditions.
  - n, the number of trails is indefinitely large i.e. n →∞

## Poisson Distribution

- p, the constant probability of success for each trial is indefinitely small i.e. p → 0
- np = m (say) is finite.

- The probability function of random variable X following Poisson Distribution is P(X=x) = p(x) = $(m^x / x!) * e^{-m}$ x = 0, 1, 2,…
  - where X = the number of successes (occurrence of the event)
  - e = 2.71828 [The base of the system of natural logarithm] and x! = x(x-1)(x-2)…3.2.1

## Poisson Distribution

| **Values of variables (x)** | **0** | **1** | **2** | **3** | **….** | **total** |
|---|---|---|---|---|---|---|
| Prob. P(X=x)= p(x) | $e^{-m}$ | $e^{-m} *m$ | $(e^{-m} *m^2 )$ /2! | $(e^{-m} *m^3 )$ / 3! | | 1 |

## Constants Of Poisson Distribution

$$\text{Mean} \quad \mu = \lambda$$

$$\text{Variance} \, \sigma^2 = \lambda$$

$$\text{S.D.} = \sqrt{\lambda}$$

$$p(k, \lambda) = \frac{\lambda^k e^{-\lambda}}{k!} \quad k = 0,1,2,.....$$

## Using Poisson Tables

| X | λ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 0.10 | 0.20 | 0.30 | 0.40 | **0.50** | 0.60 | 0.70 | 0.80 | 0.90 |
| 0 | 0.9048 | 0.8187 | 0.7408 | 0.6703 | 0.6065 | 0.5488 | 0.4966 | 0.4493 | 0.4066 |
| 1 | 0.0905 | 0.1637 | 0.2222 | 0.2681 | 0.3033 | 0.3293 | 0.3476 | 0.3595 | 0.3659 |
| **2** | 0.0045 | 0.0164 | 0.0333 | 0.0536 | **0.0758** | 0.0988 | 0.1217 | 0.1438 | 0.1647 |
| 3 | 0.0002 | 0.0011 | 0.0033 | 0.0072 | 0.0126 | 0.0198 | 0.0284 | 0.0383 | 0.0494 |
| 4 | 0.0000 | 0.0001 | 0.0003 | 0.0007 | 0.0016 | 0.0030 | 0.0050 | 0.0077 | 0.0111 |
| 5 | 0.0000 | 0.0000 | 0.0000 | 0.0001 | 0.0002 | 0.0004 | 0.0007 | 0.0012 | 0.0020 |
| 6 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0001 | 0.0002 | 0.0003 |
| 7 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |

Example: Find P(X = 2) if λ = 0.50

$$P(X = 2) = \frac{e^{-\lambda}\lambda^X}{X!} = \frac{e^{-0.50}(0.50)^2}{2!} = 0.0758$$

## Multinomial Distribution

- The binomial distribution is generalized as follows. Suppose the sample space of an experiment is partitioned into say mutually exclusive events $A_1$, $A_2$, $A_2$, ….., $A_s$ with respective probabilities $P_1$, $P_2$, $P_3$, …., $P_s$ (hence $P_1 + P_2 + P_3 + \ldots + P_s = 1$).
- Theorem : In n repeated trials, the probability that $A_1$ occurs $K_1$ times, $A_2$ occurs $K_2$ times and $A_s$ occurs $K_s$ times is equal to

$$\frac{n!}{k_1!k_2!k_3!\ldots k_s!} p_1^{k_1} p_2^{k_2} p_3^{k_3} \ldots\ldots p_s^{k_s}$$

  where $k_1 + k_2 + \ldots\ldots + k_s = n$
- The above numbers from the so-called multinomial distribution since they are precisely the terms in the expansion of $(p_1+p_2+\ldots.+p_s)$

## Continuous Distribution

- These distribution are associated with continuous variables.
- A continuous variable defined over a given range, may take any of the intermediate values. It is always written as an approximate values. For ex. Weight, height etc.
- The continuous variables are generally represented by a smooth curve.
- The cumulative distribution of a continuous variables is also a smooth curve.
- Normal distribution is one of the continuous distribution.

## Normal Distribution

- The most important continuous probability distribution used in the entire statistics is the normal distribution.
- Its graph, called the normal curve, is a bell shaped curve that extends indefinitely in both directions, coming closer and closer to the horizontal axis without even reaching it.
- The mathematical equation of normal curve was developed by De-Moivre in 1733.
- The normal distribution is often referred to as the Gaussian distribution in honour of Karl F. Gauss (1777-1855) who also derived the equation from the study of errors in repeated measurements of the same quantity.

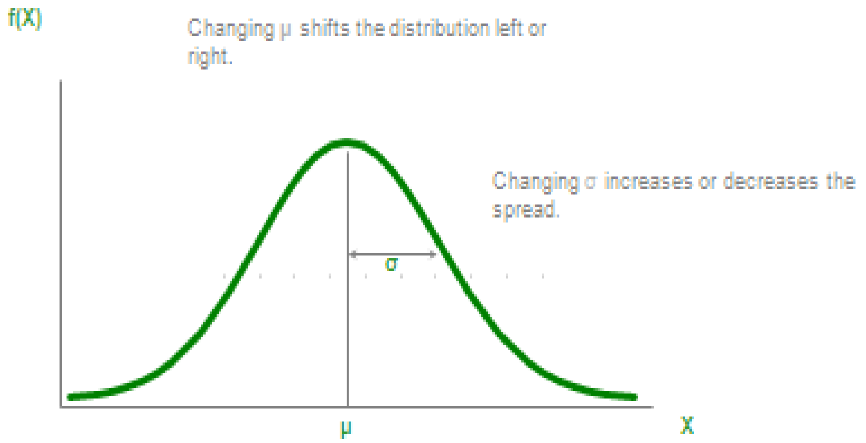
## Normal Distribution

- Definition : A continuous random variable X is said to be normally distributed if it has the probability density function represented by the equation:

$$\boxed{p(X) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(1/2)[(X-\mu)/\sigma]^2}}$$

  - $-\infty \leq x \leq \infty$
  - Where $\mu$ and $\sigma$, the mean and the standard deviation are known as two parameters and $\Pi$ = 3.14159 e = 2.7183 are two constant.

## Normal Distribution



Changing μ shifts the distribution left or right.

Changing σ increases or decreases the spread.

## Normal Distribution

- Now, standard normal distribution or Z – distribution is

$$p(Z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(1/2)Z^2}$$

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

  - Where
  - e = the mathematical constant approximated by 2.71828
  - π = the mathematical constant approximated by 3.14159
  - Z = any value of the standardized normal distribution
  - The Z distribution always has mean = 0 and standard deviation = 1

## Fitting A Normal Curve

- There are two main objects of fitting a normal curve to sample data.
  - To provide a visual device for judging whether or not the normal curve is a good fit to the sample data, and
  - To use the smoothed normal curve, instead of the irregular curve representing the sample data, to estimate the characteristics of the population.

## Methods Of Fitting

- There are two methods of fitting a normal curve.
  - Method of ordinates,
  - Method of area.

Minal Shah

# Measure Of Dispersion

Computer Oriented Numerical and Statistical Methods

## Outline

- Introduction
- Some Definitions
- Measure of Dispersion
- Range
- Quartile Deviation
- Mean Deviation
- Standard Deviation
- Coefficient of Variation
- Standard Deviation and Normal Curve
- Empirical Relationships

## Introduction

- An average does not tell the full story. It is hardly fully representation of a mass unless we know the manner in which the individual items scatter around it. A further description of the series is necessary if we are to gauge how representative the average is .

  –G. Simpson & F. Kafka
- An average is a single value which represents a set of values in a distribution.
- It is a central value which typically represents the entire distribution.
- Dispersion, on the other hand, indicates the extent to which the individual values fall away from the average or the central value

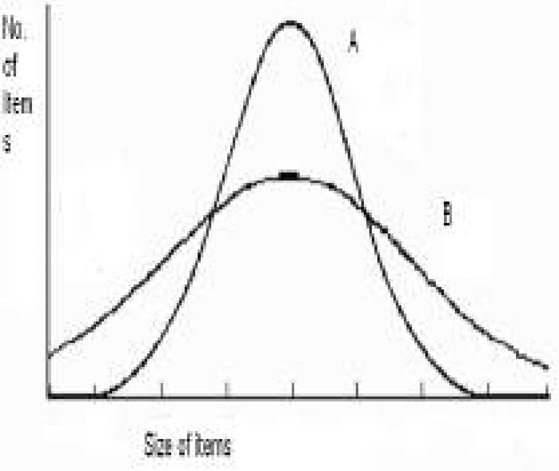
## Introduction

- This measure brings out how two distributions with the same average value may have wide differences in the spread of individual values around the central values

# Introduction

| Size of Item | No. of items (frequency) | |
|---|---|---|
| | A | B |
| 0 – 5 | 1 | 0 |
| 5 – 10 | 2 | 3 |
| 10 – 15 | 3 | 4 |
| 15 – 20 | 5 | 5 |
| 20 – 25 | 6 | 8 |
| 25 – 30 | 5 | 5 |
| 30 – 35 | 3 | 4 |
| 35 – 40 | 2 | 3 |
| 40 – 50 | 1 | 0 |

- Both have the same mean but the spread-out values



# Introduction

- Average is useful to study the state of output, sales, profits, etc, not its variability.
- But to study temperature, the price of shares, rainfall etc. dispersion is found to be more important.
- Dispersion is an important measure sought for describing the character of variability in data.
- While an average discovers the representative value, dispersion finds out how individual values fall apart, on an average, from the representative value.
- The average was derived from the actual values but dispersion is known by averaging the deviations in individual values from some representative value and therefore called am average of the second order.

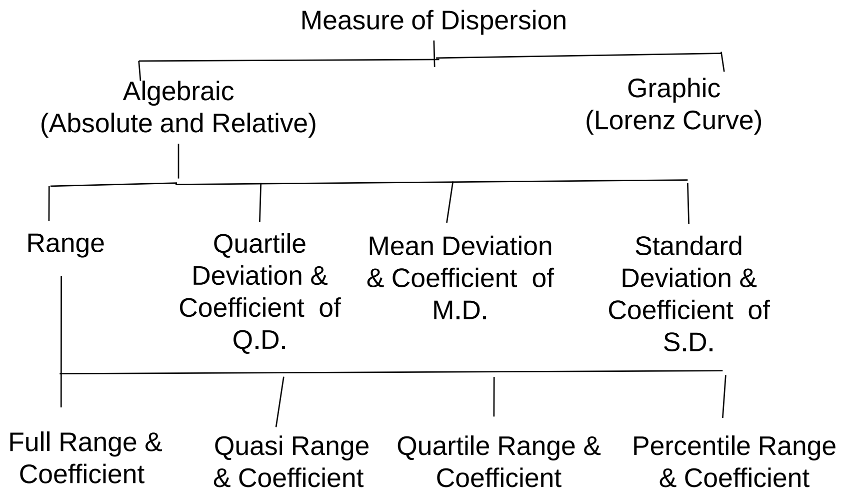
## Some Definitions

- Dispersion is measure of the extent to which the individual items vary. ----- L. R. Connor.
- Dispersion or spared is the degree of the scatter or variation of the variables about a central value. ------ B. C. Brooks & W. F.L. Dick
- The degree to which numerical data tend to spread about an average value is called the variation or dispersion of the data. ------ Spiegel
- Dispersion is the measure of the variations of the items. ------ A. L. Bourey

## Uses / Significance of Measure of Variation

- To judge the reliability of central tendency.
- To compare two or more series with regards to their variability.
- To control the variability itself
- To facilitate the use of other statistical measure
- To control quality
- To analysis of time series.

## Measure of Dispersion


Measure of Dispersion

Algebraic
(Absolute and Relative)

Graphic
(Lorenz Curve)

Range

Quartile
Deviation &
Coefficient of
Q.D.

Mean Deviation
& Coefficient of
M.D.

Standard
Deviation &
Coefficient of
S.D.

Full Range &
Coefficient

Quasi Range
& Coefficient

Quartile Range &
Coefficient

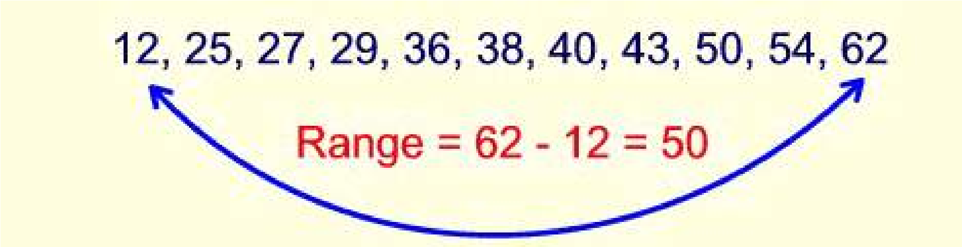
Percentile Range
& Coefficient

## Range

- The range method is based on any two boundary values of a distribution.
- It is not concerned with the rest of the values nor with the concentration on individual values.
- The following are the various types of range.
  - Full Range
  - Quasi Range
  - Inter-Quartile or Quartile Deviation
  - Percentile Range

## Full Range

- This is defined as the difference between the largest /highest and the smallest/lowest values in the distribution.
- Range (R ) = $X_{max} – X_{min}$ or Range (R ) = $X_n – X_1$
- where $X_n$ is the last item and $X_1$ is the first item of a series arrange in an ascending order of magnitude.



12, 25, 27, 29, 36, 38, 40, 43, 50, 54, 62

Range = 62 - 12 = 50

## Full Range Example

- The two sets below have the same mean and median (7). Find the range of each set.

| | | | | | |
|---|---|---|---|---|---|
| Set A | 1 | 2 | 7 | 12 | 13 |
| Set B | 5 | 6 | 7 | 8 | 9 |

- Solution
- Range of Set A: 13 – 1 = 12.
- Range of Set B: 9 – 5 = 4.

## Quasi – Range

- Quasi – Range refers to the difference between the values leaving the extreme values.
- If we leave two extreme values it will be $R = X_{n-1} – X_2$ where n is the total number of items. $X_{n-1}$ is the last but one item and $X_2$ is next to the first item.
- If more than two extreme values are left out, the expression will be $R = X_{n-2} – X_3$

## Coefficient Of Range

- The coefficient of range is derived by dividing a given range by sum of the values of the two boundary values taken into account for calculating range.

$$\text{Coefficient of range} = \frac{\text{Highest value} - \text{Lowest value}}{\text{Highest value} + \text{Lowest value}}$$

$$= \frac{X_n - X_1}{X_n + X_1}$$

- Coefficient of range is more suitable than range for comparison purpose.

## Inter-Quartile Range Or Quarter Deviation (Q.D.)

- Q.D. is a measure of dispersion based on the upper quartile ($Q_3$) and the lower quartile ($Q_1$).
- It is also called semi-inter-quartile range because it represents the average difference between two quartiles. Quarter Deviation (Q.D.) = ($Q_3 – Q_1$) / 2
- Q.D. as defined above is only an absolute measure of dispersion.
- For comparative studies of variability of two distributions we need an absolute measure which is known as coefficient of quartile deviation and is given by

$$\text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

## Percentile Range

- Percentile range is based on the values of any two percentile.
- For example percentile range ($P_{90} – P_{10}$) is based on values of the 90th and 10th percentile.

## Mean Deviation (M.D.)

- The range and quartile deviation are positional measure of dispersion and are based on the position of certain items in a distribution.
- The M.D. or average deviation is a measure of dispersion that is based on all items.
- The M.D. is the arithmetic mean of the deviations of the individual values from the average of the given data.
- The average which is frequently used in computing the mean deviation is mean or median. Also only the absolute values of the dispersion are used.

## Mean Deviation (M.D.)

- “Average deviation is the average amount of scatter, of the items in a distribution from either the mean or the median, ignoring the signs of the deviation. The average that is taken of the scatter is an A.M., which accounts for the fact that this measure is often called the M.D.”
- M.D. is denoted by $\delta$ (delta). The sign of average taken is deviation is used as subscript i.e.
  - $\delta_{\bar{x}}$ (mean deviation from mean)
  - $\delta_{Md}$ (mean deviation from median)
  - $\delta_{Mo}$ (mean deviation from mode)

## Coefficient Of Mean Deviation

- M.D. calculated by any measure of central tendency is an absolute measure.
- When it is divided by the average used for calculating it, we get coefficient of mean deviation which will give a relative measure of dispersion suitable for comparing two or more series which are expressed in different units or expressed in the same units but of different order of magnitude.

1. Coefficient of M.D. = M.D. / Mean or Median or Mode
   if the result is desired in percentage, then
2. Coefficient of Mean variation = (M.D. / Mean or Median or Mode) * 100

## Methods of Computation

- If $X_1$, $X_2$, …., $X_n$ are n given observations, then the M.D. about an average A, say is given by
  M.D. (about an average A ) $= \frac{1}{n}\sum | X - A |$

$$= \frac{1}{n}\sum | d |$$

where |d| = |X – A|
A is any one of the average mean (M), median (Md) or Mode (Mo)

## Methods of Computation (Steps)

1. Calculate the average A of the distribution by the usual method.
2. Take the deviation d = X – A of each observation from the average A.
3. Ignore the negative signs of the deviations, taking all the deviation to be positive to obtain the absolute deviations. |d| = |X – A|
4. Obtain the sum of the absolute deviations obtain in step(3).
5. Divide the total obtained in step(4) by n, the number of observations. The result gives the value of the mean deviation about the average A.

## Mean Deviation for Group Data

- In case of frequency distribution or grouped or continuous frequency distribution, mean deviation about an average A is given by M.D. (about the average A) = 1 / N * Σf*|X – A| = 1 / N * Σf*|d|
  - Where X is the value of the variable or it is the mid value of the class interval (in case of grouped or continuous frequency distribution) f is the corresponding frequencies, N = Σf is the total frequency and |X – A| is the absolute values of the deviation d = (X – A) of the given values of X from the average A (mean, median, mode)

# Mean Deviation for Group Data (Steps)

1. Calculate the average A of the distribution by the usual method.
2. Take the deviation d = X – A of each observation from the average A.
3. Ignore the negative signs of the deviations, taking all the deviation to be positive to obtain the absolute deviations. |d| = |X – A|
4. Multiply the absolute deviation |d| = |X – A| by the corresponding frequency f to get f*|d|
5. Divide the total in step (4) by N, the total frequency.

- The resulting value is the mean deviation about the average A.

# Standard Deviation (S.D.)

- The concept of S.D. was first suggest by Karl Pearson in 1893.
- It may be defined as the positive square root of the arithmetic mean of the square deviations of given observations from their arithmetic mean.
- In short S.D. may be defined as "Root-Mean-Square-Deviation from mean."
- It is denoted by $\sigma$ (sigma).
- It is by far the most important and a widely used measure of studying dispersion.
- For a set of N observations $X_1, X_2, \ldots, X_N$ with mean $\overline{x}$
- Deviation from mean : $(X_1 - \overline{X}), (X_2 - \overline{X}), \ldots\ldots, (X_N - \overline{X})$

# Standard Deviation (S.D.)

- Square – Deviations from mean :

$$(X_1 - \overline{X})^2, (X_2 - \overline{X})^2, \ldots\ldots, (X_N - \overline{X})^2$$

- Mean Square Deviation from mean :

$$\tfrac{1}{N}[(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \ldots\ldots + (X_N - \overline{X})^2]$$

$$= \frac{1}{N} \sum (x - \bar{x})^2$$

- Root-mean square deviation from mean i.e.

S. D. ($\sigma$) = $\sqrt{\frac{1}{N} \sum (x_i - \bar{x})^2}$

- The square of S.D. i.e. $\sigma^2$ is known as variance

$\therefore$ varaince = (S.D.)$^2$

# Standard Deviation (S.D.)

- In case the data is grouped in the form of frequency distribution, the individual measurements of the variable or the mid-value of the class intervals (as the case may be) are weighted by the corresponding frequency.
- For e.g. if $X_1$, $X_2$,…. Are the individual measurements and if $f_1$, $f_2$, …. are the corresponding frequencies, then $X_1$ is weighted by $f_1$, $X_2$ by $f_2$ and so on.
- $\therefore$ S.D. for the frequency distribution is

$$\sigma = \sqrt{\frac{1}{N} \sum f * (x - \bar{x})^2} \ ; \ N = \sum f$$

## Comparison Between M.D. And S.D.

| M.D. | S.D. |
|---|---|
| Deviation are calculated from mean, median or mode. | These are calculated from the arithmetic mean only. |
| The algebraic sign have to be ignored- only values of deviation are taken. | Since the deviations are squared the plus and minus signs need not be omitted. |
| It is based on simple average of the sum of absolute deviations. | It is based on the square root of the average of the squared deviation. |

## Comparison Between M.D. And S.D.

| M.D. | S.D. |
|---|---|
| It is simple to calculate when mean is a round number. The short-cut method is somewhat cumbersome. | This is somewhat complex because of squaring of the deviations but it is suitable in all cases- whether the mean is a round number or a fraction, since a short-cut method is also available. |
| It lacks mathematical prosperities since only absolute values are considered. | It is mathematically sound on account of the fact that algebraic signs are not ignored. |

## Computation Of S.D.

- Ungrouped Data :
  - Direct Method
  - Short-cut method
- Direct Method :

a) The procedure of computing S.D. from ungrouped data is given below:

1) Obtain the arithmetic mean of the given data.
2) Obtain the deviation of each value from the arithmetic mean, or $x = X - \overline{X}$
3) Square each deviation to make it positive, or $x^2$

Minal Shah

## Computation Of S.D.

- Direct Method :

4) Obtain sum of the deviations squared, or $\Sigma x^{2\cdot}$.
5) Find the variance ($\sigma^2$) by dividing the sum by the number of observations (n) in the data i.e. $\sigma^2 = \Sigma x^2 / n$
6) Extract the square root of the variance to find S.D.

$$\sigma = \sqrt{\frac{\sum x^2}{n}}$$

Minal Shah

## Computation Of S.D.

- Direct Method :

b) The second direct method of computing the S.D. is the one where the values of the variable are used directly and no deviations need to be calculated. The formula is

$$\sigma = \sqrt{\frac{\sum X^2}{n} - \overline{X}^2}$$

$$= \sqrt{\frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2}$$

- i.e S.D. is the square root of the average of the sum of the squared values minus the square of the average of all the values.

## Computation Of S.D (Short Cut Method)

1. One of the value, generally the value in the middle, is taken as assumed or working mean (A).
2. Obtain the deviation of each item from the assumed average, dx = X – A, and take the total of the deviations, Σdx.
3. The deviations are squared up and totaled to obtain $\Sigma dx^2$.
4. The standard deviation is obtained by using any of the following formula

$$(a)\ \sigma = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2}$$

$$(b)\ \sigma = \sqrt{\frac{\sum d^2}{n} - \left(\overline{X} - A\right)^2}$$

## Computation Of S.D (Grouped Data [Direct Method])

1. Obtain the A.M. of the given data.
2. Find the deviation of the item (in case of discrete series) or the mid-points of each class interval (in-case of continuous series) from the A.M. or

$$x = X - \overline{X} \ \ or \ \ x = m - \overline{X}$$

3. Obtain the total of deviations squared for each class $fx^2$.
4. Obtain the sum of the deviations squared or $\Sigma fx^2$.
5. $\Sigma fx^2$ is divided by the number of items. Then extract the square root of the quotient to obtain the S.D.

$$\sigma = \sqrt{\frac{\sum fx^2}{N}} \ ; N = \sum f$$

## Computation Of S.D (Grouped Data [Short-Cut Method])

- The short-cut method of computing the S.D. for grouped data is basically the same as the short-cut method for ungrouped data.
- The only difference is $d^2$ and d values in each class are multiplied by corresponding frequency of that class

$$\therefore \sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \ ; N = \sum f$$

## Computation Of S.D (Grouped Data [Step-Deviation Method])

- In order to simply further the calculation work, some common factor may be taken out from all the deviations.
- The formula for computing S.D. stands modified as follows :

$$\sigma = h * \sqrt{\frac{\sum fd'^2}{N} - \left(\frac{\sum fd'}{N}\right)^2} \quad ; N = \sum f$$

- Where d' = (X-A)/h (for discrete series) d' = (m-A)/h (for continuous series) and h is a common factor, A is the assumed average and m is the mid-value of the class-interval

## S.D Of Combined Series

- If two groups contains $n_1$ and $n_2$ observations with mean $\overline{X_1}$ *and* $\overline{X_2}$ and S.D. $\sigma_1$ and $\sigma_2$ respectively, the S.D. ($\sigma_{12}$) of the combined group is given by

$$\sigma_{12} = \sqrt{\frac{n_1 * (\sigma_1^2 + d_1^2) + n_2 * (\sigma_2^2 + d_2^2)}{n_1 + n_2}}$$

$$\text{where } \sigma_{12} = \text{combinedS.D. of the two groups}$$

$$d_1 = \overline{X_{12}} - \overline{X_1} \text{ and } d_2 = \overline{X_{12}} - \overline{X_2}$$

$$and\ \overline{X_{12}} = \frac{n_1 * \overline{X}_1 + n_2 * \overline{X}_2}{n_1 + n_2}$$

## S.D (Combined Group Example)

- The mean and the S.D. of a sample of size 10 were found to be 9.5 and 2.5 respectively. Later on an additional observation became available. This was 15.0 and was included in the original sample. Find the mean and S.D. of the 11 observations.

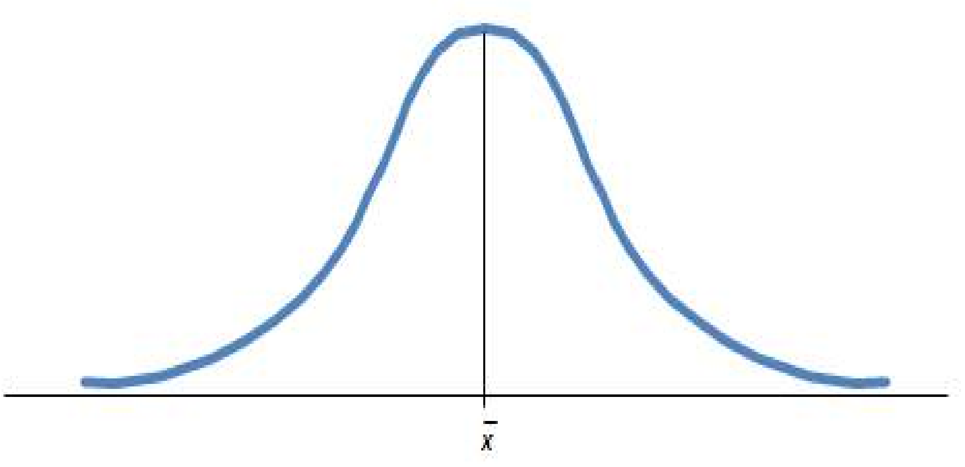
## S.D And Normal Curve

- Most distributions are symmetrical bell-shape.
- The frequency curved formed from such distributions may be regarded as approximations to an important well known curve known as the ‘normal curve’.
- Mean ± S.D. will indicate the range within which a given percentage of values of the distributions are likely to fall i.e. nearly 68.27% will lie within mean ± 1 S.D., 95.45% within mean ± 2 S.D. and 99.73% within mean ± 3 S.D. under the normal curve.

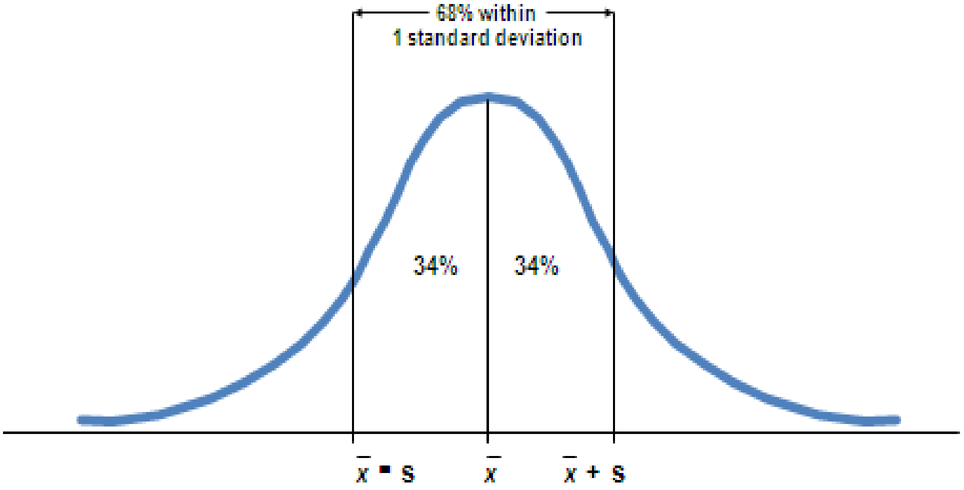
## Relationship Among 3 measure of Variability (For a Normal Distribution)

| **Measure of Variability** | **% of item included within given range of mean $\overline{X}$** | | | **Size of measure of variability to S.D. at** |
|---|---|---|---|---|
| | ± One deviation | ± two deviation | ± three deviation | |
| Quartile Deviation | 50.0 | 82.3 | 95.7 | 0.6748 |
| Mean Deviation | 57.5 | 88.9 | 98.3 | 0.7979 |
| Standard Deviation | 68.27 | 95.45 | 99.73 | 1.0000 |

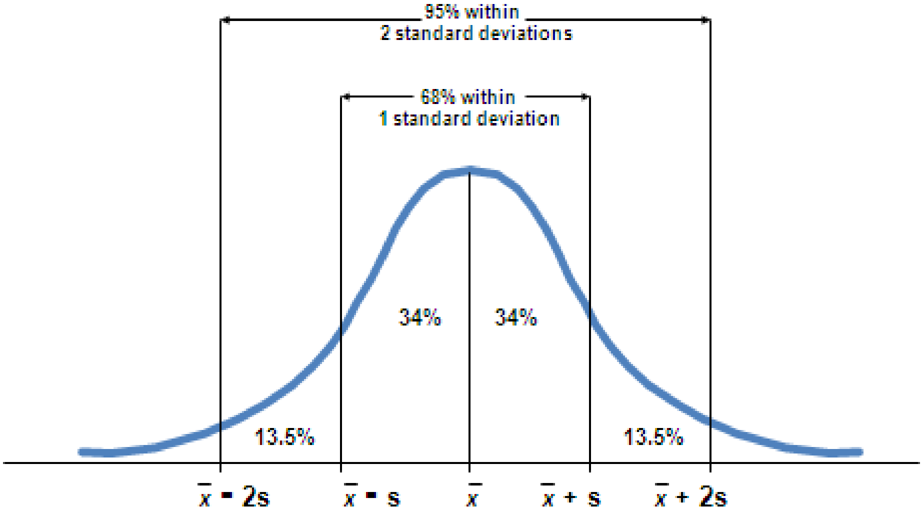
## Relationship Among 3 measure of Variability (For a Normal Distribution)



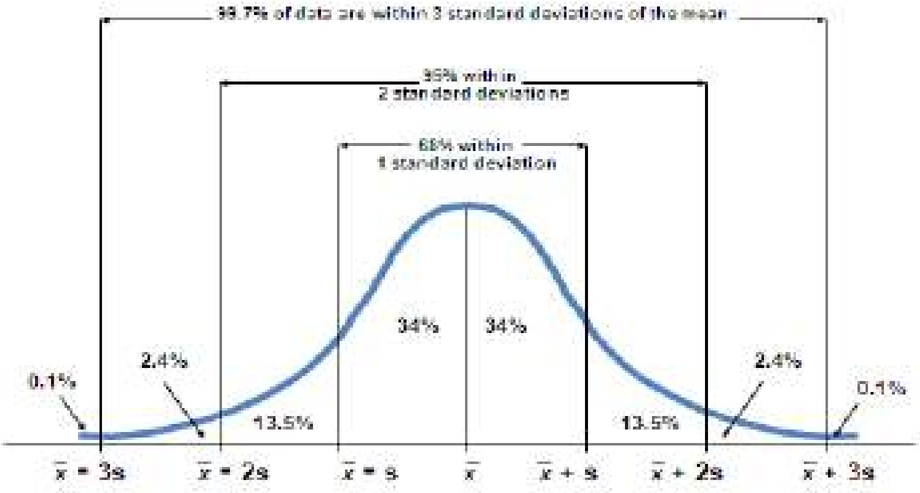
## Relationship Among 3 measure of Variability (For a Normal Distribution)



## Relationship Among 3 measure of Variability (For a Normal Distribution)



## Relationship Among 3 measure of Variability (For a Normal Distribution)



## Empirical Relationship

1. Quartile Deviation ≈ 2/3 S.D. ≈ 0.6745σ
2. M.D. ≈ 4/5 S.D. ≈ 0.7979σ
3. Q.D. ≈ 5/6 M.D. = 0.8333 M.D.
4. Range = 6 * S.D. or 6σ

- These can further also be expressed as

5. 3 Q.D. ≈ 2 S.D.
6. 5 M.D. ≈ 4 S.D.
7. 6 Q.D. ≈ 5 M.D.

- S.D. ensures highest degree of reliability and Q.D. the lowest 4 S.D. ≈ 5 M.D. ≈ 6 Q.D.

Minal Shah

# Statistical Inference

Computer Oriented Numerical and Statistical Methods

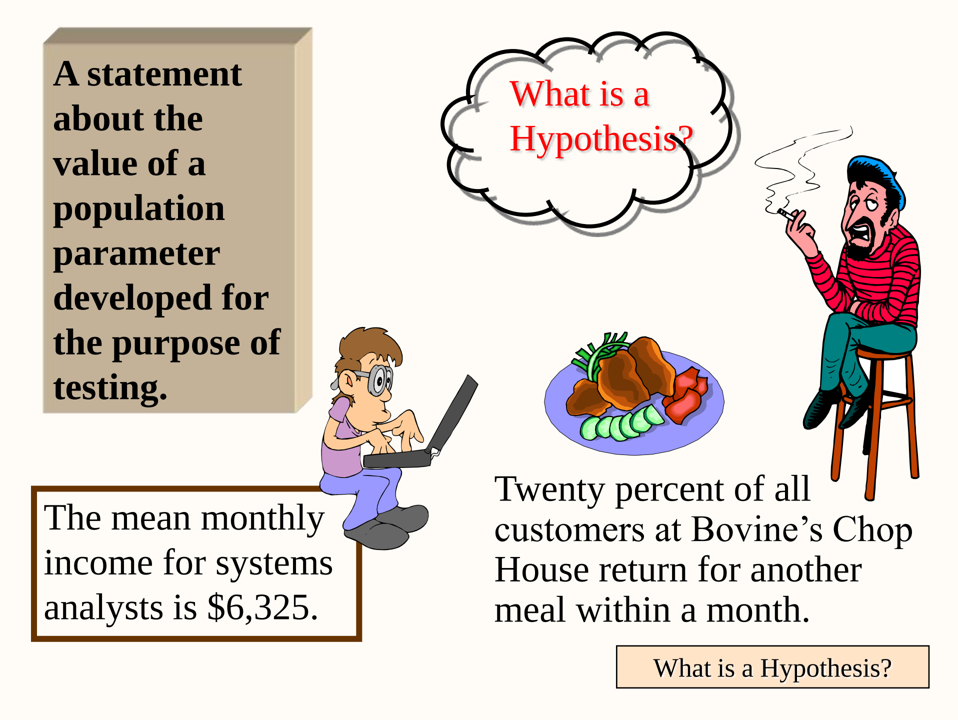
- **Statistical Inference** is branch of statistics which is concerned with using probability concept to deal with uncertainty in decision making
- Statistical inference treats two different classes of problems
  - **Hypothesis Testing :** To test some hypothesis about parent population from which the sample is drawn
  - **Estimation :** To use the statistics obtained from the sample as estimate of the unknown parameters of the population from which the sample is drawn

- Hypothesis Testing begins with an assumption, called a hypothesis
- A **hypothesis** in statistics is simply a quantitative statement about a population.
- In order to make statistical decisions, we make an certain assumptions about the population parameters to be tested.
- These ***assumptions*** are known as hypothesis

**Hypothesis Testing**

- There can be several types of hypotheses
- **For example** : The average marks of the 100 students of a class and may get the result as 65% we are now interested in testing the hypothesis that the sample has been drawn from a population with average marks 70%
- A coin may be tossed 100 times and we may get heads 75 time and tails 25 times, we are now interested in testing the hypothesis that the coin is unbiased

**Hypothesis Testing**



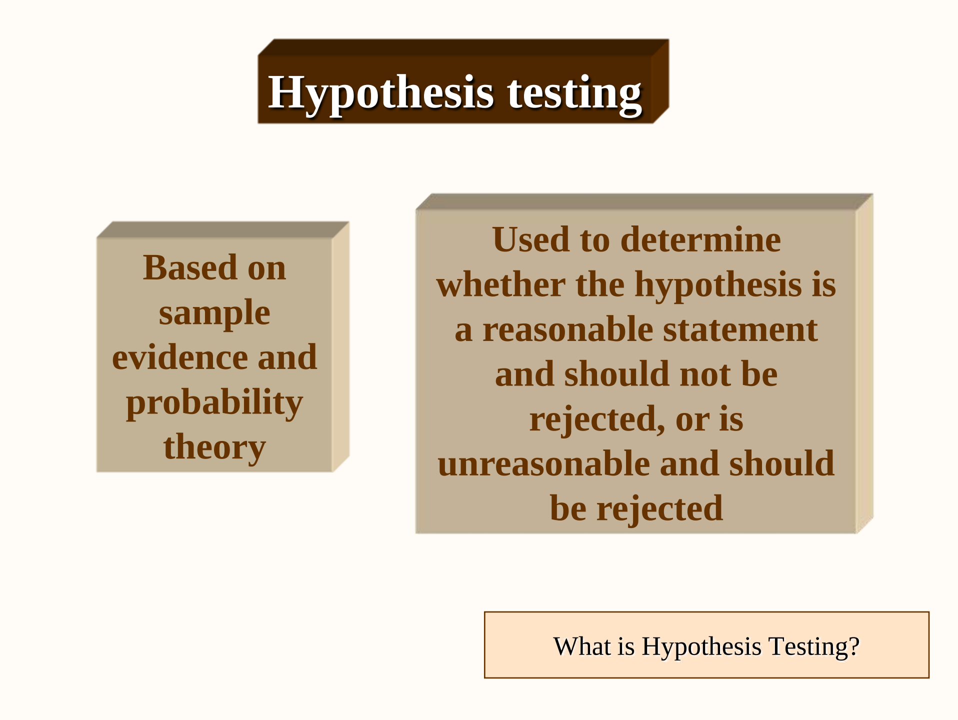
What is a Hypothesis?

**A statement about the value of a population parameter developed for the purpose of testing.**

The mean monthly income for systems analysts is $6,325.

Twenty percent of all customers at Bovine's Chop House return for another meal within a month.

What is a Hypothesis?

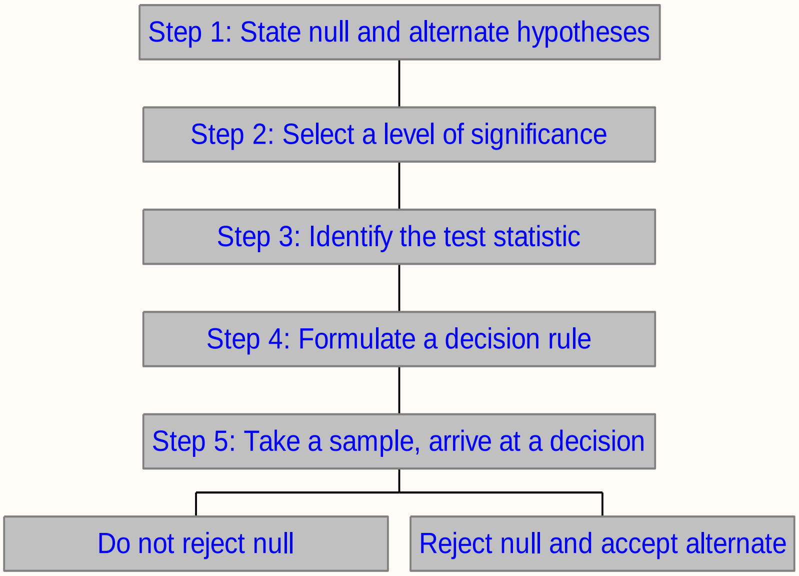


## Hypothesis testing

**Based on sample evidence and probability theory**

**Used to determine whether the hypothesis is a reasonable statement and should not be rejected, or is unreasonable and should be rejected**

What is Hypothesis Testing?



Step 1: State null and alternate hypotheses

Step 2: Select a level of significance

Step 3: Identify the test statistic

Step 4: Formulate a decision rule

Step 5: Take a sample, arrive at a decision

Do not reject null

Reject null and accept alternate

Hypothesis Testing

## Step One: State the null and alternate hypotheses

**Null Hypothesis $H_0$**

**A statement about the value of a population parameter**

**Alternative Hypothesis $H_1$:**

**A statement that is accepted if the sample data provide evidence that the null hypothesis is false**

## Type I Error

Rejecting the null hypothesis when it is actually true (α).

## Level of Significance

The probability of rejecting the null hypothesis when it is actually true; the level of risk in so doing.

## Type II Error

Accepting the null hypothesis when it is actually false (β).

Step Two: Select a Level of Significance.

- **Defines the unlikely values of the sample statistic if the null hypothesis is true**
  - Defines rejection region of the sampling distribution
- Is designated by α , (level of significance)
  - Typical values are 0.01, 0.05, or 0.10
- Is selected by the researcher at the beginning
- Provides the critical value(s) of the test

Level of Significance, α

## Step Two: Select a Level of Significance.

| Null Hypothesis | Researcher<br>Accepts $H_o$ | Researcher<br>Rejects $H_o$ |
|---|---|---|
| $H_o$ is true | Correct decision | Type I error ($\alpha$) |
| $H_o$ is false | Type II Error ($\beta$) | Correct decision |

Risk table

## Test statistic

A value, determined from sample information, used to determine whether or not to reject the null hypothesis.

Examples: $z$, $t$, $F$, $\chi^2$

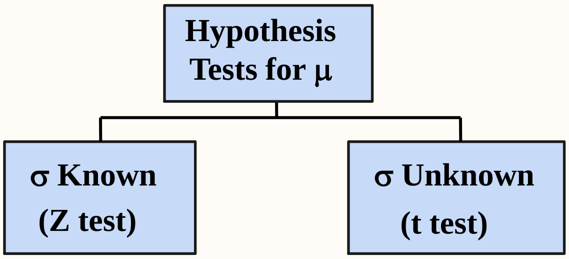
## *z* Distribution as a test statistic

$$\boxed{z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}}$$

The *z* value is based on the sampling distribution of X, which is normally distributed when the sample is reasonably large (recall Central Limit Theorem).

Step Three: Select the test statistic.

## Hypothesis Tests for the Mean



Hypothesis Tests for μ

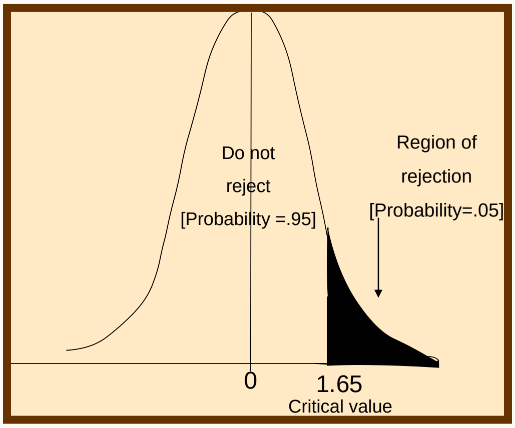
σ Known (Z test)

σ Unknown (t test)

## Step Four: Formulate the decision rule.

**Critical value:** The dividing point between the region where the null hypothesis is rejected and the region where it is not rejected.

**Sampling Distribution Of the Statistic *z*, a Right-Tailed Test, .05 Level of Significance**



## *p*-Value

The probability, assuming that the null hypothesis is true, of finding a value of the test statistic at least as extreme as the computed value for the test

### Decision Rule

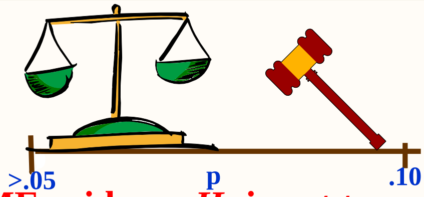
If the *p*-Value is larger than or equal to the significance level, α, $H_0$ is not rejected.

If the *p*-Value is smaller than the significance level, α, $H_0$ is rejected.

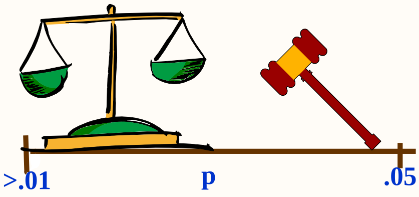
Calculated from the probability distribution function or by computer

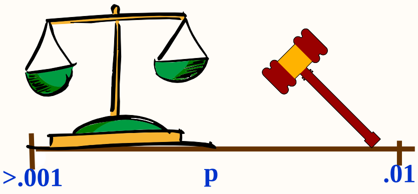
Using the p-Value in Hypothesis Testing

## Interpreting p-values



SOME evidence $H_o$ is not true

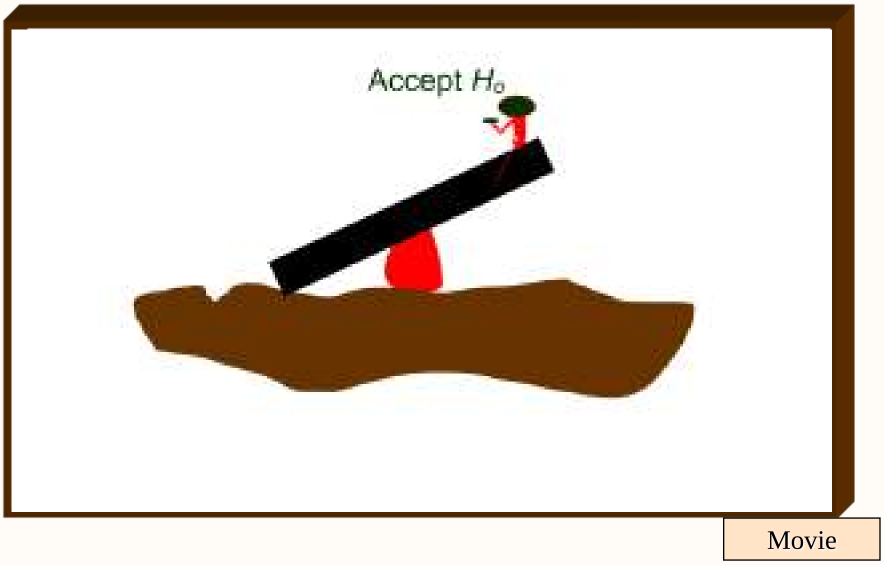




STRONG evidence $H_o$ is not true

VERY STRONG evidence $H_o$ is not true

## Step Five: Make a decision.



Movie

## Level of Significance and the Rejection Region

Level of significance = $\alpha$

Represents critical value

Rejection region is shaded

$H_0$: $\mu = 3$

$H_1$: $\mu \neq 3$

Two-tail test

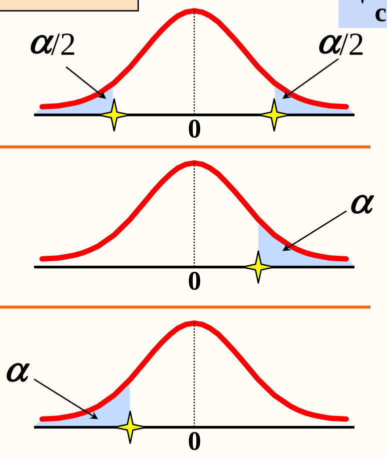
$H_0$: $\mu \leq 3$ $H_1$: $\mu > 3$

Upper-tail test

$H_0$: $\mu \geq 3$

$H_1$: $\mu < 3$

Lower-tail test



## Test for the population mean from a large sample with population standard deviation known

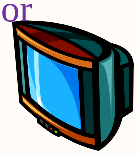
$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Testing for the Population Mean: Large Sample, Population Standard Deviation Known

## Hypothesis Testing Example

**Test the claim that the true mean # of TV sets in US homes is equal to 3.**
**(Assume σ = 0.8)**

1. State the appropriate null and alternative hypotheses
   - $H_0$: μ = 3 $H_1$: μ ≠ 3 (This is a two-tail test)
2. Specify the desired level of significance and the sample size
   - Suppose that α = 0.05 and n = 100 are chosen for this test



## Hypothesis Testing Example

(continued)

3. Determine the appropriate technique
   - σ is known so this is a Z test.
4. Determine the critical values
   - For α = 0.05 the critical Z values are ±1.96
5. Collect the data and compute the test statistic
   - Suppose the sample results are

     $n = 100$, $\bar{X} = 2.84$ ($\sigma = 0.8$ is assumed known)

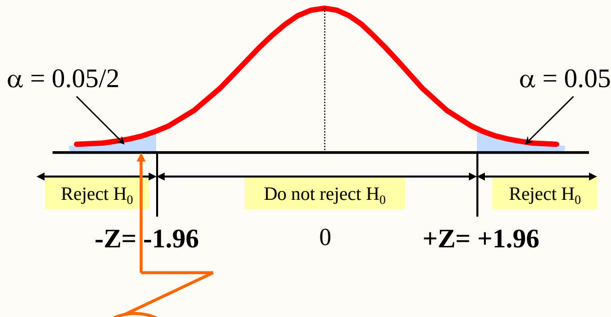
So the test statistic is:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{2.84 - 3}{\frac{0.8}{\sqrt{100}}} = \frac{-.16}{.08} = -2.0$$

## Hypothesis Testing Example

(continued)

❍ 6. Is the test statistic in the rejection region?



Reject $H_0$ if $Z < -1.96$ or $Z > 1.96$; otherwise do not reject $H_0$

Here, Z = -2.0 < -1.96, so the test statistic is in the rejection region and conclude that there is sufficient evidence that the mean number of TVs in US homes is not equal to 3

## Example: Upper-Tail Z Test for Mean (σ Known)

A phone industry manager thinks that customer monthly cell phone bills have increased, and now average over $52 per month. The company wishes to test this claim. (Assume $\sigma = 10$ is known)

Form hypothesis test:

$H_0$: $\mu \leq 52$   the average is not over $52 per month

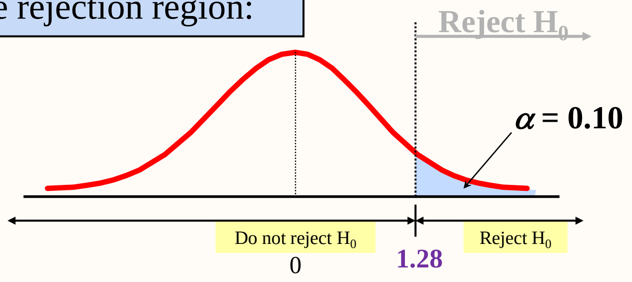
$H_1$: $\mu > 52$   the average is greater than $52 per month (i.e., sufficient evidence exists to support the manager's claim)

## Example: Find Rejection Region

(continued)

- Suppose that $\alpha = 0.10$ is chosen for this test

Find the rejection region:



Reject $H_0$ if $Z > 1.28$

# Example: Test Statistic

*(continued)*

Obtain sample and compute the test statistic

Suppose a sample is taken with the following results:
$n = 64$, $\bar{X} = 53.1$ ($\sigma$=10 was assumed known)

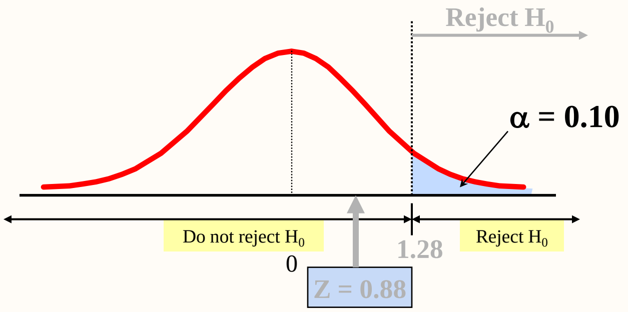
- Then the test statistic is:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{53.1 - 52}{\frac{10}{\sqrt{64}}} = 0.88$$

# Example: Decision

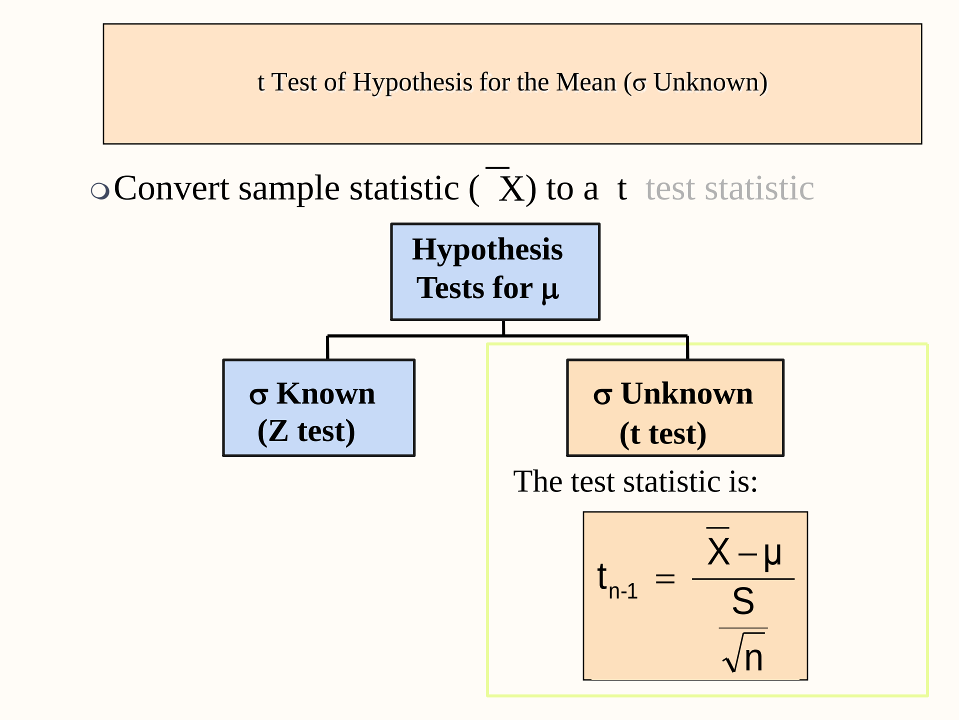
*(continued)*

Reach a decision and interpret the result:



**Do not reject $H_0$ since $Z = 0.88 \le 1.28$**

i.e.: there is not sufficient evidence that the mean bill is over $52



## t Test of Hypothesis for the Mean (σ Unknown)

- Convert sample statistic ( $\bar{X}$) to a t test statistic

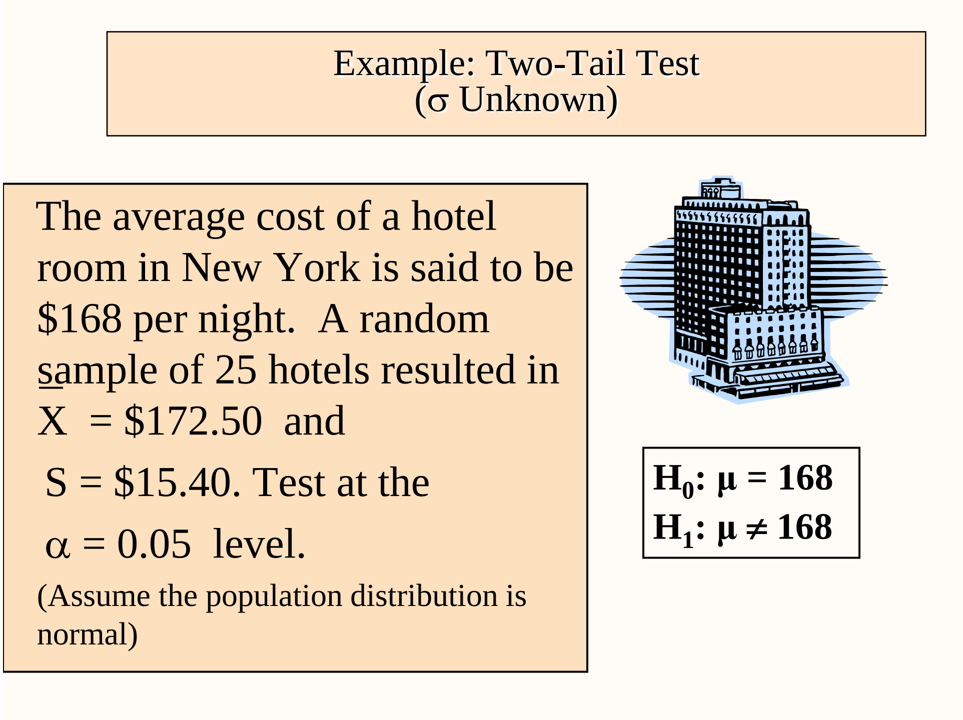
Hypothesis Tests for μ

σ Known (Z test)

σ Unknown (t test)

The test statistic is:

$$t_{n-1} = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$



## Example: Two-Tail Test (σ Unknown)

The average cost of a hotel room in New York is said to be \$168 per night. A random sample of 25 hotels resulted in $\bar{X} = \$172.50$ and

S = \$15.40. Test at the

$\alpha = 0.05$ level.

(Assume the population distribution is normal)

**$H_0$: μ = 168**
**$H_1$: μ ≠ 168**

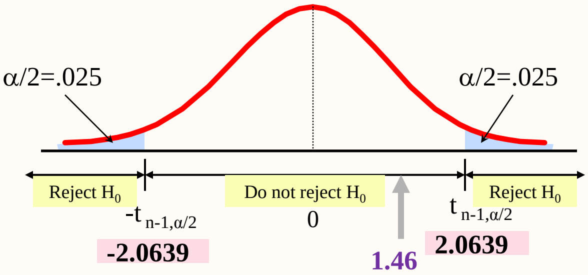
## Example Solution: Two-Tail Test

$H_0$: $\mu = 168$

$H_1$: $\mu \neq 168$

- $\alpha = 0.05$
- n = 25
- $\sigma$ is unknown, so use a t statistic
- Critical Value:

$t_{24} = \pm\ 2.0639$



$$t_{n-1} = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{172.50 - 168}{\frac{15.40}{\sqrt{25}}} = 1.46$$

**Do not reject $H_0$:** not sufficient evidence that true mean cost is different than $168

# Probability

Computer Oriented Numerical and Statistical Methods

## Outline

- Introduction
- Meaning
- Basic Definitions
- Types
- Basic Probability Rules

## Introduction

- So far we have studied the methods of collection, description and analysis of data.
- This does not cover the entire operational field of statistics.
- Modern statistics has also to provide for :
  1. Estimation of population ‘parameters’ on the basis of sample ‘statistics’
  2. Drawing inferences about the sample ‘statistics’ on the basis of population ‘parameters’
  3. Testing of hypothesis with regards to (1) and (2) above and

## Introduction

4. Decision-making under risk and uncertainty be estimating the degree of risk and the likely effect on business objectives in terms of pay off and expected values of a decision.

- The theory of probability helps in all these areas.
- Probability helps a person to make ‘educated guesses’ on matters, where either full facts are not known or there is uncertainty about the outcome.
- The decision-makers always face some degree of risk while selecting a particular decision (course of action or strategy) to solve a decision problem.

## Introduction

- It is because each strategy can lead to a number of different possible outcomes (or results).
- Thus it is necessary for the decision-makers to enhance their capability of grasping the probabilistic situation so as to gain a deeper understanding of the decision problem and base their decision on rational considerations.

## Meaning

- Probability means the number of occasion that a particular event is likely to occur in a large population of events.
- The particular event may be expressed positively where the event is likely to happen or negatively where the event is not likely to happen.

## Meaning

- Broadly, there are three possible states of expectations :
  - Certainty
  - Impossibility
  - Uncertainty
- The probability theory describe certainty by 1, impossibility by 0 and the various grades of uncertainties by coefficients ranging between 0 and 1

## Probability in Everyday Life

- Possible, it will rain tonight.
- Probably you will catch the train.
- There is a high chance of your getting the job in August.
- This year’s demand for the product is likely to exceed that of the last year’s.

# Definitions

- **Probability** : It means the number of occasion that a particular event is likely to occur in a large population of events.
  - The particular event may be expressed positively where the event is likely to happen or negatively where the event is not likely to happen.

# Definitions

- **Random Experiment or Trial**: An experiment is said to be a random experiment (or trial or act or operation or process), if it's out-come can't be predicted with certainty.
- Example :
  - If a coin is tossed, we can't say, whether head or tail will appear. So it is a random experiment.
  - Drawing a card from a pack of cards

## Definitions

- **Sample Space (Possible outcomes)**: The set of all possible out-comes of an experiment is called the sample space. It is denoted by 'S' or 'U' and its number of elements are n(s).
- **Example:**
  - In throwing a dice, the number that appears at top is any one of 1,2,3,4,5,6. So here: S ={1,2,3,4,5,6} and n(s) = 6
  - In the case of a coin, S={Head,Tail} or {H,T} and n(s)=2.
- The elements of the sample space are called sample point or event point.

## Definitions

- **Event**: Every subset of a sample space is an event. It is denoted by 'E'.
  - The empty set ∅ is called impossible event and the sample space U is called certain event.
  - Clearly E is a sub set of S
- **Example:**
  - In throwing a dice S={1,2,3,4,5,6}, the appearance of an event number will be the event E={2,4,6}.
- **Simple event** : An event, consisting of a single sample point is called a simple event.
- **Example**:
  - In throwing a dice, S={1,2,3,4,5,6}, so each of {1},{2},{3},{4},{5} and {6} are simple events.

## Definitions

- **Compound event:** A subset of the sample space, which has more than one element is called a mixed event (compound event).
- Example:
  - In throwing a dice, the event of appearing of odd numbers is a compound event, because E={1,3,5} which has '3' elements.

## Definitions

- **Equally likely events**: Events are said to be equally likely, if we have no reason to believe that one is more likely to occur than the other. OR The outcomes are said to be equally likely or equally probable if none of them is expected to occur in preference to other.
- Example:
  - When a dice is thrown, all the six faces {1,2,3,4,5,6} are equally likely to come up.

# Definitions

- **Exhaustive events**: When every possible out come of an experiment is considered.
- **Example:**
  - A dice is thrown, cases 1,2,3,4,5,6 form an exhaustive set of events.
- **Collectively Exhaustive events**: The total number of possible outcomes of a random experiment is called the collective exhaustive events.
- **Example:**
  - A dice is thrown, cases 1,2,3,4,5,6 form an exhaustive set of events and number of cases is 6.
  - In toss of a single coin exhaustive number of cases is 2

# Definitions

- **Complementary events**: The set of all elements of the sample space U except the elements of event A is called the complementary event A. It is denoted by A' or Ā it means that 'the event A does not occur.
- **Union events**: Let A and B be two events. The set of elements which are either in A or B is called the union event of event A and B. it is denoted by $A \cup B$. $A \cup B$ means the event 'either A occurs or B occurs'.
- **Intersection events**: Let A and B be two events. The set of elements which are in A and in B is called the intersection event of event A and B. it is denoted by $A \cap B$. $A \cap B$ means the event in which A and B occurs simultaneously.

## Definitions

- **Differnce events**: Let A and B be two events. The set of elements all elements which are in A but not in B is called the difference event of event A and B. it is denoted by A - B. A - B means the event in which 'the event A occurs but the event B does not occurs'.
- **Mutually exclusively events /disjoint event** : If the intersection event of two events is the impossible event then these two events are said to be mutually exclusive. Thus it is $A \cap B = \varnothing$ then A and B are mutually exclusive events.
- **Example:**
  - When a coin is tossed, the event of occurrence of a head and the event of occurrence of a tail are mutually exclusive events.

## Definitions

- **Dependent or Independent events**: Two or more events are considered to be independent if the occurrence of one event is no way affects the occurrence of the other.
- **Example :**
  - Tossing a coin a trial is not affected by the result of the previous trial
- If the occurrence of one event influences the occurrence of the other event, the events are said to be dependent events.
- **Example :**
  - If a card is drawn from a pack of shuffled cards, and not replaced before drawing the second card, then the second card drawn is dependent on the first one.

## Definitions

- **Probability Set Function**: Let U be a finite sample space and let S(U) be its power set. Let P: S(U) → R be a set function satisfying the following postulates.
  1. $P(A) \geq 0$ for every $A \in S(U)$
  2. $P(U) = 1$
  3. If A and B are mutually exclusive events, then $P(A \cup B) = P(A) + P(B)$

  Then function P is said to be probability set function on S(U) and the real number P(A) is said to be the probability of event A

## Definitions

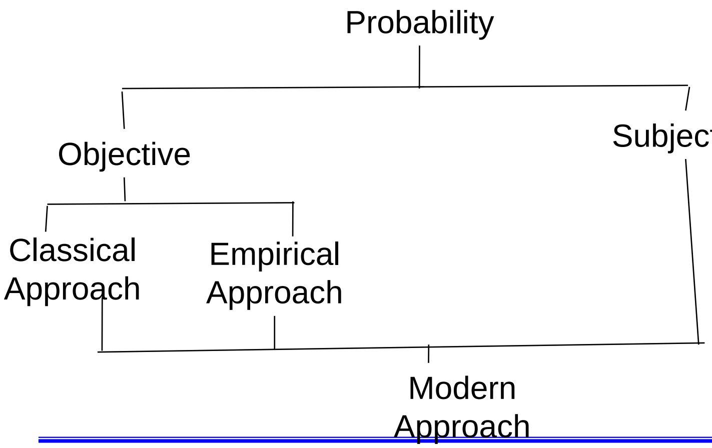
- **Elementary event**: If $U = \{x_1, x_2, \ldots, x_n\}$ is a sample space, then the single element subset $\{x_1\}, \{x_2\}, \ldots, \{x_n\}$ of U are called elementary events or primary events.
- If the probability of each primary event is same, then the primary event are called equi-probable
- If the above n primary events are equi-probable then the probability of each primary event is 1/n.

## Definitions

- **Probability of an event / Classical definition of probability:** If 'S' be the sample space, then the probability of occurrence of an event 'E' is defined as:
- $P(E) = n(E)/N(S) = \dfrac{\text{(number of elements in 'E')}}{\text{(number of elements in sample space 'S')}}$
- **Example:**
  - Find the probability of getting a tail in tossing of a coin.
- Solution:
  - Sample space S = {H,T} and n(s) = 2
  - Event 'E' = {T} and n(E) = 1 therefore P(E) = n(E)/n(S) = ½
- Note: This definition is not true, if (a) The events are not equally likely. (b) The possible outcomes are infinite.

## Types

- The following is the broad classification of the concepts used in probability



## Classical Approach

- If a random experiment results in N exhaustive mutually exclusive and equally likely outcomes out of which m are favourable to the happening of an event A, then the probability of occurrence of A, usually denoted by P(A) is given by : P(A) = m / N
- Example : What is the chance of getting king in a draw from the pack of 52 cards?
- Solution : Total number of cases that can happen = 52 No. of favourable case = 4

  ∴ Probability of drawing a king = 4 / 52 = 1/ 13

## Empirical Approach

- Empirical concept the probability of an event ordinary represents the proportion of times, under identical circumstances, the outcome can be expected to occur.
- The value refers to the event’s long run frequency of occurrence.
- The main assumptions are :
  - The experiments or observations are random. As there is no bias in favour or any outcome all elements enjoy equal chance of selection.
  - There are a large number of observations.

## Empirical Approach

- “If the experiment be repeated a large number of times under essentially identical conditions, the limiting value of the ratio of the number of times the event A happens to the total number of trails of the experiments as the number of trails increases indefinitely, is called the probability of the occurrence of A.”

## Empirical Approach (Example)

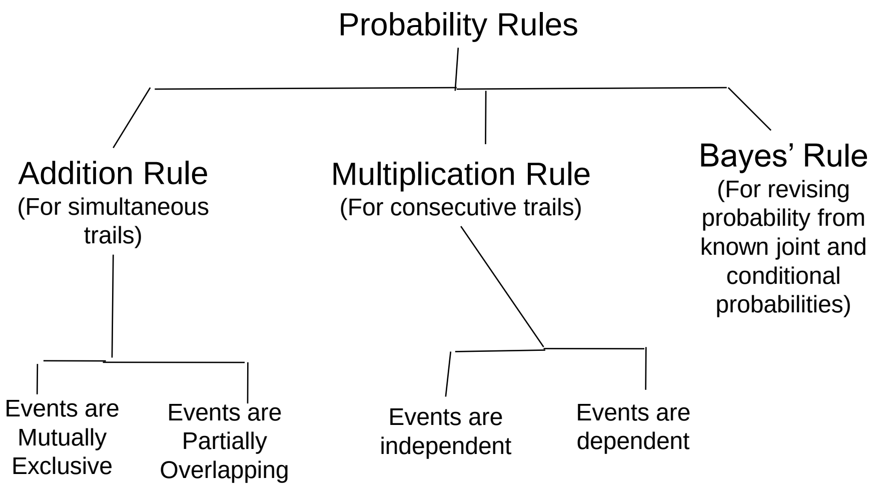
- A foreman in a factory examines the lots of 100 parts each after an interval of half hour during the day and records the number of defective parts. The day’s record of 16 lots reveals the following number of defective parts.

| No. of Defective parts | No. of lots |
|---|---|
| 0 | 1 |
| 1 | 4 |
| 2 | 5 |
| 3 | 3 |
| 4 | 2 |
| 5 | 1 |

## Subjective Probability

- It measures the confidence that an individual has in the truth of a particular proposition. It is bound to vary with person to person and is therefore called subjective probability.

## Basic Probability Rules



Probability Rules

Addition Rule
(For simultaneous trails)

Events are Mutually Exclusive

Events are Partially Overlapping

Multiplication Rule
(For consecutive trails)

Events are independent

Events are dependent

Bayes' Rule
(For revising probability from known joint and conditional probabilities)

## Addition Rule (Mutually Exclusive Events)

- If A and B are mutually exclusive events, then
  $P(A \cup B) = P(A) + P(B)$
  In other words : P(A or B) = P(A) + P(B)
- Example
  The probability that a company executive will travel by train in 2/3 and the he will travel by plane is 1/5. the probability of his travelling by train or plane is :
- Solution :
  P(T or P) = P(T) + P(P)
  = 2/3 + 1/5 = 13/15
  the probability of not travelling by either train or plane =
  1- P(T or P) = 1 – 13 / 15 = 2 /15.

Minal Shah

## Addition Rule (Not Mutually Exclusive Events)

- If A and B are any two events, then
  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  In other words : $P(A \text{ or } B) = P(A) + P(B) - P(A \cap B)$
- Generalization : it can be shown that for any 3 events A, B, C
  $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$
- In general, for any events $A_1, A_2,\ldots, A_n$ we have
  $P(A_1 \cup A_2 \cup \ldots \cup A_n) = P(A_1) + P(A_2) + \ldots.. + P(A_n)$ – (sum of probabilities of all possible intersections taken two at a time) + (sum of probabilities of all possible intersections taken three at a time) + ….. + $(-1)^{n-1}P(A_1 \cap A_2 \cap \ldots\ldots \cap A_n)$

Minal Shah

## Addition Rule (Not Mutually Exclusive Events)

- Deduction : If the two events A and B are mutually exclusive i.e. disjoint, then the addition rule of probability reduces to $P(A \cup B) = P(A) + P(B)$
- In general, for any events $A_1, A_2, \ldots, A_n$ we have $P(A_1 \cup A_2 \cup \ldots \cup A_n) = P(A_1) + P(A_2) + \ldots.. + P(A_n)$

## Multiplication Rule (Events Are Independent)

- The probability that both independent events A and B will occur is $P(A \cap B) = P(A) * P(B)$
  In other words : P(A and B) = P(A) * P(B)

## Multiplication Rule (Events Are Dependent)

- If event A and B are so related that the occurrence of B is affected by the occurrence of A, then A and B are called dependent events. The probability of event B depending on the occurrence of event A is called conditional probability and is written as P(B / A) which may be read, “the probability of B given A.”
- The probability that both the dependent events A and B will occur is $P(A \cap B) = P(A) * P(B/A)$
  In other words : P(A and B) = P(A) * P(B/A)

## Conditional Probability

- If it is given that a particular event B has already occurred then the probability of occurrence of event A is the conditional probability of A. It is denoted by the symbol P(A/B). If P(B) > 0 then

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

- If P(B) = 0 then B does not occur at all and the question of defining P(A/B) does not arise.
- If the events are independent, the happening of B shall not affect the probability of A and therefore

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

## Joint And Marginal Probability

- Joint probability is the probability of the occurrence of two or more events.
- Normally the probabilities can be expressed in a 2x2 table or joint probabilities table.
- Example : Two events A and B, there are four joint probabilities possible as explained below :

| Joint Event | Sample points belonging to the joint event | Probability |
|---|---|---|
| $A \cap B$ | $n(A \cap B)$ | $n(A \cap B) / n(s)$ |
| $A \cap B'$ | $n(A \cap B')$ | $n(A \cap B') / n(s)$ |
| $A' \cap B$ | $n(A' \cap B)$ | $n(A' \cap B) / n(s)$ |
| $A' \cap B'$ | $n(A' \cap B')$ | $n(A' \cap B') / n(s)$ |

## Bayes’ Rule

- Joint and marginal probabilities can be used to revise the probability of a particular event in the light of additional information.
- Example we have two box containing defective and non-defective items.
- One item is picked at random from any one of the boxes and is found defective, and now we might like to known the probability that it came from box one.
- To answer questions of this sort, we use Bayes’ Rule which may be considered as an application of conditional probabilities.

## Bayes' Rule

- The popularity of the theorem has been mainly because of its usefulness in revising a set of old probabilities called prior probabilities (derived subjectively or objectively) in the light of additional information made available and derive a set of new probabilities called the posterior probabilities.
- Although Bayes' Rule may be applied to more than two mutually exclusive and exhaustive events, we shall for the sake of simplicity confine generally to the application of Bayes' rule for two mutually exclusive and exhaustive events.

## Bayes' Rule

- We know that the marginal probabilities are the sum of the two relevant joint probabilities as indicated below.
  - $P(A) = P(A \cap B) + P(A \cap B')$
  - $P(B) = P(A \cap B) + P(A' \cap B)$
- We can restate them in terms of conditional and marginal probabilities as follows
  - $P(A) = P(A /B ) * P(B) + P(A / B') * P(B')$
  - $P(B) = P(B / A) * P(A) + P(B / A' ) * P(A')$
- Now recollect the original formulation of conditional probabilities viz $P(A/B) = P(A \cap B) / P(B) = P(B \cap A) / P(B)$ which can be written as

## Bayes’ Rule

$$P(A/B) = \frac{P(B/A) * P(A)}{P(B/A) * P(A) + P(B/\overline{A}) * P(\overline{A})}$$

- Similarly

$$P(B/A) = \frac{P(A/B) * P(B)}{P(A/B) * P(B) + P(A/\overline{B}) * P(\overline{B})}$$

- Proceeding on the same lines, other conditional probabilities viz P(A’/B) and P(B/A’) can be revised.

## Bayes’ Rule

- Generalization : If an event B can only occur in conjunction with one of the n mutually exclusive and exhaustive events $A_1$, $A_2$, ….., $A_n$ and if B actually happens, then the probability that it was preceded by the particular event $A_i$ (i = 1, 2, …., n) is given by
$P(A_i / B) = P(A_i \cap B) / P(B)$   i = 1, 2, …., n
and $P(B) = P(A_1 \cap B) + P(A_2 \cap B) +….. P(A_n \cap B)$

Minal Shah

# Measure Of Central Tendency

Computer Oriented Numerical and Statistical Methods

## Outline

- Introduction
- Some Definitions
- Types Of Average
- Related Positional Measures
- Empirical Relationships Between Average
- Choice Of A Suitable Average

## Introduction

- The limit of ingenuity of human mind requires that the entire mass of unwieldy data should not only be compressed in tabular form but its chief characteristics should be represented by a single figure which summarizes and represent he characteristics or general significance of data comprising a set of unequal values.
- This single figure is called an average.
- A concise numerical description also enables us to form a mental image of data and interpret its significance.
- This single value is the point of location around which individual values cluster and therefore called the measure of location.
- Since this single value has a tendency to be somewhere at the centre and within the range of all values it is also known as the measure of central tendency.

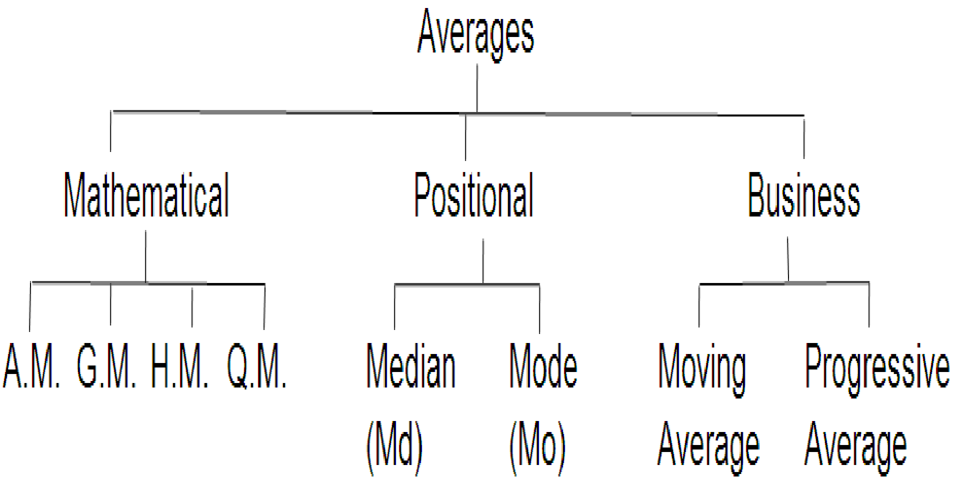
## Some Definitions

- Averages are statistical constants which enable us to comprehend in a single effort the significance of the whole. –A.L.Bowely
- An average is a single value selected from a group of values to represent them in some way a value which is supposed to stand for whole group of which it is part, as typical of all the values in the group. – A.E. Waugh
- A measure of central tendency is a typical value around which other figures congregate. – Simpson and Kafea
- An average stands for the whole group of which it forms a part yet represents the whole. – A. E. Waugh
- Averages / Measure of central tendency / measure of location / measure of central location.

## Objectives

- To get a single value.
- To facilitate comparison.
- To trace precise relationship.
- To know about the universe from a sample.
- To help in decision-making.

## Types of Averages



Averages

Mathematical

A.M. G.M. H.M. Q.M.

Positional

Median (Md)

Mode (Mo)

Business

Moving Average

Progressive Average

## Arithmetic Mean ( $\bar{x}$ )

- The arithmetic average or mean usually denoted by $\bar{x}$ of a set of observations is the sum of the values of all observations in a series (Σx) divided by the number of items (N) contained in the series.
- If $x_1 + x_2 + x_3 + \ldots\ldots.+ x_n$ are the given n observations then

$$\bar{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$

$$Mean = \frac{Sum\ of\ the\ items}{Number\ of\ the\ items} = \frac{\sum X}{N}$$

## Arithmetic Mean ( $\bar{x}$ )

- if the observations $x_1 + x_2 + x_3 + \ldots\ldots.+ x_n$ are repeated $f_1 + f_2 + f_3 + \ldots\ldots+ f_n$ times, then we have

$$(\bar{x}) = \frac{\sum f_i x_i}{\sum f_i}\ i = \frac{f_1 x_1 + f_2 x_2 + f_3 x_3 + \ldots\ldots + f_n x_n}{f_1 + f_2 + f_3 + \ldots\ldots + f_n}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{N}, where\ N = \sum_{i=1}^{n} f_i$$

- In case of continuous or grouped frequency distribution, the value of x is taken as the mid-point of the corresponding class.

## Calculation of Arithmetic Mean (Individual Observations)

- Steps:

1. Add together all the values of the variables X and obtain the total i.e ∑X
2. Divide this total by the number of the observations i.e. N

$$\bar{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$

## Arithmetic Mean ( $\bar{x}$ )

- A.M. is very simple and therefore, commonly used in business and economics.
- Whenever there is a mention of average income, profit, wages, output, the reference is to the arithmetic mean unless there are some qualifying words suggesting some other type of average

## Characteristics of Arithmetic Mean

- If the number of items and their A.M. are both known, the aggregate or the sum of items can be obtained by multiplying the average by the number of items i.e.

$$\overline{x} = \frac{\sum x}{N} \text{ or } N\overline{x} = \sum x$$

## Characteristics of Arithmetic Mean

- If an observation equal to the mean is excluded, the A.M. of the remaining observations remains unchanged. Similarly, if an observation equal to the mean is added to the series, the average of the total observations remains unchanged.
- If a wrong figure is taken in the computation of A.M., the correction can be made without repeating the entire calculation.

## Characteristics of Arithmetic Mean

- The A.M. has two important mathematical properties which provides further mathematical analysis easy and which have made its use more popular than any other type of average.
  - The algebraic sum of the deviations of a given set of individual observations from the A.M. is always zero. Symbolically $\sum (X - \overline{X}) = 0$
  - The sum of squares of deviations of a set of observations is the minimum when deviations are taken from the A.M. symbolically $\sum (X - \overline{X})^2$ is least

## Characteristics of Arithmetic Mean

- If each of the value of a variate X is increase (or decreased) by a constant b, the A.M. also increase (or decreased) by the same amount. And when the values of X are multiplied by a constant say k, the A.M. is also multiplied by the same amount k.

## Short-cut Method Of Computing A.M.

- Here an assumed or an arbitrary average (indicated by ‘A’ or ‘a’) is used as the basis of calculation of deviation from individual values.
- The formula is $\overline{X} = A + \frac{\sum d}{N}$
  - Where A = the assumed mean or arbitrarily selected value
  - d = the deviation of each value from the assumed mean i.e. d = ( X – A)
  - ∑d / N is the correction for the difference between the actual average and the assumed average. The ∑d / N = 0 if the assumed average is equal to the actual average.

## Short-cut Method Of Computing A.M. For Grouped Data

1. Select the assumed mean A.
2. Find the deviation of each class mid-point from the assumed mean in original units of data, such as Rs. , cm, and so on.
   - Let d = the deviation in original units
   - m = each class mid-point
   - ∴ d = m – A
3. Multiply each deviation d by the frequency in the class to obtain the total deviations of the class or fd
4. Add these products to obtain the total deviations of all items included in the distribution, or ∑fd by the sum of the frequencies (∑f or N) to obtain the correction factor

$$\frac{\sum fd}{\sum f} \quad or \quad \frac{\sum fd}{N}$$

## Short-cut Method Of Computing A.M. For Grouped Data

5. Add the correction factor to the assumed mean to obtain exact mean of the grouped data.

$$\overline{X} = A + \frac{\sum fd}{\sum f} \quad or \quad \overline{X} = A + \frac{\sum fd}{N}$$

## Calculation of A.M. For Discrete Data

- Direct method :
  - The formula for computing mean is

$$\overline{X} = \frac{\sum fx}{\sum f} \quad or \quad \overline{X} = \frac{\sum fx}{N}$$

  - Where f = frequency
  - x = the variable in question
  - N = total number of observations i.e. ∑f

## Calculation of A.M. For Discrete Data

- Direct method steps:
  1. Multiply the frequency of each row with the variable and obtain the total ∑fx
  2. Divide the total obtained by step (1) by the number of observation i.e. total frequency.

## Calculation of A.M. For Discrete Data

- Short-cut method :
  - The formula for computing mean is

$$\overline{X} = A + \frac{\sum fd}{\sum f} \quad or \quad \overline{X} = A + \frac{\sum fd}{N}$$

  - Where f = frequency
  - A = assumed mean
  - d = x - A
  - N = total number of observations i.e. ∑f

## Calculation of A.M. For Discrete Data

- Short - cut method steps:
  1. Take an assumed mean
  2. Take deviations of the variables x from the assumed mean and denote the deviation by d
  3. Multiply the deviation with the respective frequency and obtain the total ∑fd
  4. Divide the total obtained by step (3) by the number of observation i.e. total frequency plus the assumed mean.

$$\overline{X} = A + \frac{\sum fd}{\sum f} \quad or \quad \overline{X} = A + \frac{\sum fd}{N}$$

## Calculation of A.M. For Continuous Data

- Direct method :
  - The formula for computing mean is

$$\overline{X} = \frac{\sum fm}{N}$$

  - Where f = frequency
  - m = mid-point of each class
  - N = total frequency i.e. ∑f

## Calculation of A.M. For Continuous Data

- Direct method steps:
  1. Obtain the mid-point of each class and denoted it by m.
  2. Multiply these mid-points by the respective frequency of each class and obtain the total ∑fm.
  3. Divide the total obtained in step (2) by the sum of the frequency i.e. N

## Calculation of A.M. For Continuous Data

- Short-cut method :
  - The formula for computing mean is

$$\overline{X} = A + \frac{\sum fd}{N}$$

  - Where f = frequency
  - m = mid-point of each class
  - A = assumed mean
  - d = deviation of mid-points from the assumed mean i.e. d = m - A
  - N = total frequency i.e. ∑f

## Calculation of A.M. For Continuous Data

- Short-cut method steps:
  1. Take an assumed mean
  2. Obtain the mid-point of each class and denoted it by m.
  3. From the mid-point of each class deduct the assumed mean.
  4. Multiply these deviations by the respective frequency of each class and obtain the total ∑fd.
  5. Apply the formula $\overline{X} = A + \frac{\sum fd}{N}$

## Calculation of A.M. For Continuous Data

- Step- deviation method :
  - In case of grouped or continuous frequency distribution, with class intervals of equal magnitude. The calculation are further simplified by taking d’ = d / h = (m – A) / h
  - Where f = frequency
  - m = mid-value of each class
  - A = assumed mean
  - d = deviation of mid-points from the assumed mean i.e. d = m – A
  - h = is the common magnitude of the *class-i*ntervals
  - N = total frequency i.e. ∑f

$$\overline{X} = A + \frac{\sum fd'}{N} * h$$

## Calculation of A.M. For Continuous Data

- Step-deviation method steps:
  1. Compute d' = (m – A) / h. Where A being any assumed mean, h is the common magnitude (class-interval) of the class. Algebraic sign + or – are to be taken with deviations.
  2. Multiply d' by the corresponding frequency f to get fd'
  3. Find the sum of the products obtained in step (2) to get ∑fd'.
  4. Divide the sum obtained in step (3) by N, the total frequency.
  5. Add A to the value obtained in step (4)

## Advantages of Step-Deviation Method

- All three methods for continuous series gives same answer.
- The direct method though the simplest, involves more calculations when mid-points and frequencies are very large in magnitude. In this case step-deviation method would be far simpler.

## Weighted A.M.

- For calculating simple A.M. , we suppose that all the values of size of items in the distribution have equal importance.
- In case some items are more important than others, a simple average computed is not representative of the distribution.
- In such cases proper weight age has to be given to the various items, the weight attached to each item being proportional to the importance of the item in the distribution.
- The term 'weight' stands for the relative importance of different items.

## Weighted A.M.

- The formula for the weighted A.M. is given by
  - Direct Method : $$\overline{x_w} = \frac{w_1x_1 + w_2x_2 + ....+w_nx_n}{w_1+w_2+...+w_n} = \frac{\sum wx}{\sum w}$$
  - Short cut method: $$\overline{x_w} = Aw + \frac{\sum wd}{\sum w}$$
    - Where Aw = assumed (weighted mean)
    - ∑wd = sum of the product of the deviations from the assumed mean (Aw) multiply by their respective weights.
  - In case of frequency distribution $$\overline{x_w} = Aw + \frac{\sum w(fx)}{\sum w}$$

## Weighted A.M.

- Weights may be either actual or arbitrary i.e. estimated.

1. Simple A.M. shall be equal to the weighted A.M. if the weights are equal. $\overline{x} = \overline{x_w}$
2. Simple A.M. shall be less than the weighted A.M. if and only if greater weights are assigned to greater values and smaller weights are assigned to smaller values. $\overline{x} < \overline{x_w} \;\; if \;\; (w_2 - w_1)(x_1 - x_2) < 0$

## Weighted A.M.

3. Simple A.M. is greater than the weighted A.M. if and only if smaller weights are attached to the higher values and greater weights are assigned to smaller values. $\overline{x} > \overline{x_w} \;\; if \;\; (w_2 - w_1)(x_1 - x_2) < 0$
4. It may be noted that weighted A.M. is specially useful in problems relating to:
   - Construction of index number
   - Standardized birth and death rates
   - Comparison of results of two or more universities where number of students differ

## Geometric Mean

- The G.M. of a series containing N observations is the $N^{th}$ root of the product of the value.
- Ungrouped data :

$$G.M. = \sqrt[n]{\text{the product of n value}}$$

$$G.M. = \sqrt[n]{x_1 * x_2 * \ldots * x_n}$$

- When the number of observations exceed two, the computation are simplified through the use of logarithms, the above formula may be written as

$$\log G.M. = \frac{1}{n}\log(x_1, x_2, \ldots, x_n)$$

$$\log G.M. = \frac{1}{n}[\log x_1 + \log x_2 + \ldots + \log x_n]$$

## Geometric Mean

$$\log G.M. = \frac{1}{n}\log(x_1, x_2, \ldots, x_n)$$

$$\log G.M. = \frac{1}{n}[\log x_1 + \log x_2 + \ldots + \log x_n]$$

- Taking antilog on both sides we have

$$G.M. = \text{Antilog } [\frac{1}{n}\sum \log x]$$

- In discrete series $G.M. = \text{Antilog } [\frac{\sum f * \log x}{N}]$
- In continuous series $G.M. = \text{Antilog } [\frac{\sum f * \log m}{N}]$

## Calculation Of G.M. (Individual observations)

$$G.M. = \text{Antilog } [\frac{\sum \log x}{N}]$$

- Steps:
  - Take the logarithms of the variable X and obtain the total ∑logX.
  - Divide ∑logX by N and take the antilog of the value so obtained. This gives the value of G.M.

## Calculation Of G.M. (Discrete Series)

$$G.M. = \text{Antilog } [\frac{\sum f * \log x}{N}]$$

- Steps:
  - Find the logarithms of the variable X
  - Multiply these logarithms with the respective frequencies and obtain the total ∑f*logX.
  - Divide ∑f*logX by total frequency and take the antilog of the value so obtained. This gives the value of G.M.

## Calculation Of G.M. (Continuous Series)

$$G.M. = \text{Antilog}\ [\frac{\sum f * \log m}{N}]$$

- Steps:
  - Find out the mid-points of the classes and take their logarithms.
  - Multiply these logarithms with the respective frequencies and obtain the total ∑f*logm
  - Divide total obtained in step 2 by the total frequency and take the antilog of the value so obtained. This gives the value of G.M.

## Harmonic Mean

- H.M. is defined as the reciprocal of the A.M. of the reciprocal of the given observations.

$$H.M. = \frac{N}{(\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n})} = \frac{N}{\sum 1/x}$$

## Calculation of H.M. (Individual Observations)

$$H.M. = \frac{N}{(\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n})} = \frac{N}{\sum 1/x}$$

- Where $x_1$, $x_2$, …, $x_n$ etc refers to the various items

## Calculation of H.M. (Discrete Series )

$$H.M. = \frac{N}{\sum f * \frac{1}{x}} = \frac{N}{\sum \frac{f}{x}}$$

- Steps:
  - Take the reciprocal of the various items of the variable x
  - Multiply the reciprocal by the frequencies and obtain the total $\sum(f * 1/x)$
  - Substitute the values of N and $\sum(f * 1/x)$ in the above formula

## Calculation of H.M. (Continuous Series )

$$H.M. = \frac{N}{\sum \frac{f}{m}}$$

- Steps:
  - The reciprocal of the mid values of the class intervals (m) are found
  - Multiply the reciprocal with the respective class frequencies and obtain the total ∑(f * 1/m)
  - The total of the product is divided by the total number of items and the reciprocal of the resultant figure is the H.M. = Reciprocal [(∑(f * 1/m))/N]

## Positional Average

- These averages are based on the position of a given observations in a series, arranged in an ascending or a descending order.
- The magnitude or the size of the values does not matter as was in the case of earlier averages.
- Ex. Median and Mode

## Median

- The median, as the name suggests, is the middle value of a series arranged in any of the order of the magnitude.
- The middle value in the case of a series with odd number of items can be easily located e.g. the 6th value if the number of items is 11.
- But in case the total number is even, say 10 there will be two middle values, viz, 5th and 6th and in which case the mean of the two middle values shall constitute the median.
- The median is just 50th percentile value below which 50% of the values in the sample fall.

## Median

- It splits the observations into two halves.
- The median is that value of the variable which divides the group into two equal parts, one part comprising all values greater, and the other, all values less than median -------- L.R. Cannor
- The central values of the distribution, a value such that, greater and smaller values occur with equal frequency, is known as median. ---- Kenney and Keeping

## Computation Of Median

- In calculation of median, there are two stages
  - Arrange the data in ascending or descending order of magnitude (Both arrangement gives same answer)
  - The search for the middle item which is indicated by (N+1)/2 or N/2th item determined on the basis of the total number of items. The value of this middle item is the median value

## Computation Of Median (Ungrouped Data/ Individual Observations)

- In case of ungrouped data, median is the middle value of the series arranged in either ascending or descending order.
- Taking the total items equal to N items, median value (abbreviated as Md) is the value of the (N+1)/2th item.

## Computation Of Median (Grouped Data)

- In grouped distribution, values are associated with frequencies.
- Grouping can be in the form of a discrete frequency distribution or continuous frequency distribution.
- Whatever may be the mode of a distribution, cumulative frequencies have to be calculate to know the total number of items.

## Computation Of Median (Discrete Series)

1. Arrange the data in ascending or descending order of magnitude.
2. Find out the cumulative frequencies [less than type]
3. Apply the formula : Median = size of (N+1)/2.
4. Now look at the cumulative frequency column and find that total which is either equal to (N+1)/2 or next higher to that and determine the value of the variable corresponding to it. That gives the value of median.

## Computation Of Median (Continuous Series)

- The method is the same as shown for the discrete frequency distribution. The two additional points are:

1. In a continuous distribution the value of N/2th item is taken and not of (N+1)/2th item. The reason is that in a continuous distribution only N+1 class limit can make N class division.
2. Having discovered the class in which the median value lies, the exact value has to be interpolated from the relevant class or on the basis of the position of N/2th value in the total frequency concentration in that class. The formula used for this purpose is as follows:

## Computation Of Median (Continuous Series)

$$\text{Median} = l_1 + \frac{\frac{N}{2} - C}{f} * h$$

where $l_1$ = real lower limit of the median class

N/2 = item whose value has to be interpolated

C = cumulative frequency of class preceding the median class

f = frequency of the median class

h = size of the class interval in the median class

## Usefulness of Median

- The median is useful for distribution containing open-end intervals.
- It is used when we require a measure of location which is not affected by high or low value item, and when we wish to measure the change in different sets of distribution which move in a similar direction in similar manner.

## Mode

- The mode refers to that value in a distribution which occurs most frequently (greatly frequency)
- Mode is the value occurring most frequency in a set of observations and around with other items of the set cluster most densely.
- Its importance is very great in marketing studies where a manager is interested in knowing about the size which has the highest concentration of items. [Ex. Size of shoe]
- It is not affected by extreme values.

## Computation Of The Mode (Ungrouped Data/ Individual Observations)

- The data is placed in the form of an array so that items having the same values can be identified and quickly counted, the value of that item which occurs most frequently is the modal value.

## Computation Of The Mode (Grouped Data Discrete Series)

- In uni-modal distribution where the highest concentration is in a single discrete value there should not be any difficult in locating modal value or a class-interval containing this value by inspection.
- Some difficulty arise when nearly equal concentrations are found in two or more neighbouring values. There are two ways of dealing with this situations.
  - In a large majority of cases it shall be possible to make a choice of one value by taking the totals of 3 values, the value with highest concentration and it two neighbouring values in case of the competing cases. The central values of the group which yield higher total should be selected.

# Computation Of The Mode (Discrete Series)

1. We prepare a grouping table with 6 columns
2. In column I we write down the frequency against the respective items.
3. In column II, the frequency is grouped in twos, starting from the top. Their totals are found out and the highest total is marked.
4. In column III, the frequency is again grouped in two, leaving the first frequency. Highest total is again marked.
5. In column IV, the frequency is grouped in three starting from the top and their totals found out with highest frequency is marked.

# Computation Of The Mode (Discrete Series)

6. In column V, the frequencies are again grouped in three, leaving the first frequency. The totals are found out and the highest total is marked.
7. In column VI, leaving the first and second frequency, group is done in threes. After finding their total, the highest total is marked.
8. In analysis table, in order to find out the item which repeats largest number of items, grouping in analyzed.

## Computation Of The Mode (Continuous Series Interpolation Formula)

- The exact value of mode in the case of continuous frequency distribution can be obtained by the following formulae: $\text{Mode} = l_1 + \frac{\Delta_1}{\Delta_1 + \Delta_2} * h$

where $l_1$ = the real lower limit of the modal class
h = magnitude of the modal class
$\Delta_1 = f_1 – f_0$
$\Delta_2 = f_1 – f_2$
$f_1$ = frequency of the modal class
$f_0$ = frequency of the class preceding the modal class
$f_2$ = frequency of the class succeeding the modal class

## Computation Of The Mode (Continuous Series Interpolation Formula)

$$\text{Mode} = l_1 + \frac{\Delta_1}{\Delta_1 + \Delta_2} * h$$

the above formula can be written as

$$\text{Mode}(Mo) = l_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} * h$$

## Empirical Relationship Between Average

- Karl Pearson’s formula
  - Mean = Mode – 3(Mean – Median)
  - Mean = Mode + 3/2 (Median - Mode) = ½(3Median – Mode)
  - Mode = 3Median – 2 Mean
  - Median = Mode + 2/3(Mean – Mode) = 1/3(2Mean + Mode)
- Unless the values of all items in a distribution are identical the G.M. will be smaller than the A.M. and the H.M. shall be smaller than both G.M. and A.M.
  - A.M. ≥ G.M. ≥ H.M.

## Further Partition Of Series

- Just as a median divides the distribution into two parts, there are other positional measures which partition a series into still smaller parts say 4, 10 or 100.
- The value which divides the series into a number of equal parts are called partition value
- The median divides the lower 50% of a set of data from the upper 50% of a set of data.

## Quartiles

- The quartiles divide data sets into fourths or four equal parts.
- There are 3 quartiles.
- The 1st quartile, denoted $Q_1$, divides the bottom 25% the data from the top 75%. Therefore, the 1st quartile is equivalent to the 25th percentile.
- The 2nd quartile divides the bottom 50% of the data from the top 50% of the data, so that the 2nd quartile is equivalent to the 50th percentile, which is equivalent to the median.
- The 3rd quartile divides the bottom 75% of the data from the top 25% of the data, so that the 3rd quartile is equivalent to the 75th percentile.

## Deciles

- These are the values which divides the total number of observations into 10 equal parts.
- Obviously there are 9 deciles, $D_1$, $D_2$, $D_3$, ….., $D_9$
- These are called as first deciles, second deciles etc. Also $D_1 < D_2 < …. < D_9$; $D_5 = Q_2 =$ Median

## Percentiles

- The percentile value divide the distribution into 100 parts each containing 1 percent of the observations.
- A division in so many parts is used only when there are a considerable number of observations, at least one thousand, otherwise there would be too few observations in each division to be significant.
- In general, the ***k*th percentile**, denoted $P_k$ , of a set of data divides the lower *k*% of a data set from the upper (100 – *k*) % of a data set.
- In particular $P_{10} = D_1$; $P_{20} = D_2$; $P_{30} = D_3$
  $P_{25} = Q_1$; $P_{50} = Q_2 = D_5$ = Median ; $P_{75} = Q_3$
  $P_1 \le P_2 \le \ldots\ldots \le P_{99}$

## Computation Of Partition Values (Discrete Series)

- The quartile, decile and percentile are determined by the same technique which is used in the computation of median.
- For quartile the value of (N+1)/4th item, for deciles, the value of (N+1)/10th item and for percentile the value of (N+1)/100th item is used.
- Now for say 2nd and 3rd quartiles, we have

$$2*\left(\frac{N+1}{4}\right)^{th}\ item$$

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

## Computation Of Partition Values (Discrete Series)

- Similarly for 3rd and 4th deciles, we have

$$3*(\frac{N+1}{10})^{th}\ item$$

$$4*(\frac{N+1}{10})^{th}\ item$$

- Again say 33rd and 44th percentiles, can be computed using formula

$$33*(\frac{N+1}{100})^{th}\ item$$

$$44*(\frac{N+1}{100})^{th}\ item$$

## Computation Of Partition Values (Continuous Series)

- In case of a continuous series N+1 would be replaced by N. The formula for interpolation will be

$$l_1 + \left(\frac{kN \quad - C}{f}\right) * h$$

  - Where kN is the item number depending on the type of partition sought e.g. it is 3N/10 or 33N/100 for 3rd decile and 33rd percentile, etc.
  - C is the cumulative frequency of the class previous to the class in which 3N/10th or 33N/100th item lies.
  - 'f' is the frequency of the relevant class referred to above.
  - 'h' is the size of class interval of the relevant class referred above. .

## Computation Of Quartiles

1. Compute N/4 where N= Σf
2. Find out the cumulative frequency (less than) just greater than N/4.
3. The corresponding class contains $Q_1$ and the value of $Q_1$ is obtained by the interpolation formula:

$$l_1 + \left(\frac{\frac{N}{4} - C}{f}\right) * h$$

- Where $l_1$ is the lower limit of the class containing $Q_1$
- f is the frequency of the class containing $Q_1$
- h is the magnitude of the class containing $Q_1$
- C is the cumulative frequency of the class preceding the

## Computation Of Quartiles

- Similarly to compute $Q_3$ we obtain cumulative frequency (less than) just greater than 3N/4 item, and the value of $Q_3$ is given by the formula

$$l_1 + \left(\frac{\frac{3N}{4} - C}{f}\right) * h$$

## Computation Of Deciles

1. Compute the i^th decile Di (i = 1, 2, …, 9) we find out cumulative frequency just greater than i * N/10 and the corresponding class contains Di and its value is obtained using the interpolation formula:

$$D_i = l_1 + \left(\frac{\frac{iN}{10}-C}{f}\right) * h; \text{ (I =1,2,…,9)}$$

- Where $l_1$ is the lower limit of the class containing $D_i$
- f is the frequency of the class containing $D_i$
- h is the magnitude of the class containing $D_i$
- C is the cumulative frequency of the class preceding the class containing $D_i$.

## Computation Of Percentiles

1. Compute the i^th percentile Pi (i = 1, 2, …, 99) we find out cumulative frequency just greater than i * N/100 and the corresponding class contains Pi and its value is obtained using the interpolation formula:

$$P_i = l_1 + \left(\frac{\frac{iN}{100}-C}{f}\right) * h; \text{ (i =1,2,…,99)}$$

- Where $l_1$ is the lower limit of the class containing $P_i$
- f is the frequency of the class containing $P_i$
- h is the magnitude of the class containing $P_i$
- C is the cumulative frequency of the class preceding the class containing $P_i$.

Minal Shah

**Mute ur call**

# Collection Of Data

Computer Oriented Numerical and Statistical Methods

## Outline

- Collection Of Data
- Primary and Secondary Data

## Introduction

- A statistical inquiry is nothing but a systematic search for truth.
- It seeks some authentic answers to a problem which is quantifiable and therefore amenable to statistical treatment.
- Statistical inquiry, like many other scientific inquires has to pass through the following four stages:
  - Observation
  - Laying down of hypothesis
  - Prediction
  - Verification

# Planning Of Statistical Inquiry

- Purpose
- Scope of enquiry
- Definition of terms
- Preparation of dummy reports
- Laying down hypothesis.
- Type of enquiry:
  - Official, semi-official or unofficial
  - Initial or repetitive
  - Confidential or non-confidential
  - Direct or indirect
  - Regular or ad-hoc
  - Primary or secondary
  - Census or sample.

# Collection of Data

- Source of information : Internal & External
- Degree of precision required
- Primary and secondary data.
- Statistical data may be classified into primary and secondary depending upon the nature of data and mode of collection.
- The data gathered by actual observation, measurements and count and direct recording during the course of an investigation is called primary data.
- This is called primary because it is collected from the very source where the information got generated.
- But any data detached from the original source, one which is reprocessed for one’s own purpose and published by agency other than which originally gathered it, becomes secondary data.

## Collection of Data

- The difference between the primary and the secondary data is only one of degree of detachment with the original source.
- The data which is primary in the hands of one may become secondary in the hands of others.
- A primary source usually has more detailed information particularly on the procedures followed in collecting and compiling the data.
- It may be noted that a given source may be partly primary and partly secondary.

## Use Of Primary Source Wherever Possible

- The secondary source may contain mistakes due to errors in transcription made when the figures were copied from the primary source.
- The primary source frequently includes definitions of terms and units used.
- The primary sources often includes a copy of the schedule and a description of the procedure used in selecting the sample and in collecting the data.
- Primary source usually shows data in greater detail.

## Secondary Data Offers The Following Advantages

- It is highly convenient to use information which someone else has compiled. There is no need for printing data collection forms, hiring enumerators, editing and tabulating the results, etc. Researchers alone or with some clerical assistance may obtained information from published records compiled by someone else.
- If secondary data are available they are much quicker to obtain than primary data.
- Secondary data may be available on some subjects where it would be impossible to collect primary data. For example, census data cannot be collected by an individual or research organization, but can only be obtained from Government publications.

## Major Problems Encountered In Using Secondary Data

- The difficult of finding data which exactly fit the need of the present project.
- Finding data which are sufficiently accurate.

## Methods of Collecting Primary Data

- Direct personal interviews
- Indirect oral interviews
- Information from correspondents or local agents
- Mailed questionnaires and schedules
- Schedules (questionnaires) to be filled in by enumerators.

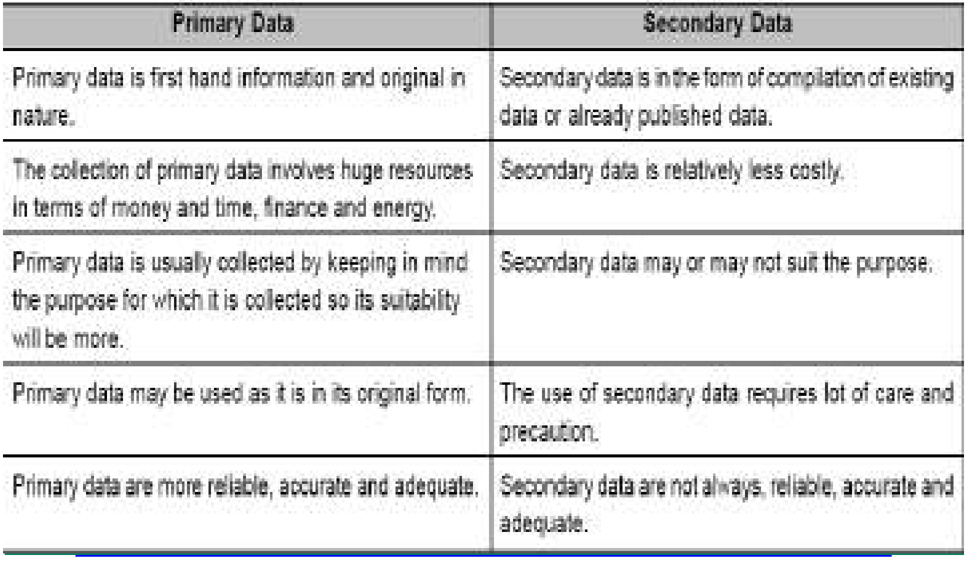
## Sources Of Secondary Data

- Published sources
  - Official publications of central government
  - Publication of semi-government organization
  - Publication of research institution
  - Publication of commercial and financial institutions
  - Newspapers and periodicals
  - Reports of various committees and commissions appointed by the government
  - Publication of international bodies.
- Unpublished source
  - Documents prepared for purpose of registration, applications for permits, licenses, loans etc.
  - Record relating to internal activities of institution.

## Secondary Data

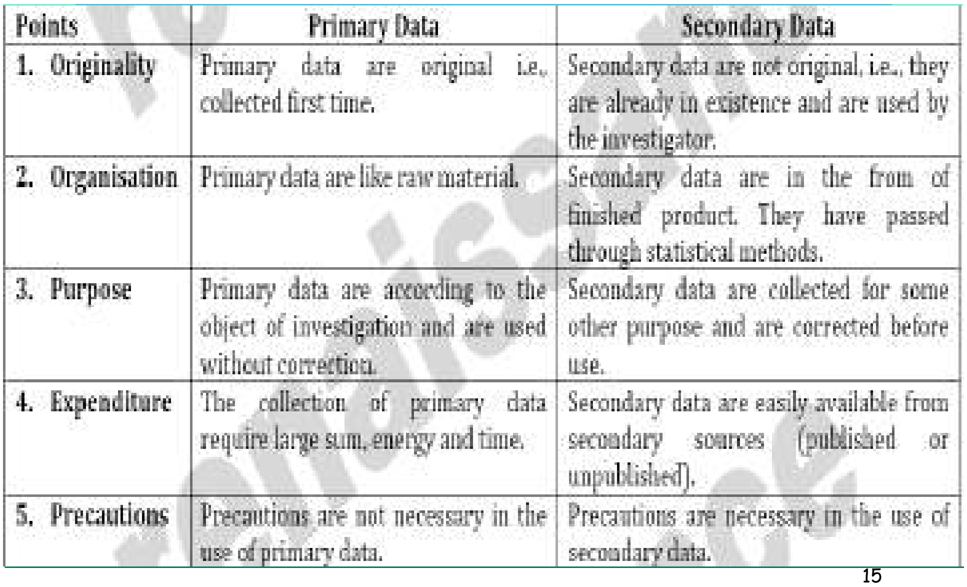
- Thus, before using such data, the investigator should consider the following aspects:
  - Suitability
  - Adequacy
  - Reliability

## Distinction between Primary and Secondary Data



| Primary Data | Secondary Data |
|---|---|
| Primary data is first hand information and original in nature. | Secondary data is in the form of compilation of existing data or already published data. |
| The collection of primary data involves huge resources in terms of money and time, finance and energy. | Secondary data is relatively less costly. |
| Primary data is usually collected by keeping in mind the purpose for which it is collected so its suitability will be more. | Secondary data may or may not suit the purpose. |
| Primary data may be used as it is in its original form. | The use of secondary data requires lot of care and precaution. |
| Primary data are more reliable, accurate and adequate. | Secondary data are not always, reliable, accurate and adequate. |

## Difference between Primary and secondary data



| Points | Primary Data | Secondary Data |
|---|---|---|
| 1. Originality | Primary data are original i.e., collected first time. | Secondary data are not original, i.e., they are already in existence and are used by the investigator. |
| 2. Organisation | Primary data are like raw material. | Secondary data are in the from of finished product. They have passed through statistical methods. |
| 3. Purpose | Primary data are according to the object of investigation and are used without correction. | Secondary data are collected for some other purpose and are corrected before use. |
| 4. Expenditure | The collection of primary data require large sum, energy and time. | Secondary data are easily available from secondary sources (published or unpublished). |
| 5. Precautions | Precautions are not necessary in the use of primary data. | Precautions are necessary in the use of secondary data. |

# Frequency Distribution

Computer Oriented Numerical and Statistical Methods

## Outline

- Classification of data
- Objective of classification
- Types / Mode of Classification
- Frequency distribution
- Basic components of a frequency distribution
- Cumulative frequency
- Tabulation

## Introduction

- In any statistical investigation, once the data is collected and edited. “*the first task of the statisticians is the organization of the figures in such a form that their significance, for the purpose in hand, may be appreciated, that comparison with masses of similar data may be facilitated, and that further analysis may be possible.”*
- This is done through classification and tabulation .
- Classification refers to the determination of various class, categories or group heads in which the whole data shall be distributed.

## Introduction

- Tabulation refers to actual sorting and placing of the data in well-designed and systematic tables according to a given mode of classification.
- Tabulation thus is concerned with systematic arrangement and presentation.
- Classification is necessary and always precedes tabulation but after the determination of class categories the mode of presentation, only may take many forms.

## Objective of Classification

- To condense the mass of data.
- To enable grasp of data.
- To prepare the data for tabulation
- To study the relationships.
- To facilitates comparison.

## Rules of Classification

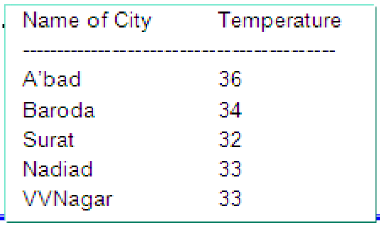
- Exhaustive
- Mutually Exclusive
- Suitability
- Stability
- Homogeneity
- Flexibility

# Types / Modes of Classification

- Modes of classification refers to the class categories or designations into which the data could be sorted out and tabulated.
- These categories depends generally on the nature of the data but more particularly on the purpose for which the data is being sought.
- Some common modes of classification are:
  - Geographical i.e. area-wise or region – wise
  - Chronological, temporal or historical, i.e., with respect to occurrence of time.
  - Qualitative i.e. by character or by attribute.
  - Numerical, quantitative, or by magnitudes.

# Geographical Classification

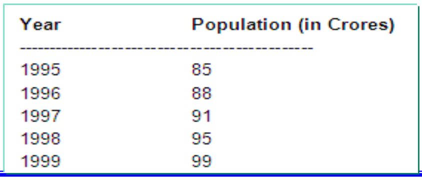
- It is a classification based on geographical regions.
- If the existing political boundaries are taken as the basis, the classification may be done by states, districts or talukas.
- The listing of individual entries in a geographical classifications may be done in an alphabetic order or according to size or value to lay more emphasis on the important area or region.
- Example :



| Name of City | Temperature |
|---|---|
| A'bad | 36 |
| Baroda | 34 |
| Surat | 32 |
| Nadiad | 33 |
| VVNagar | 33 |

# Chronological Classification

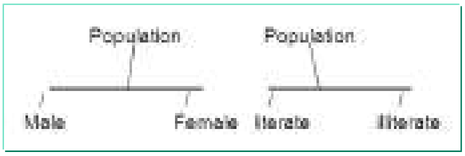
- When statistical data are classified according to the time of its occurrence, the type of classification is known as chronological classification.
- Data regarding sales of firm, population, imports and exports etc. are always subjected to chronological classification.
- Example :



| Year | Population (in Crores) |
|---|---|
| 1995 | 85 |
| 1996 | 88 |
| 1997 | 91 |
| 1998 | 95 |
| 1999 | 99 |

# Qualitative Classification

- When the data are classified according to some qualitative phenomenon which are not capable of quantitative measurement like beauty, honesty, employment, intelligence, occupation, literacy etc. the classification is termed as qualitative classification or descriptive with respect to attributes.
- In qualitative classification the data are classified according to the presence or absence of the attributes in given data.
- Example :



## Quantitative Classification

- If the data are classified on the basis of phenomenon which is capable of quantitative measurement like age, height, weight, production, income, prices, sales, profits, expenditure, etc. it is termed as quantitative classification.
- The quantitative phenomenon under study is known as variable and hence the classification is also sometimes called classification by variables

## Variables

- Variable : A variable in statistical methods stands for any measureable quantity which can assume a range of numerical values with certain limits. Example age, weight, height, price, wages are all variables.
- A variables may be classified into discrete and continuous.
- Discrete variable : A discrete variable is characterized by jumps and gap between one value and next. i.e. it takes integral values depending upon the variables under study.
  - **Example :** Number of students, number of rooms in hostel.

## Variables

- Continuous variable : Those variable which can take all possible values (integrals as well as fractional) in a given specified range are termed as continuous variables.
  - **Example** : Age of student, height, weight, distance.

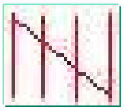
## Frequency Distribution

- A frequency distribution is a series where a number of observations with similar or closely related values are put in separate bunches or groups, each group being in order or magnitudes in a series.
- A classification according to the number possessing same value of the variable
- Frequency distribution is a statistical table which shows the set of all distinct values of the variable arranged in order of magnitude, either individually or in groups with their corresponding frequencies side by side
- The frequency distribution has two parts :
  - On its left there are sizes or magnitudes of values.
  - On its right the number of time a value or a group of values has repeated.

## Frequency Distribution Tables

- A **frequency distribution table** consists of at least two columns - one listing categories on the scale of measurement (X) and another for frequency (f).
- In the X column, values are listed from the highest to lowest, without skipping any.
- For the frequency column, tallies are determined for each value (how often each X value occurs in the data set). These tallies are the frequencies for each X value.
- The sum of the frequencies should equal N.

## Frequency Distribution Tables

- Tally Marks :
  - It helps in the preliminary construction of frequency distribution.
  - It facilitates counting the frequency of a value of a variate in a systematic manner.
  - The distinct values of the variate are written down in ascending (or descending) order in a column.
  - Scan the data and insert a tally mark in appropriate box. For counting, we use tally marks |||| and the fifth tally mark is entered as  by crossing diagonally the four tally marks already entered.
  - Tallies are usually marked in bunches of five.
  - It is not in final presentation of a frequency distribution.

## Discrete or Ungrouped Frequency Distribution

- The type of representation of the data in the above examples is called discrete or ungrouped frequency distribution.
- In this form of distribution, the frequency refers to a given discrete values, the number of rooms in a house, the number of house in locality, the number of companies registered in India. etc.

## Continuous Frequency Distribution

- In this form of frequency distribution the frequencies refers to groups of values.
- It can take fractional values.
- Discrete variable can be presented in the form of continuous frequency distribution when discrete distribution is likely to be too long and unwieldy to handle and somewhat odd in presentation.
- Thus the steps in preparing the grouped frequency distribution are:
  - Determining the class intervals.
  - Recording the data using tally marks.
  - Finding frequency of each class by counting the tally marks.

## Basic Components of Frequency Distribution

- **Class interval (or class):**
  - A large number of observations varying in a wide range are usually classified in several groups according to the size of values.
  - Each of these groups defined by an interval known as class interval is specified by two extreme values called the class limits, the smaller one being termed as the lower limit and the larger one the upper limit of the class.
- **Class-limits:** The maximum and minimum values of a class-interval are called upper class limit and lower class-limit respectively

## Basic Components of Frequency Distribution

- **Type of Class interval :**
  - The lower class limits are represented by $l_1$ and the upper class limit by $l_2$.

**Class limits in any of the following forms**

| A | B (Exclusive type) | C (Inclusive type) | D (open-end classes) |
|---|---|---|---|
| 50 and under 60 | 50-60 | 50-59 | Below 60 |
| 60 and under 70 | 60-70 | 60-69 | 60-70 |
| 70 and under 80 | 70-80 | 70-79 | 70-80 |
| 80 and under 90 | 80-90 | 80-89 | 80-90 |
| 90 and under 100 | 90-100 | 90-99 | 90-100 |
| 100 and under 110 | 100-110 | 100-109 | 101 & above |

## Basic Components of Frequency Distribution

- **Class-mark, or, Mid-value:** The class-mark, or, mid-value of the class-interval lies exactly at the middle of the class-interval.
  - It lies half way between the class limits or the class boundaries.
  - Class mark = (lower class limit + upper class limit) /2
  - Class mark = (lower class boundary + upper class boundary) /2
  -

| Class limit | class boundaries | class mark |
|---|---|---|
| 50 – 59 | 49.5 ---- 59.5 | 55 |

## Basic Components of Frequency Distribution

- **Width (or size) of the class interval:** The difference between the lower class boundary and upper class boundary (not class limits) is called width or size of a class and denoted by 'h'.
  - Width of the class (h) = upper class boundary - lower class boundary
  - Generally preferable to have classes of equal width.
- **Class Frequency** : The number of observations falling within a particular class interval is called frequency or simply frequency.
- Frequency density : It is the frequency of a class per unit of width and indicates the concentration of frequency in a class Frequency density = class frequency/ width of the class

## Guidelines for Constructing a Frequency Distribution

- There should be between 5 and 20 classes.
- The class width should be an odd number.
- The classes must be mutually exclusive.
- The classes must be continuous.
- The classes must be exhaustive.
- The class must be equal in width.
- **Range** : The difference between the largest and smallest observation in the given data gives the range.
- Size of class intervals = range / number of classes

  $= \text{range} / 1 + 3.322 \log_{10} N$

  where N is number of observations.

## Procedure for Constructing a Grouped Frequency Distribution

$\text{Log}_a a = 1$

| **No. of x** | **0.0001** | **0.001** | **0.01** | **0.1** | **1** | **10** | **100** | **1000** |
|---|---|---|---|---|---|---|---|---|
| X written as power of 10 | $10^{-4}$ | $10^{-3}$ | $10^{-2}$ | $10^{-1}$ | $10^{0}$ | $10^{1}$ | $10^{2}$ | $10^{3}$ |
| Log of x to the base 10 | -4 | -3 | -2 | -1 | 0 | 1 | 2 | 3 |

## Constructing  of a Grouped Frequency Distribution

- Range : The difference between the largest and the smallest observations in the given data gives the range.
- Number of class intervals : Divide the range into a suitable number  of classes intervals basically depends on
  - The number of items to be classified.
  - The magnitude of the items.
  - The accuracy desired.
  - The ease of calculation for further processing of the data
- The general rule is to have six to fifteen classes, the choice of actual number will depend on the number of observations and the size of the class interval desired.

## Procedure for Constructing a Grouped Frequency Distribution

- Find the highest and lowest value.
- Find the range.
- Select the number of classes desired.
- Find the width by dividing the range by the number of classes and rounding up.
- Select a starting point (usually the lowest value); add the width to get the lower limits.
- Find the upper class limits.
- Find the boundaries.
- Tally the data, find the frequencies and find the cumulative frequency.

## Cumulative Frequency Tables

- The cumulative frequency table of a set of data is a table which indicates the sum of the frequencies of the data up to a required level. It can be used to determine the number of items that have values below a particular level.
- The cumulative frequencies is of **‘less than**’ type will represent the total frequency of all values less than and equal to the class values to which it relates.
- The cumulative frequencies is of **‘more than’** type will represent the total frequency of all values more than and equal to the class values to which it relates.
- A frequency distribution showing the cumulative frequencies against values of variables systematically arrange increasing (or decreasing) order is known as cumulative frequency distribution

## Tabulation Of Data

- **Meaning:**
- Tabulation means, a systematic presentation of numerical data in columns and rows in accordance with some salient features or characteristics.
- Columns are vertical arrangement and rows are horizontal arrangement.
- Croxton and Cowden state that, ”either for one’s own use or for the use of others, the data must be presented in a suitable form”.
- It facilitates comparison and also facilitates analysis.

## Tabulation Of Data

- **Definitions:**
- Tables are a means of recording in permanent form the analysis that is made through classification and of placing things that are similar and should be compared.
- The purpose of a table is to summarize a mass of numerical information and to present it in the simplest form consistent with the purpose for which it is to be used.

## Objectives (purposes): of Tabulation

- Tabulation is a medium of communication of great economy and effectiveness for which ordinary prose is inadequate.
- In addition to its function in simple presentation the statistical table is a useful tool of analysis.
- The main objectives of tabulation are:
  - To clarify the object of investigation
  - To simplify complex data.
  - To clarify the characteristics of data
  - To present facts in the minimum of space

## Objectives (purposes): of Tabulation

- To facilitate comparison
- To detect errors and omission in the data
- To depict trend and tendencies of the problem under consideration
- To facilitate statistical processing
- To help reference

## Importance of Tabulation

- Under tabulation, data is divided into various parts and for each part there are totals and sub totals. Therefore, relationship between different parts can be easily known.
- Since data are arranged in a table with a title and a number so these can be easily identified and used for the required purpose
- Tabulation makes the data brief. Therefore, it can be easily presented in the form of graphs.
- Tabulation presents the numerical figures in an attractive form.
- Tabulation makes complex data simple and as a result of this, it becomes easy to understand the data.
- This form of the presentation of data is helpful in finding mistakes.

## Importance of Tabulation

- Tabulation is useful in condensing the collected data.
- Tabulation makes it easy to analyze the data from tables.
- Tabulation is a very cheap mode to present the data. It saves time as well as space.
- Tabulation is a device to summaries the large scattered data. So, the maximum information may be collected from these tables.

## Difference Between Classification And Tabulation:

- Classification and tabulation are important processes in statistical investigation.
- Through these processes, the collected data are summarized and put in a systematic order.
  - Both classification and tabulation are important for statistical information. First the data are classified, then, they are presented in tables. Classification is the basis for tabulation.
  - Tabulation is a mechanical function of classification, because in tabulation classified data are placed in columns and rows.
  - Classification is a process of statistical analysis; tabulation is a process of presenting data in a suitable structure.

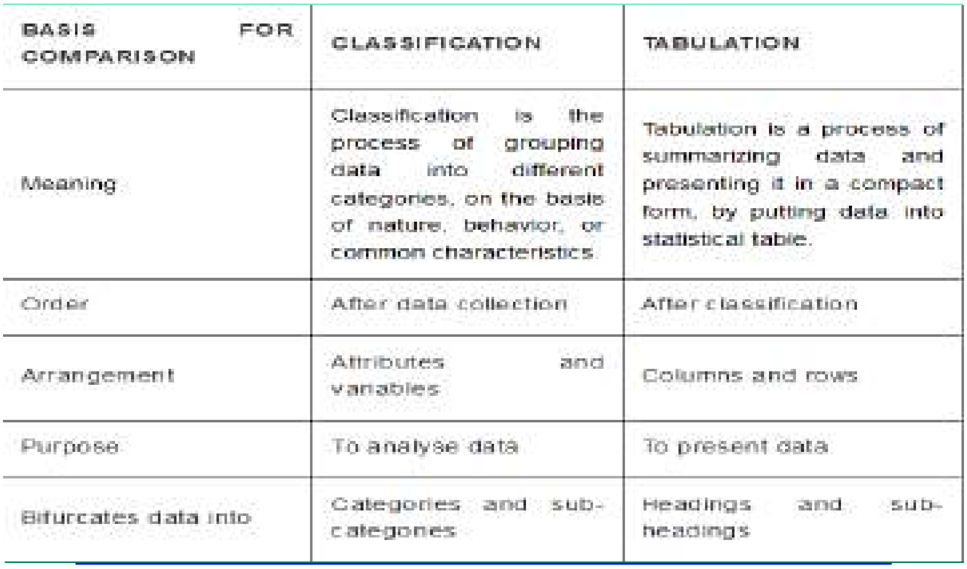
## Difference Between Classification And Tabulation:

- Classification and tabulation are important processes in statistical investigation.
- Through these processes, the collected data are summarized and put in a systematic order.
  - Both classification and tabulation are important for statistical information. First the data are classified, then, they are presented in tables. Classification is the basis for tabulation.
  - Tabulation is a mechanical function of classification, because in tabulation classified data are placed in columns and rows.
  - Classification is a process of statistical analysis; tabulation is a process of presenting data in a suitable structure.

## Limitation of Tabulation

- Tables contain only numerical data. They do not contain details.
- qualitative expression is not possible through tables.
- Tables can be used by experts only to draw conclusions. Common men do not understand them properly.

## Classification Vs Tabulation



| BASIS FOR COMPARISON | CLASSIFICATION | TABULATION |
|---|---|---|
| Meaning | Classification is the process of grouping data into different categories, on the basis of nature, behavior, or common characteristics. | Tabulation is a process of summarizing data and presenting it in a compact form, by putting data into statistical table. |
| Order | After data collection | After classification |
| Arrangement | Attributes and variables | Columns and rows |
| Purpose | To analyse data | To present data |
| Bifurcates data into | Categories and sub-categories | Headings and sub-headings |

# Graphical Presentation of Frequency Distribution

Computer Oriented Numerical and Statistical Methods

## Outline

- Diagrammatic Presentation
  - Meaning
  - Objective
  - Deficiencies/ restrictions
  - Rules for constructing a diagram
  - Types of diagram
- Graphic Presentation
  - Meaning
  - Advantages
  - Types of graphs
- Significance of diagrams and graphs
- Comparison of tabular and diagrammatic presentation
- Difference between diagrams and graphs

## Introduction

- Although tabulation is very good technique to present the data, but diagrams are an advanced technique to represent data.
- As a layman, one cannot understand the tabulated data easily but with only a single glance at the diagram, one gets complete picture of the data presented.
- According to M.J. Moroney, "diagrams register a meaningful impression almost before we think.
- Diagrams refers to the various types of devices such as bars, circles, maps, pictorials, cartograms etc. which are strictly speaking not graphic device.

## Rules For Making A Diagram

- Diagrammatic presentation of a statistical table is simple and effective as photographic memory will last long in the mind than any other form.  The construction of a diagram is an art, which can be acquired through practice.
- However, the following guide line will help in making them more effective:
  - **Heading:** every diagram must have suitable title. The title, in bold letters, conveys the main facts depicted by the diagram. If needed, sub-headings can also be given.  It must be brief, self-explanatory and clear.
  - **Size:**  The size of the diagram should neither be too big nor too small.  It must match with the size of the paper.  It should be in the middle of the paper

## Rules For Making A Diagram

– **Length and Breadth:** An appropriate proportion should be maintained between length and breadth. Lutz has suggested that proportions of length and breadth should be 2:1, 1:4 or 4:1. If it is so, the diagram looks attractive. Care should be taken to ensure that the diagram does not look ugly.

– **Drawing:** Since impression is needed it should be drawn neatly and accurately with the help of drawing instruments. Each diagram should also be numbered for ready reference.

– **A proper Scale:** A proper Scale must be chosen for the diagram to look attractive and create a visual impact on the reader. It must suit the space available. Accuracy should not be sacrificed to attractiveness.

## Rules For Making A Diagram

– **Selection of appropriate diagram:** The most important point is the selection of proper diagram to present a set of figures. All types of diagrams are not suitable for all types of data. A wrong selection of the diagram will distort the true characteristics of the phenomenon to be presented and might lead to very wrong and misleading interpretations.

– **Right method:** C.W. Lowe writes," The important point, which must be borne in mind at all times, is that the pictorial presentation, chosen for any situation, must depict the true relationship and point out the proper conclusion. Use of an inappropriate chart may distort the facts and mislead the reader.

– **Index:** When many items are shown in a diagram, through different colors, dotting, crossings etc an index must be given for identifying and understanding the diagrams.

## Rules For Making A Diagram

- **Sources:** If the data presented have been acquired from some external source, the fact should be indicated at the bottom of the diagram.
- **Simplicity:** Diagram should be very simple. It must be so simple that even a lay man who does not have the knowledge of mathematical or statistical background, can understand the diagram. If the data is very more diagrams can be used to represent the data. Too much of information presented in a diagram will be confusing. Therefore, it is suggested to draw several simple diagrams which are more effective than a complex one.

## Limitations of Diagrammatic Presentation

- Diagrams do not present the small differences properly.
- These can easily be misused.
- Only artist can draw multi-dimensional diagrams.
- In statistical analysis, diagrams are of no use.
- Diagrams are just supplement to tabulation.
- Only a limited set of data can be presented in the form of diagram.
- Diagrammatic presentation of data is a more time consuming process.
- Diagrams present preliminary conclusions.
- Diagrammatic presentation of data shows only on estimate of the actual behavior of the variables.

## Bar Chart

- Bar chart is the simplest way to represent a data.
- In consists of rectangular bars of equal width.
- The space between the two consecutive bars must be the same.
- Bars can be marked both vertically and horizontally but normally we use vertical bars.
- The height of bar represents the frequency of the corresponding observation.
- Every bar graph has a:
  - title
  - scale
  - axis
  - labels on both axis
  - bars that show data

## How to construct Bar Graphs

- On a graph, draw two lines perpendicular to each other, intersecting at 0.
- The horizontal line is x-axis and vertical line is y-axis.
- Along the horizontal axis, choose the uniform width of bars and uniform gap between the bars and write the names of the data items whose values are to be marked.
- Along the vertical axis, choose a suitable scale in order to determine the heights of the bars for the given values. (Frequency is taken along y-axis).
- Calculate the heights of the bars according to the scale chosen and draw the bars.

## Types of Diagrams

- Line Diagrams :
  - In these diagrams only line is drawn to represent one variable. These lines may be vertical or horizontal. The lines are drawn such that their length is the proportion to value of the terms or items so that comparison may be done easily.
- Simple Bar Diagram :
  - Like line diagrams these figures are also used where only single dimension i.e. length can present the data. Procedure is almost the same, only one thickness of lines is measured. These can also be drawn either vertically or horizontally. Breadth of these lines or bars should be equal. Similarly distance between these bars should be equal. The breadth and distance between them should be taken according to space available on the paper.

## Types of Diagrams

- Multiple Bar Diagrams :
  - The diagram is used, when we have to make comparison between more than two variables. The number of variables may be 2, 3 or 4 or more. In case of 2 variables, pair of bars is drawn. Similarly, in case of 3 variables, we draw triple bars. The bars are drawn on the same proportionate basis as in case of simple bars. The same shade is given to the same item. Distance between pairs is kept constant.
- Sub-divided Bar Diagram :
  - The data which is presented by multiple bar diagram can be presented by this diagram. In this case we add different variables for a period and draw it on a single bar as shown in the following examples. The components must be kept in same order in each bar. This diagram is more efficient if number of components is less i.e. 3 to 5.

## Types of Diagrams

- Percentage Bar Diagram :
  - Like sub-divide bar diagram, in this case also data of one particular period or variable is put on single bar, but in terms of percentages. Components are kept in the same order in each bar for easy comparison.
- Duo-directional Bar Diagram :
  - In this case the diagram is on both the sides of base line i.e. to left and right or to above or below sides.
- Broken Bar Diagram :
  - This diagram is used when value of some variable is very high or low as compared to others. In this case the bars with bigger terms or items may be shown broken.

## Simple Bar Diagram

- It represents only one variable.
- For example sales, production, population figures etc. for various years may be shown by simple bar charts.
- Since these are of the same width and vary only in heights ( or lengths ), it becomes very easy for readers to study the relationship.
- Simple bar diagrams are very popular in practice.
- A bar chart can be either vertical or horizontal; vertical bars are more popular.

## Sub - divided Bar Diagram

- While constructing such a diagram, the various components in each bar should be kept in the same order.
- A common and helpful arrangement is that of presenting each bar in the order of magnitude with the largest component at the bottom and the smallest at the top.
- The components are shown with different shades or colors with a proper index.

## Multiple Bar Diagram

- This method can be used for data which is made up of two or more components.
- In this method the components are shown as separate adjoining bars.
- The height of each bar represents the actual value of the component.
- The components are shown by different shades or colors.
- Where changes in actual values of component figures only are required, multiple bar charts are used.

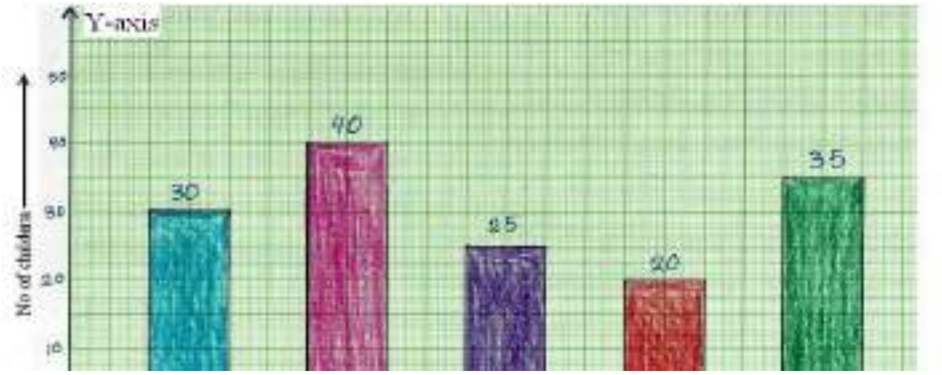
## Deviation Bar Diagram

- Deviation bars are used to represent net quantities - excess or deficit i.e. net profit, net loss, net exports or imports, swings in voting etc.
- Such bars have both positive and negative values.
- Positive values lie above the base line and negative values lie below it..

## Bar Graphs (example)

- The following data gives the information of the number of children involved in different activities.

| Activities | Dance | Music | Art | Cricket | Football |
|---|---|---|---|---|---|
| No. of Children's | 30 | 40 | 25 | 20 | 35 |



## Two – Dimensional  Diagram

- Rectangle
- Square
- Circle
- Pie

## Pie Chart/Graph

- It is a circular graph which is used to represent data. In this :
  - Various observations of the data are represented by the sectors of the circle.
  - The total angle formed at the centre is 360°.
  - The whole circle represents the sum of the values of all the components.
  - The angle at the centre corresponding to the particular observation component is given by

    $$\frac{\text{Value of the component}}{\text{Total value}} \times 360^{\circ}$$

  - If the values of observation/components are expressed in percentage, then the centre angle corresponding to particular observation/component is given by
- $$\frac{\text{Percentage value of component}}{100} \times 360^{\circ}$$

## How to construct Pie Chart/Graph

- Find the central angle for each component using the formula given on the previous slide.
- Draw a circle of any radius.
- Draw a horizontal radius.
- Starting with the horizontal radius, draw radii, making central angles corresponding to the values of respective components.
- Repeat the process for all the components of the given data.
- These radii divide the whole circle into various sectors.
- Now, shade the sectors with different colours to denote various components.
- Thus, we obtain the required pie chart.

## Pictographs

- The expression or illustration regarding the different information about any object or objects or activities through pictures or picture symbols is called pictorial representation or pictograph.
- Some basic ideas of pictorial representation or pictograph, often related types of symbols or pictures are used to represent a specific number of objects.
- For example, a symbol may represent 1 or 10 or 100 or 1000 or any other number of related objects.
- The symbol/picture used is very simple, clear and self explanatory.
- The quantity that each symbol represents is indicated clearly in the representation, i.e. the scale is mentioned clearly.

## Pictographs

- Examples of objects whose pictures or symbols are used are: different kinds of animals like birds, insects, men, women, boys, girls, fruits like mangoes, grapes, oranges, apples, trees, cars, scooters, bicycles, plants, etc.
- The symbol or pictures may be colored. So such colors should be used which are very common, like red, blue, green, yellow, etc.
- Only the pictures of common fruits are used such as banana, apples, guava, orange, grapes, mangoes, etc.

## How to make Pictographs

1. First of all, the related data is collected. For example, to form the pictograph showing the number of boys and girls in different classes and in the school as a whole, the number of boys and girls in different classes are counted.
   – As, I (35B, 20G), II (30B, 10G), III (10B, 30G), IV (25B, 35G), V (20B, 20G) [B for boys, G for girls]
2. Different symbols are selected for different types of data. As, for the symbol of a boy and a girl, we select a boy’s face and a face with long hair for a girl.
3. A framework containing columns and rows is made according to necessity.
4. The related symbols are made at proper places. The symbol must be clear, simple and recognizable. They should be self-explanatory.

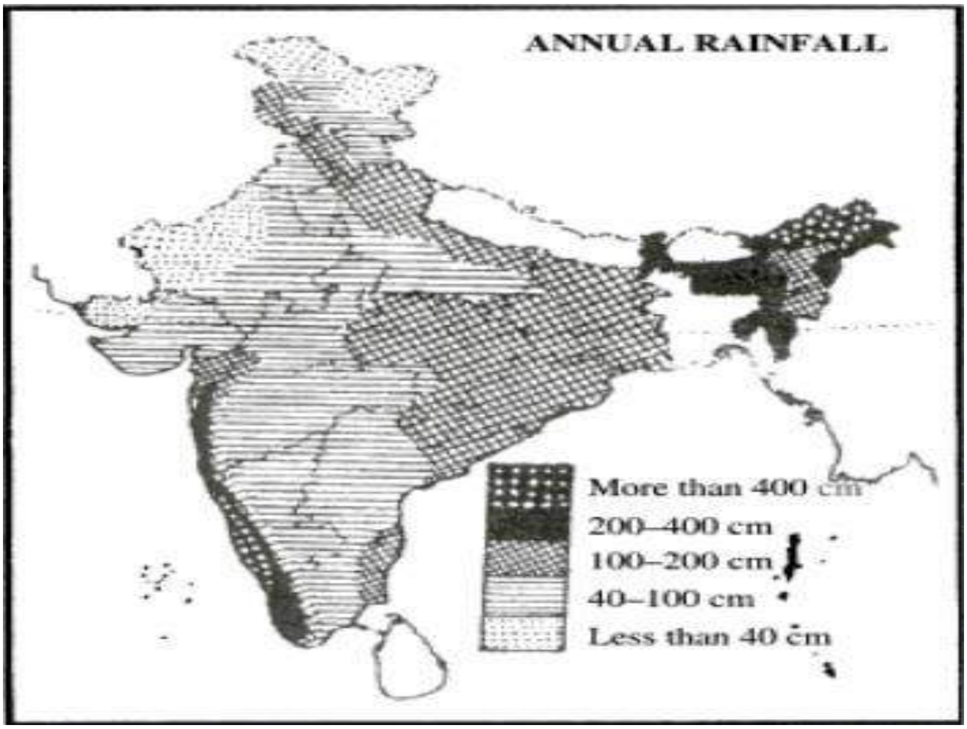
## How to make Pictographs

5. One symbol may represent many, i.e., 1, 5, 10, 20, etc., units. The quantity that each symbol represents is clearly indicated in the pictograph.
6. A pictograph has a title and is labeled.

- Every pictograph has a:
  - title
  - pictures or symbols
  - labels
  - key

## Cartogram

- When statistical data are presented in the form of maps, it is known as cartogram. Cartograms is also called statistical map.
- In cartogram the regional distribution of data is shown by the use of map. It is more informative and effective.
- Five properties of areas may be inferred from a traditional undistorted map:
  - Size , Shape , Distance, Direction, Contiguity
- These properties help us recognise which area is which.
- If size is distorted, then so must at least one other property
- Objective is to minimise distortions of other properties to permit areas to be recognised



## Graphic Presentation of Data Introduction

- A graph refers to the plotting of different values of the variables on a graph paper which gives the movement or a change in the variable over a period of time.
- Diagrams can present the data in an attractive style but still there is a method more reliable than this.
- Diagrams are often used for publicity purposes but are not of much use in statistical analysis.
- Hence graphic presentation is more effective and result oriented.
- According to A. L. Boddington, "The wandering of a line is more powerful in its effect on the mind than a tabulated statement; it shows what is happening and what is likely to take place, just as quickly as the eye is capable of working."

## Advantages Of Graphs

- The presentation of statistics in the form of graphs facilitates many processes in economics. The main uses of graphs are as under:
- Attractive and Effective presentation of Data:
  - The statistics can be presented in attractive and effective way by graphs.
  - A fact that an ordinary man can not understand easily, could understand in a better way by graphs.
  - Therefore, it is said that a picture is worth of a thousand words.
- Simple and Understandable Presentation of Data:
  - Graphs help to present complex data in a simple and understandable way.
  - Therefore, graphs help to remove the complex nature of statistics.

## Advantages Of Graphs

- Useful in Comparison:
  - Graphs also help to compare the statistics.
  - IF investment made in two different ventures is presented through graphs, then it becomes easy to understand the difference between the two.
- Useful for Interpretation:
  - Graphs also help to interpret the conclusion.
  - It saves time as well as labour.
- Remembrance for long period:
  - Graphs help to remember the facts for a long time and they cannot be forgotten.
- Helpful in Predictions:
  - Through graphs, tendencies that could occur in near future can be predicted in a better way.

## Limitations Of Graph

- Limited Application: Graphic representation is useful for a common man but for an expert, its utility is limited.
- Lack of Accuracy: Graphs do not measure the magnitude of the data. They only depict the fluctuations in them.
- Subjective: Graphs are subjective in character. Their interpretation varies from person to person.
- Misleading Conclusions: The person who has no knowledge can draw misleading conclusions from graphs.
- Simplicity: Graph should be as simple as possible
- Index: In order to show many items in a graph, index for identification should be given.

## How To Choose A Scale For A Graph

- The scale indicates the unit of a variable that a fixed length of axis would represent.
- Scale may be different for both the axes.
- It should be taken in such a way so as to accommodate whole of the data on a given graph paper in a lucid and attractive style.
- Sometimes data to be presented does not have low values but with large terms.
- We have to use the graph so as it may present the given data for comparison even.

## Types of Graphs

- There are two types of graphs.
  - Time series Graphs.
  - Frequency Distribution Graphs.
- Time series graphs may be of one variable, two variables or more variables graph.
- Frequency distribution graphs present
  - Histograms
  - Frequency Polygons
  - Frequency Curves and
  - Ogives / Cumulative Frequency Curves

## Line Graph

- The data which changes over a period of time can be displayed through a line graph.
- In line graph:
  - Points are plotted on the graph related to two variables
  - Points are joined by the line segments.

## How To Construct A Line Graph

- On a graph, draw two lines perpendicular to each other intersecting at O.
- The horizontal line is x-axis and vertical line is y-axis.
- Mark points at equal intervals along x-axis and write the names of the data items whose values are to be marked.
- Along the y-axis, choose an appropriate scale considering the given values.
- Now, make the points.
- Join each point with the successive point using a ruler. Thus, a line graph is obtained

## Graphs of Frequency Distributions

- The methods used to represent a grouped data are :-
  - Histogram
  - Frequency Polygon
  - Frequency Curve
  - Ogive or Cumulative Frequency Curve

## Histogram

- It is defined as a pictorial representation of a grouped frequency distribution by means of adjacent rectangles, whose areas are proportional to the frequencies.
- To construct a Histogram, the class intervals are plotted along the x-axis and corresponding frequencies are plotted along the y – axis.
- The rectangles are constructed such that the height of each rectangle is proportional to the frequency of the that class and width is equal to the length of the class.
- If all the classes have equal width, then all the rectangles stand on the equal width.

## Histogram

- In case of classes having unequal widths, rectangles too stand on unequal widths (bases).
- For open-classes, Histogram is constructed after making certain assumptions. As the rectangles are adjacent leaving no gaps, the class-intervals become of the inclusive type, adjustment is necessary for end points only.

## Construction Of Histogram When Class Intervals Have Unequal Widths

- Here a correction for unequal width class interval is to be done. The correction consists of finding a frequency density that is the frequency for the unequal width class divided by the width of that class, that is,
- Frequency density = $\dfrac{\text{frequency}}{\text{Width of the class-interval}}$
- The histogram constructed using these frequency densities for the unequal width class-intervals.

## Histogram (Remarks)

- Grouped (Not Continuous) Frequency Distribution :
  - It should be clearly understood that histogram can be drawn only if the frequency is continuous.
  - In case of grouped frequency distribution, if classes are not continuous, they should be made continuous by changing the class limits into class boundaries and then rectangles should be erected on the continuous classes so obtained.
- Mid –points are given :
  - Only mid-values of different classes are given.
  - Then the given distribution is converted into continuous classes by ascertaining the upper and lower limits of the carious classes under the assumption that the frequency is uniformly distributed throughout the class intervals.

## Histogram (Remarks)

- Discrete frequency distribution
  - Histograms, may sometimes also be used to represent discrete frequency distribution by regarding the given values of the variables as the mid-points of continuous classes and then proceeding as explained in two above.
- Open –end classes
  - Histograms can’t be constructed for frequency distribution with open- end classes unless we assume that the magnitude of the first open class is same as that of the succeeding (second) class and the magnitude of the last open class is same as that of the preceding (i.e. last but one) class.
- Histogram may be used for the graphic location of the value of mode.

## Difference Between Histogram And Bar Diagram

- A histogram is a 2D (area) diagram where both the width (base) and the length (height of the rectangle) are important whereas bar diagram is 1D diagram in which only length (height of the bar) matters while width is arbitrary.
- In histogram the bars (rectangles) are adjacent to each other where as in bar diagram proper spacing is given between different bars.
- In a histogram the class frequency is represented by the area of the rectangle while in bar diagram they are represented by the height of the corresponding bar.

## Frequency Polygon

- A curve is super imposed on a histogram by joining the mid-points of the top of the consecutive rectangles.
- Generally it is used when the class-intervals have a common width.
- The polygon is closed at the base by extending it on both its sides ( ends ) to the midpoints of two hypothetical classes, at the extremes of the distribution, with zero frequencies.
- The area of the polygon is same as the area of histogram.
- It is an improvement to histogram as it provides a continuous curve showing the gradient of rise and fall in data.

## Frequency Polygon

- On comparing the Histogram and a frequency polygon, you will notice that, in frequency polygons the points replace the bars ( rectangles ). Also, when several distributions are to be compared on the same graph paper, frequency polygons are better than Histograms

## Frequency Distribution (Curve)

- Frequency distribution curves are like frequency polygons.
- In frequency distribution, instead of using straight line segments, a smooth curve is used to connect the points.
- Its objective is to present graphically the area covered by histogram in a more symmetrically fashion.
- The frequency curve for the above data is shown as:

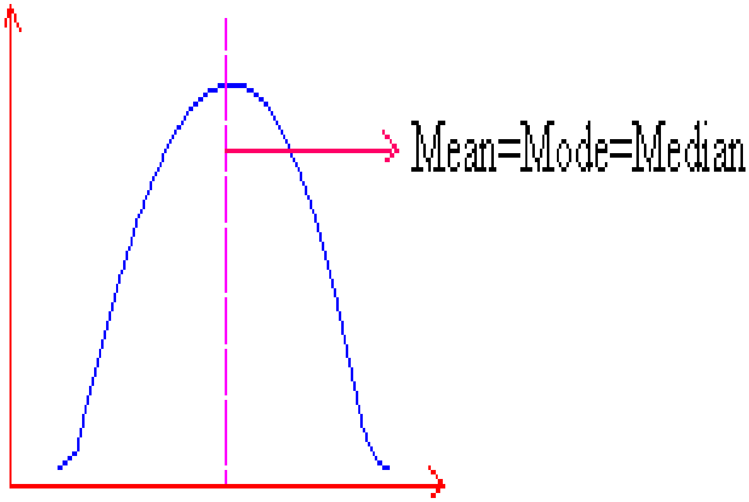
## Frequency Distribution (Curve)

- Shape of Distribution Curves:-
  - There are four types of distribution curves.
  - (i) Symmetrical or bell-shaped
  - (ii) Moderately symmetrical or skew
  - (ii) J-shaped and
  - (iv) U-shaped.

## Symmetrical Distribution

- As the name suggests, if the distribution is symmetrical, the curve rises gradually, reaches a maximum then falls equally gradually.
- The curve looks like a bell.
- This curve is symmetrical about the maximum frequency.
- For a perfectly symmetrical distribution, the mean, the mode and the median all coincide.
- This curve displays perfect symmetry, that is, its left half is the mirror image of the other right half.
- A bell-shaped or mound-shaped curve is also known as the normal curve, giving it special properties.
- This is an ideal situation and therefore, rarely found in practice.

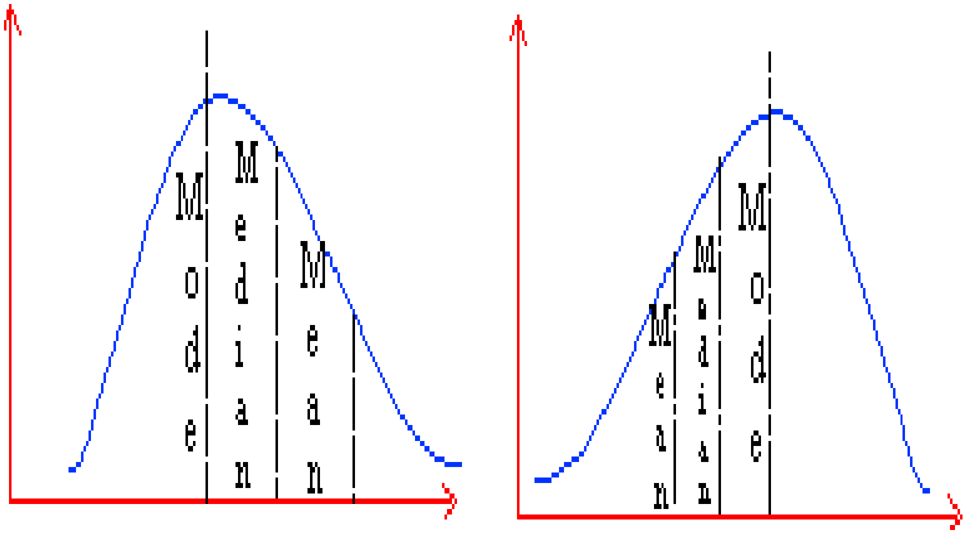
## Symmetrical Distribution



## Moderately Symmetrical Or Skew

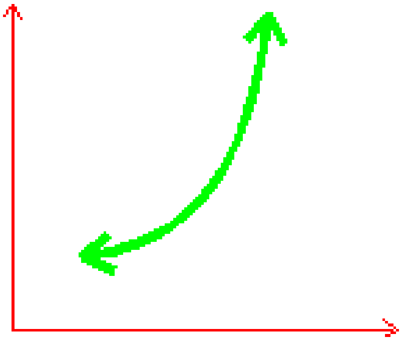
- The curve is symmetrical but only moderately.
- If it rises rapidly, reaches a maximum and then falls slowly, it is a called a positively skewed curve.
- If it rises slowly, reaches a maximum and then falls rapidly, it is called a "negatively skewed curve".
- For these curves; the mean, the median and the mode do not coincide.
- In much of the economic and social phenomenon, we come across such curves.

## Moderately symmetrical or Skew



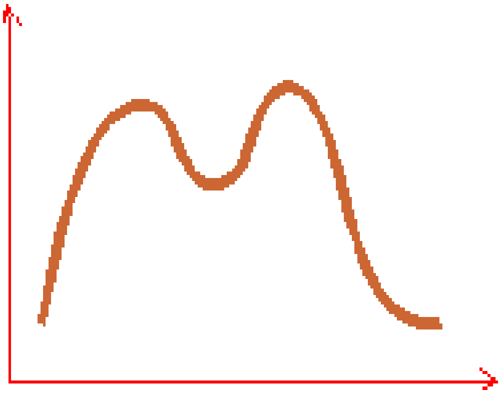
## J-Shaped Curve

- These curves are so called, as they have the shape of a 'J' or a' U'.
- The distributions are extremely asymmetrical.



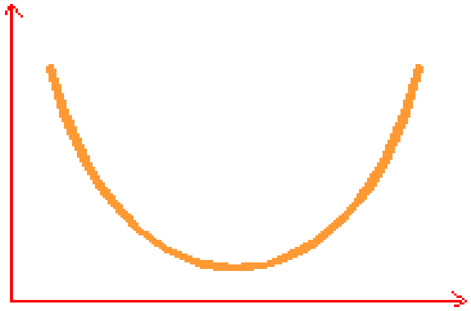
## Biomodal Curve

- Unlike a normal curve ( symmetric distribution), a bimodal curve has two peaks or humps



## U-Shaped Curve

- When a distribution is highly asymmetric we get this type of curve (U - type).
- U-Shaped curve exhibits the maximum frequencies at the ends of the range and the minimum towards the center.



Minal Shah

## Ogives or Cumulative Frequency Curves

- When frequencies are added, they are called cumulative frequencies.
- The curve obtained by plotting cumulating frequencies is called a cumulative frequency curve or an ogive ( pronounced ojive ).
- To construct an Ogive:-

1. Add up the progressive totals of frequencies, class by class, to get the cumulative frequencies.
2. Plot classes on the horizontal ( x-axis ) and cumulative frequencies on the vertical ( y-axis).
3. Join the points by a smooth curve. Note that Ogives start at (i) zero on the vertical axis, and (ii) outside class limit of the last class. In most of the cases it looks like 'S'. Note that cumulative frequencies are plotted against the 'limits' of the classes to which they refer.

Minal Shah

## Ogives or Cumulative Frequency Curves

- Less than Ogive:-
  - To plot a less than ogive, the data is arranged in ascending order of magnitude and the frequencies are cumulated starting from the top.
  - It starts from zero on the y-axis and the lower limit of the lowest class interval on the x-axis.
- Greater than Ogive:-
  - To plot this ogive, the data are arranged in the ascending order of magnitude and frequencies are cumulated from the bottom.
  - This curve ends at zero on the y-axis and the upper limit of the highest class interval on the x-axis.

## Uses of Ogives or Cumulative Frequency Curves

- Certain values like median, quartiles, deciles, percentile, deviation, coefficient of skewness etc. can be located using ogives.
- It can be used to find the percentage of items having values less than or greater than certain value.
- Ogives are helpful in the comparison of the two distributions.

Minal Shah

# Iterative Methods
## Computer Oriented Numerical and Statistical Methods

Minal Shah

## Outline

- Iterative methods :
  - Bisection
  - False-Position
  - Newton-Raphson

Minal Shah

## Types of Equations

- Linear Equations : The equation in which power of the unknown quantity is one is called linear equation. The equation in which the power of the unknown is two is called quadratic equation.
- Non-linear Equations : Most of the equation having more power of unknowns or involving sin, log function are non-linear equations. Ex. $x^2$ -3x =15 , x – cosx = 4

## Kinds of Equations

- They are classified on the basis of unknown quantity or power.
  - An equation which contains the first power only of an unknown quantity is called simple / linear equation [e.x. x – 2 = 5 here the power of x is 1]
  - If the power of the unknown quantity in an equation is 2 then it is called a quadratic equation [e.x. $x^2$ – 2x = 15 here the power of x is 2]
  - Some times two linear equations contains two unknown quantities. Also in order to find two unknown quantities we must have two linear equations. Such equations are called simultaneous equation [e.x 2x + y = 5; x + 3y = 8]

## Kinds of Equations

- Non-linear equations
  - Most of non-linear equations can be solved algebraically.
  - The solution obtained by algebraic manipulation is known as algebraic solution or an analytical solution
  - There are many non-linear equation that cannot be solved algebraically for example $2^x - x - 3 = 0$ which seems very simple but cannot be solved algebraically.
  - This solutions will be numerical not algebraic, and are called numerical solutions.
  - Types of non- linear equations are
    - Polynomial
    - Transcendental

## Polynomial Equations

- A polynomial has the general form

  $a_n x^n + a_{n-1} x^{n-1} + a_{n-2} x^{n-2} + \ldots + a_2 x^2 + a_1 x + a_0 = 0$ where $a_n \neq 0$
- It is $n^{th}$ degree polynomial in x and has n roots. These roots may be
  - Real and different
  - Real and repeated
  - complex

## Transcendental Equations

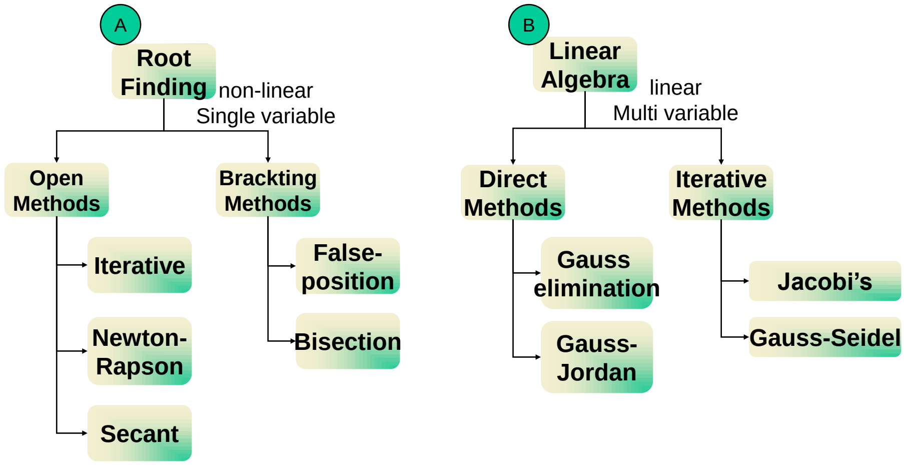
- A non-polynomial equation is called transcendental equations.
- Examples
  - $xe^x - x\sin x = 0$
  - $e^x\cos x - 3x = 0$
  - $2^x - x - 3 = 0$
- A transcendental equation may have finite/ infinite numbers of roots or may not have any roots at all.

## Convergence Notation

A sequence $x_1, x_2, \ldots, x_n, \ldots$ is said to **converge** to $x$ if to every $\varepsilon > 0$ there exists $N$ such that:

$$|x_n - x| < \varepsilon \quad \forall n > N$$

# Equation Solving



A

Root Finding

non-linear

Single variable

Open Methods

Iterative

Newton-Raphson

Secant

Brackting Methods

False-position

Bisection

B

Linear Algebra

linear

Multi variable

Direct Methods

Gauss elimination

Gauss-Jordan

Iterative Methods

Jacobi’s

Gauss-Seidel

# Iterative Method

- The Latin word ‘Iterate’ means to ‘repeat’
- It is also known as trial and error methods, are based on idea of successive approximations.
- They start with one or more initial approximations to the root and obtain a sequence of approximations by repeating a fixed sequence of steps till the solution with reasonably accuracy is obtained.
- Iterative method generally gives one root at a time.
- Iterative methods are very cumbersome and time-consuming for solving non-linear equations manually.

# Iterative Method

- However they are best suited for use on computers, due to following reasons:
  - Iterative methods can be concisely expressed as computational algorithms.
  - It is possible to formulate algorithms that can handle class of similar problems. For e.g. an algorithm can be developed to solve a polynomial equation of degree n
  - Round – off errors are negligible in iterative methods as compared to direct methods.

# Steps involved in an Iterative Method

- To develop an algorithm which is a step by step procedure for solving the problem
- An initial guess or initial estimates is made for the variable or variables of the solution.
  - The initial estimates should be reasonable.
  - Success in the solution depends on the choice of proper initial values for the variables.
- Using the algorithm developed, better and better estimates are obtained in the successive iterations.
- The iteration process is stopped when an acceptable solution is obtained, based on some reasonable criteria for stopping the iteration process.

## Bracketing / Interpolation Methods

- In bracketing methods, the method starts with an interval that contains the root and a procedure is used to obtain a smaller interval containing the root.
- Two estimates of the roots are made- one giving a positive value for the function f(x) and other a negative value for the function f(x). Since the value of f(x) would be zero at the root.
- It means the root is effectively bracketed between these two values..
- By proper choice, the gap between the two estimates of the roots is reduced further and further so as to arrive at very small gap between the two estimates successively.
- Examples of bracketing methods:
  - Bisection method
  - False position method

## OpenEnd / Extrapolation / Successive Approximation Methods

- In the open methods, the method starts with one or more initial guess points. In each iteration, a new guess of the root is obtained.
- Open methods are usually more efficient than bracketing methods.
- They may not converge to a root.
- Examples of open end methods:
  - Netwon –Raphson Method
  - Secant Method.

## System of Linear Equations

A set of *n* equations and *n* unknowns

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n = b_1$$

$$a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \ldots + a_{2n}x_n = b_2$$

$$\vdots \qquad\qquad \vdots$$

$$a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \ldots + a_{nn}x_n = b_n$$

## Bisection / Binary Chopping / Half- Interval/ Midpoint / Bolzano / Interval –Halving Method.

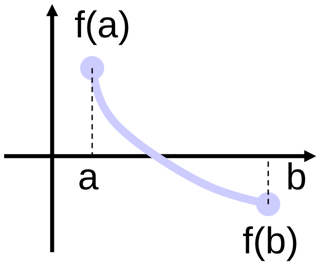
- This method is based on the theorem which states that if a function f(x) is continuous between a and b and f(a) and f(b) are of opposite signs, then there exists atleast one root between a and b
- Here f(a) is negative and f(b) is positive
- The root lies between a and b and its approximation is given by $x_0 = (a + b)/2$
- If $f(x_0) = 0$ then $x_0$ is the root.

## Intermediate Value Theorem

- Let f(x) be defined on the interval [a,b].

- Intermediate value theorem:

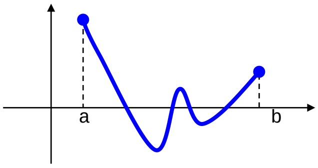
  if a function is continuous and f(a) and f(b) have different signs then the function has at least one zero in the interval [a,b].



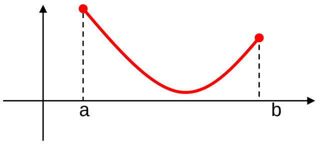
## Examples

- If f(a) and f(b) have the same sign, the function may have an even number of real zeros or no real zeros in the interval [a, b].

- Bisection method can not be used in these cases.



The function has four real zeros



The function has no real zeros

## Bisection Method

**Assumptions:**

Given an interval [a,b]

f(x) is continuous on [a,b]

f(a) and f(b) have opposite signs.

These assumptions ensure the existence of at least one zero in the interval [a,b] and the bisection method can be used to obtain a smaller interval that contains the zero.

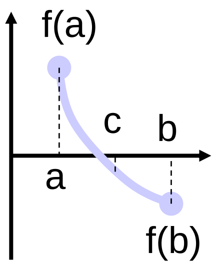
## Bisection Algorithm

**Assumptions:**

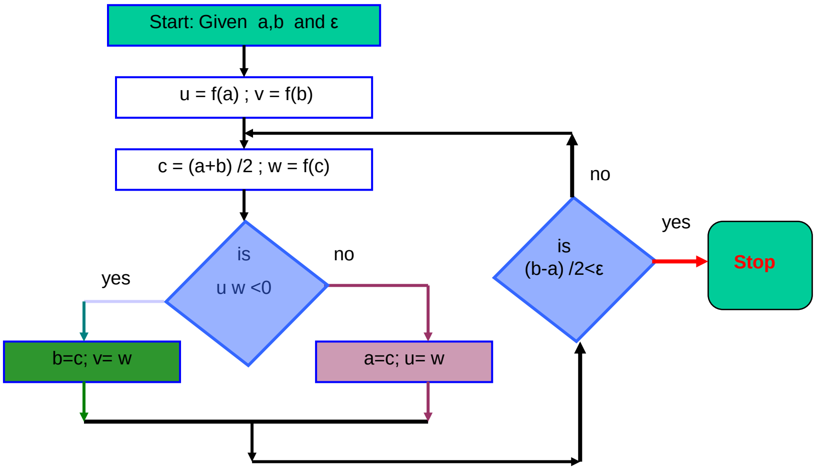
- f(x) is continuous on [a,b]
- f(a) f(b) < 0

**Algorithm:**

```
Loop
  1. Compute the mid point  $c=(a+b)/2$
  2. Evaluate $f(c)$
  3. If  $f(a) f(c) < 0$  then  new interval [a, c]
     If  $f(a) f(c) > 0$  then  new interval [c, b]
End loop
```



# Flow Chart of Bisection Method



# Stopping Criteria

Two common stopping criteria

1. Stop after a fixed number of iterations
2. Stop when the absolute error is less than a specified value

How are these criteria related?

$c_n$ : is the midpoint of the interval at the $n^{th}$ iteration

( $c_n$ is usually used as the estimate of the root).

r : is the zero of the function.

After $n$ iterations :

$$|error| = |r - c_n| \le E_a^n = \frac{b-a}{2^n} = \frac{\Delta x^0}{2^n}$$

## Bisection Method

**Advantages**

- Simple and easy to implement
- One function evaluation per iteration
- The size of the interval containing the zero is reduced by 50% after each iteration
- The number of iterations can be determined a priori
- No knowledge of the derivative is needed
- The function does not have to be differentiable

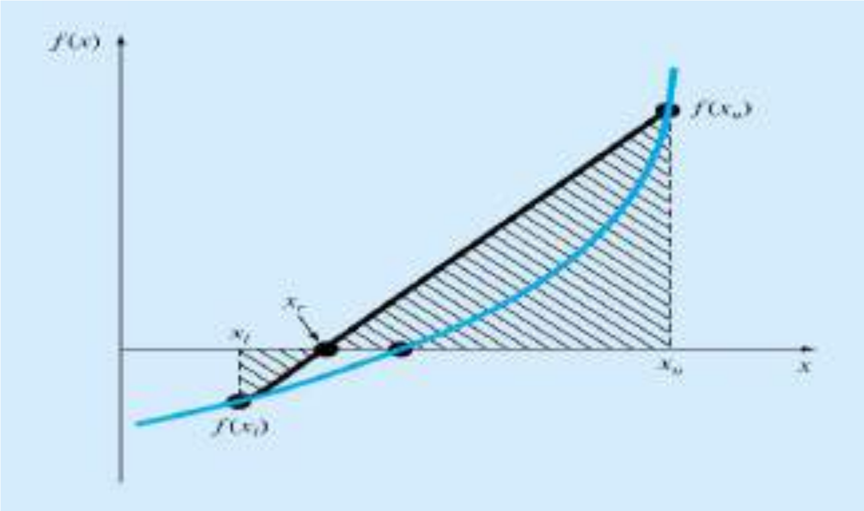
**Disadvantage**

- Slow to converge
- Good intermediate approximations may be discarded

## False- Position /Regula – Falsi / Linear Interpolation / Method Of Chords

- Here, we choose two points $x_n$ and $x_{n-1}$ such that $f(x_n)$ and $f(x_{n-1})$ are of opposite signs.
- Intermediate value property suggests that the graph of $y = f(x)$ crosses the $x$-axis between these two points and therefore, a root say lies between these two points.
- Thus, to find a real root of $f(x) = 0$ using Regula-Falsi method, we replace the part of the curve between the points $A[x_n, f(x_n)]$ and $B[x_{n-1}, f(x_{n-1})]$ by a chord in that interval and we take the point of intersection of this chord with the $x$-axis as a first approximation to the root.

## False- Position /Regula – Falsi / Linear Interpolation / Method Of Chords



## False- Position Method

- Start with two approximation to root say $x_1$ and $x_2$ for which f(x) has opposite sign.
- Compute $x_3$ as

$$x_3 = \frac{x_1 f(x_2) - x_2 f(x_1)}{f(x_2) - f(x_1)}$$

- There are 3 possibilities
  - If $f(x_3) = 0$ then we have a root as $x_3$
  - If $f(x_1)$ and $f(x_3)$ are of opposite sign, then the root lies in the interval $[x_1, x_3]$ . Thus $x_2$ is replaced by $x_3$ and the iterative procedure is repeated.
  - If $f(x_1)$ and $f(x_3)$ are of same sign, then the root lies in the interval $[x_3, x_2]$ . Thus $x_1$ is replaced by $x_3$ and the iterative procedure is repeated.
  - Terminate the process when the size of the search interval becomes less than prescribed tolerance.

## Bisection Method Vs False- Position Method

- Difference between bisection and regula falsi method

| Bisection | Regula -Falsi |
|---|---|
| If bisect the given interval $[x_1,x_2]$ on x -axis | It intersects the x- axis by a straight line joining the points $(x_1, f(x_1))$ and $(x_2, f(x_2))$ |
| It only depends on the sign change of f(x), but it does not depend on the values of f(x) | It only depends on both the sign change of f(x) and values of f(x) |
| Usually it obtains a repeated zero (i.e. root) accurately | It cannot obtain a repeated zero accurately |
| It detects both the zero and the jump discontinuity in a given interval | It detects mainly the zero but it may fail to detect jump discontinuity |

## Bisection Method Vs False- Position Method

- Similarity between bisection and regula falsi method

1. Activation occurs when there is a sign change in f(x) for the points $x_1$ and $x_2$.
2. They do not accurately find the number of repetitions of a root.
3. These methods are applicable only for obtaining real roots of a real function, not for complex roots.
4. When the curve y = f(x) approaches and touches the x- axis, these methods gives an indication of the approach but they fail to detect the root (i.e. the touching point)

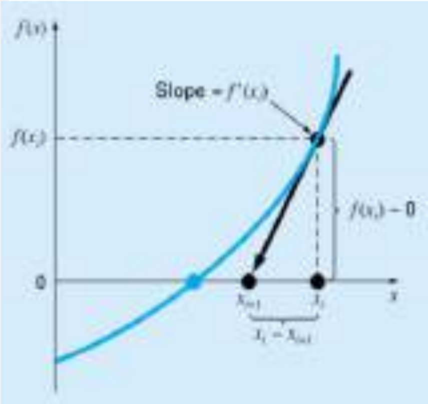
## Newton-Raphson Method/ Netwon’s Method of Tangents

- Most widely used formula for locating roots.
- It is one of the fastest iterative methods.
- Can be derived using **Taylor series** or the geometric interpretation of the slope in the figure

$$f'(x_i) = \frac{f(x_i)-0}{(x_i - x_{i+1})}$$

rearrange to obtain :

$$\boxed{x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}}$$



## Newton-Raphson Method

- Given an initial guess of the root ***x*₀**, Newton-Raphson method uses information about the function and its derivative at that point to find a better guess of the root.
- Assumptions:
  - ***f(x)*** is continuous and the first derivative is known
  - An initial guess ***x*₀** such that ***f’(x*₀*)≠0*** is given

## Newton-Raphson Method

- Start with the arbitrary point x0
- Determine $f(x_0)$, f ‘ $(x_0)$
- Determine

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

- Stop the iterative cycle when two successive values of Xi are nearly equal with a prescribed tolerance
  $|x_0 - x_1| < \varepsilon_1$ or $|f(x_1)| < \varepsilon_2$

## Advantage of Newton’s Method

- It as quadratic convergence. It converges fast at the cost of slightly increase labour in less number of iteration.
- Convergence is assured.

## Disadvantage of Newton’s Method

- For every iteration, $f(x^{(k)}) = f'(x^{(k)})$ have to be evaluated.
- If the initial guess of the root is far from the root the method may not converge.
- Since $f'(x^{(k)})$ occurs in the denominator of the expression for $x^{(k+1)}$ this poses a problem. If $f'(x^{(k)}) = 0$ ***or nearly zero***
- Newton’s method converges linearly near multiple zeros { $f(x) = f'(x) = 0$ }. In such a case, modified algorithms can be used to regain the quadratic convergence.

## Summary

| Method | Pros | Cons |
|---|---|---|
| Bisection | - Easy, Reliable, Convergent<br>- One function evaluation per iteration<br>- No knowledge of derivative is needed | - Slow<br>- Needs an interval [a,b] containing the root, i.e., f(a)f(b)<0 |
| Newton | - Fast (if near the root)<br>- Two function evaluations per iteration | - May diverge<br>- Needs derivative and an initial guess x0 such that f’(x0) is nonzero |

Minal Shah

**Mute ur call**

# Introduction to Statistics

Computer Oriented Numerical and Statistical Methods

# Outline

- Introduction
- Definition
- Functions of statistics
- Limitations

*“Statistical thinking will one day be as necessary for effective citizenship as the ability to read and write”* – H.G.Wells

# Introduction

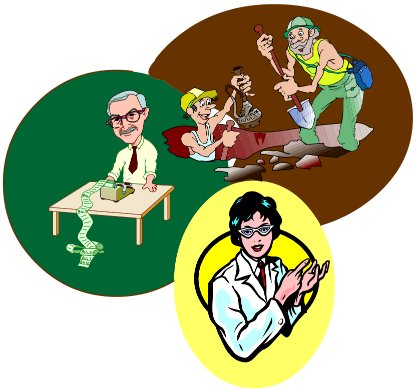
- The subject-matter of statistics has to do a great deal with the ‘science of statecraft’.
- The very word ‘statistics’ is said to have been derived from, say the Latin ‘*status*’, Italian ‘*statista*’, German ‘*statistik*’ or French ‘*statistique*’ all referring to the political state.
- In the past only population, its wealth or poverty was concern as statistics.
- Statistics in modern times is not a mere tool of state administration; it has become a fact of day-to-day life.

# Introduction

- 'Statistics' is being used both as singular noun and a plural noun
  - As plural noun, it stood for data.
  - While as a singular noun, it represented a method of study based on analysis and interpretation of facts.
- The word statistics may mean any one of the following:
  - Numerical statements of facts or simply data.
  - Scientific methods to help analysis and interpretation of data.
  - A measure based on sample observations.
- But, only the first two of these being more relevant to general purposes.

# Who Uses Statistics?

Statistical techniques are used extensively by marketing, accounting, quality control, consumers, professional sports people, hospital administrators, educators, politicians, physicians, and many others.



## Introduction

- Statistics is concerned with reduction of data or with obtaining correct facts from figures.
- Statistics deals with populations.
- Statistics deals with variation.
- Statistics deals with only numerically specified populations.
- A single **figure** is not called as statistics.
  e,g, Sales of computers in shop A is 300.
  - This is not a statistical statement.
- But sales of a computers in three shops A, B and C are 300, 400 and 200 respectively.
  - This statement will be called a **Statistics.**

## Examples

- Production statistics is compiled for judging the progress of business firm (here ‘statistics’ has been used for data)
- Statistics helps in simplification, analysis and presentation of data (here ‘statistics’ has been used to represent statistical methods)
- Statistics derived from a small representative group taken from the whole lot use for drawing inference about the characteristics of the whole (here ‘statistics’ represents measure based on sample observations)

## Definition

- It is based under two main heads :
  - Statistics as data
  - Statistics as method
- Statistics as data
  - Statistics are measurements, enumerations or estimation of natural or social phenomena, usually systematically arranged, analyzed and presented as to exhibit important inter-relationships among them. ------A.M. Tuttle
- Statistics as method
  - The science which deals with the collection, classification and tabulation of numerical facts as a basis for explanation, description and comparison amongst phenomena.

## Definition

- Statistics as method
  - Statistics is the science and art of handling aggregate facts observing, enumeration, recording, classifying and other wise systematically treating them. ------Harlow
  - Statistics may be defined as the collection, presentation , analysis and interpretation of numerical data. ------ Croxton and Cowden.

## Characteristics Of Statistical As Data

- They must relate to the aggregate of facts
- They are affected to a marked extent by a multiplicity of causes
- They are numerically expressed.
- They should be enumerated or estimated.
- They should be collected with reasonable standard of accuracy.
- They should be collected in a systematic manner.
- They must be relevant to the purpose.
- They should be placed in relation to each other.

## Distinction Between The Two

| Statistics as data | Statistics as method |
|---|---|
| It is quantitative | It is an operational technique |
| It is often in the raw state | It helps in processing the raw data |
| It is descriptive in nature | It is basically a tool of analysis |
| It provides material for processing. Unprocessed data does not help in decision - making | The processing is done by the scientific methods of analysis and interpretation. |
| As it is, it would not make much sense without application of the tools of analysis | Tools of analysis will be idle without the facts available for making use of such tools |

## Distinction Between The Two

| Statistics as data | Statistics as method |
|---|---|
| Surely the choice of tools will depends on the nature of data | The nature of the data to be collected will also depend on the tool sought to be used for processing |

## Functions Of Statistics

- To present facts in a proper form
- To simply wieldy and complex data to make them easily understandable.
- To help classification of data.
- To provide technique for making comparisons.
- To enlarge individual experience
- To formulate policies in different fields.
- To study relationships between different phenomena.
- To indicate trend behavior.
- To measure uncertainty.
- To test a hypothesis.
- To draw valid inferences.

## Limitations

- Statistics does not study individuals.
- Statistics does not study qualitative phenomena.
- Statistics results are true only on an average.
- Statistical laws are not exact.
- Statistic does not reveal the entire story.
- Statistical relations do not necessarily bring out the cause and effect relationship between phenomena.
- Statistics is collected with a given purpose and cannot be indiscriminately applied to any situations.
- Statistics is liable to be misused.

## Levels of Measurement

There are four levels of data

**Nominal**

**Ordinal**

**Interval**

**Ratio**

## Nominal data

### Nominal level

Data that is classified into categories and cannot be arranged in any particular order.

## Levels of Measurement

Nominal level variables must be:

### Mutually exclusive

An individual, object, or measurement is included in only one category.

### Exhaustive

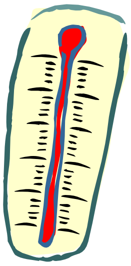
Each individual, object, or measurement must appear in one of the categories.

**Ordinal level**: involves data arranged in some order, but the differences between data values cannot be determined or are meaningless.

During a taste test of 4 soft drinks, Coca Cola was ranked number 1, Dr. Pepper number 2, Pepsi number 3, and Root Beer number 4.

## Interval level

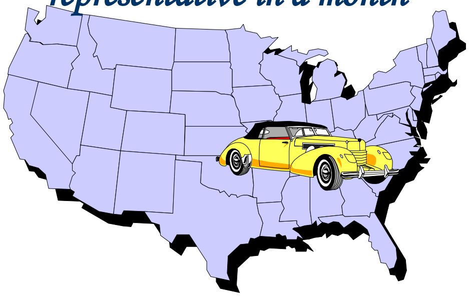
Similar to the ordinal level, with the additional property that meaningful amounts of differences between data values can be determined. There is no natural zero point.



*Temperature on the Fahrenheit scale.*

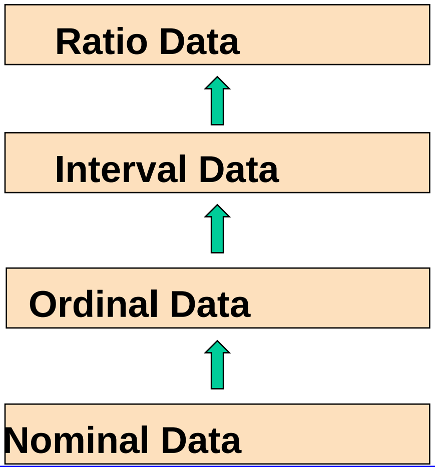
**Ratio level:** the interval level with an inherent zero starting point. Differences and ratios are meaningful for this level of measurement.

*Miles traveled by sales representative in a month*



*Monthly income of surgeons*

# Levels of Measurement and Measurement Scales



Differences between measurements, true zero exists — **Ratio Data** — **Highest Level** (Strongest forms of measurement)

Differences between measurements but no true zero — **Interval Data**

**Higher Levels**

Ordered Categories (rankings, order, or scaling) — **Ordinal Data**

Categories (no ordering or direction) — **Nominal Data** — **Lowest Level** (Weakest form of measurement)

# Levels of Measurement and Measurement Scales

| | | EXAMPLES: |
|---|---|---|
| **Ratio Data** | Differences between measurements, true zero exists | Height, Age, Weekly Food Spending |
| ↑ | | |
| **Interval Data** | Differences between measurements but no true zero | Temperature in Fahrenheit, Standardized exam score |
| ↑ | | |
| **Ordinal Data** | Ordered Categories (rankings, order, or scaling) | Service quality rating, Standard & Poor's bond rating, Student letter grades |
| ↑ | | |
| **Nominal Data** | Categories (no ordering or direction) | Marital status, Type of car owned |