Chapter 1 : Statistical Description of Data

History of Statistics :

* **Latin word** - Status, **Italian word** - Statista, **German word** - Statistik, **French word** - Statistique.
* Qualitative characterstic is known as **attribute**.

Two types of data :

(i) Primary Data (ii) Secondary Data

Collection of Primary Data :

(i) Interview Method :

(a) Personal interview : Ex. Natural calamity (b) Indirect Interview : Ex, Rail Accident.

(c) Telephonic interview : less consistent and has a wide coverage. Amount of non-reponse is maximum.

(ii) Mailed Questionnaire Method : Well drafted and covering all aspect of the problem under consideration. Amount of non-response is maximum in this method.

(iii) Observation Method : Best method for data collection but time consuming, laborious and covers only a small area. **(iv) Questionnaires filled and sent by enumerators.**

Collection of Secondary Data :

(a) International Sources (b) Goverment Sources (c) Private and Quasi- government sources

(d) Unpublished sources of various research institutes, researchers etc.

Presentation of Data :

(i) Chrononlogical or Temporal or Time series Data : Non frequency group.

(ii) Geographical or Spatial Series Data : Non frequency group.

(iii) Qualitative or Ordinal Data : Frequency group. (iv) Qunatitaive or Cardinal Data : Frequency group.

Mode of Presentation of Data :

(a) Textual Presentation : In the form of paragraph.

(b) Tabular Presentation :

(i) Caption : Upper part of the table, describes columns and subcolumns.

(ii) Box - head : Entire upper part of the table which includes columns and subcolumns numbers, unit(s) of measurement along with caption.

(iii) Stub : Left part of the tables provinding the description of rows.

(iv) Body : Main part of the table that contains the numerical figures. (v) Foot note : Source of table.

(c) Digramaatic repesentation :

(i) Line diagram or historiagram : Time series exhibit wide range of fluctuations : Logarthimic or ratio chart. Multiple line chart and Multiple axis chart : Representing two or more related time series data.

(ii) Bar Diagram : Horizontal bar diagram for qualitative data, Vertical Bar diagram for quantiatative data or time series data.Multiple or Grouped bar diagram to compare related series. (iii) Pie chart.

Range : Range = Largest observation – Smallest observation. No. of class interval x Class length = Range.

Class limit : Minimum value and maximum value the class interval may contain.The minimum value is known as lower class limit (LCL) and the maximum value is known as upper class limit (UCL).

Class Boundary : Class boundaries may be defined as the actual class limit of a class interval.For exclusive series class limits coincides with class limits but for inclusive series like 10-19, 20-29, 30-39, ...

$LCB = LCL - \frac{D}{2}$, where $UCB = UCL + \frac{D}{2}$, here D = 1So, LCB for first interval = 9.5, UCB = 19.5

Width or size of class interval : Difference between UCB and LCB of class interval.

Frequency Density : $\frac{\text{Class Frequency}}{\text{Class Length}}$ **Relative Frequency :** $\frac{\text{Class Frequency}}{\text{Total Frequency}}$

Graphical Representation of a Frequency Distribution : (i) Histogram or Area Diagram : Mode
(ii) Frequency Polygon (iii) Ogives (less than and More than) or Cumulative frequency graphs : Median and Quartiles.

Frequency Curve : (a) Bell shaped (b) U - shaped (c) J-shaped (d) Mixed curve

Chapter 2 : Measure of Central Tendancy and Dispersion

Central Tendancy : (i) Arithmetic Mean (ii) Median (iii) Mode. Also besides these mainly three, Geometric Mean and Harmonic mean are other measure of central tendancy.

Arithmetic Mean : $\bar{x} = \frac{\sum x_i}{n}$ or $\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$ [discrete or ungrouped series]

$\bar{x} = A + \frac{\sum f_i d_i}{N} \times h$ where $d_i = \frac{x_i - A}{h}$, A = assumed mean, h = class length [Continuous series]

Properties of AM : (i) Algebraic sum of deviations of a set of observations from their AM is zero i.e $\sum (x_i - \bar{x}) = 0$.

(ii) AM is affected due to change of origin and scale. $\bar{y} = a + b\bar{x}$ (iii) Combined AM = $\frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$

(iv) Very much affected by sampling fluctuations (v) Cannot be used for open end classification.

Median : It is a positional average. Median = $\left(\frac{N+1}{2}\right)^{th}$ Observation [discrete and ungrouped series]

Median = $\ell + \frac{\frac{N}{2} - C}{f} \times h$ [Continuous series] where ℓ = Lower limit of median class, N = Total frequency,

h = length of median class, C = cumulative freq. of the class preceding the median class, f = freq. of median class.

Properites of Median : (i) Median is affected due to change of origin and scale. $y_{Me} = a + bx_{Me}$

(ii) Sum of absolute deviations is minimum when the deviations are taken from Median.

(iii) Not much affected by sampling fluctuations (iv) Most appropiate measure for an open end classification.

Quartiles : First quartile $Q_1 = \left(\frac{N+1}{4}\right)^{th}$ Obs., Third Quartile $Q_3 = 3\left(\frac{N+1}{4}\right)^{th}$ Obs. [discrete and ungrouped series]

First quartile $Q_1 = \ell + \frac{\frac{N}{4} - C}{f} \times h$, Third Quartile $Q_3 = \ell + \frac{\frac{3N}{4} - C}{f} \times h$ [Continuous series]

Mode : Mode = $\ell + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$ where ℓ = Lower limit of modal class, f_1 = Frequency of modal class,

f_0 = Frequency of class preceding the modal class, f_2 = Frequency of class suceeding the modal class, h = length of modal class.

For Modereately skewed distribution : Mean – Mode = 3 (Mean – Median) or Mode = 3 Median – 2 mean.

Median is affected due to change of origin and scale. $y_{Mo} = a + bx_{Mo}$ and is also affected by sampling fluctuations.

Geometric Mean : G = $(x_1 \times x_2 \times x_3 ... x_n)^{\frac{1}{n}}$ or G = $\left(x_1^{f_1} \times x_2^{f_2} \times x_3^{f_3} ... x_n^{f_n}\right)^{\frac{1}{N}}$

Properties :

(i) $\log G = \frac{1}{r}\sum \log x$ (ii) If z = xy, then GM of z = (GM of x) × (GM of y) (ii) If z = $\frac{x}{y}$, then GM of z = $\frac{\text{GM of x}}{\text{GM of y}}$

Harmonic Mean : H = $\frac{n}{\sum(1/x_i)}$ and for a grouped distribution H = $\frac{n}{\sum(f_i/x_i)}$ Combined HM = $\frac{n_1+n_2}{\frac{n_1}{H_1}+\frac{n_2}{H_2}}$

Relation between AM, GM and HM : For two numbers x and y, AM = $\frac{x+y}{2}$, GM = $\sqrt{xy}$ and HM = $\frac{2xy}{x+y}$.

So, G^2 = AH and $A \geq G \geq H$.

Dispersion :

(i) Absolute Measure of dispersion (having units) :

(a) Range: (i) Range = L – S

(ii) Corffiecient of Range = $\frac{L-S}{L+S} \times 100$

(iii) Range remains unaffected due to change of origin but affected due to change of scale.If x and y are related as y = a + bx then Range of y is given by $R_y = |b|\, R_x$.

(b) Mean Deviation : (i) $\frac{1}{n}\sum|x_i - A|$ (About Mean), $\frac{1}{n}\sum|x_i - M|$ (About Median), $\frac{1}{n}\sum|x_i - Z|$ (About Mode)

(ii) Coefficient of MD about mean = $\frac{\text{MD}}{\text{Media}}$ × 100 , Coefficient of MD about median = $\frac{\text{MD}}{\text{Median}}$ × 100

Coefficient of MD about mode = $\frac{\text{MD}}{\text{Mode}}$ × 100

(iii) Mean Deviation remains unaffected due to change of origin but affected due to change of scale.If x and y are related as y = a + bx then Mean Deviation of y is given by $MD_y = |b|\, MD_x$.

(iv) Mean deviation takes its minimum value when the absolute deviation are taken from median.

(c) Standard Deviation : SD = $\sqrt{\frac{\sum(x_i-\bar{x})^2}{n}}$ or $\sqrt{\frac{\sum x_i^2}{n} - \bar{x}^2}$

Coefficient of variation = $\frac{\text{SD}}{\text{Mean}} \times 100$ [To check consistency of data]

(i) Standard Deviation remains unaffected due to change of origin but affected due to change of scale.If x and y are related as y = a + bx then Standard Deviation of y is given by $SD_y = |b|\, SD_x$.

Var (y) = b^2 Var (x)

(ii) Combined SD = $\sqrt{\frac{n_1(s_1^2+d_1^2)+n_2(s_2^2+d_2^2)}{n_1+n_2}}$ with $d_1 = x_1 - \bar{x}$ and $d_2 = x_2 - \bar{x}$

If $\bar{x}_1 = \bar{x}_2$, then Combined SD = $\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1+n_2}}$

(iii) SD of two numbers = $\frac{1}{2}$ (Range)

(iv) SD of first n natural numbers = $\sqrt{\frac{n^2-1}{12}}$

(d) Quartile Deviation : Quartile deviation of semi-interquartile range = $\frac{Q_3-Q_1}{2}$

Coefficient of quartile deviation = $\frac{Q_3-Q_1}{Q_3+Q_1}\times 100$ = $\frac{QD}{\text{Median}}\times 100$

Quartile deviation is the best measure of dispersion for the open end classifications.

PROBABILITY

Definition of Probability : Bernoullli and Laplace. It is also termed ar Priori definition.

Probability = $\frac{\text{Total no. of favourable outcomes}}{\text{Total no. of possible outcomes}}$, $0 \le P(A) \le 1$

If P(A) = 0, Impossible event, If P(A) = 0, Certain or sure event.

Odds in favour of A = $\frac{P(A)}{P(\bar{A})}$, Odds against A = $\frac{P(\bar{A})}{P(A)}$

Mutually Exclusive Events : $P(A \cap B) = 0$, Exhaustive Events : $P(A \cup B) = 1$

If A, B and C are mutually exclusive, exhaustive and Equally likely than P(A) = P(B) =P(C) and P(A) + P(B) + P(C) = 1.

Statistical definition of Probability : $P(A) = \lim_{n\to\infty}\frac{F_A}{n}$ where event A occurs F_A times.

Information of Playing cards :

	Playing Cards	Red = 26		Black = 26	
		Heart	Diamond	Spade	Club
	Ace = 4	1	1	1	1
Face Cards = 12	King = 4	1	1	1	1
	Queen = 4	1	1	1	1
	Jack = 4	1	1	1	1
	No. 2 to No. 10	9	9	9	9
	Total	13	13	13	13
Total Cards = 52					

1. **Additiion Theorem of Probability :** If **A** and **B** are two events associated with a random experiment, then **$P(A \cup B) = P(A) + P(B) - P(A \cap B)$**

   If A and B are mutually exclusive events then **$P(A \cap B) = 0$** y **$P(A \cup B) = P(A) + P(B)$**

Verbal description of the event	Equivalent set theoretic notation
Not A	$\bar{A}$
A or B (at least one of A or B)	$A \cup B$
A and B	$A \cap B$
A but not B	$A \cap \bar{B}$
Neither A nor B	$\bar{A} \cap \bar{B}$
At least one of A, B or C	$A \cup B \cup C$

2. **ConditionalProbability :**

P (A / B) = Probability of occurrence of **A** given that **B** has already occurred and **P(B)** $\neq$ **0**.

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

3. **Compound Theorem of Probability or Multiplication Theorem of Probability :** If **A** and **B** be two events associated with a random experiment, then

P (A $\cap$ B) = P (A) P (B / A), if **P(A)** $\neq$ **0**. or, **P (A $\cap$ B)** = **P (B) P (A / B)**, if **P(B)** $\neq$ **0**.

4. **Independent Event : P (A $\cap$ B) = P (A) P (B)**.

5. **Probability Distribution : Expectation or Mean =** $\sum P_i x_i$, Variance = $\sum P_i x_i^2 - (\sum P_i x_i)^2$, SD = $\sqrt{\text{Variance}}$

Chapter 5 : Theoretical Distribution

1. Discrete Probability Distribution (probability mass function) : (a) Binomial Distribution (ii) Poisson Distribution

2. Continuous Probability Distribution (probability density function) :

(a) Normal Distribution (b) Chi-square Distribution (c) Student -t distribution (d) F - distribution.

(A) Binomial Distribution (Method of Moments) : Trials - Independent, No. of trials - Finite.

$P(X = r) = {}^nC_r\, p^r\, q^{n-r}$ Where p =sucess and q = failure, p + q = 1.

Properties :

(i) Mean of binomial distribution = np, Variance of binomial distribution = npq and SD = $\sqrt{npq}$

(ii) Variance < Mean and Maximum variance = $\frac{n}{4}$ when p = q = $\frac{1}{2}$.

(iii) Binomial distibution is known as bi parametric distribution as it has two parameters n and p.

(iv) Binomial distribution may be unimodal or bi-modal. μ_0 = largest integer contained in (n + 1) p is (n + 1)p is a non - integer and (n + 1) p and (n + 1) p – 1 if (n + 1) p is an integer.

(B) Poisson Distribution : (i) n : No. of trials is indefinetly large as $n \to \infty$ (ii) p, the probaility of success for each trial is indefinetly small i.e. $p \to 0$ (iii) np = m is finite.

$P(X = r) = \frac{m^r e^{-m}}{r!}$ where r = 0, 1, 2, ... ∞ e = 2.71828, e^{-1} = 0.3678

M is known as parameter of poisson distribution and m > 0.

Properties :

(i) Uniparametric distribution m.   (ii) Mean = m, variance = m

(iii) Poisson distribution may be unimodal or bi-modal depending upon the calue of parameter m.

μ_0 = largest integer contained in m if m is a non - integer and m and m – 1 if m is an integer.

(C) Normal or Gaussian Distribution :

Probability density function $X \sim N(\mu, \sigma^2)$ is given by f (x) = $\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\bar{x}-\mu)^2}{2\sigma^2}}$ for $-\infty < x < \infty$

Properties :

(i) It is a biparametric distribution and is characterised by two parameters μ and σ^2.

(ii) For the normal distribution Mean = Median = Mode and 4 SD = 5 MD = 6 QD. So, SD > MD > QD.

(iii) $Q_1 = \mu - 0.675\,\sigma$ and $Q_3 = \mu + 0.675\,\sigma$. So, Q.D. = $0.675\,\sigma$

(iv) Inflexion Points : $\mu - \sigma$ and $\mu + \sigma$

CENTRAL MOMENTS OF A PROBABILITY DISTRIBUTION

DEFINITION :

Central Moments is the moment of a probability distribution of a random variable about the random variables mean.

First Central Moment E (x – «) : Always 0.

Second Central moment E (x – «)^2 : Variance.

Third Central Moment E (x – «)^3 : Skewness.

Fourth Central Moment E (x – «)^3 : Kurtosis.

	First Central Moment	Second Central Moment	Third Central Moment	Fourth Central Moment
Binomial Distribution	0	npq	npq (q – p)	npq [1 + (3n – 6) pq]
Poisson Distribution	0	m	m	$3m^2 + m$
Normal Distribution	0	$\pm^{2}$	0	$3\pm^{4}$

Correlation and Regression

Bivariate Data : Two types of univariate distribution can be obtained :

(a) Marginal Distribution : For a m × n classification of bivariate data, maximum number of marginal distribution is 2.

(b) Conditional Distribution : For a m × n classification of bivariate data, maximum number of conditional distribution is m + n.

Correlation Analysis :

(i) Positive correlation : Ex. Height and Weight, Profit and investment, Age of insured person and premium etc.

(ii) Negative correlation : Ex. Price and demand, profit of insurance company and number of claims etc.

Measure of correlation :

(a) Scatter Diagram

(b) Karl Perason's product moment correlation coefficient :

$$r = \frac{Cov(x,y)}{s_x s_y} \text{ where Cov (x, y) = } \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n} = \frac{\sum x_i y_i}{n} - \bar{x}\,\bar{y},\ s_x = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n}} = \sqrt{\frac{\sum x_i^2}{n} - \bar{x}^2}$$

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}},\ r = \frac{\sum(d\bar{x}.d\bar{y})}{n s_x s_y}, \text{ where } d\bar{x} = x - \bar{x},\ d\bar{y} = y - \bar{y}$$

Properties of Correlation Coefficient :

1. The coefficient of Correlation is a unit free measure.

2. The coefficient of correlation remain invariant under a change of origin and scale of the variables but depends on the sign of scale factors. $u = \frac{x-a}{b}$, $v = \frac{y-c}{d}$, then $r_{xy} = \frac{bd}{|b||d|} r_{uv}$.

3. The coefficient of correlation always lies between – 1 and + 1, including both the limiting values. $-1 \le r \le 1$.

(c) Spearman's Rank Correlation Coefficient : $r_R = 1 - \frac{6\sum d^2}{n(n^2-1)}$ where d = difference in the rank

Sum of difference of rank are always zero.

(d) Coefficient of Concurrent Deviations : $r_c = \pm\sqrt{\pm\frac{(2c-m)}{m}}$ where c = no. of concurrent deviations, m = n – 1.

Probable Error : P.E. = 0.6745 $\frac{1-r^2}{\sqrt{n}}$, Standard Error = $\frac{1-r^2}{\sqrt{n}}$

If r < P.E. there is no significant correlation between the population variable.

If $r \geq 6$ P.E. there is a significant correlation between the population variable.

Regression : Based on method of least squares.
Regression equations are of two types :

(i) Regression equation y on x : $y = a + b_{yx}x$, where $b_{yx} = r\frac{s_y}{s_x} = \frac{\text{cov}(x,y)}{s_x^2}$ and $a = \bar{y} - b_{yx}\bar{x}$ .Also, $y - \bar{y} = b_{yx}(x - \bar{x})$.

(ii) Regression equation x on y : $x = a + b_{xy}y$, where $b_{xy} = r\frac{s_x}{s_y} = \frac{\text{cov}(x,y)}{s_y^2}$ and $a = \bar{x} - b_{xy}\bar{y}$ .Also, $x - \bar{x} = b_{xy}(y - \bar{y})$.

Properties of Regression Lines :

(i) The regression coefficient remains unchanged due to shift of origin but change due to shift of scale.

$u = \frac{x-a}{p}$, $v = \frac{y-c}{q}$ So, $b_{yx} = \frac{q}{p}r_{vu}$ and $b_{xy} = \frac{p}{q}r_{uv}$

(ii) The two lines of regression intersect at the point ($\bar{x}$, $\bar{y}$) where $\bar{x}$ and $\bar{y}$ are the mean.

(iii) $r = \pm\sqrt{b_{yx} \times b_{xy}}$ The sign of the correlation coefficient is the common sign of two regression coefficients.

(iv) Two lines of regressions become identical when r = – 1 and r = 1.

(v) If regressions are perpendicular to each other than r = 0.

Coefficient of determination = $\frac{\text{Explained Variance}}{\text{Total Variance}} = r^2$.

Coefficient of non-determination = $\frac{\text{Unexplained Variance}}{\text{Total Variance}} = 1 - r^2$.