Statistics-1

* Basics :

→ 2 branches : (a) Descriptive statistics
(b) Inferencial statistics

→ Cross-sectional data : data collected at a point of time
Time-series data : data collected over a period of time

→ scales of measurement / data types :

categorical	Nominal scale	→	named categories
	ORdinal scale	→	ordered categories
Numerical	Interval scale	→	Numeric but no absolute zero
	Ratio scale	→	Numeric with all arithmetic

* Categorical Data :

→ Frequency distribution table / Relative frequency table

→ Charts : (a) Pie chart
(b) Bar chart
(c) Pareto chart (frequency sorted bar chart)

→ mode : Variable/ Category with max. frequency
median : Variable/ category of middle observation

→ continuous data category :
20 - 30 : 20 is included but 30 does not.
Upper limit is 30 and lower limit is 20.
10 is class width. (i.e. $30-20=10$)
25 is class mark. (i.e. $\frac{20+30}{2}=25$)

* Numericale Data :

→ Sample size : n | Sample mean : $\bar{x}$ | Sample S.D. : S

Population size : N | Population mean : μ | Population S.D. : σ

→ Mean : $\dfrac{\sum f_i m_i}{\sum f_i}$ or $\dfrac{\sum x_i}{n}$

→ Adding constant : $y_i = x_i + c \Rightarrow \bar{y} = \bar{x} + c$

multiplying with constant : $y_i = c x_i \Rightarrow \bar{y} = c\bar{x}$

→ Median : middle value in ordered List

(a) n is odd : $\frac{n+1}{2}$ th observation.

(b) n is even : average of $\frac{n}{2}$ th and $\frac{n}{2}+1$ th observation.

→ Adding constant : New median = Old median + C

multiplying with constant : New median = C x old median

→ mode : most frequently occuring value

→ Adding constant : New mode = Old mode + c

multiplying with constant : New mode = c x old mode

→ Range : Diff. of largest & smallest value

→ Variance : Let $x_i - \bar{x}$ is deviation.

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}$$

$$\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \dots + (x_N - \mu)^2}{N}$$

→ Adding constant : New variance = old variance

multiplying with constant : New variance = c^2 x old variance

→ Standard Deviation : Positive square root of variance

→ Adding constant : New S.D. = Old S.D.

multiplying with constant : New S.D. = C × Old S.D.

→ measure of centrle tendency : Mode, Median & Mean

measure of dispersion/ spread : Variance & S.D.

→ Percentile : x percentile means x%. of data are $\leq$ it and ;

(100 − x) %. of data are $\geq$ it.

→ First / Lower Quartile (25 th perentile) (Q_1)

Second / Median Quartile (50 th percentile) (Q_2)

Third/ upper Quartile (75 th percentile) (Q_3)

* Five-Number Summary :

1. minimum
2. Q1
3. Q2
4. Q3
5. maximum

$IQR = Q_3 - Q_1$

* Scatter Plot :

→ x-axis : Explanetory / Independent variable

y-axis : Responsive / Dependent variable

→ Gives insight about association

→ Covairance and corelation are measure of linear association.

* Covariance :

→ Population's covariance $= \dfrac{\sum_{i=1}^{N} (x_i - \mu_x)(y_i - \mu_y)}{N}$

Sample's covariance $= \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$

→ If covariance $> 0 \Rightarrow$ Positive association

covariance $< 0 \Rightarrow$ Negative association

→ Covariance has unit i.e. unit of x_i × unit of y_i

* Correlation :

→ Pearson correlation $r = \dfrac{\text{covariance}}{\text{S.D. of } x_i \times \text{S.D. of } y_i}$

$\therefore r = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}}$ Range (r) is $[-1, 1]$

-1 ——— 0 ——— $+1$

strong ← weak → strong

→ Correlation coefficient does not have any units.

→ r^2 or R^2 gives goodness of straight line / curve fitting. $R^2 \in [0, 1]$

* Association b/w numeric and categorical variables :

→ Point-Bi-Serial correlation coefficient r_{PS}

$$r_{PS} = \left(\frac{\overline{Y}_0 - \overline{Y}_1}{S_x}\right)\sqrt{P_0 P_1}$$

where; P_0 = Proportion of categorical variable coded with 0
P_1 = Proportion of categorical variable coded with 1
$\overline{Y}_0$ = mean of categorical variable coded with 0
$\overline{Y}_1$ = mean of categorical variable coded with 1
S_x = S.D. of numerical variable

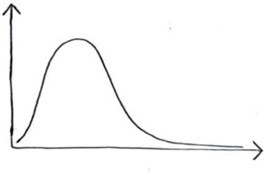
* Infinite series :

→ $$\sum_{n=0}^{\infty} ar^n = \frac{a}{1-r} \quad \text{where } |r| < 1$$

→ $$\sum_{n=0}^{\infty} \left(\frac{x^n}{n!}\right) = e^x$$

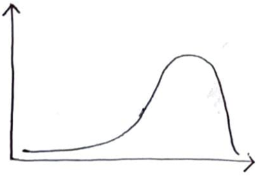
* Distributions :

(a) Right skewed
(Positively skewed)



(b) Symmetric

(c) Left skewed
(Negatively skewed)



* Permutation :

→ Definition : Ordered arrangement of all or r objects from n.

$${}^{n}P_{r} = \frac{n!}{(n-r)!} \quad \text{(without repetation)}$$

$$= n^{r} \quad \text{(with repetation)}$$

→ Non-distinct permutation :

$${}^{n}P_{n} = \frac{n!}{P_1! P_2! \ldots P_K!}$$ where P_i = group of same objects

→ Circular permutation :

cw & ccw are considered different = $(n-1)!$

cw & ccw are considered same = $\frac{(n-1)!}{2}$

* Combination :

→ Definition : Selecting / choosing r objects from n.

$${}^{n}C_{r} = \frac{{}^{n}P_{r}}{r!} \quad \text{or} \quad \frac{n!}{r!\,(n-r)!}$$

→ Identity : ${}^{n}C_{r} = {}^{n}C_{n-r}$

$${}^{n}C_{r} = {}^{n-1}C_{r} + {}^{n-1}C_{r-1}$$

$$\sum_{r=0}^{n} {}^{n}C_{r} = 2^{n}$$

* Playing Cards :

♣ Clubs (13)	♥ Heart (13)
♠ Spade (13)	♦ Diamond (13)
26 black	26 red

Total 52

* Axiometic Probability :

→ $P(E)$ is in accord with ; $(E_1 \cap E_2 = \phi)$

(i) $0 \le P(E) \le 1$

(ii) $P(S) = 1$

(iii) For mutually exclusive / disjoint events $E_1, E_2, \dots E_n$

$$P(\cup E_i) = \Sigma P(E_i)$$

→ General properties :

(a) $P(E^c) = 1 - P(E)$

(b) $P(\phi) = 0$

(c) $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$

→ other approches of probability are ;

- Classical / Apriori / Theoretical approch
- Relative frequency / Aposteriori / Impirical approch
- Subjective approch

* Independent Event :

→ Two event E & F are independent if and only if

(i) $P(E \cap F) = P(E) \times P(F)$

→ Three event E, F, & G are independent if and only if

(i) $P(E \cap F \cap G) = P(E) \times P(F) \times P(G)$

(ii) $P(E \cap F) = P(E) \times P(F)$

(iii) $P(F \cap G) = P(F) \times P(G)$

(iv) $P(G \cap E) = P(G) \times P(E)$

→ If E & F are independent, then following are also independent.

(i) E and F^c

(ii) E^c and F

(iii) E^c and F^c

* Conditional Probability :

→ Probability of E conditioned on F :

$$P(E/F) = \frac{P(E \cap F)}{P(F)} \quad ; \; P(F) > 0$$

→ If E and F are independent then $P(E/F) = P(E)$.

→ multiplication rule :

- $P(E \cap F) = P(E) \cdot P(F|E)$

- $P(E_1 \cap E_2 \cap E_3 \cap \ldots \cap E_n) = P(E_1) \cdot P(E_2|E_1) \cdot P(E_3|E_1 \cap E_2) \ldots P(E_n | E_1 \cap E_2 \ldots \cap E_{n-1})$

* Total Probability :

→ Let's $F_1, F_2, \ldots, F_K$ are mutually exclusive and exhaustive then for any event E;

$$P(E) = \sum_{i=1}^{K} P(E/F_i) \cdot P(F_i)$$

* Bay's Rule :

→ Let's $F_1, F_2, \ldots, F_K$ are mutually exclusive and exhaustive then for any event E;

$$P(F_i | E) = \frac{P(E|F_i) \cdot P(F_i)}{\sum_{i=1}^{K} P(E|F_i) \cdot P(F_i)}$$

* Random Variable :

→ Definition : The quontities of interest or real valued functions defined on the sample space are known as random variable.

→ Discrete random variable
Continuous random variable

* Probability Mass Function (PMF) :

→ PMF for discrete random variable defined as;

$P(x_i) = P(X = x_i)$ (i.e. probability of occuring x_i)

→ General properties :

$P(x_i) \geq 0$

$\sum_{i=1}^{\infty} P(x_i) = 1$

* Cumulative Distribution Function (CDF) :

→ CDF for discrete random variable defined as;

$F(a) = P(X < a)$ (where $x_1 < x_2 < \ldots < x_n$)

→ For discrete random variable CDF is step function.

$F(X_1)$	0
$F(X_2)$	$0 + P(X_1)$
$F(X_3)$	$0 + P(X_1) + P(X_2)$
$\vdots$	
$F(X_n)$	$P(X_1) + P(X_2) + \ldots + P(X_{n-1})$
$F(X_{n+1})$	1

* Expectation of Random Variable :

→ Expectation of Random variable x_i where $i = 1$ to n is;

$$E(X) = \sum_{i=1}^{\infty} x_i P(x_i)$$

→ It is 'Long run Average' value of random variable.

* Variance of Random Variable :

→ Variance of Random variable x_i where $i=1$ to n is;

$$V(X) = E((X-\mu)^2) \quad \text{where } \mu = E(X)$$

$$\therefore V(X) = E(X^2) - E(X)^2$$

* Standard Deviation of Random variable :

→ It is positive square root of variance of random variable.

* Proposition Rules :

→ $E(aX+b) = a\,E(X) + b$

$V(aX+b) = a^2\,V(X)$

$SD(aX+b) = a\,SD(X)$

→ $E(X) + E(Y) = E(X+Y)$ is always true

→ $V(X+Y) = V(X) + V(Y)$ is only true when X & Y are independent.

→ $SD(X+Y) = \sqrt{V(X)+V(Y)}$ is only true when X & Y are independent.

* Bernoulli Random Variable :

→ A random variable that takes either 0 or 1.

X	0	1
$P(X=x_i)$	$1-p$	p

$E(X) = p$

$V(X) = p(1-p)$

* Uniformly Distributed Random variable :

→ A random variable that takes values 1 to n.

X	1	2	......	n
$P(X=x_i)$	$1/n$	$1/n$	......	$1/n$

$$E(X) = \frac{n+1}{2}$$

$$V(X) = \frac{n^2-1}{12}$$

* Hypergeometric Random Variable :

→ Let's there is two category in population of size N. If size of category 1 is m, then size of another category is N−m. Choosing a sample of size n is having i from category 1 then;

$$P(X=i) = \frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}} \quad , \; i = 0,1,2,\dots n$$

Here, X is called hypergeometric random variable.

→ $$E(X) = \frac{nm}{N}$$

$$V(X) = \frac{nm}{N}\left[\frac{(n-1)(m-1)}{(N-1)} - \frac{nm}{N} + 1\right]$$

* Binomial Random Variable :

→ Independent and identically distributed (IID) Bernouli random variables is called Binomial Random variable because it's PMF is binomial.

→ Let n Bernouli trails are performed with probability of success in each trial p. Let X denotes the no. of successes in n trlals then;

$$P(X=i) = \binom{n}{i} p^{i} (1-p)^{n-i} \qquad (X \sim B(n,p))$$

* Graph of Binomial Distribution :

(i) $p < 0.5$ and n small.

(ii) $p > 0.5$ and n small.

(iii) $p = 0.5$

(iv) n large.

Approches symmetry.

* Expectation & Variance of Binomial Distribution :

→ $E(X) = np$

$V(X) = np(1-p)$

→ $p = 1 - \frac{V(X)}{E(X)}$ (only for Bernouli & Binomial Random variables.)

* Straight Line Fit by Min(SSE) :

→ Square Sum Error (SSE) $= \sum (y_i - (mx_i + c))^2$

→ $m = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$

$c = \bar{y} - m\bar{x}$