Statistics-1

Week-1

- Statistics
    - Descriptive Statistics
    - Inferential Statistics (Drawing conclusion from a sample - Probability)
- Population & Sample
- Unstructured data & Structured data
- Cases (Observations) => Rows
- Variables => Columns
- Types of Data
    - Categorical (Qualitative) data
    - Numerical (Quantitative) data
    - Cross-sectional & Time Series data
- Scales of measurement
    - Nominal Scale (Labels or names)
    - Ordinal Scale (Ranked or Ordered) — Categorical data
    - Interval Scale (Numerical values of fixed unit of difference) — Numerical data
    - Ratio Scale (True zero exists & Ratios possible)

Week-2

- Frequency distribution for Categorical
    - Relative frequency
    - Charts
        - Pie chart
        - Bar chart
        - Pareto chart
    - Area principle
    - Misleading graphs

        - Violating area principle
        - Truncated graphs (baseline is not zero)
        - Indicating y-axis break
        - Round off errors
    - Measures of Central tendency
        - Mode
            - Longest bar in bar chart
            - Widest slice in pie chart
            - First category in Pareto chart
            - Bimodal and multimodal data
        - Median (Ordinal data)
            - If no. of cases are odd, then median is $\left(\frac{n+1}{2}\right)$ value in the ordered list
            - If no. of cases are even, then median is the $\left(\frac{n}{2}\right)$ & $\left(\frac{n}{2}+1\right)$ values in the ordered list

Week-3

- Organizing Numerical data
    - Organizing Discrete data (count of something)
    - Organizing Continuous data (measurement of something)
        - No. of classes (5 to 20)
        - Lower class limit
        - Upper class limit
        - Class width (difference of two lower class limit)
        - Class mark (midpoint value of a class)
        - Class interval [a,b)
    - Histogram
    - Stem & Leaf diagram
    - Measures of Central tendency
        - Mean (average) => $\bar{x}$
            - Sample mean ($\bar{x}$) = $\frac{x_1+x_2+x_3+---x_n}{n}$

- Population mean (μ) = $\frac{x_1+x_2+x_3+---x_N}{N}$
- Mean for grouped data (discrete single value data)

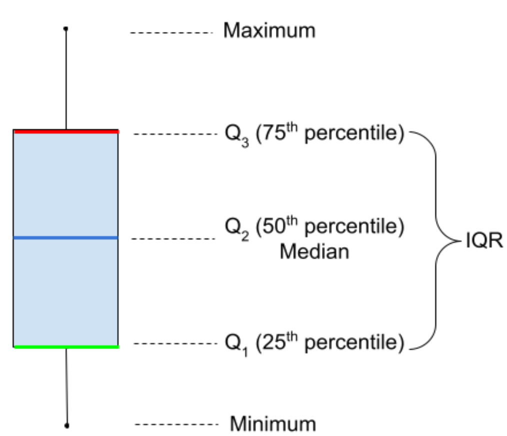
  $\bar{x} = \frac{f_1x_1+f_2x_2+f_3x_3+---f_nx_n}{n}$
- Mean for grouped data (continuous data)

  $\bar{x} = \frac{f_1m_1+f_2m_2+f_3m_3+---f_nm_n}{n}$, m_1, m_2, m_3 --- m_n are midpoints of the class
- Adding a constant ($\bar{y} = \bar{x}$+c)
- Multiplying a constant ($\bar{y} = \bar{x}$ c)
- Sample mean is sensitive to outliers

- Median (Ordered list)
  - If no. of cases are odd, then median is $\left(\frac{n+1}{2}\right)$ value in the ordered list
  - If no. of cases are even, then median is the average of $\left(\frac{n}{2}\right)$ & $\left(\frac{n}{2}+1\right)$ values in the ordered list
  - Adding a constant ($y_i = x_i$ +c)
  - Multiplying a constant ($y_i = x_i$ c)
  - Sample median is **not** sensitive to outliers
- Mode
  - Adding a constant ($y_i = x_i$ +c)
  - Multiplying a constant ($y_i = x_i$ c)

- Measures of dispersion or variance or spread
  - Range (max-min)
    - Range is sensitive to outliers
  - Variance
    - Sample variance (S^2) = $\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+ ..... + (x_n-\bar{x})^2}{n-1}$
    - Population variance (σ^2) = $\frac{(x_1-\mu)^2+(x_2-\mu)^2+ ..... + (x_n-\mu)^2}{N}$
    - Adding a constant (new variance = old variance)
    - Multiplying a constant (new variance = c^2 x old variance)

- Standard deviation
  - $S = \sqrt{Variance} = \sqrt{\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+ \ldots\ldots + (x_n-\bar{x})^2}{n-1}}$ for sample
  - $S = \sqrt{Variance} = \sqrt{\frac{(x_1-\mu)^2+(x_2-\mu)^2+ \ldots\ldots + (x_n-\mu)^2}{N}}$ for population
  - Adding a constant (new SD = old SD)
  - Multiplying a constant (new SD = c x old SD)
- Percentile (ordered data)
  - If 'np' not an integer, then percentile is the smallest integer greater than 'np' value in the ordered data
  - If 'np' is an integer, then percentile is average of 'np' & 'np+1' values in the ordered data
  - 50th percentile is the median
- Quartiles
  - Minimum
  - 25^{th} percentile is first quartile (Q_1)
  - 50^{th} percentile is second quartile (Q_2) or median
  - 75^{th} percentile is third quartile (Q_3)
  - Maximum
- Interquartile Range (IQR)
  - IQR = Q_3 - Q_1
- Outliers
  - Outliers < Q_1 - 1.5*IQR
  - Outliers > Q_3 + 1.5*IQR
- Boxplot



Week-4

- ➔ Association between categorical variables
  - ➔ Contingency table
    - ➔ Relative frequencies
      - ➔ Row relative frequencies
      - ➔ Column relative frequencies
    - ➔ If row/column relative frequencies are same for all rows/columns, then two variables are not associated with each other
    - ➔ If row/column relative frequencies are different for some rows/columns, then two variables are associated with each other
    - ➔ Stacked bar chart
      - ➔ 100% stacked bar chart
- ➔ Association between numerical variables
  - ➔ Scatter plot
  - ➔ Describing association
    - ➔ Direction (pattern trend up or down)
    - ➔ Curvature (pattern is linear or curve)
    - ➔ Variation (tightly clustered or variable)
    - ➔ Outliers
  - ➔ Measuring association
    - ➔ Covariance
      - ➔ Sample covariance, cov(x,y) = $\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}$
      - ➔ Population covariance, cov(x,y) = $\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{N}$
    - ➔ Correlation
      - ➔ Correlation Coefficient, ‘r’ is given by

$$r = \frac{cov(x,y)}{S_x S_y}$$

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

- ➔ 'r' is always lies in between -1 & +1

➔ Association between Categorical & Numerical Variables

- ➔ Point Bi-serial correlation coefficient (r_{pb}) is given by

$$r_{pb} = \left(\frac{\overline{Y_0}-\overline{Y_1}}{S_x} \right) \sqrt{P_0 P_1}$$

- ➔ $\overline{Y_0}$ = Mean of that particular '0' coded values
- ➔ $\overline{Y_1}$ = Mean of that particular '1' coded values
- ➔ S_x = Standard deviation of numerical variable
- ➔ P_0 = Probability of '0' coded among total categorical variables
- ➔ P_1 = Probability of '1' coded among total categorical variables
- ➔ $P_0 P_1 = \frac{n_0}{(n-1)} \frac{n_1}{n}$ for sample
- ➔ $P_0 P_1 = \frac{n_0}{N} \frac{n_1}{N}$ for population

Week-5

➔ Permutations & Combinations

- ➔ Addition rule of counting
  - ➔ If an action A occur in n_1 different ways, another action B occur in n_2 different ways, then total no. of occurrence of actions **A or B** is **n_1+n_2**
- ➔ Multiplication rule of counting

- If an action A occur in n_1 different ways, another action B occur in n_2 different ways, then total no. of occurrence of actions **A and B** is $\mathbf{n_1 \times n_2}$
- Permutations (ordered arrangement)
  - Permutations when objects are distinct
    - When repetition not allowed

$$^{n}P_{r} = \frac{n!}{(n-r)!}$$

      - $^{n}P_{0} = 1$
      - $^{n}P_{n} = n!$
    - When repetition is allowed

$$n^r = \underbrace{n \times n \times n \times \ldots\ldots \times n}_{\text{r times}}$$

  - Permutations when objects are not distinct
    - For 'n' objects, when 'p' of them are one kind

$$\frac{n!}{p!}$$

    - For 'n' objects, when 'p_1' is one kind, 'p_2' is second kind and so on

$$\frac{n!}{p_1!p_2!\ldots.p_k!}$$

- Circular permutations
  - When clockwise and anticlockwise are different

$$(n-1)!$$

  - When clockwise and anticlockwise are same

$$\frac{(n-1)!}{2}$$

- Combinations (no ordered arrangement)

$$^{n}C_{r} = \frac{n!}{(n-r)!r!}$$

$$n_{C_r} r! = n_{P_r}$$

$$n_{C_r} = (n-1)_{C_{r-1}} + (n-1)_{C_r}$$

- No. of ways of distributing ‘n’ identical things into ‘r’ different boxes ($x_1+x_2+x_3+....+x_r = n$)

$$(n+r-1)_{C_{r-1}}$$

- Drawing lines in a circle
  - If the line segment has no direction, then the lines can be drawn in a circle for ‘n’ points are

$$n_{C_2}$$

  - If the line segment has direction, then the lines can be drawn in a circle for ‘n’ points are

$$n_{P_2}$$

Week-6

- Random experiment
  - Any process that produces an outcome
- Sample space (Ω or S)
  - Collection of all possible outcomes
    - Mutually exclusive or Disjoint Events
      - If E∩F = Φ, then E and F are disjoint events
    - Exhaustive
- Events
  - Subset of the sample space
- Null event (Φ)
  - Event without any outcomes
- Properties of probability
  - Equally likely outcomes

- Three main interpretations of probability
  - Classical (Apriori or theoretical)
    - For ‘n’ equally likely outcomes of sample space ‘S’, and for ‘m’ outcomes of an event ‘E’, then P(E) = $\frac{m}{n}$
  - Relative frequency (Aposteriori or empirical)
    - If n(E) is the no. of times ‘E’ occurs in ‘n’ repetitions of experiment, then P(E) = $\lim_{n \to \infty} \frac{n(E)}{n}$
  - Subjective
    - Probability measures an individual’s degree of belief in the event (best guess)
- Probability Axioms
  - $0 \leq P(E) \leq 1$
  - P(S) = 1
  - For disjoint events, $P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$
- General properties of probability
  - $P(E^C) = 1 - P(E)$
    $P(E) + P(E^C) = 1$
    $P(E \cup E^C) = 1 = P(S)$
  - $P(\Phi) = 0$
- Addition rule of Probability
  - $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$ for E_1 & E_2 are not disjoint
  - $P(E_1 \cup E_2) = P(E_1) + P(E_2)$ for E_1 & E_2 are disjoint

Week-7

- Joint probabilities
  - Displayed in cells of contingency table
- Marginal probabilities
  - Displayed in margins of contingency table
- Conditional probabilities

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

- Multiplication rule

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

- Independent events
  - When P(E|F) = P(E), then E & F are said to be independent events

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

  - If E & F are independent, then E & F^C are also independent
  - If E & F are independent, then E^C & F are also independent
  - If E & F are independent, then E^C & F^C are also independent
  - If any 3 events are independent if and only if
    - $P(E \cap F \cap G) = P(E) \times P(F) \times P(G)$
    - $P(E \cap F) = P(E) \times P(F)$
    - $P(E \cap G) = P(E) \times P(G)$
    - $P(F \cap G) = P(F) \times P(G)$
- Law of total probability

$$P(E) = P(E \cap F) \cup P(E \cap F^c)$$

$$P(E) = P(F)P(E|F) + P(F^c)P(E|F^c)$$

- Baye's rule

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

$$P(E|F) = \frac{P(E) \cdot P(F|E)}{P(E) \cdot P(F|E) + P(E^c) \cdot P(F|E^c)}$$

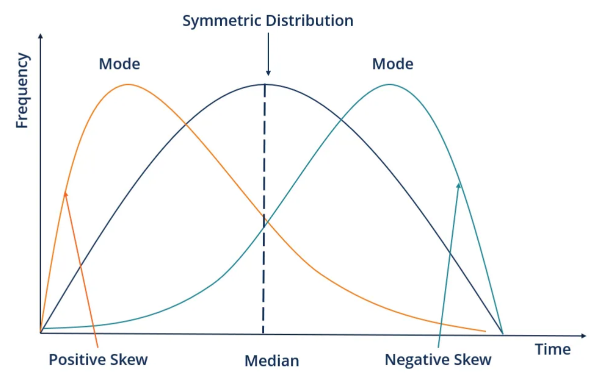
Week-8

- Random variable
- Types of random variable
  - Discrete random variable
  - Continuous random variable
- Probability mass function (p.m.f)

- $P(x_i) = P(X=x_i)$ for i=1,2,3,4. . . . . n

X	x_1	x_2	x_3	...	...	x_n
$P(X=x_i)$	$P(x_1)$	$P(x_2)$	$P(x_3)$	...	...	$P(x_n)$

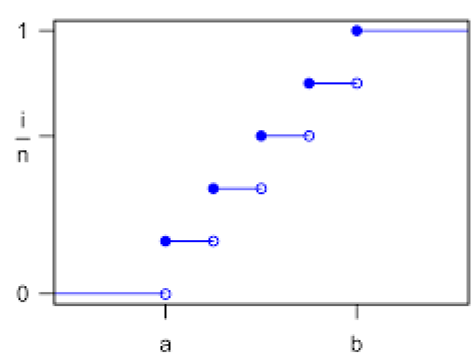
- Properties of p.m.f
  - $P(x_i) \geq 0$ for i = 1,2,3,. . . . n
  - $P(x) = 0$ for all other values of x
  - $\sum_{i=1}^{n} P(x_i) = 1$
- Graph of p.m.f
  - Positive or Right skewed distribution
  - Negative or Left skewed distribution
  - Symmetric distribution
  - Uniform distribution



- Cumulative distribution function (c.d.f)
  - $F(a) = P(X \leq a)$

X	1	2	3	4
$P(X=x_i)$	1/4	1/2	1/8	1/8

- $$F(a) = \begin{cases} 0, & a < 1 \\ \frac{1}{4}, & 1 \leq a < 2 \\ \frac{3}{4}, & 2 \leq a < 3 \\ \frac{7}{8}, & 3 \leq a < 4 \\ 1, & 4 \leq a \end{cases}$$



Week-9

- Expectation of a Random variable

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

  - E(X) is considered as the "long run average"
  - Properties
    - $E(aX + b) = aE(X) + b$
    - $E(X + Y) = E(X) + E(Y)$
- Variance of a Random variable

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

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

  - Properties
    - $V(aX + b) = a^2 V(X)$
    - $V(X + Y) = V(X) + V(Y)$ (only if X & Y are independent)
- Standard deviation of Random variable

$$SD(X) = \sqrt{V(X)}$$

  - Properties
    - $SD(aX + b) = aSD(X)$
- Bernoulli Random Variable
  - P.m.f of Bernoulli is given by

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

  - $E(X) = (1 \times p) + (0 \times (1 - p)) = p$
  - $V(X) = p(1 - p)$
- Discrete Uniform Random variable
  - P.m.f of Discrete Uniform is given by

X	1	2	3	...	n
$P(X=x_i)$	1/n	1/n	1/n	...	1/n

- ➔ $E(X) = \frac{n+1}{2}$
- ➔ $E(X^2) = \frac{(n+1)(2n+1)}{6}$
- ➔ $V(X) = \frac{n^2-1}{12}$

➔ Hypergeometric Random Variable

$$P(X = x_i) = \frac{(m_{C_i}) \times (N-m)_{C_{n-i}}}{N_{C_n}}$$ for i=0,1,2,. . . . n

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

Week-10

➔ Binomial Random Variable

- ➔ Independent and identically distributed bernoulli trials (iid)
  - ➔ For 'n' independent Bernoulli trials, each trial probabilities will be either 'p' for 'success' or '1-p' for 'failure'. If 'X' is a Random variable with no. of successes that occur in 'n' trials, then 'X' is said to be a Binomial Random variable.
- ➔ P.m.f of Binomial distribution is given by

$$P(X = i) = n_{C_i} \times p^i \times (1-p)^{(n-i)}$$

- ➔ Graph of p.m.f
  - ➔ If **p<0.5**, then Binomial is **Right Skewed** for small 'n'
  - ➔ If **p=0.5**, then Binomial is **Symmetric** for small 'n'
  - ➔ If **p>0.5**, then Binomial is **Left Skewed** for small 'n'
  - ➔ For large 'n', the Binomial distribution tends to **Symmetric**

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