

'STATISTICS' Formula Sheet :-

ADITYA

$$\text{Sample mean } \bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$$

$$\text{Population mean } \bar{\mu} = \frac{x_1 + x_2 + \dots + x_N}{N}$$

Adding a constant \Rightarrow old mean + constant = new mean

Multiplying a constant \Rightarrow old mean \times constant = new mean

Mean :- For grouped data [frequency is given]

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + \dots + f_nx_n}{n} = \frac{\sum_{i=1}^n f_i x_i}{n}$$

$$\text{where } n = f_1 + f_2 + f_3 + \dots + f_n$$

Mean :- For grouped data [Class interval is given]

$$\bar{x} = \frac{f_1m_1 + f_2m_2 + \dots + f_nm_n}{n} = \frac{\sum_{i=1}^n f_i m_i}{n}$$

where, $n = f_1 + f_2 + \dots + f_n$ and m = mid point of interval
e.g. $m[60-70] = 65$.

Mean is sensitive to outliers.

Mode :- observation with highest frequency

[Most frequent value of the data set]

Adding a constant \Rightarrow old mode + c = new mode

Multiplying a constant \Rightarrow old mode \times c = new mode

Median :- Middle value of the data set

[where data is in a ordered form]

$$\text{Median} [\text{no. of observations is odd}] = \left[\frac{n+1}{2} \right]^{\text{th}} \text{observation}$$

Median [no. of observations is even] = average of / mean of

$$\text{Median} [\text{no. of observations is even}] = \left[\frac{(n)}{2}^{\text{th}} + \frac{(n+1)}{2}^{\text{th}} \right] \text{observation}$$

Adding a constant \Rightarrow old median + c = new median

Multiplying a constant \Rightarrow old median \times c = new median.

Median is not sensitive to outliers.

Range \Rightarrow Max. value of data set - Min. value of data set.

Adding a constant \Rightarrow old range = new range

Multiplying a constant \Rightarrow old range \times c = new range.

Range is sensitive to outliers.

Variance :- Variability / Spread of data set.

$$\text{Population Var}(\sigma^2) = (x_1 - \mu)^2 + (x_2 - \mu)^2 + \dots + (x_n - \mu)^2$$

Variance of n observations with N observations

$$\text{Sample Var}(s^2) = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \dots + (x_n - \mu)^2}{n-1}$$

Adding a constant \Rightarrow old variance = new variance.

Multiplying a constant \Rightarrow old variance $\times c^2$ = new variance.

Standard deviation :- Measure of spread of data in the same unit as original data.

$$SD = \sqrt{\text{Variance}}$$

Adding a constant \Rightarrow old SD = new SD

Multiplying a constant \Rightarrow old $SD \times C$ = new SD.

Percentile :- For computing percentile, first we have to arrange the data in increasing order

n = total no. of observations , p = Percentile $\frac{100}{}$

$$\text{Computing percentile} = np$$

If np is an integer then the average of $(np^{\text{th}} + (np+1)^{\text{th}})$ observation is the required percentile value.

If np is not an integer then the smallest integer greater than np . The data value in that position is the required percentile.

Ex:- 38, 35, 61, 68, 66, 70, 68, 47, 79, 58

Increasing order :- 35, 38, 47, 58, 61, 66, 68, 68, 70, 79

For 25th percentile , $n=10$, $p=0.25$

$$\Rightarrow np = 10 \times 0.25 = 2.5 \text{ [not an integer]}$$

So, the smallest integer greater than 2.5 is 3 then

25th percentile = 3rd observation = 47.

For , 50th percentile , $n=10$, $p=0.5$, $np=5$ [Integer]

$$\text{So, } 50\text{th percentile} = \frac{5^{\text{th}} \text{ obser.} + 6^{\text{th}} \text{ obser.}}{2} = \frac{61+66}{2} = \underline{63.5}$$

The Five Number Summary :-

Minimum

Q_1 :- First Quartile or Lower quartile - 25th percentile

Q_2 :- Second Quartile or Median - 50th percentile.

Q_3 :- Third Quartile or upper quartile - 75th percentile

Maximum

The Interquartile Range :-

$$\text{IQR} = Q_3 - Q_1 = 75^{\text{th}} \text{ perc.} - 25^{\text{th}} \text{ perc.}$$

Outliers :- $Q_3 + 1.5 \text{ IQR} < \text{Outliers}, Q_1 - 1.5 \text{ IQR} > \text{Outliers}$

Row relative frequency for Contingency table :-

Divide each cell frequency in a row by its row total.

Column relative frequency for Contingency table :-

Divide each cell frequency in a column by its column total.

Covariance :- quantifies the strength of the linear association between two numerical variables.

$$\text{Population Cov}(x, y) = \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$$

$$\text{Sample Cov}(x, y) = \sum_{i=1}^{n-1} (x_i - \bar{x})(y_i - \bar{y})$$

Correlation :- The correlation measure always lies between -1 and +1.

$$\rho_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \times \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{\text{Cov}(x, y)}{S_x S_y}$$

where, S_x = standard deviation of x and S_y = SD of y .

Association between categorical and numerical variables :-

Point Bi-Serial Correlation Coefficient :-

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

Where, S_x = standard deviation of numerical variable

Y_0 = mean value of group of data associated with 0

Y_1 = mean value of group of data associated with 1.

P_0 = no. of observations associated with 0

total no. of observations

P_1 = no. of observations associated with 1

total no. of observations

When σ_{pb} is closer to 0 \rightarrow no association

When σ_{pb} is closer to -1 \rightarrow negatively associated (strongly)

When σ_{pb} is closer to 1 \rightarrow positively associated.

Addition rule of counting :-

If an action A can occur in n_1 different ways and another action B can occur in n_2 different ways then the total number of occurrences of the action A or B is $n_1 + n_2$

Multiplication rule of counting :-

Assume same situation as before then the total number of occurrences of the action A and B together is $n_1 \times n_2$.

Permutation :- Choosing and arranging

The number of possible permutations of r objects from a collection of n distinct objects is given

by $n \times (n-1) \times \dots \times (n-r+1)$ or

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

$$* {}^n P_0 = \frac{n!}{n!} = 1$$

$$* {}^n P_1 = n$$

$$* {}^n P_n = n!$$

When repetition is allowed then the number of possible permutations of r objects from a collection of n distinct objects is

$$n \times n \times n \dots \text{times} = n^r$$

Permutation when objects are not distinct.

The number of permutations of n objects when p of them are one of kind and rest distinct is equal

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

The number of permutations of n objects where p_1 is of one kind, p_2 is of second kind, and so on p_k of k^{th} kind is given by $\frac{n!}{p_1! p_2! \dots p_k!}$

Circular Permutation

The number of ways n distinct objects can be arranged in a circle (clockwise and anticlockwise are different) is equal to

$$(n-1)!$$

The number of ways n distinct objects can be arranged in a circle (clockwise and anticlockwise are same) is equal to $\frac{(n-1)!}{2}$

Combination :- Selection. :- Order is not important

The number of possible combinations of r_1 objects from a collection of n distinct objects is denoted by

$${}^n C_{r_1} = \frac{n!}{r_1!(n-r_1)!}$$

$$* {}^n C_{r_1} = {}^n C_{n-r_1} = \frac{n!}{(n-r_1)! r_1!}$$

$$* {}^n C_n = 1, {}^n C_0 = 1$$

$$* {}^n C_{r_1} = {}^{n-1} C_{r_1-1} + {}^{n-1} C_{r_1}; 1 \leq r_1 \leq n$$

$$* {}^{n-p} C_{r_1-p} = \frac{(n-p)!}{(r_1-p)!(n-r_1)!} \rightarrow \text{When } p \text{ particular things are always included.}$$

$$* {}^n P_{r_1} = {}^n C_{r_1} \times r_1!$$

Probability :-

Classical :-

For sample space S , in which there are n equally likely outcomes, and the event E consist of exactly m of these outcomes probability of E that is

$$P(E) = \frac{m}{n}$$

* $0 \leq P \leq 1$

* $P(S) = 1$, where S is sample space.

* For a sequence of mutually disjoint events, E_1, E_2, \dots

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$$

If E_1 and E_2 are disjoint events, then

$$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

* $P(E^c) = 1 - P(E)$. * $P(E \cup E^c) = 1$

If E_1 and E_2 are not disjoint events then,

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

Conditional Probability :-

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

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

When Events E and F are independent then.

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

$$P(E|F) = P(E)$$

Any three 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 :-

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

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

Baye's Rule :-

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

Probability Mass function | Cumulative Distribution Function

PMF	X	1	2	3	4
	$P(x=x_i)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{8}$	$\frac{1}{8}$

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

CDF

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{1024}{3125} & 0 \leq x < 1 \\ \frac{3125}{3125} & 1 \leq x < 2 \\ \frac{2304}{3125} & 2 \leq x < 3 \\ \frac{3125}{3125} & 3 \leq x < 4 \\ \frac{2944}{3125} & 2 \leq x < 3 \\ \frac{3104}{3125} & 3 \leq x < 4 \\ \frac{3124}{3125} & 4 \leq x < 5 \\ 1 & x \geq 5 \end{cases} \xrightarrow{\text{pmf}} P(x=x)$$

Expectation of a Random variable :- long run average

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

$$* E(ax+b) = aE(x) + b$$

$$* E(X+Y) = E(X) + E(Y)$$

Variance of a Random variable :-

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

$$\text{Var}(X) = E(X^2) - E(X)^2$$

$$* V(ax+b) = a^2 V(X)$$

$$* V(X+Y) = V(X) + V(Y) \quad [\text{only if } X \text{ and } Y \text{ are independent}]$$

Standard deviation of a Random variable :-

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

$$* SD(ax+b) = a SD(X).$$

Expectation of a Bernoulli random variable :-

$$E(X) = p \quad [0x(1-p) + 1xp]$$

Variance of a Bernoulli random variable :-

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

Discrete uniform random variable :-

pmf	X	1	2	---	n
P(X=x_i)	$\frac{1}{n}$	$\frac{1}{n}$	---	$\frac{1}{n}$	

$$E(X) = \frac{(n+1)}{2} = \sum_{i=1}^n x_i p(x_i)$$

$$E(X^2) = (n+1)(2n+1)$$

6

$$V(X) = \frac{n^2 - 1}{12} \quad [\text{Variance of Urv}]$$

Hypergeometric random variable :-

$$P(X = x_i) = \frac{\binom{m}{x_i} \times \binom{N-m}{n-x_i}}{\binom{N}{n}} \text{ for } i=0, 1, 2, \dots, n$$

* $E(X) = \frac{nm}{N}$

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

Binomial Random Variable :-

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

* $E(X) = p + p + p + \dots \text{ n times} = np.$

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

Continuous Random Variables :- probability density function

$$P(a \leq X \leq b) = \int_a^b f(x) dx = P[X \in [a, b]] \quad \leftarrow$$

Cdf :- $F(a) = P(X \leq a) = \int_{-\infty}^a f(x) dx$

* $P(a \leq X \leq b) = P(a < X < b)$

$$* E(x) = \int x f(x) dx.$$

$$* V(x) = \int (x - E(x))^2 f(x) dx.$$

Uniform distribution $U(a,b)$:-

$$f(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{otherwise} \end{cases}$$

Standard uniform distribution :- $U(0,1)$

$$f(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

Verify $f(x)$ is a pdf

$$* (x) \geq 0, \text{ for } 0 < x < 1$$

$$* \int_{-\infty}^{\infty} f(x) dx = \int_0^1 f(x) dx = 1$$

Cumulative distribution of Uniform distribution :-

For $X \sim U(a,b)$

$$F(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } x \in [a,b] \\ 1 & \text{for } x \geq b \end{cases}$$

Expectation of $X \sim U(a,b)$

$$E(x) = \frac{b+a}{2}$$

Variance of $X \sim U(a,b)$

$$\text{Var}(x) = (b-a)^2$$

Exponential distribution :- For some $\lambda > 0$

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Cumulative distribution function of Exponential distribution:-

$$F(a) = 1 - e^{-\lambda a}$$

Expectation :- $x \sim \exp(\lambda)$

$$E(x) = \frac{1}{\lambda}$$

Variance :- $x \sim \exp(\lambda)$

$$\text{Var}(x) = \frac{1}{\lambda^2}$$

BY - ADITYA DHAR DWIVEDI