Page No. :
Date : / /

:- STATISTICS : Formula Sheet :- ADITYA

Sample mean $\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n-1}$

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

Adding a constant ⇒ old mean + constant = new mean
Multiplying a constant ⇒ old mean X constant = new mean

Mean :- For grouped data [frequency is given]

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

where $n = f_1 + f_2 + f_3 + \dots + f_n$

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

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

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 heighest frequency
[Most frequent value of the data set]

Adding a constant ⇒ old mode + c = new mode
Multiplying a constant ⇒ old mode X c = new mode.

Page No. :
Date : / /

Median :- Middle value of the data set
[where data is in a ordered form]

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

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

$\left[\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th}\right]$ observation.

Adding a constant ⇒ old median + C = new median
Multiplying a constant ⇒ old median × C = new median.

Median is not sensitive to outliers.

Range ⇒ Max. value of data set − Min. value of data set.

Adding a constant ⇒ old range = new range
Multiplying a constant ⇒ old range × C = new range.

Range is sensitive to outliers.

Variance :- Variability / Spread of data set.

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

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

Adding a constant ⇒ old variance = new variance.
Multiplying a constant ⇒ old variance × c^2 = new variance

Page No. :
Date : / /

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

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

Adding a constant ⇒ old SD = new SD
Multiplying a constant ⇒ old SD × C = new SD.

Percentile :- For computing percentile, first we have to arrange the data in increasing order.
n = total no. of observations , $p = \frac{\text{percentile}}{100}$

Computing percentile = np

If np is an integer then the average of $(np^{th} + (np+1)^{th})$ obser. 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
⇒ np = 10 × 5 = 2.5 [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]
So, 50th percentile = $\frac{5^{th} \text{ obser.} + 6^{th} \text{ obser.}}{2} = \frac{61+66}{2} = 63.5$

Page No. :
Date : / /

The Five Number Summary :-

Minimum
Q_1 :- First Quartile or lower quartile - 25^{th} percentile.
Q_2 :- Second Quartile or Median - 50^{th} percentile.
Q_3 :- Third Quartile or upper quartile - 75^{th} percentile.
Maximum.

The Interquartile Range :-

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

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

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.

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

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

Page No. :
Date : / /

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

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

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

Association between categorical and numerical variables :-

Point Bi-Serial Correlation Coefficient :-

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

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.

$$P_0 = \frac{\text{no. of observations associated with 0}}{\text{total no. of observations}}$$

$$P_1 = \frac{\text{no. of observations associated with 1}}{\text{total no. of observations}}$$

When r_{pb} is closer to 0 → no association
When r_{pb} is closer to -1 → negatively associated (strongly)
When r_{pb} is closer to 1 → positively associated

Page No. :
Date : / /

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 $n_1 \times n_2$.

Permulation :- Choosing and arranging
The number of possible permulations of r objects from a collection of n distinct objects is given by $n \times (n-1) \times \dots \times (n-r+1)$ or

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

* $^nP_0 = \frac{n!}{n!} = 1$

* $^nP_1 = n$

* $^nP_n = n!$

When repetation is allowed then the number of possible permulations of r objects from a collection of n distinct objects is

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

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

Page No. :
Date : / /

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 objects from a collection of n distinct objects is denoted by

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

* $^nC_r = {}^nC_{n-r} = \frac{n!}{(n-r)!\,r!}$

* $^nC_n = 1 \quad , \; ^nC_0 = 1$

* $^nC_r = {}^{n-1}C_{r-1} + {}^{n-1}C_r \; ; \; 1 \le r \le n$

* $^{n-p}C_{r-p} = \frac{(n-p)!}{(r-p)!\,(n-r)!}$ → When p particular things are always included.

* $^nP_r = {}^nC_r \times r!$

Page No. :
Date : / /

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 E. (disjoint) events, $E_1, E_2 \ldots$

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

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.

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

Page No. :
Date : / /

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 :-

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

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

↳ Cdf : $F(a) =$

$$\begin{cases} 0 & a < 1 \\ 1/4 & 1 \le a < 2 \\ 3/4 & 2 \le a < 3 \\ 7/8 & 3 \le a < 4 \\ 1 & 4 \le a \end{cases}$$

Cdf

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{1024}{3125} & 0 \le x < 1 \\ \frac{2304}{3125} & 1 \le x < 2 \\ \frac{2944}{3125} & 2 \le x < 3 \\ \frac{3104}{3125} & 3 \le x < 4 \\ 3124/3125 & 4 \le x < 5 \\ 1 & x \ge 5 \end{cases}$$

→ pmf

$$P(x = x) \begin{cases} 1024/3125 & x = 0 \\ 1280/3125 & x = 1 \\ 640/3125 & x = 2 \\ 160/3125 & x = 3 \\ 20/3125 & x = 4 \\ 1/3125 & x = 5 \end{cases}$$

Page No. :
Date : / /

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 :-

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

* $V(aX+b) = a^2 V(X)$
* $V(X+Y) = V(X) + V(Y)$ [only if X and Y are independent]

Standard deviation of a Random variable :-

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

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

Expectation of a Bernouli random variable :-

$$E(X) = p \quad \boxed{0 \times (1-p) + 1 \times p}$$

Variance of a Bernouli random variable :-

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

Discrete uniform random variable :-

pmf

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

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

Page No. :
Date : / /

$$E(x^2) = \frac{(n+1)(2n+1)}{6}$$

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

Hypergeometric random variable :-

$$P(X = x_i) = \frac{{}^{m}C_i \times {}^{N-m}C_{n-i}}{{}^{N}C_n} \quad \text{for } i = 0, 1, 2, \ldots n$$

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

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

Binomial Random Variable :-

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

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

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

Continuous Random Variables :- probability density function

$$P(a \le X \le b) = \int_a^b f(x)\,dx = P[X \in [a,b]].$$

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

* $P(a \le x \le b) = P(a < x < b)$

Page No. :
Date : / /

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

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

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

* $(x) \geq 0$, 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 $X \sim U(a,b)$

$$Var(X) = \frac{(b-a)^2}{12}$$

Page No. :
Date : / /

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