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