*) Step by step solution to a question is called "Algorithm"
(or)
problem

*) Graphical way of expressing the ~~flow~~ algorithm is called "Flowchart"

*) Pseduo Code is a way of representing the flowchart using words (or) building logic which almost resembles the actual code but without proper syntax. ~~[illegible]~~ ~~order~~

One, who cannot understand Vedanta, should uplift himself step by step under Guru's guidance.

Sun
Jan 10

→ Individual :- Data which is unique in the data set which will be taken as ref.

→ Mean :- Average of all the data.

→ Median :- Middle value of all the data after sorting it in ascending order. For ODD number of values, you get exact middle value
For EVEN — " — , you need to take avg of middle two values

→ Mode :- Most number of occurences.

2016 Sun Jan 17

→ Marginal Distribution :-

is a frequency / relative frequency distribution of either the row / column variable.

→ Conditional Distribution :-

is a distribution of values for one variable that exists when you specify the other values of other variables.

(or)

An event occurring in the presence of a second event.

→ Interquartile Range (IQR) :- Range b/w 25^{th} percentile & 75^{th} percentile

Mon Jan 18

→ Bar Graph :- is a graphical representation of "categorical data". Plotted to each "individual data"

→ Histogram :- is a graphical representation of "quantitative data"

Here the data is grouped into continuous ranges of equal width

It represents how many times a data is repeated

→ Stem & Leaf :- "Stem" contains "Left" part of data

"Leaf" contains "Right" part of data

19 Tue Jan 2016

*) Standard deviation measures how far the data is spread out from the mean.

*) If the distribution is "left skewed", then the mean would be < median

*) If the distribution is "right skewed", then the mean would be > median

*) In box plot, "Interquartile Range (IQR)" is the difference b/w left vertical line & right vertical line of the box

*) For "Normal Distribution", standard deviation is a good measure of spread

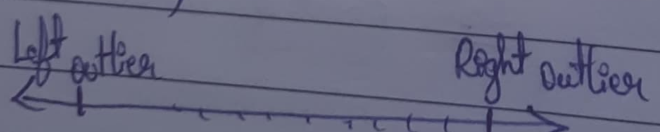
For "Flat Distribution", standard deviation is not a good measure of spread

20 Wed Jan

*] Removing an Outlier :-

(lowest outlier)
from left :- Mean ↑ Median ↑ (by small margin)

from right :- Mean ↓ Median ↓ (changes by small margins)
(Highest outlier)



Statistics

→ Descriptive Statistics :- Describing the data without giving all of the data (or) using smaller set of numbers

→ Inferential Statistics :- Giving Inferences on the data, giving judgement (or) conclusions on the data

→ Average :- "typical" (or) "middle" ⇒ (Measure of) "Central" Tendancy.

↳ Arithmetic Mean : $\frac{\Sigma(\text{all numbers})}{\text{Total no. of data}}$ (μ)

↳ Median : Sort the values in ascending order, & then the middle value of the sorted values will be median

For ODD no. of data ⇒ Median : Middle value

For EVEN no. of data ⇒ Median : [Mean of middle two value]

↳ Mode : The most common value in the data set.

- There can be 1 mode, 0 mode (or) multiple modes

Range :- Diff b/w MAX & MIN value

↳ $\boxed{\uparrow \text{Range} \quad \uparrow \text{Spread of data}}$

Mid-Range :- Avg of MAX & MIN value

23 Sat Jan 2016

→ Inter Quartile Range [IQR] :-

- Measure of "statistical dispersion", which is the spread of data

- Also be called "Mid spread" / "Middle 50%" / "Fourth Spread" / "H-spread"

- Also defined as the difference b/w 75th percentile & 25th percentile of the data

⇒ How to find : i) Sort the data in ascending order.
ii) Find the "median value" (50th percentile)
iii) Find the median value on either side of the 1st median value
(25th percentile → on left) (75th percentile → on right)
iv) Get the difference b/w 75th percentile & 25th percentile.

Those who are proud of their devotion are not true devotees.

24 Sun Jan

Measure of Spread / Dispersion :-

1] Variance :- (σ^2)

$$\boxed{\sigma^2 = \frac{\sum (x_i - \text{Mean})^2}{\# \text{ of values}}}$$

*) Higher the variance value, more will be the spread of data

2] Standard Deviation :- (σ)

$$\boxed{\sigma = \sqrt{\text{variance}} = \sqrt{\sigma^2}}$$

*) ↑ std. deviation ↑ spread of data

- SD can't be -ve
- SD close to 0 indicates that data points tends to be close to mean

2016

Mon Jan 25

Eg: Find SD of $\{1, 4, 7, 2, 6\}$

$$\mu = \frac{1+4+7+2+6}{5} = \frac{20}{5} = \underline{\underline{4}}$$

$$\sigma^2 = \frac{(1-4)^2 + (4-4)^2 + (7-4)^2 + (2-4)^2 + (6-4)^2}{5}$$

$$= \frac{9+0+9+4+4}{5} = \frac{26}{5} = \underline{\underline{5.2}}$$

$$SD = \sqrt{\sigma^2} = \sqrt{5.2} = \underline{\underline{2.28}}$$

- For Symmetrical Datasets :- Mean & Std. Deviations are good

To recognize divinity in man and to serve humanity unselfishly is true devotion to God.

- For Skewed Datasets :- Median & IQR are better. (They are robust to the outliers)

Tue Jan 26

- Alternate formula for Variance (σ^2)

$$\boxed{\sigma^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - \mu^2}$$

→ Population :- Complete data set is considered to be population

→ Sample :- Sample is sub-set of population
Collection of small data made out of population

27 Wed Jan 2016

Sample mean → $\boxed{\bar{x} = \frac{\sum_{i=1} x_i}{n}}$

Sample Variance → (un-biased) $\boxed{S_{n-1}^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$

Sample Std. Deviation → $\boxed{S_{n-1} = \sqrt{S_{n-1}^2}}$

But due to non-linear functionality of square root, it induces bias into the solution.

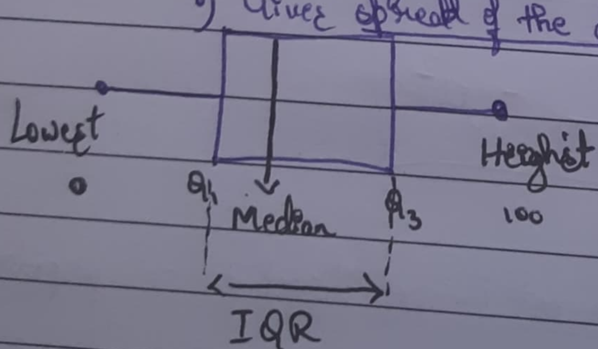
∴ The sample std deviation is biased.

- Unbiased estimation :- Expected value of the quantity obtained by dividing the observed sample variance by a correction factor.
  → That the estimator is equal to the true value within the population

Faith in God should unite, not break up communities.

28 Thu Jan

* Whisker & Box plots :-
  - Gives lowest & highest value
  - Median of the entire range
  - IQR for the data
  - Gives "spread of the data"



Outlier :- Data $\begin{cases} < Q_1 - 1.5 * IQR \\ > Q_3 + 1.5 * IQR \end{cases}$

2016

7. Mean Absolute Deviation (MAD)

Fri Jan 29

Avg distance b/w each data point & the mean

Gives idea about the variability in a dataset.

$$\boxed{MAD = \frac{\sum_{i=1}^{n} |x_i - m(x)|}{n}}$$

$m(x) \rightarrow$ Arithmetic mean of data set

$x_i \rightarrow$ Data value in set

$n \rightarrow$ No. of data values

Unit Test

4 rounds → _ _ _ _

97
(mean & median)

Forget yourself, like a small child playing in the lap of nature.

Sat Jan 30

22, 23, 23, 25, 27

~~median~~

$m(x) = 23$.

$$\frac{|22-23| + |23-23| + |23-23| + |25-23| + |27-23|}{5} = \frac{1+0+0+2+4}{5}$$

$$= \frac{7}{5} = 1\frac{2}{5} = 1.4$$

4, 5, 10, 13

$$\frac{5+10}{2} = \frac{15}{2} = 7.5$$

High Outliers > $Q_3 + 1.5\ IQR$

$73.5 + 1.5(73.5 - 64.5)$

$73.5 + 13.5$

$= 87$

⊛. Normal Distribution :- [(or) Gaussian Distribution]

2016 Mon Feb 01

→ **Z-scores** :- Statistical measurement that describes a value's relationship to the mean of a group of values.

- It is measured in terms of Std. deviations from mean.
- If Z-score is 0, then it indicates that the data point is identical to the mean score.

$$\boxed{Z = \frac{x - \mu}{\sigma}}$$

x → Observed Value
μ → Mean of the sample
σ → Std. deviation of sample

- How many "Std. deviation [σ]" away from "Mean [μ]"

Z =
+ve → Value is ~~to the~~ above mean
-ve → Value is below mean

0 → Value is identical to mean

Tue Feb 02

$3 < Z$, $-3 > Z$ → Data point is considered to be unusual [Outliers]

⊛. ~~When~~ the parameters of central tendency & Dispersion, changes for :

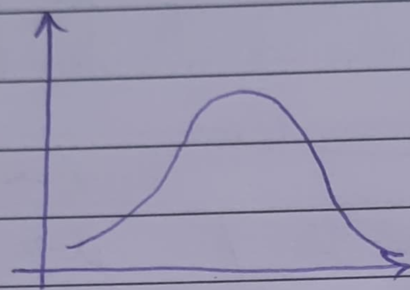
i) Shifting the data value [Adding / Subtracting] → Central Tendency shifts accordingly; Dispersion doesn't shift.

ii) Scaling the data values [Multiplying / Dividing] → Central Tendency scales accordingly; Dispersion also scales accordingly

03 Wed Feb 2016

v). **Normal Distribution** :- is a probability distribution that is symmetric about mean, showing the data near the mean are more frequent in occurance than data far from mean.

"bell curve"



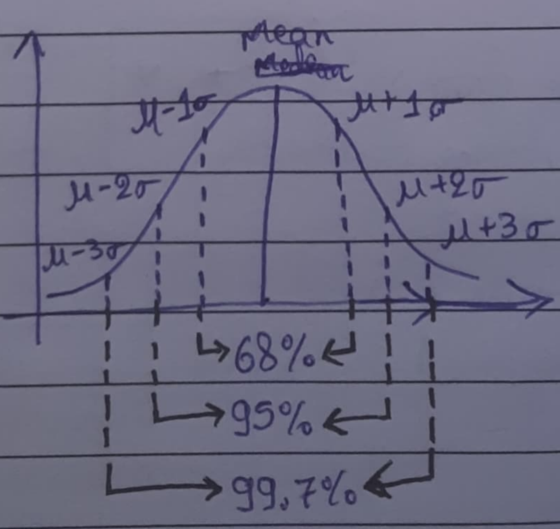
*). **Density Curve** :- Graphical representation of numerical distribution where the outcomes are continuous.

> Like birds and animals humans also must obey the clock of nature by waking up early.

04 Thu Feb

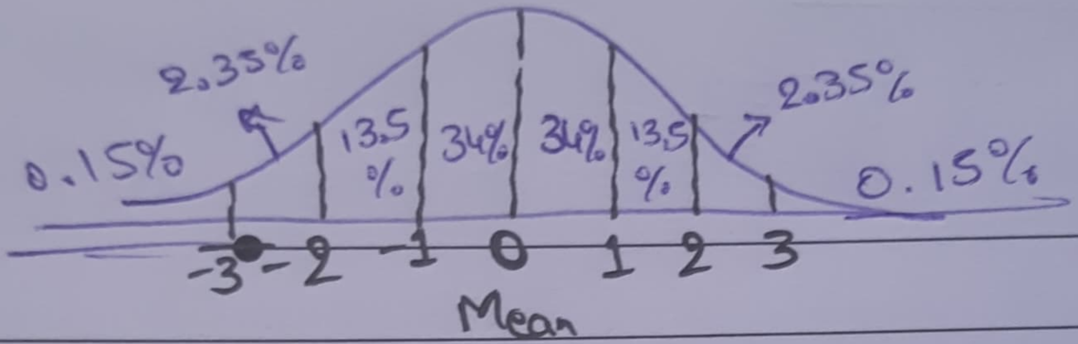
*). **Empirical Rule** :- (or) "68 - 95 - 99.7 % Rule"

- Applies only to "normal distribution"
- Gives the probability percentage of finding a value in the curve area.



7 Sun
Feb

*) Normal Distribution



7. Scatter Plot :-

2016 Tue Feb 09

Kinds of relationships

$$\underbrace{\begin{pmatrix} \text{Linear} \\ \text{Non-linear} \end{pmatrix}}_{\text{Form}} + \underbrace{\begin{pmatrix} \text{Strong} \\ \text{Weak} \end{pmatrix}}_{\text{Strength}} + \underbrace{\begin{pmatrix} \text{Positive} \\ \text{Negative} \end{pmatrix}}_{\text{Direction}}$$

Linear → Able to draw a straight line connecting the dots

Non-linear → Not able to ——— " ———

Strong → Most of the dots are on the line (or) nearer to the line

Weak → Most of the dots are away from the line (or) Less no. of dots are on the line

+ve → ~~If one variable ↑, the other variable ↑~~

-ve → ~~If one variable ↑, the other variable ↓~~

Wed Feb 10

Correlation :- Any statistical relationship, whether casual (or) not between two random variables (or) bivariate data

i) Bivariate data :- Data which depends on two variables

Correlation coefficient :- (r)

Measures the direction & strength of a linear relationships.

$$\boxed{r = \frac{1}{n-1} \sum \left[Z_{x_i} \cdot Z_{y_i} \right]} \qquad \boxed{-1 \le r \le 1}$$

$r \approx \begin{cases} -1 & \rightarrow \text{Strong -ve linear} \\ 0 & \rightarrow \text{Weak / No relationship} \\ +1 & \rightarrow \text{Strong +ve linear} \end{cases}$

11 Thu Feb 2016

$$\text{Slope} = \frac{\text{Change in y-axis}}{\text{Change in x-axis}}$$

+ve ⇒ +ve linear trend
-ve ⇒ -ve linear trend

Linear eq^n

$y = mx + b$ — m: slope, b: y-intercept

Regression line Equation

$$\boxed{\hat{y} = (\text{Slope} \cdot x) + (\text{y-intercept})}$$

Linear Regression :- Process of drawing a line through data in a scatter plot.

The line summarizes the data, which is useful when making predictions.

→ The line has to go through the "middle" of the points

12 Fri Feb

*]. Residual :- is a measure of how well a line fits an individual data plot.

Distance b/w the actual dot & the predicted dot on the regression line is the "residual".

- for Data points above the line, residual will be +ve
- for data points below the line, residual will be -ve

- Residual value closer to 0, the better the fit

Eqⁿ for regression line,

$$\hat{y} = \overset{\text{slope}}{m}x + \underset{\text{y-intercept}}{b}$$

We know, that the regression line must pass through the mean of the data values i.e., $(\bar{x}, \bar{y})$

$\therefore$ $\boxed{\text{slope } m = r \frac{S_y}{S_x}}$ → r → correlation coefficient when $r = 1/-1$

$\boxed{\text{y-intercept, } b = \bar{y} - r \frac{S_y}{S_x} \bar{x}}$

*] Residual Plot:- ~~Plot above~~

Drawn against the residual values for each data points.

- If ~~the~~ more no. of points above $(y=0)$ on residual plot, then the model i.e. under ~~over~~estimating y

Do not take Guru for granted just because He speaks to you freely.
Realize that His powers are immense and hidden.

- If ~~the~~ more no. of points below $(y=0)$ on residual plot, then the model i.e. over ~~under~~ estimating y

- If there are equal no. of points on both sides, then it is overestimating just as often as it underestimating y.

*]. R^2 :- r^2 measures how much prediction error is eliminated when we use least-square regression

- Gives what percent of the prediction error in the y-variable is eliminated when we use least-square regression on x-variable

$\boxed{r^2 \rightarrow \text{Coefficient of determination}}$

Total Squared Error b/w points & line,

$$\boxed{SE_{Line} = \sum_{i=1}^{n} (R_i)^2}$$

R → Residual
(Actual − Predicted)
[y − (mx + b)]

- How much / what % of total variation in y is described by variation in x

Total variation in y

$$\boxed{SE_{\bar{y}} = \sum_{i=1}^{n} (y_i - \bar{y})^2}$$

- "SE_{Line}" gives the total variation is not described by regression line (or) by variation in x

Family quarrels disturb the minds of children and spoil their future.

- To find out the % of variation not described by regression line (or) by variation in x

$$\boxed{\frac{SE_{Line}}{SE_{\bar{y}}}}$$

- $\boxed{1 - \frac{SE_{Line}}{SE_{\bar{y}}}}$ → Gives the % of variation described by regression line (or) by variation in x

↳ Coefficient of Determination (or) r^2

if SE_{Line} is small → Line is a good fit
↳ r^2 close to 1

SE_{Line} is huge → Line is not a good fit
↳ r^2 close to 0

2016 Wed Feb 17

7. Standard deviation of residuals (or) Root - Mean - Square Error [RMSE]

$$\boxed{\sqrt{\frac{\sum_{i=1}^{n} (r_i)^2}{(n-1)}}}$$

Typical prediction error

→ Gives how much the model disagrees the data

→ Lower the value, the better the fit

⇒ Minimizing squared error to regression line:-

$$y = mx + b.$$

→ $(\bar{x}, \bar{y})$ lies on the line

Clever manipulation of others for selfish benefit hurts you in the end.

→ $\left(\frac{\overline{x^2}}{\bar{x}}, \frac{\overline{xy}}{\bar{x}}\right)$ on the "best" line

Thu Feb 18

$$\boxed{m = \frac{\bar{x}\bar{y} - \overline{xy}}{(\bar{x})^2 - (\overline{x^2})}}$$

$$\boxed{b = \bar{y} - m\bar{x}}$$

Correlation v/s Causality:-

19 Fri Feb 2016

Causality is that one event is the result of the occurrence of the other event.

Correlation is a relation b/w two variables. If value of one variable changes so does the other variable.
But it's not necessarily mean that the change in one variable is the cause of the change in other variable
"Correlation is weaker than Causation"

*] Techniques for Random Sampling & avoid biasing

Sampling → Random ⇒ Simple ; Stratified ; Clustered ; Systematic
Sampling → Not Random ⇒ Voluntary ; Convenience

20 Sat Feb

If samples are not random then they → This introduces bias

wording biasing ; Response bias ;
↓ Influencing the decision
↓ May be noresponse bias (or) May have complete response with not completly true with opinions

→ Sampling methods refer to how we select members from the population to be in study

Not Random :- (Bad ways to sample)

⇒ Convenience :- Chooses a sample that is readily available in some non-random way

⇒ Voluntary :- Puts out a request for members of a population to join the sample & people decides whether to join (or) not.

⇒ ~~Good~~

Random :- (Good ways to sample)

⇒ Simple :- Every member & set of members has an equal chance of being included in the sample

[Technology, Random number generators ----]

⇒ Stratified :- Population is first split into groups. And same number ~~of people~~ of data will be choosen randomly from each group randomly

⇒ Cluster :- Population is split into groups. And groups get selected randomly and all members of selected group will ~~take part~~ be included in the sample

⇒ Systematic :- Members of population are put in some order. A starting point is selected at random & every n^{th} member from the starting point is included in the sample

23 Tue Feb 2016

Types of Statistical Study

① Sample Study :- To estimate a certain parameter

② Observational Study } To compare two parameters

③ Experimental Study }

① Sample study :- Taking some random sample from the large population

② Observational study :- We measure / survey members of sample without trying to affect them
↳ [Cannot report Casual relationships]

③ Experimental study :- We divide the population into groups based on some control condition & apply one group with treatment & other without treatment.
→ Also called as controlled experiment

Uncontrolled desire for sensual pleasures leads you away from God.

→ Designed to suggest Causation.

24 Wed Feb

	Random Sampling	Not Random Sampling
Random assignment	Can determine Casual relationship in population. This design is relatively rare in real world	Can determine Casual relationship in that sample only. This design where most experiments would fit.
No Random assignment	Can detect relationships in population, but cannot determine causality. This design is where many surveys & observational studies would fit	Can detect relationships in that sample only, but cannot determine causality. This design is where many unscientific surveys & polls would fit

Probability

2016 Thu Feb 25

- How likely something is going to happen

- Analysis of events governed by probability is called Statistics

$$\boxed{P(\text{of an event}) = \frac{\text{No. of ways it can happen}}{\text{Total no. of outcomes}}} \qquad \boxed{0 \leq P \leq 1}$$

→ That are equally likely to come

$$P = \begin{cases} 0 \Rightarrow 0\% \text{ chance of occuring} \\ 1 \Rightarrow 100\% \text{ chance of occuring} \end{cases}$$

Sample space :- Set of all possible outcomes.

The mind turned inwards is the Self; turned outwards, it is ego and the world.

Sample size → Total no. of all possible outcomes

Fri Feb 26

Set :- Collection of objects.

$\cap$ → Intersection [AND]

$\cup$ → Union [OR]

Difference :- Remove the elements from first set [or Left Set] which are present in the second set [or Right set].

Eg: $A = \{1,2,3,4,5,6,7\}$ $B = \{1,5,7,9\}$

$A - B = \{2,3,4,6\}$ "B subtracted from A"

(or)

"Relative Complement of set B in set A"

$A \backslash B$ → Elements in set A not in set B (or)

Relative Compliment of B in A (or)

B subtracted from A.

$\boxed{A \backslash A \rightarrow \phi}$ → A subtracted from A gives Null set / Empty set

27 Sat Feb 2016

$\mathbb{R}$ → Set of all real numbers

$\mathbb{Q}$ → Set of all rational numbers

$\mathbb{Z}$ → Set of all integers

$\in$ → Belongs to $\notin$ → Not belongs to

U → Universal Set ' → Compliment

$\subseteq$ → Subset [If all elements in one set is a member of other]

$\subsetneq$ → Strict subset [Same like subset, but vice-versa doesn't hold]
(or)
Proper subset

$\supseteq$ → Superset [Contains all elements from the other set]

Only Prayer does not grant liberation.Grace and guidance from Guru are also required.

28 Sun Feb

•) When there is over lap in the data sets,

$\boxed{P(A \text{ OR } B) = P(A) + P(B) - P(A \text{ AND } B)}$

$\boxed{P(A \cup B) = P(A) + P(B) - P(A \cap B)}$

$\boxed{P(A \text{ AND } B) = P(A) + P(B) - P(A \text{ OR } B)}$

$\boxed{P(A \cap B) = P(A) + P(B) - P(A \cup B)}$

•) When there is no overlapping in data sets, then

$\boxed{P(A \text{ AND } B) = 0}$ → $\boxed{P(A \cap B) = 0}$

$\therefore$ $\boxed{P(A \text{ OR } B) = P(A) + P(B)}$ → $\boxed{P(A \cup B) = P(A) + P(B)}$

Mutually Exclusive :- Sets are said to be mutually exclusive, then there are common data b/w them

Independent Events :-

2016 Mon Feb 29

Events whose previous results won't affect their future results.

- $\boxed{P(\text{Event happening}) = 1 - P(\text{Event Not happening})}$
  $\boxed{P(\text{Event Not happening}) = 1 - P(\text{Event happening})}$

- $\boxed{P(\text{Event happening } n \text{ times in row}) = [P(\text{Event happening one time})]^n}$

when A & B are independent

- $\boxed{P(A\cdot B) = P(A) \cdot P(B)}$ $\boxed{P(A \text{ AND } B) = P(A) \cdot P(B)}$ $\boxed{P(A|B) = P(A) \quad P(B|A) = P(B)}$

$P(A+B) \Rightarrow$ A & B are independent events

Probability of event A & Probability of event B, both are happening.

Share your love and friendship with plants, birds, and animals.

Notes $\Rightarrow$ $\boxed{P(\text{atleast 1 success}) = 1 - P(\text{all failures})}$
$\boxed{P(\text{atleast 1 failure}) = 1 - P(\text{all success})}$

Dependent Events :- Events whose previous result will affect their future result

General Multiplication Rule, $\boxed{P(A \text{ AND } B) = P(A) \cdot P(B|A)}$

$P(B|A)$ - "Given"

Probability B occurs given that A already occured

(or)

$\boxed{P(A \text{ AND } B) = P(B) \cdot P(A|B)}$

$$P(A \text{ AND } B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$$

Notes

Conditions, $P(A) < P(A|B)$ $\quad$ $P(B) < P(B|A)$

2016

$$\boxed{P(A|B) = \frac{P(B|A) * P(A)}{P(B)}}$$

Permutation & Combination

2016 Tue Mar 01

Permutation is an act of arranging objects (or) numbers in order.

Combinations are the way of selecting objects (or) numbers from group of objects (or) collections, in such a way the order of the ~~matter~~ objects does not matter

$$\boxed{{}^nP_r / nPr / P(n,r) = \frac{n!}{(n-r)!}}$$

$\boxed{0! \to 1}$

$\boxed{{}^nP_n = n!}$

$\boxed{{}^nP_1 = n}$

$\boxed{{}^nP_0 = 1}$

Glitter and glamour do not give peace and harmony. Truth and simplicity do.

Wed Mar 02

$$\boxed{{}^nC_r / nCr / C(n,r) = \frac{{}^nP_r}{r!} = \frac{n!}{r!\,(n-r)!}}$$

$\boxed{{}^nC_n = 1}$ $\boxed{{}^nC_0 = 1}$ $\boxed{{}^nC_1 = n}$

- For a fair coin flipped 'n'-times,

Total possible outcomes $= 2^n$

Getting 'k' exactly in n flips $= \frac{n!}{(n-k)!\,k!} = nC_k$

$P(\text{event}) = \frac{nC_k}{2^n}$

03 Thu Mar. 2016

Random Variables (Quantifying a random process to the possible outcomes)

It is a numerical description of the outcome of a statistical experiment.

- Represented using "Capital" letters.

Eg:-

$$X = \begin{cases} 1 \to \text{if heads} \\ 0 \to \text{if tails} \end{cases}$$

P(X head)

$$Y = \begin{Bmatrix} \text{Sum of all upward facing dice} \\ \text{after rolling 7 times} \end{Bmatrix}$$

Problems are blessings in disguise when they make you stronger and wiser.

$P(Y \text{ even})$; $P(Y \leq 30)$

04 Fri Mar

Types of random variables:-

↳ Discrete :- Distinct / Separate value [Can Count the no. value taken]

↳ Continuous :- Any value in interval.

Mean of discrete random variable,
(or)
Expected Value

$$\boxed{\sum_{i=1}^{n} x_i \cdot P(x_i)}$$

$x_i \to$ Distinct value

$P(x_i) \to$ Probability of x_i

↳ This is the value to be expected in longer runs

2016

Sat Mar 0

Variance of discrete random variable,

$$\boxed{Var(X) = \sum_{i=1}^{n} (x_i - E(x))^2 \cdot P(x_i)}$$

$x_i \to$ Distinct value

$E(x) \to$ Expected Value of the random variable

$P(x_i) \to$ Probability of the corresponding distinct value

$$\boxed{Var(x) = E((x-\mu_x)^2) = \sigma_x^2}$$

Standard deviation of discrete random variable,

$$\boxed{\sigma_X = \sqrt{Var(X)}}$$

Sun Mar 0

- Probability for a continuous value in the continuous random variable will be the "Area Under the Curve"

07 Mon Mar 2016

*]. Shifting in Random Variable,
(Adding / Subtracting)

Data points, Mean shifts accordingly
• Standard Deviation doesn't shifts

*] Scaling in Random Variables,
(Multiplication / Division)
Everything (Data points, Mean, Standard deviation) scales accordingly

=> Sum of two different ~~ra~~ random variable,
Mean of

$$\boxed{E(X+Y) = \mu_{X+Y} = \mu_X + \mu_Y}$$

08 Tue Mar

=> Diff of ~~to~~ mean of two diff random variables

$$\boxed{E(Y-X) = \mu_{Y-X} = \mu_Y - \mu_X}$$

-> Variances of random variables

$$Var(X+Y) = Var(X) + Var(Y)$$

$$Var(X-Y) = Var(X) + Var(Y)$$

X & Y are independant

Variance remains same whether the two variables are being added / subtracted

$\therefore \sigma^2_{-Y} = Var(-Y) = E\left((-Y - E(-Y))^2\right)$

2016 Wed Mar 09

$$E\left(\left(-Y - E(-Y)\right)^2\right) = E\left((-1)^2\left(Y + E(-Y)\right)^2\right)$$

$$E\left(\left(Y + E(Y)\right)^2\right) = \sigma_Y^2$$

$$\boxed{E(-Y) = -E(Y)}$$

- Even when we subtract two random variables, we still add their variances; subtracting two variables increases the overall variability in the outcomes.

→ Standard Deviation.

$$\boxed{S_{(X+Y)} = \sqrt{\sigma^2_{(X+Y)}}}$$

- We can't add two different SD of random variables directly

Binomial Variables

11 Fri Mar 2016

Type of discrete random variable, which satisfies the following:

i) Made up of independent trials
ii) Each ~~con~~ trial can be classified as either ~~two~~ success (or) failure
iii) Fixed # of trials
iv) Probability of success on each trial is constant

→ 10% rule of independance

if ~~10%~~ sample ≤ 10% of population, then we can assume approximate independence

Guru shows by example how to surrender to God.

12 Sat Mar

Binomial Distribution:-

is calculated by multiplying the probability of success raised to the power of the number of successes & the probability of failure raised to the power of diff b/w number of trials & number of successes.

$$\boxed{P_x = \binom{n}{x} p^x q^{n-x}}$$

$x \rightarrow$ # of times for a specific outcome within trials

$P_x \rightarrow$ Binomial probability

$n \rightarrow$ # of trials

$\binom{n}{x} \rightarrow$ # of combinations

$p \rightarrow$ Probability of success on a single trial

$q \rightarrow$ Probability of failure on a single trial $[q = 1-p]$

→ Binomial distribution with parameters n & p is the discrete probability ~~function~~ distribution of the no. of successes in a sequence of n-independent experiments.

2016 Sun Mar 13

→ binom pdf :- Binomial Probability Distribution Function

binompdf(n,p,x): Give the probability that exactly x successes occur during n trials where the probability of success on a given trial is equal to p

→ [Summation of each binom pdf for single value]

→ binom cdf :- Binomial Cumulative Distribution Function

binom(n,p,x) : Gives the probability the x successes (or) fewer occur during n trials where the probability of success on a given trial is equal to p

Expected Value of ~~Random~~ ~~Binomial~~ Binomial Variable. $\boxed{E(X) = np}$

n → trials
p → probability of success

Mon Mar 14

*] Bernouli Distribution :-

is a special case of Binomial distribution where a single trial is conducted. [∴ n = 1]

Expected value → $\boxed{\text{mean}, = p}$

p → probability of success
(1-p) → probability of failure

Variance, $\boxed{\sigma^2 = p(1-p)}$

Standard Deviation, $\boxed{\sigma = \sqrt{p(1-p)}}$

→ Expected Value of binomial variable

$$\boxed{E(X) = np}$$

$n \rightarrow$ No. of trials $\quad p \rightarrow$ Probability of success

15 Tue Mar 2016

→ Variance of binomial variable,

$$\boxed{\sigma^2 = np(1-p)}$$

Sacrifice pleasures, earnings, and freedom for parents who sacrificed them for you.

16 Wed Mar

Geometric Random Variables

Random Variables which doesn't have the finite # trials but followes other rules :

- Made up of independent trials
- Trials classified as success/failure.
- Probability of success remains constant.

- Expected Value of Geometric variable,

$$\boxed{E(X) = \frac{1}{p}}$$

Poisson Distribution

it a discrete probability distribution that expresses the probability of the given number of events occuring in a fixed interval of time if these events occur with a known constant mean rate & independently of the time since the last event.

$$\boxed{\frac{\lambda^{k} e^{-\lambda}}{k!}}$$

$\lambda \rightarrow$ Expectation of events

$k \rightarrow$ events

$e \rightarrow$ Euler's number

$\boxed{e = 2.71828\ldots}$

$$\boxed{\lambda = E(x) = Var(x)}$$

Sampling Distribution

Probability distribution of a statistic that

Comes from choosing random samples of a given population

*) A statistic is an unbiased estimator of a parameter when the mean of its sampling distribution is equal to value of parameter

*) Mean of sample proportion

$$\boxed{\mu_{\hat{p}} = \frac{\mu_x}{n} = \frac{\cancel{n}p}{\cancel{n}} = p}$$

$p \rightarrow$ Probability of the event happening in population.

Be gentle with every being that shares this home of God with you.

*) Standard deviation of sample proportion

$$\sigma_{\hat{p}} = \frac{\sigma_x}{n} = \frac{\sqrt{np(1-p)}}{n}$$

$$= \sqrt{\frac{\cancel{n}p(1-p)}{n^{\cancel{2}}}}$$

$$\boxed{\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}}$$

$\rightarrow$ Normal Conditions for sample distribution of sample proportion,

$$\left.\begin{matrix} np \geq 10 \\ \& \\ n(1-p) \geq 10 \end{matrix}\right\} \approx \text{normal}$$

$n \rightarrow$ # of samples

$p \rightarrow$ probability of population proportion

If not, based on p value (closer to 0 (or) 1), the hump will be closer to that value & skewed in opp

Central Limit Theorem

03 Sun Apr 2016

It states that, under appropriate conditions, the distribution of a normalized version of the sample mean converges to standard normal distribution

- This holds even if the original variables themselves are not normally distributed

- Sample sizes must be ≥ 30 for getting a normal distribution for sample means

Sample means - $\uparrow$ Samples $\rightarrow$ more normal $\rightarrow$ $\downarrow$ Standard deviation

Your real birthday is the day you meet your spiritual Guru who removes your ignorance and gives you a fresh start in life.

04 Mon Apr

Variance of sample mean distribution

$$\boxed{\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}}$$

$\sigma^2 \rightarrow$ Variance of population distribution

$n \rightarrow$ # of samples

- Standard error of the mean (or) Standard deviation of mean

$$\boxed{\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}}$$

*]. Sample mean $\mu_{\bar{x}}$ will be the same as population mean

$$\boxed{\mu_{\bar{x}} = \mu_x}$$

∵ of central limit theorem for more sample means

Confidence interval:-

Displays the probability that a parameter will fall b/w a pair of values around the mean

This measures the degree of uncertainty or certainty in a sampling method.

Margin Error:- The degree of error in results received from random sampling surveys.

Larger the margin of error, less confidence one should have on the polls/surveys

- Confidence intervals don't make estimates about upcoming values of sample statistics
- CI don't show the distribution of sample data;
- CI give plausible estimates for population parameters.

→ Changing Confidence levels,
  ↳ ↑ Confidence levels, ↑ margin of error resulting in wider interval
  ↳ ↓ Confidence levels, ↓ Margin of error resulting in narrower interval

07 Thu Apr 2016

*) A larger margin of error produces a wider confidence interval that is more likely to contain the parameter of interest
(Increased confidence)

*) A smaller margin of error produces a narrower confidence interval that is less likely to contain the parameter of interest
(Decreased confidence)

*] Conditions for Valid Confidence Intervals.

① Must be "Random Samples"

Even a meticulously chosen stone patiently suffers countless blows from the sculptor before it becomes a deity. A disciple, before attaining divinity has to suffer similar blows from the Guru.

08 Fri Apr

② "Normal" Conditions MET 

i.e., there should be atleast 10 success [$np \geq 10$]
& 10 failures [$n(1-p) \geq 10$]

③ "Independence" of samples [$n \leq 10\%$] [10 % Rule]

(or) sampling with replacement

sample size is less than 10% of population then the samples are assumed to be independance

⇒ If the samples failed to meet any of these conditions, then the inference on that proportion which is based on either confindence interval (or) significance test gets invalid.

$$\boxed{\text{statistic} \pm z^*\left(\begin{array}{c}\text{Std. deviation}\\ \text{of mean}\end{array}\right)}$$

2016 | Sat Apr 09

This gives the confidence interval of the portion.

Here, statistic $\longrightarrow$ Mean (or) Sample proportion

z^* $\longrightarrow$ Critical Value

$\left(\begin{array}{c}\text{Std. deviation}\\ \text{of mean}\end{array}\right)$ $\longrightarrow$ Margin of Error. (or) Std. Error

* The confidence interval is exactly half from the mean on both sides of mean

* ↑ sample size, ↓ Margin of Error

Smaller sample size will have higher MOE than higher sample size

t-statistics

We use t-table in place of z-table

Conditions for inference on a mean with t-values

① Random Samples
② Sampling distribution of $\bar{x}$ (or Sample mean) should be approximately Normal.
  ↳ Possible, if population is normal (or)
  ↳ if the sample size is ~~20~~ ≥ 30.

③ Individual Observation need to be "Independent".
  ↳ 10% Rule
  ↳ With replacement method.

Confidence interval $\boxed{\bar{x} \pm t^*\left(\frac{s}{\sqrt{n}}\right)}$ $\boxed{\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}}$

$\bar{x}$ → Mean of the sample
t^* → Critical t value
s → Std. Error / Std. deviation of ~~mean~~ sample
n → # observations / samples

It should Std deviation of population. Since it won't be there sometimes, then use Std. deviation of sample

For finding t^*

→ Degrees of freedom $\boxed{df = n-1}$

→ In table, while looking for df see under the confidence % value that you are looking for

2016 Wed Apr 13

In t-table,

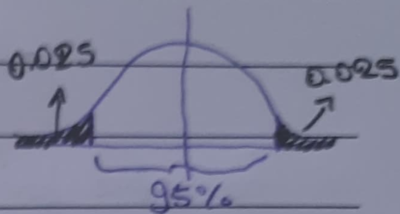
one-tail prob is that the % of space under normal curve left on any side for not considering in confidence %.

two-tail prob is that combined % space under the normal curve left out without considering in confidence %

Ex: for confidence level 95%,

one-tail prob = 0.025

two-tail prob = 0.05



→ Paired Data :- We make two observations on the same individual. When we work with paired data, we're typically interested in the "difference b/w each pair".

Thu Apr 14

*] CI of mean gives us a range of plausible values for the population mean.

*]. If a CI interval does not include a particular value, we can say that it is not likely that the particular value is the true population mean.

*] Even if the particular value is within the interval, we shouldn't conclude that the population means equal that specific value

15 Fri Apr 2016

t-statistic

$$\boxed{t^* = \frac{\bar{x} - \mu}{\left(\frac{S_x}{\sqrt{n}}\right)}}$$

$\bar{x} \rightarrow$ Sample mean

$\mu \rightarrow$ Population mean

$S_x \rightarrow$ Std. deviation of sample

$n \rightarrow$ # samples

→ t-statistic Confidence Interval,

$$-t_{conf} < t < +t_{conf}$$

-ve value for confidence level "conf"

+ve value for confidence level "conf"

Significance Tests

2016

Sun Apr 17

⇒ **Hypothesis test** :-

is a method of statistical inference used to decide whether the data sufficiently support a particular hypothesis.

- It involves a calculation of a test statistic.
- You put your assumptions about a population parameter to the test.
- Used to estimate relationship b/w 2 statistical variables.

~~Null Hypothesis~~ :- We design a significance test to evaluate the strength of the evidence against NULL hypothesis.

While you are alive, preserve your life-energy by doing Pranayama, breath control.

⇒ **Statistical Test** :-

Mon Apr 18

It needs 3 things

i) It needs Data
ii) It needs NULL (or) Primary Hypothesis [H_0]
   [for reject (or) fail to reject]
iii) It needs an Alternative Hypothesis [H_a]
   [Based on this alternative hypothesis, we decide to reject (or) fail to reject the NULL Hypothesis]

→ Alternative hypothesis is opposite to the NULL hypothesis. for 2 groups. For more than 2 groups, there will be more than 1 alternative hypothesis.

→ Here we don't accept any hypothesis, we just decide to reject / fail to reject

⇒ Null hypothesis should always contain a statement of equality.
In other words, it is a statement of "no difference".

19 Tue Apr 2016

$\boxed{H_0 : \text{parameter} = \text{value}}$

⇒ Alternate hypothesis can take

H_a: parameter $\neq$ value
H_a: parameter $<$ value
H_a: parameter $>$ value

→ We'll check how many are far (or) farther than population mean

*] It is always infered to the "population"

*] Rejecting an hypothesis is based on the "significance level" (α) we set during the test. If the observed outcome is less than significance level, then we reject the hypothesis

Make room in your heart for enlightenment. Sweep out hypocrisy, doubt, hate, greed, and anger.

20 Wed Apr

⇒ p-values of significance test

When we can't able to reject the hypothesis, but don't have enough evidance to accept the hypothesis too, then we calculate the "p-values" i.e., probability values for the samples taken.

p-value is a "conditional probability"

$\boxed{p(\text{sample statistic} \mid H_0 \text{ is true})}$

We calculate the p for sample given that NULL hypothesis is True

~~What is the probability of obtaining~~ a sample statistic as extreme (or) more than the one observed by random chance alone

2016

if $p <$ significance value (α) → We reject H_0 & accept H_a

$p \geq$ significance value (α) → We ~~don't~~ fail to reject H_0 but doesn't mean we accept H_0

Thu Apr 21

* Commonly used $\boxed{\alpha = 0.05}$

* p-value composed of 3 parts:

Add them to get p-value
i) Probability random chance would result in observation.
ii) Probability of observing something else that is equally rare
iii) Probability of observing something rare (or) more extreme

Be fearless. Have firm faith that God will protect you.

⇒ Type I & Type II Errors :-

Fri Apr 22

We might be wrong while comparing the p-values with significance value. That's when the errors come into play.

	H_0 true	H_0 false
Reject H_0	Type - I error	Correct Conclusion
Failed to reject H_0	Correct Conclusion	Type - II error

Type I Error:-
Rejecting NULL hypothesis even though it is True
Accepting H_a, when H_0 is True

Type II Error:-
Not Rejecting NULL hypothesis even though it is not True
Accepting H_0, when H_0 is false

⇒ **Power :-**

Probability of avoiding a Type II error.

$$\boxed{P(\text{Rejecting } H_0 \mid H_0 \text{ false})}$$

$= 1 - P(\underbrace{\text{Not rejecting } H_0 \mid H_0 \text{ false}}_{\text{Type II error}})$

$= P(\text{Not making Type-II error})$

→ ↑ α, ↑ Power ↑ probability of Type-I error.

*** → ↑ n, ↑ Power

→ True Parameter far from H_0, ↑ Power

→ Less variability, ↑ Power

*). ↑ Power, lesser will be the probability of making Type-II error

*) ↓ Power, ~~less~~ More will be the probability of making Type II error

*) ↓ α ↓ Probability of Type I error

Summary of Error probabilities & significance values

Significance Value (α)	Probability of Type-I error	Probability of Type-II error	Power
Low	Less	More	Decrease
High	Higher	Less	Increases

*) The significance values are adjusted by looking the dangerousness of errors.
[α will change based on which error is more danger]

⇒ Z-test & significance test (Test about a population proportion)
This is done against "population"

•) Conditions for Z test,
- Random
- Normal [$n\hat{p} \geq 10$ & $n(1-\hat{p}) \geq 10$]
- Independant [with replacement (or) 10% Rule]

$n \rightarrow$ # samples
$p \rightarrow$ Proportion of population

$H_0 : p = \ldots$

In significance test, the standard deviation,

$$\boxed{\sigma = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

27 Wed Apr **2016**

Z-statistics,

$$\boxed{Z = \frac{\hat{p} - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}}}$$

$\sigma_{\hat{p}}$ ← $\sqrt{\frac{P_0(1-P_0)}{n}}$

$\hat{p}$ → Prob of sample

P_0 → Prob of Null hypothesis

n → # samples

·) Finding p-value,

for finding P-value, we use H_a hypothesis, & based on the condition it is decided whether the probability comes from above test statistic / below / ~~above~~ from two-sided.

for
- ≠ → two sided.
- > → above test statistic
- < → below test statistic

Do not let your desires be dictated by addiction or envy.

28 Thu Apr

for two sided,

lower tail

P-value $= 2 \cdot P(Z < \text{value})$

for above statistic:

P-value $= P(Z > \text{value})$

for below statistic

P-value $= P(Z < \text{value})$

⇒ t-test & significance test :-

2016

Fri
Apr **29**

~~Test against the~~

Test about a "population mean"

- Conditions for t-test,

→ Random

→ ~~A~~ Independence [With Replacement sampling (or) 10% Rule]

→ Normal [Parent population NORMAL (or) $n \geq 30$ from Central limit theorem (or) Sample symmetric, with no outliers]

Standard deviation, for population mean,

$$\boxed{\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}}$$

σ → Population standard deviation.

if we don't have population std. dev then we use sample std. dev $S_{\bar{x}}$

$$\boxed{\sigma_{\bar{x}} = \frac{S_{\bar{x}}}{\sqrt{n}}}$$

Listen to the elderly. Amidst their repetitions, some gems may emerge.

t-statistics,

$$\boxed{t = \frac{\bar{x} - \mu}{S_{\bar{x}}/\sqrt{n}}}$$

$\bar{x}$ → Sample mean

μ → Population mean

S_x → Std. deviation of sample

n → # samples

Sat
Apr **30**

- Finding P-value is |||lar to what we did ~~for the~~ with z-statistics. But the slight change here is that we look for the t-value along the data against the degrees of freedom ($df = n-1$)

Loge (Logarithms) :- Logs isolate the exponents.

- It can take any number as its base

8	$\log_2(2^3)$	3
7		
6		2
5	$\log_2(2^2)$	1
4	$\log_2(2^1)$	0
3	$\log_2(2^0)$	-1
2		-2
1		-3
0		

$\log_2$ scale

$1 = 2^0$ $\quad \frac{1}{2} = 2^{-1}$

$2 = 2^1$ $\quad \frac{1}{4} = 2^{-2}$

$4 = 2^2$ $\quad 1/8 = 2^{-3}$

$8 = 2^3$

$$\boxed{\log_2(0) = -\text{Infinity}}$$

$$\boxed{\log_x(x) \rightarrow 1}$$

$$\boxed{\log_x(x^n) \rightarrow n}$$

* log scale is used when the data values are huge

* If any dataset becomes "normal" (bell curve) after log is applied then it is said to be "log normally distributed"

$\sin\theta = \frac{opp}{Hyp}$; $\cos\theta = \frac{adj}{Hyp}$; $\tan\theta = \frac{opp}{adj} = \frac{\sin\theta}{\cos\theta}$

pythagoras theorem $\Rightarrow Hyp^2 = opp^2 + Adj^2$

A/B Testing :-

Just like Experimental Study with ~~[illegible]~~ more than one random sample.

Modified Z-score

Individual value

Median

Median Absolute Deviation, $\dfrac{\sum_{i=1}^{n} |x_i - M(x)|}{n}$

$n \rightarrow$ # of values

We have mean AD there we use mean

Modified Z score $= 0.6745 * \dfrac{x_i - M(x)}{\text{Median AD}}$

⇒ Chi - Squared Distribution

$\chi^2_{\#}$ → Chi Squared Distributed Random

→ Degrees of freedom.

↓ (n-1)

(# of squared samples from a random variable involved in the addition)

The chi squared distribution with degrees of freedom is the distribution of a sum of the squares of independent standard normal random variables

$$\boxed{\chi^2 = \sum_{i=0}^{n} \frac{(O_i - E_i)^2}{E_i}}$$

χ^2 → Chi Square Test Statistic

O_i → Observed frequency

E_i → Expected frequency

degrees of freedom = (n - 1)

How to look at chi squared distribution table,

We look for the probability value against our degrees of freedom & look for specific significance level. (or) for the calculated value, by doing which, we get the p-value

Once we get the value, we compare it with the calculated value,

if table value > Calculated value
we reject H_0

table value < Calculated value
we fail to reject H_0

if $p > \alpha$
fail to reject

$p < \alpha$
reject

Conditions for Chi-Squared Distribution

03 Tue May | 2016

- Random
- Large Counts i.e., Expected frequency ≥ 5
- Independance

For Contingency table,

degrees of freedom = (rows - 1) * (columns - 1)

$$\text{Expected Count} = \frac{(\text{row total})(\text{Column total})}{\text{Table total}}$$

When you hear nonsense, flush it with forgetfulness.

04 Wed May

Conditions for inference on population regression line

[LINER]

Linear :- Actual linear relationship b/w x & y

Independence :- 10% Rule (or) Sampling with replacement

Normal :- Normal distribution

Equal Variance :- Each of distribution must have equal variance / std deviation

Random

2016 ANOVA

Thu May 05

- Analysis Of Variance

* "Grand mean ($\bar{\bar{X}}$)" is the "mean of entire data"
* Grand mean is same as "mean of the means of individual groups"

→ Total Sum of Squares,

$$\boxed{SST = \sum_{i=1}^{n} (x_i - \bar{\bar{x}})^2} \quad \bar{\bar{X}} \rightarrow \text{Grand Mean}$$

→ Degrees Of Freedom, $\boxed{[(\#\,\text{Rows}) * (\#\,\text{Columns})] - 1}$

No one watches you with as much concern as your Guru.

Fri May 06

* ANOVA is a statistical test used to analyse the difference b/w mean of more than two groups

↳ One-way ANOVA uses one independent variable

↳ Two way ANOVA uses two independent variable

→ Sum of Squares Within,

07 Sat May 2016

SSW = Sum of squared difference b/w a value & its sample mean for all values

$\boxed{\text{degree of freedom} = \#\text{Columns} * (\#\text{Rows} - 1)}$

→ Sum of Squares Between,

SSB = Sum of squares difference b/w a value & the grand mean, the mean of all values regardless of sample, for all values

$$\boxed{\sum_{i=1}^{n} (\overline{x}_i - \overline{\overline{x}})^2}$$

$\overline{x}_i$ → Mean of sample i

$\overline{\overline{x}}$ → Grand mean.

The world's delicacies give indigestion. Guru's teachings energize your soul.

$\boxed{\text{degrees of freedom} = \#\text{Columns} - 1}$

08 Sun May

*].

$$\boxed{\begin{aligned} SST &= SSW + SSB \\ df_{SST} &= df_{SSW} + df_{SSB} \end{aligned}}$$

*]. F-statistic

$$\boxed{F\text{-statistic} = \frac{SSB / df_{SSB}}{SSW / df_{SSW}}}$$

if F-statistic is,

>1, More Variations b/w samples

<1, More Variations within sample

Critical F :- F_c is found using F-distribution with specified significance (α) value

if F-statistic < Critical F-value

↳ Fail to reject Null hypothesis

F-statistic > Critical F-value

↳ Reject Null hypothesis & accept the alternate hypothesis