ALL ABOUT STATISTICS

Lets denote Population as (N) and sample (n)

1] What all are the sampling techniques, describe about it?

→ a] Simple random sampling :- Every member of a Population (N) has an equal chance of being selected for your sample (n)
eg :- Exit Poll, survey

b] Stratified Sampling :-

Strata → Layers → Clusters → groups

making a groups or clusters as per the requirements.

eg :- Gender → Male, Female ; Education degree → High school, degree, master, Phd

c] Systematic sampling :- Select every n^{th} individual out of Population (N) eg :- every 5^{th} individual selection for credit card selling.

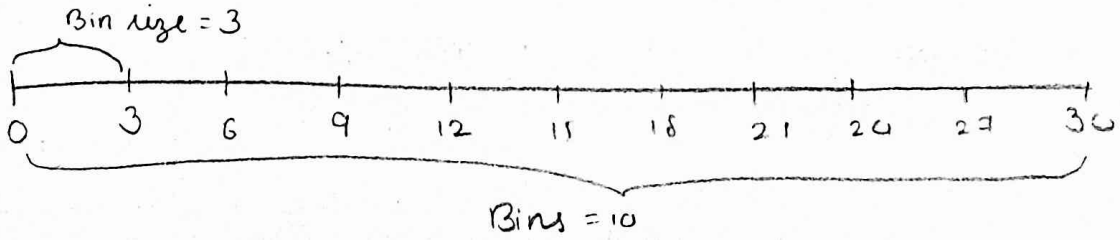
d] Convience sampling :- Only those who are interested in the survey will only Participate.

2] Construction of Histogram ?

→ a] sort the numbers

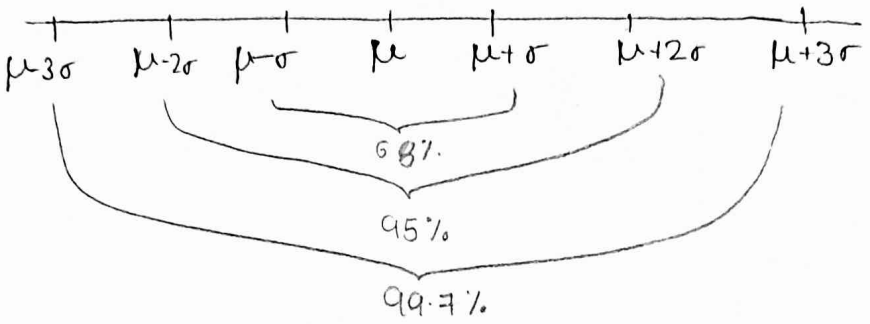
b] Bins → No. of groups = 10 (from diagram)

c] Bin size → $\frac{\text{Max} - \text{Min}}{\text{Bins}}$ (if not starting two.) or $\frac{\text{Max}}{\text{Bins}}$ (if starts from 0) = 3 (from diagram)



3] Empirical Rule of Normal distribution

→



1st Standard deviation = 68%.

2nd standard deviation = 95%.

3rd standard deviation = 99.7%.

4] From Gaussian distribution → Standard Normal distribution

$X \approx$ Gaussian distribution (μ, σ) → $Y \approx$ standard Normal distribution $(\mu = 0, \sigma = 1)$

$$Z\text{-score} = \frac{x_i - \mu}{\sigma/\sqrt{n}}$$ where $\frac{\sigma}{\sqrt{n}}$ = standard error, n = sample size

$$= \frac{x_i - \mu}{\sigma}$$ where $n = 1$

5] What is p-value?

→ <u>P-value</u> are used to make a decision about a hypothesis test. P-value is the minimum significant level at which you can reject the null hypothesis. The lower the p-value, the more likes you reject the null hypothesis.

6] How can we relate standard deviation and variance?

→ Standard deviation refers to the spread of your data from the mean. Variance is the average degree to which each point

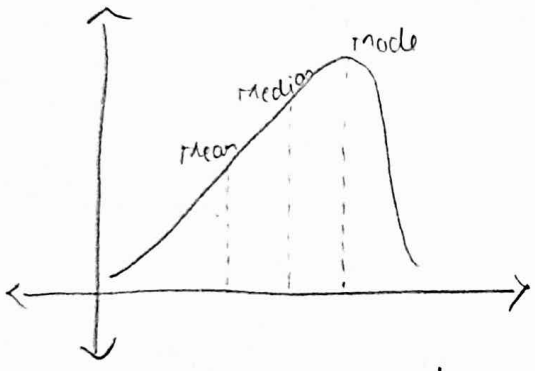
differ from the mean i.e the average of all data points
We can relate standard deviation and variance because it is the square root of variance

4] What is central limit theorem?

→ The central limit theorem states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement, then the distribution of sample means will be approximately normally distributed.

5] Relation between Mean, Median, Mode?

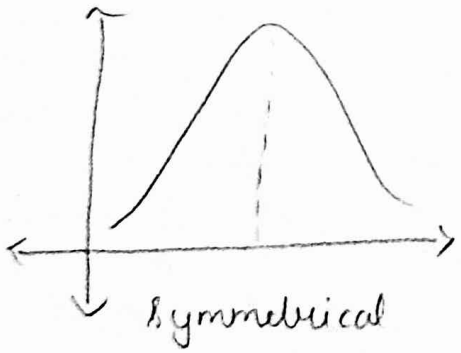
→



Negatively skewed.

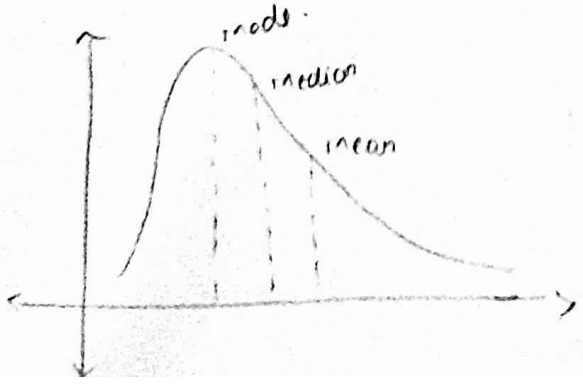
Mean < Median < Mode.

due to large number of skewness in left hand side mean is larger on it



Symmetrical

Mean = Median = Mode

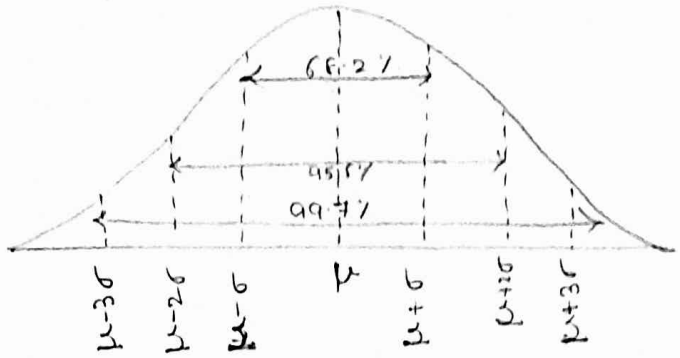


Mean > Median > Mode.

It is rightly skewed data in Right hand side large number of data is found so mean is high on it

9] What is gaussian distribution ?

→



10] What is difference between standardization and Normalization?

→ In standardization :- $\mu = 0, \sigma = 1$ by z-score = $\frac{x_i - \mu}{\sigma}$

In Normalization :- [lower scale, Higher scale] is defined by user

min max scaler : $\frac{x - x_{min}}{x_{max} - x_{min}}$

eg [0, 1] → its defined by user.

11] What is KDE Plot?

→ a] A Non- Parametric way of estimation to get Probability density function

b] Basically used for data smoothing

c] Non - Parametric – Less restriction, Less assumption

Mathematically :-

$$\hat{f}(x,h) = \frac{1}{nh} \sum_{i=1}^{n} K\{(x - x_i)/h$$

$x_i = \{x_1, x_2, \ldots x_n\}$

K = kernel function. (Non- Negativity, integrates to 1)

h = smoothing Parameter (bandwidth)

d] Histogram itself a non-parametric But KDE and Histogram is almost similar. But KDE is more interpretable, regressive, stability to show distribution of data

12] Why sample variance is divided by $1/n-1$?

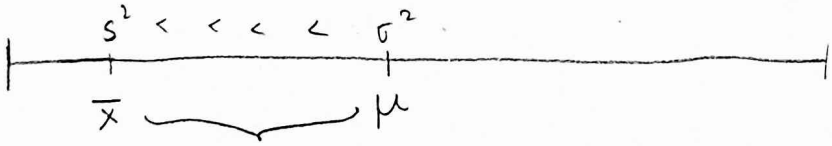
$$S^2 = \frac{\sum_{i=1}^{n} (x - \bar{x})^2}{n-1}$$

→ a] $S^2 = \frac{\sum_{i=1}^{n} (x - \bar{x})}{n}$ < b] $S^2 = \frac{\sum (x - \bar{x})}{n-1}$

↓ since n is large, s^2 is small

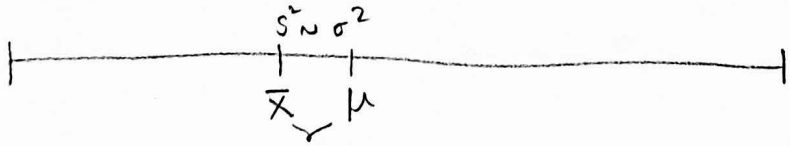
↓ since n-1 is small comparatively, s^2 is large

b] while doing estimation in both the formulas with Population mean, sample mean, Population variance, sample variance.



distance will be large if we apply $\frac{\sum (x - \bar{x})}{n}$

But



distance will be almost same if we apply $\frac{\sum (x - \bar{x})}{n-1}$

13] Why to use Z-score?

→ To bring the feature in a ~~same~~ same scale we use Z-score formula

$$Z\text{-score} = \frac{x_i - \mu}{\sigma/\sqrt{n}} \quad \text{where } \frac{\sigma}{\sqrt{n}} = \text{standard error.}$$

$$= \frac{x_i - \mu}{\sigma} \quad \text{where } n = 1$$

14] What is Pearson Correlation Coefficient and Spearman Rank Correlation Coefficient?

→ a] Pearson Correlation coefficient

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

i] Strength :- How strongly it is correlated.

ii] direction of Relationship

iii] $-1 \leq \rho \leq 1$

b] Spearman Rank Correlation coefficient?

$$r_s = \frac{Cov[R(x), R(y)]}{\sigma[R(x)] \times \sigma[R(y)]}$$

$R(x)$ = Rank of X

$R(y)$ = Rank of Y

a] Non-linear data.

b] Actually correlation between x and y

Its more efficient than Pearson correlation coefficient

15] What is the difference between a Parameter and hyperparameter?

→ Model parameter are estimated from data automatically and model hyperparameter are set manually and are used in process to help estimate model parameters.

16] What is the difference between Estimation, Estimator and Estimate

→ Estimation – Process.
Estimator – Formula
Estimate – Final numerical value

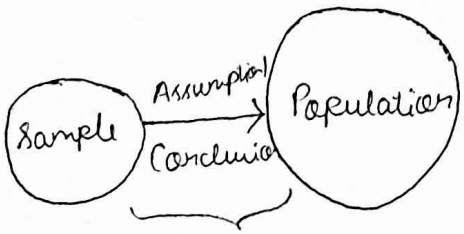
Estimation :- A Process in which we obtain the values of unknown population parameters with the help of sample data.

Estimator :- It is a rule, formula or function that tells how to calculate an estimate i.e to estimate the value of a Population parameter.

Estimate – An estimate is the numeric value of the estimator.

17] What is Hypothesis Testing?

→



The Procedure is called hypothesis testing.

18] What is Confidence interval, significance value, p-value?

→ a] Confidence Interval :- It is the probability that a population parameter will fall between a set of values for a certain proportion of times.

for eg C.I = 95%.

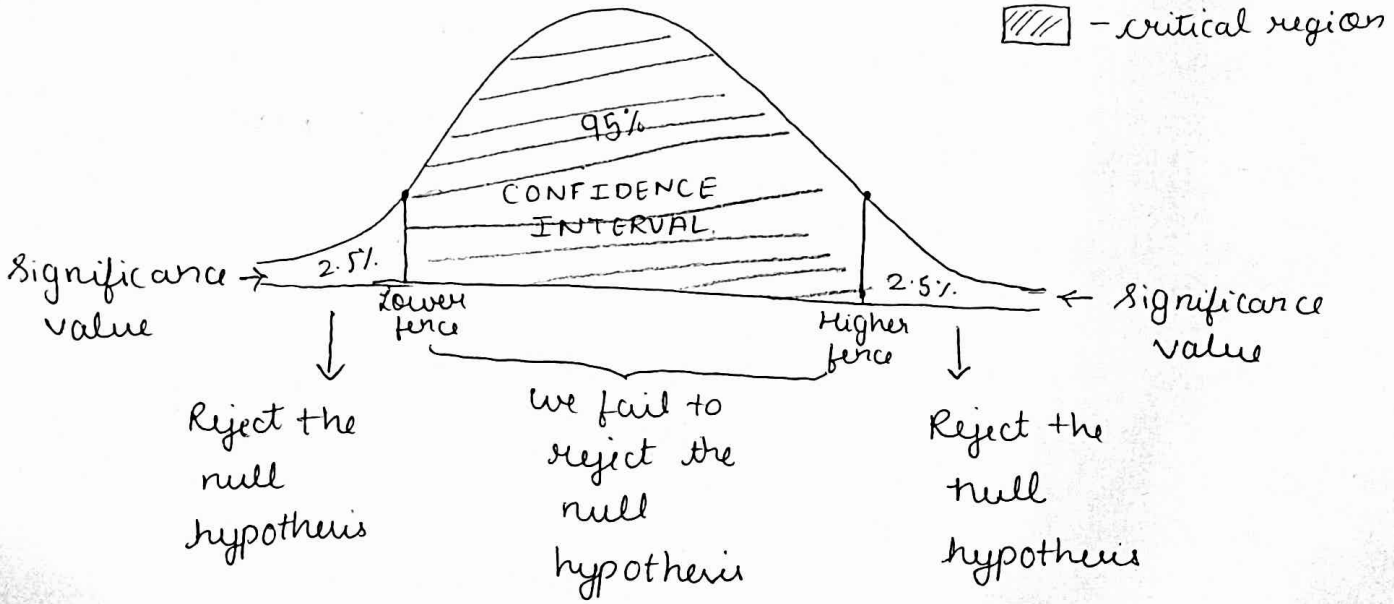
b] Significance level :- S.V = 1 - C.I = 1 - 0.95 = 0.05

Confidence interval and significance level are done by the domain expert.

c] P-value :- Tells us how likely it is that your data could have occurred under the null hypothesis.

eg :- $P \leq \alpha$ Reject Null hypothesis

$P > \alpha$ fail to reject null hypothesis.



19] Hypothesis testing Z-test ?

→ a] It is used when ($n \geq 30$)

b] To determine the difference between the Population mean (μ) and sample mean ($\bar{x}$)

c] Population S.D (σ) is known.

d] Used to compare the mean of two samples.

e] One sample : $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

two sample : $z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

20] Hypothesis testing t-test ?

→ a] It is used when ($n < 30$)

b] To determine the difference between sample mean ($\bar{x}$) and Population mean (μ)

c] Population S.D (σ) is not known

d] Used to compare the mean of two samples.

e] One sample $t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$

two samples $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$