LECTURE - 37

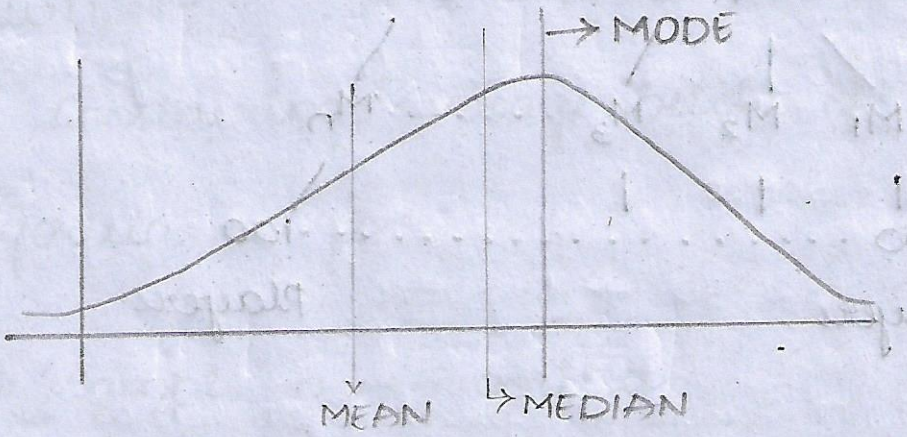
MEAN, MEDIAN AND MODE GETS SHIFTED IN A DISTRIBUTION:

If we have large values, the mean gets attracted to those large values.

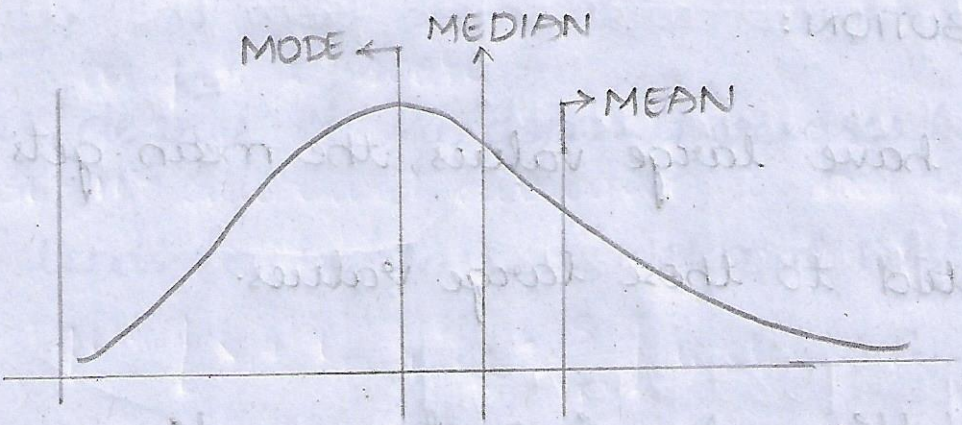
→ the outliers has impact on median.

→ so, the mean will be shifted towards right because of the outliers.

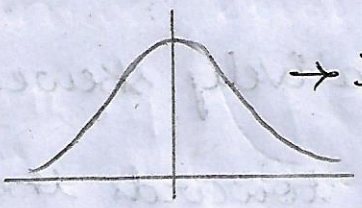
*If the distribution is negatively skewed, the mean will be shifted towards the left.



* If the distribution is positively skewed, the mean will be shifted towards the right.

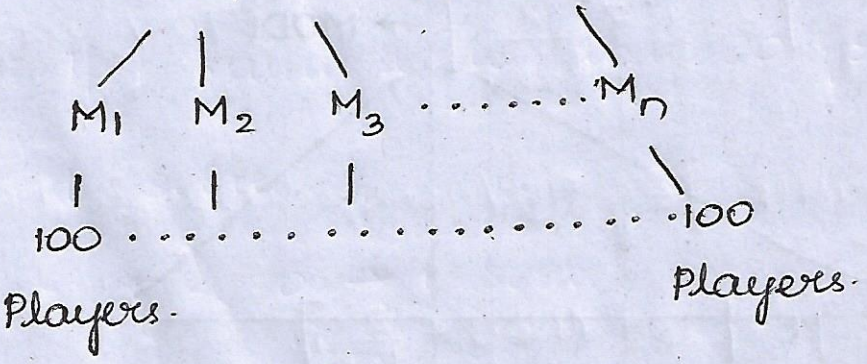


Then the relationship is mean < median < mode



→ In this case, mean ; median and mode are equal.

QUESTION: PUBG TOURNAMENT



SOLUTION:

⇒ In average, the damage is 800, which is mean damage.

⇒ Variance (σ^2) = 400

⇒ $\sigma = 20$.

⇒ $z = 1.20$

⇒ $\frac{D - \mu}{\sigma} = 1.20$

⇒ $\frac{D - 800}{20} = 1.20$

⇒ $D = 1.20 \times 20 + 800$

⇒ $D = 824$

QUESTION: ~~AVERAGE~~ HEIGHT OF ALL GORILLAS.

SOLUTION:

Considering average height of all gorillas as μ.

1. Sample $n \Rightarrow 5$ gorillas.
2. $\bar{x}$

3.

CASE-1: Standard deviation is given.

$$\mu = \left[\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}\right] \text{ with } \_\_\% \text{ confidence}$$

as per z^* score

CASE-11: Standard deviation is not given

$$\mu = \left[\bar{x} \pm t_{n-1, \alpha/2} \cdot \frac{s}{\sqrt{n}}\right] \text{ with } \_\_\% \text{ confidence as per t score}$$

where $\sigma/\sqrt{n}$ is the standard error of standard ~~error~~ sampling distribution.

→ z^* and t-score depends on sample size.

For sample size n:

CASE-1: If 'n' value is less than 30, go with t-score

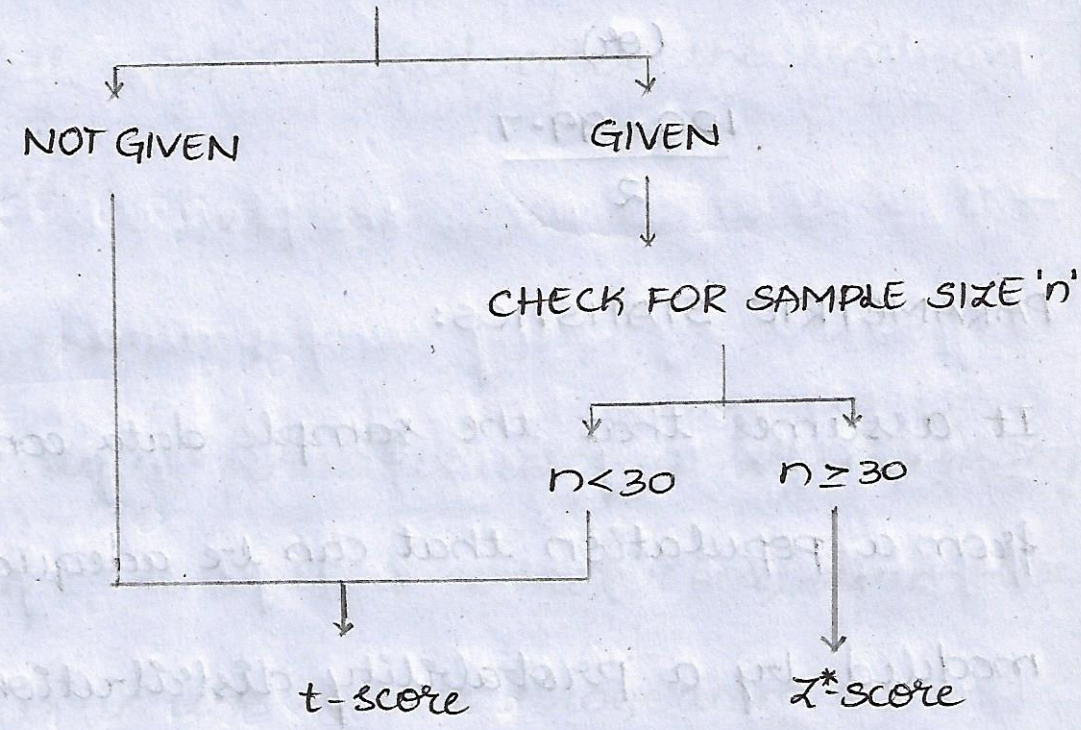
$\Rightarrow$ $n < 30 \rightarrow$ t-score.

CASE-11: If 'n' value is greater or equal to 30 go with z^*-score.

$\Rightarrow n > 30$ or $n \geq 30 \rightarrow z^*$-score

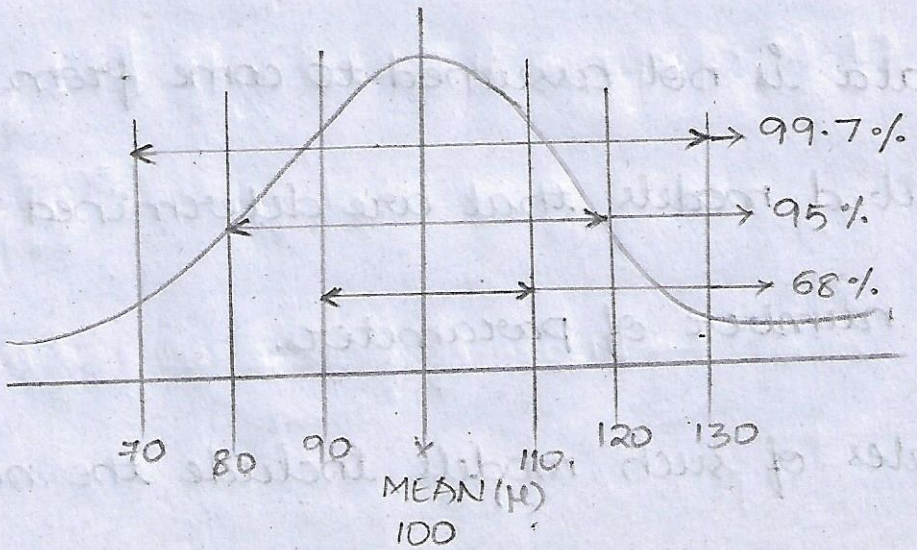
↳ LOOKS LIKE A NORMAL DISTRIBUTION

Check σ value:



QUESTION: $X \sim N(100, 10)$ what is the probability of $X > 130 \rightarrow P(X > 130)$?

SOLUTION:



∴ The probability of $\hat{x}$ is 99.7%.

$\Rightarrow P(X > 130) \rightarrow 99.7\%$

The rest probability is $100\% - 99.7\% = 0.03\%$

(or)

$$\frac{100 - 99.7}{2}$$

PARAMETRIC STATISTICS:

It assumes that the sample data comes from a population that can be adequately modeled by a probability distribution that has a fixed set of parameters.

$$N(\mu, \sigma)$$

NON-PARAMETRIC STATISTICS:

The data is not assumed to come from a prescribed models that are determined by a small number of parameters.

Examples of such models include the normal

distribution model and the linear regression model.

* The NON-parametric statistics are very powerful and invented after the modern computers.

BOOTSTRAPPING:

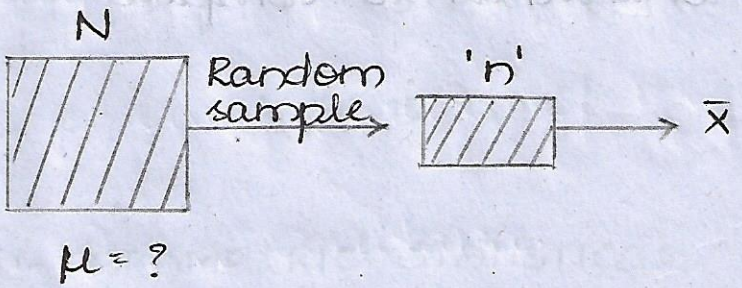
A technique for producing a self-compiling compiler - that is, a compiler (or) assembler written in the source programming language that it intends to compile.

HYPOTHESIS TESTING:

A testing where an analyst tests an assumption regarding a population parameter.

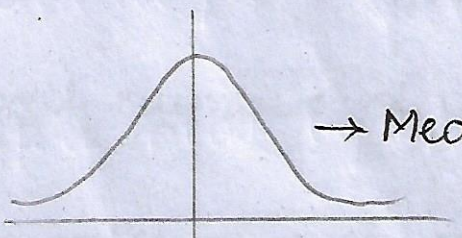
→ It is used to assess the plausibility of a hypothesis by using sample data.

→ Such data may come from a larger population, or from a data-generating process.



From CI (Confidence Interval), $\mu = [\bar{x} \pm \_\_]$

comes from CLT



→ Mean of entire distribution

$\mu_{\bar{x}} \approx \mu$ (roughly equal to)

It defeates CLT → When σ is given

CLAIM: It is about the entire population looking at a very small sample.

Is the claim correct or not?

→ We are going to test the claims of the population is called as hypothesis testing.

The Hypothesis is classified into 2 types.

They are

1. NULL HYPOTHESIS: STATUS QUO.

The statement is assumed to be True.

It suggests that no statistical relationship and significance exists in a set of given single observed variable, between two sets of observed data and measured phenomena.

→ Represented by H_0

→ It is often an initial claim that is

based on previous analyses or specialized knowledge.

→ Also states that a population parameter such as the mean, standard deviation, etc is equal to hypothesized value.

2. ALTERNATIVE HYPOTHESIS: BOLD CLAIM

The Contradictory statement.

It is one in which a difference (or an effect) between two or more variables is anticipated by the researchers; that is, the observed pattern of the data is not due to a chance occurance.

→ The concept of alternative hypothesis is a central part of formal hypothesis testing.

→ Represented by H_1

→ The alternative hypothesis is what we might believe to be true or hope to prove true.

→ Also states that the population parameter is smaller, greater or different than the hypothesized value in the Null hypothesis.

The Hypothesis Tests are classified according to the form of the alternative hypothesis in the following way.

→ If H_1 has the form of $\mu \neq \mu_0$, then, the test is called as two-tailed test.

→ If H_1 has the form of $\mu < \mu_0$, then the test is called as a left-tailed test.

→ If H_1 has the form of $\mu > \mu_0$, then the test is called as a right-tailed test.

ONE-SIDED HYPOTHESIS:

→ Also called as Directional Hypothesis.

→ Used to determine whether the population parameter differs from the hypothesized value in a specified direction.

→ It can specify the direction to be either greater than or less than the hypothesized value.

→ A one-sided test has greater power than a two-sided test, but it can't detect whether the population parameter differs in the opposite direction.

TWO-SIDED HYPOTHESIS:

→ Also called as Non-Directional Hypothesis.

→ Used to determine whether the population parameter is greater than or less than the

hypothesized value.

→ A two-sided test can detect when the population parameter differs in either direction, but has less power than a one-sided hypothesis test.

Z-SCORE:

A numerical measurement that describes a value's relationship to the mean of a group of values.

→ Measured in terms of standard deviations from the mean.

→ If a Z-score is '0', it indicates that the data point's score is identical to the mean score.

**

→ Z-score is calculated when the standard deviation is given.

WHY DO WE USE Z-SCORE?

→ The Z-score (or) standard score is a very useful statistic because it

↳ Allows us to calculate the probability of a score occurring within our normal distribution.

↳ Enables us to compare two scores that are from different normal distributions.

t-SCORE:

A t-score (or) t-statistic is the result of applying a T-test and it represents a point from a STUDENT'S T-DISTRIBUTION with n-k degrees of freedom where 'n' is the total sample size and 'k' is the number of test groups in a A/B test.

WHY DO WE USE t-SCORE?

→ The t-distribution is a probability distribution that is used to estimate population parameters where the sample size is small and/or when the population variance is unknown.

P-VALUE:

The p-value or calculated probability of finding the observed, or more extreme results when the null hypothesis (H_0) of a study question is true.

→ The definition of 'extreme' depends on how the hypothesis is being tested.

CONFIDENCE INTERVAL:

It refers to the probability that a population parameter will fall between a

set of values for a certain proportion of times.

SIGNIFICANCE LEVEL (α):

The probability of rejecting the null hypothesis when it is True.

For Example:

→ α of 0.05 indicates 5% risk of concluding that a difference exists when there is no actual difference.

CRITICAL VALUE:

It is a point on the test distribution that is compared to the test statistic to determine whether to reject the Null hypothesis.

→ If the absolute value of your test statistic is greater than the critical value,

we can declare statistical significance and reject the null hypothesis.

ONE-TAILED TEST:

A statistical test in which the critical area of a distribution is one-sided so that the it is either greater than or less than a certain value, but not both.

→ If the sample being tested falls into the one-sided critical area, the alternative hypothesis will be accepted instead of the null hypothesis

TWO-TAILED TEST:

A statistical test method in which the critical area of a distribution is two-sided and tests whether a sample is greater than or less than a certain range of values.

→ It is used in null-hypothesis testing and testing for statistical significance.

RIGHT-TAILED TEST:

A test where the rejection region is located to the extreme right of the distribution.

→ It is conducted when the alternative hypothesis contains the condition of greater than a given quantity(x)

i.e., $H_1 > x$

→ The upper tail is stated as "right tail" since the highest numbers will appear on the right. It is because, the highest values on a number line are to the right.

LEFT - TAILED TEST:

A test where the rejection region is located to the extreme left of the distribution.

→ It is conducted when the alternative hypothesis contains the condition of less than a given quantity (x).

i.e., $H_1 < x$.

TYPE II ERROR:

The mistake of failing to reject the null hypothesis when it is false.

TYPE I ERROR:

The mistake of rejecting the null hypothes-is when it is true.

QUESTION: HYPOTHESIS TEST FOR A COIN.

SOLUTION:

COIN < UNBIASED → P(H) = 1/2 - 0.5.
COIN < BIASED TOWARDS HEAD → P(H) > 0.5

STEP-1: Figure out the test statistics

i.e.; flip the coin 5 times.

→ As we can count number of Heads, say denote as x.

STEP-2: Identify the H_0 & H_1.

⇒ H_0 = COIN IS UNBIASED

⇒ H_1 = COIN IS BIASED TOWARDS HEAD.

STEP-3: Compute the p-value.

P-value is written as $P(X=5 \mid H_0)$

| → denotes CONDITIONAL PROBABILITY

Given that, the coin is unbiased, find the probability of getting 5 heads in 5 flips is

→ $\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}$

Exactly 5 heads in 5 flips, then

→ $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{32}$

⇒ $P(X = 5 \mid H_0) = \frac{1}{32}$

⇒ $P(X = 5 \mid H_0) = 0.03 = 3\%$ chance.

∴ There is 3% of chance of getting 5 heads in 5 flips for the coin which is unbiased.

→ Basically we go with 5% p-value.

→ If anything is less than 5%, then we reject the Null Hypothesis (H_0)

⇒ In the above case, H_0 is rejected as the coin is unbiased.

→ Accept the Alternative Hypothesis (H_1).

** 5% is the threshold value and used in medical field around 2%-3%

→ If the p-value is greater than the 5%, then we say that failed to reject the Null hypothesis (H_0).