LECTURE - 34

INFERENTIAL STATISTICS:

It is a way of making inferences about populations based on samples.

→ Given information about a subset of examples, how do we draw conclusions about the full set (including other specific examples in that full set)

→ Inferential / Inferences based on principles of evidences we use sample statistics.

We will discuss,

1. Probability
2. Distributions and
3. Hypothesis testing

CENTRAL LIMIT THEOREM (CLT):

It 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 the sample means will be approximately normally distributed.

STEPS:

The steps used to solve the problem of CLT that are either involving '>', '<' or 'between' are as follows:

1. The information about the mean, population size, standard deviation, sample size and a number that is associated with '>', '<' or two numbers associated with both values for range of 'between' is identified from the problem.

2. A graph with a centre as mean is drawn.

3. The formulae $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ is used to find the z-score.

4. The z-table is referred to find the 'z' value obtained in the previous step.

5.

CASE-i: CLT '>'

Subtract the z-score value from 0.5

CASE-ii: CLT '<'

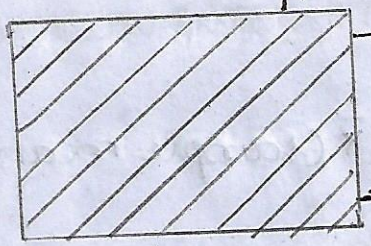
Add 0.5 to the z-score value

CASE-iii: CLT 'BETWEEN'

Step 3. is executed

6. The z-value is found along with $\bar{x}$.

The last step is common to all the three cases, that is to convert the decimal obtained into a percentage.



→ find mean/average salary
→ Entire population of India
1.3 Billion.
→ Inferential statistics → μ_{sal}

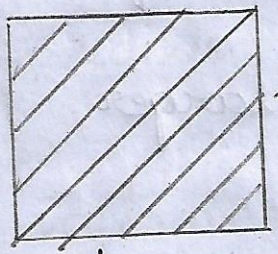
↓

$\mu_{sal} = ?$

Step-1: collect the data from all 1.3 Billion.
(no need to worry about cost or time)

Step-2: $\frac{\sum_{i=1}^{1.3\text{BILLION}} sal_i}{1.3\text{ Billion}} = \mu_{sal}$.

Step-3: Using Inferential stats, cost & time gets reduced.



10K DATA POINTS

$\frac{\sum_{i=1}^{10000} sal_i}{10,000}$

↓

mean of sample = $\bar{x}$

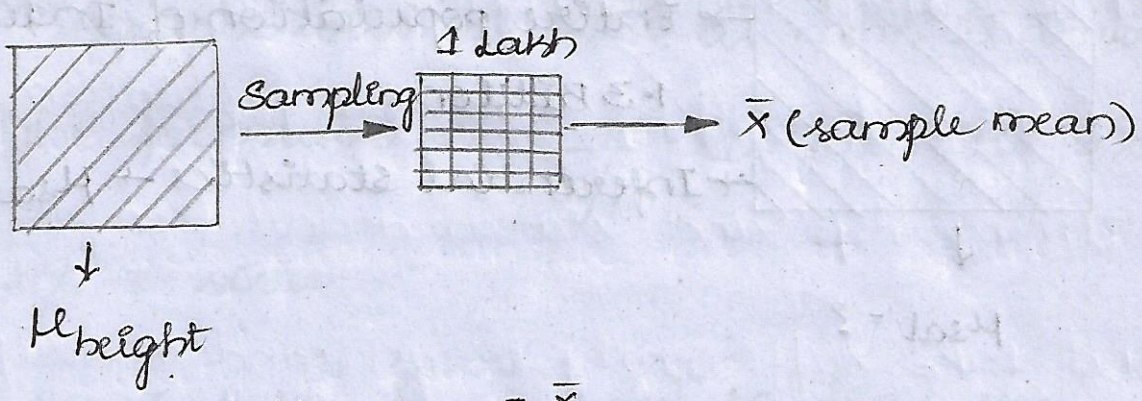
↓

μ_{sal}

μ = mean of population

σ = Standard deviation of population

s = Standard deviation of sample.



$\rightarrow \mu_{height} - \bar{x}$

$\rightarrow \boxed{\mu \approx \bar{x}} \rightarrow$ point estimate.

$\therefore$ Sampling space is biased.

SAMPLING:

It is a process used in statistical analysis in which a predetermined number of observations are taken from a larger population.

The methodology used to sample from a larger population depends on the type of analysis being performed, but it may include simple random sampling or systematic sampling.

SAMPLING TECHNIQUES:

1. CONVENIENT SAMPLING:

It is a type of non-probability sampling in which people are sampled simply because they are "convenient" sources of data for researches.

↳ In probability sampling, each element in the population has a known non-zero chance of being selected through the use of a random selection procedure.

2. VOLUNTEER SAMPLING:

It is a sampling technique where participants self-select to become part of a study because they volunteer when asked, or respond to an advert.

3. RANDOM SAMPLING:

It is a part of the sampling technique in which each sample has an equal probability of being choosen.

→ A sample chosen randomly is meant to be an unbiased representation of the total population.

→ An unbiased random sample is important for drawing conclusions.

4. STRATIFIED SAMPLING:

In a stratified sample, researchers divide a population into homogeneous subpopulations called strata (the plural of stratum) based on specific characteristics like race, gender, location, etc.

→ Every member of a population should be in exactly one stratum.

* Random sampling and stratified sampling are used frequently.

CONFIDENCE INTERVAL:

A confidence interval refers to the probability that a population parameter will fall between a set of values for a certain proportion of times.

→ This measures the degree of uncertainity or certainity in a sampling method.

INTERPRETATION OF CONFIDENCE INTERVAL:

The proper interpretation of a confidence interval is probably the most challenging aspect of this statistical concept.

The most common interpretation of the concept is as follows:

There is a 95% probability that, in the future, the true value of the population parameter (e.g., mean) will fall with in X [lower bound] and Y [upper bound] interval.

↳ In addition,

we may interpret the confidence interval using the statement below:

We are 95% confident that the interval between X[lower bound] and Y[upper bound] contains the true value of the population parameter.

↳ However, it would be inappropriate to state the following:

There is a 95% probability that the interval between X[lower bound] and

y[upper bound] contains the true value of the population parameter.