DAY 03 INTERVIEW QUESTIONS - Types of Sampling Methods, Standard Error

Q1) What do you mean by sample and sampling error? or Variation due to sampling

Ans - Sample → A sample is a selection of objects or observations taken from the population of interest.

Example of sample → A population might be all citizens of India at a given time. We wish to measure Investment by all the citizens.

So now suppose we find sample mean of number of investments.

Batch 1 = 4 (East India), Batch 2 = 6 (West India), Batch 3 = 2 (North India)

Batch 4 = 5 (South India)

Difference in Sample mean (Batches) is called Sampling Error/Variation due to sampling.

→ So whenever we estimate of population based on samples, we should not say equal/give exact values rather we should say lies between.

For ex, number of investment lies between (,) with Confidence Interval

Q2) What is Standard Error? Give an example.

Ans - Standard Error is a measure of uncertainity in sample mean. Higher the standard error, lesser we are confident.

Standard Error, $\boxed{SE(\bar{x}) = \frac{S}{\sqrt{n}}}$

S → standard deviation

n → number of samples

In one term, $\boxed{\text{Standard Error} \Rightarrow \text{Population mean} \neq \text{Sample mean.}}$

Example, we want to know the average age of Fixed Deposit Investment.

Let's take 500 random sample, average mean = 56. Confidence → little bit.

Let's take 5000 random samples, avg mean (age) = 59. Confidence → more than last

Finally, 50,000 random samples, avg mean (age) = 60. Confidence → Good.

So, higher the observation, confidence go up and standard error will decrease.

Therefore after all calculation we can sum up, we are 95% confident that average age of Fixed Deposit investment is in range of (59 to 60) / (59, 60).

Q3) What are Sampling techniques?

Ans - Sampling techniques are mainly grouped in 2 categories - i) Probability sampling
ii) Non probability sampling

i) Probability Sampling (Randomized Sampling)

- It uses randomization to make sure every element of population get equal chances to be part of sample. Also known as Random Sampling.

Types in Probability Sampling - i) Simple random sampling ii) Stratified sampling iii) Reservoir sampling.

i) Simple Random Sampling →
- Every element has an equal chances of getting selected to be part of sample.
- It is used when we don't have any kind of prior information about target population.
- For eg → Random select 20 students out of 50. P(student) = 1/50

ii) Stratified sampling →
- Every element has an equal chance of getting selected, but this technique first divides the element of population into small group or strata based on similarity.
- We need to have prior information about population to create subgroups.
- For eg → strata can be identified such as age, sex, location etc.

iii) Reservoir sampling →
- Reservoir sampling is a randomized algorithm that is used to select K out of n samples where n is very large or unknown. This algorithm select K elements with uniform probability.

2) ~~Probability~~ Non-Probability Sampling (Non-Randomization) →
- Does not rely on randomization. Outcomes may be bias.

Types - i) Convenience Sampling ii) Purposive Sampling iii) Quota Sampling

i) Convenience Sampling → Samples are selected based on availability. It is costly.

ii) Purposive Sampling → Only those elements will be choosen/selected from population which suits best for pupose of study.

iii) Quota Sampling → Elements are selected until exact properties of certain types of data is obtained or sufficient data in different categories is collected. For eg - we need more sample of women than men in a survey.

Q) How to check if the sample is adequate or not?

Ans - Kaiser - Meyer - Olkin (KMO) test measures sampling adequancy for each variable in the model. It is mostly used in Factor Analysis. The statistics is a measure of the proportion of variance among variable might be common variance.

- KMO return values from 0 and 1. We can interpret like

0.8 to 1 indicates the sampling is adequate. (very good for factor analysis)

0.5 to 0.8 Indicates sampling is not adequate, use some remedies.

0 to 0.5 indicates samples are unacceptable. There are widespread of correlations.

Q) What is appropiate sample size of my study? Criteria, Methods.

Ans - Answer depend on number of factors including 1) Purpose of study 2) Population size 3) Risk of selecting bad sample 4) Allowable Sampling error

Sample size criteria - 1) Level of precision 2) Level of Confidence/risk 3) Degree of variability in the attributes being measured

1) Level of precision - Also known as Sampling Error / Margin of error.
- It is the range in which true value of population is estimated to be
- This range is often expressed in percentage (eg ±5%)

Example, if a researcher finds 70% of student in sample has adopted a recommended practise of submitting the assignment with a precision rate of ±5%, then he/she can conclude that between 65% to 75% of students in the population have adopted the practice.

2) Confidence Interval - Also known as Risk level.
- Based on Central Limit Theorem, which means when a population is repeatedly sampled, the average value of attribute obtained by those samples is equal to true population value.
- This is expressed in percentage (eg 95%)

Example, if a 95% confidence level is selected, 95 out of 100 samples will have true population value within range of precision specified earlier.

3) Degree of Variability - Refers to the distribution of attributes in population.
- More hetrogenous (large variance) a population, large sample size required.
- Less variable (more homogenous) less variance, small sample size required.

A proportion of 50% indicates we need large sample size because 50-50. In case 20% or 90% indicates we need small sample size because remaining 80% (in case of 20%) and 90%, large population is on one side (less variance).

Methods used - 1) Cochran formula 2) Yamane formula.

1) Cochran formula, $n_0 = \frac{Z^2 pq}{e^2}$, $n_0 \rightarrow$ Sample size, $Z \rightarrow$ Z value at given confidence interval, $p \rightarrow$ proportion of attribute present in population, $q \rightarrow 1-p$, $e \rightarrow$ desired level of precision

part 1 - Infinite population

Example - Assume there is large population and we don't know variability in population. Therefore we assume, p = 0.5 (maximum variability). Furthermore, we desire a confidence level of 95% and precision of ±5% precision. Resulting sample size is:

$$n_0 = \frac{(1.96)^2 (0.5)(0.5)}{(0.05)^2} = 385$$ sample size

1.96 → z value at 95% confidence interval

$p = 0.5$ $q = 1-p = 0.5$

$e = 5\% = 0.05$

Part 2 - Finite population

$$n = \frac{n_0}{1 + \frac{(n_0 - 1)}{N}}$$

$n_0 \rightarrow$ Initial sample size calculated as per larger population criterion.

$N \rightarrow$ population size.

example - Assume our last example, our evaluation of student adoption of the recommended practise will only affect 5,000 students.

$$n = \frac{385}{1 + \frac{(385-1)}{5000}} = 358 \text{ students}.$$

② Yamane Formula → simplified formula to calculate sample size in case of finite population

$$n = \frac{N}{1 + N(e)^2},$$

n = Sample size.

N = Population size

e = level of precision.

example - Take above question of 5000 students, $n = \frac{5000}{1 + 5000(0.05)^2} = 371$ students.

e = precision = 5% = 0.05