Learning Module 7: Estimation and Inference

LOS 7a: compare and contrast simple random, stratified random, cluster, convenience, and judgmental sampling and their implications for sampling error in an investment problem

Sampling refers to the systematic process of selecting a subset or sample from a larger population. Sampling is essential because it is costly and time-consuming to analyze the whole population.

Sampling methods can be broadly categorized into: probability sampling and non-probability sampling.

In probability sampling, every population member has an equal chance of being chosen for the sample, ensuring a representative sample. In contrast, non-probability sampling depends on factors such as the sampler's judgment or data accessibility, increasing the risk of an unrepresentative sample.

Probability Sampling Methods

Simple Random Sampling

Simple random sampling means selecting a sample from a population where each element has an equal chance of being chosen. This method aims to create an unbiased sample that accurately represents the population.

Simple random sampling is appropriate when applied to a homogeneous population.

Example: Simple Random Sampling

Imagine we wish to come up with a sample of 50 CFA level I candidates out of 100,000.

One approach may involve numbering each of the 100,000 candidates, placing them in a basket, and shaking the basket to jumble up the numbers. Next, we would randomly draw 50 numbers from the basket, one after the other, without replacement.

A more scientific approach may also involve the use of random numbers where all the 100,000 candidates are numbered in a sequence (from 1 to 100,000). We may then use a computer to randomly generate 50 numbers between 1 and 100,000, where a given number represents a particular candidate who can be identified by their name or admission number.

The underlying feature of random sampling is that all elements in the population must have equal chances of being chosen.

Stratified Random Sampling

In stratified random sampling, analysts subdivide the population into separate groups known as strata (singular stratum). Each stratum comprises elements with a common characteristic (attribute) that distinguishes them from all the others. The method is most appropriate for large **heterogeneous** populations.

A simple random sample is then drawn from within each stratum and combined to form the overall, final sample that takes heterogeneity into account. The number of members chosen from any one stratum depends on its size relative to the population as a whole.

Example: Stratified Random Sampling

An advertising firm wants to determine the extent to which it needs to invigorate television advertisements in a district. The company decides to conduct a survey to estimate the mean number of hours households spend watching TV per week. The district has three distinct towns – A, B, which are urbanized, and C, located in a rural area. Town A is adjacent to a major factory where most residents work, with most having kids of school-going age. Town B mainly harbors retirees, while most people in town C practice agriculture.

There are 160 households in town A, 60 in town B, and 80 in C. Given the differences in the composition of each region, the firm decides to draw a sample of 50 households, considering the total number of families in each.

What is the number of homes that have been sampled in each town?

Solution

We have three strata: towns A, B, and C. We use the following formula to determine the number of households from each region to be included in the sample:

$$\text{Number of households in sample} = \left(\frac{\text{Number of households in the region}}{\text{Total number of households}}\right) \times \text{Required sample size}$$

Therefore, the number of households to be sampled in town A,

$$= \frac{160}{300} \times 50 = 27 \,(\text{approximately})$$

Similarly, the number of households to be sampled in town B,

$$= \frac{60}{300} \times 50 = 10$$

Finally, the firm would need $\left(\frac{80}{300} \times 50\right) = 13$ households in town C.

Advantages of Stratified Sampling over Simple Random Sampling

- Stratification is associated with a smaller estimation error compared to simple random sampling, especially when each stratum is homogeneous.

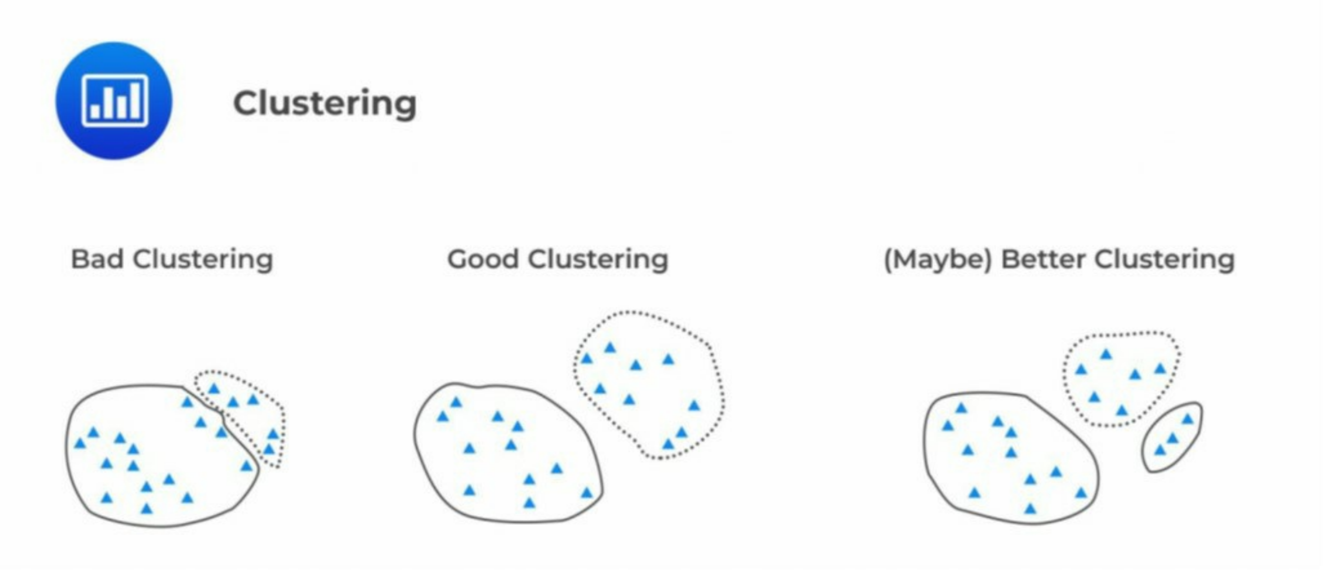
- Stratification enables analysts to estimate the population parameter, say, the mean for all the subgroups of the entire population.

Cluster Sampling

Cluster sampling involves categorizing all population elements into distinct and all-encompassing groups called clusters. Then, you can either choose a random sample of entire clusters or select a random subset from each cluster. So, there are two cluster sampling approaches:

- One-stage (or single-stage) cluster sampling: All the members in each sampled cluster are sampled.

- Two-stage cluster sampling: A simple random sub-sample of members is selected from each cluster.



Key Point Difference Between Stratified Sampling and Cluster Sampling

- In cluster sampling, a cluster serves as a single sampling unit, and only specific clusters are sampled.
- In stratified sampling, you select members from within each stratum and then draw a random sample from each stratum.

Non-Probability Sampling Techniques

Non-probability samples are selected based on judgment or the convenience of accessing data. As such, non-probability sampling depends on the researchers' sample selection skills. There are two types of non-probability sampling methods:

i. **Convenience sampling**: Researchers choose population elements based on ease of access. This method may not provide a fully representative sample, limiting sampling accuracy.
ii. **Judgmental sampling**: Researchers select elements subjectively, often based on their own knowledge and expertise. However, this approach can introduce bias and result in a

non-representative sample.

Judgmental sampling is preferred when a restricted number of people in a population possess qualities that the researcher expects from the target population.

Comparison Between Probability Sampling and Nonprobability Sampling

Method	Strengths	Weaknesses
Probability Sampling		
Simple random sampling	Easy to use	Lower precision; no assurance of representativeness
Stratified sampling	Higher precision relative to simple random sampling	Difficult to choose relevant stratification; expensive
Cluster sampling	Cost-effective and efficient	Lower precision
Non-probability Sampling		
Convenience sampling	Cost-effective and saves time; easy to use	Selection bias, sample may not accurately represent population
Judgmental sampling	Cost-effective, convenient, less time consuming	Subjective method. Selection bias, sample may not accurately represent the population.

Sampling Error and Its Implications on Investment

Sampling error refers to the difference between the observed value (results obtained from analyzing a sample of investment data) and the true values that would have been obtained from analyzing the entire population of investments.

For instance, when we take a sample to estimate a population's mean, there's typically a difference between the sample mean and the true population mean. This difference, known as sampling error, emerges due to natural variation in sampling and because we work with data from only a part of the full population.

Therefore, any conclusions or predictions drawn based on the sample data may deviate from the actual performance or characteristics of the entire investment population.

Question 1

An analyst is analyzing the spending habits of people belonging to different annual income categories. In his analysis, he creates the following different groups according to the annual family income: Less than $30,000, $31,000 – $40,000, $41,000 to $50,000, and $51,000 to $60,000. He then selects a sample from each distinct group to form a whole sample. The sampling method used by the analyst is most likely:

A. Cluster sampling.
B. Stratified sampling.
C. Simple random sampling.

Solution

The correct answer is B.

Dividing the population into different strata/groups and selecting a sample from each group is called the stratified sampling technique.

A is incorrect. In cluster sampling, each cluster is considered a sampling unit, and only selected clusters are sampled.

C is incorrect. Simple random sampling involves the selection of a sample from an entire population such that each member or element of the population has an equal probability of being picked.

Question 2

A Ph.D. student is conducting research related to her thesis, and for this purpose, she uses some students from her university to constitute a sample. The sampling method used by the analyst is *most likely*:

A. Simple random sampling.
B. Convenience sampling.
C. Judgmental sampling.

Solution

The correct answer is B.

The researcher has selected the students from her university because she can conveniently access them.

A is incorrect. Simple random sampling involves the selection of a sample from an entire population such that each member or element of the population has an equal probability of being picked.

C is incorrect. Judgmental sampling involves handpicking elements from a sample based on the researcher's knowledge and expertise.

Question 3

An analyst wants to estimate the downtime of ABC Bank's ATMs in a city for the last six months. For this purpose, he selects 20 locations or areas within the city and then selects 50% of the ATMs in each area. The sampling method used by the analyst is *most likely*:

A. Cluster sampling.
B. Stratified random sampling.
C. Simple random sampling.

Solution

The correct answer is A.

In cluster sampling, all population elements are categorized into mutually exclusive and exhaustive groups called clusters. A simple random sample of the cluster is selected, and then the elements in each of these clusters are sampled.

B is incorrect. In stratified random sampling, analysts subdivide the population into separate groups known as strata (singular–stratum), and each stratum is composed of elements that have a common characteristic (attribute) that distinguishes them from all the others.

C is incorrect. Simple random sampling involves the selection of a sample from an entire population such that each member or element of the population has an equal probability of being picked.

LOS 7b: explain the central limit theorem and its importance for the distribution and standard error of the sample mean

The central limit theorem asserts that "given a population described by any probability distribution having mean μ and finite variance σ^2, the sampling distribution of the sample mean $\bar{X}$ computed from random samples of size n from this population will be approximately normal with mean μ (the population mean) and variance $\frac{\sigma^2}{n}$ (the population variance divided by n) when the sample size n is large".

What is a Large Enough n?

The answer to this question might not be straightforward. Nevertheless, the widely accepted value is $n \geq 30$. The truth is that the value of n depends on the shape of the population involved, i.e., the distribution of X_i and its skewness.

In a non-normal but fairly symmetric distribution, $n = 10$ can be considered large enough. With a very skewed distribution, the value of n can be 50 or even more.

Standard Error of the Sample Mean

Remember that from the central limit theorem, the variance of the sample mean distribution is given by:

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

Where n is the sample size.

The standard error is the standard deviation of the statistic (sample mean).

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Formally defined, for a sample mean $\bar{X}$ computed from a sample generated by a population with standard deviation σ, the standard error of the sample mean is given by:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Where σ = Known population standard deviation.

When the population standard deviation, σ, is unknown, the following formula is used to estimate the standard error of the sample mean, also denoted as $s_{\bar{X}}$:

$$s_{\bar{X}} = \frac{s}{\sqrt{n}}$$

Where: s = Sample standard deviation.

The formula above is applicable where we do not know the population standard deviation. Note that the sample standard deviation is the square root of the sample variance, s^2, given by:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$
$$\Rightarrow s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

The standard error of the sample mean estimates the variation that would occur if you took multiple samples from the same population. While the standard deviation measures variation within one sample, the standard error estimates variation across many samples. So, standard deviation and standard error are distinct concepts.

The standard error of the sample mean gives analysts an idea of how **precisely** the sample mean estimates the population mean. A lower standard error value indicates a more precise estimation of the population mean. On the other hand, a larger standard error value indicates a less precise estimate of the population mean.

It is also important to note that the standard error becomes smaller as the sample size increases. This can be seen from its formula. This happens because increasing the sample size ultimately brings the sample mean closer to the true value of the population mean.

Example 1

In a certain property investment company with an international presence, workers have a mean hourly wage of $12 with a population standard deviation of $3. Given a sample size of 30, the standard error of the sample mean is closest to:

$$\begin{aligned}\sigma_{\bar{X}} &= \frac{\sigma}{\sqrt{n}} \\ &= \frac{3}{\sqrt{30}} = \$0.55\end{aligned}$$

If we were to draw several samples of size 30 from the employee population and construct a sampling distribution of the sample means, we would end up with a mean of $12 and a standard error of $0.55.

Example 2

A sample of 30 latest returns on XYZ stock reveals a mean return of $4 with a sample standard deviation of $0.13. The standard error of the sample mean is closest to:

$$s_{\bar{X}} = \frac{s}{\sqrt{n}}$$
$$\frac{0.13}{\sqrt{30}} = \$0.02$$

If we were to draw more samples from the population of yearly returns on XYZ stock and construct a sample mean distribution, we would end up with a mean of $4 and a standard error of $0.02.

Question

Emma Johnson wants to know how finance analysts performed last year. Johnson assumes that the population cross-sectional standard deviation of finance analyst returns is 8 percent and that the returns are independent across analysts.

The random sample size that Johnson needs if she wants the standard deviation of the sample means to be 2% is *closest to*:

A. 4.
B. 16.
C. 72.

Solution

The correct answer is B.

Remember that,

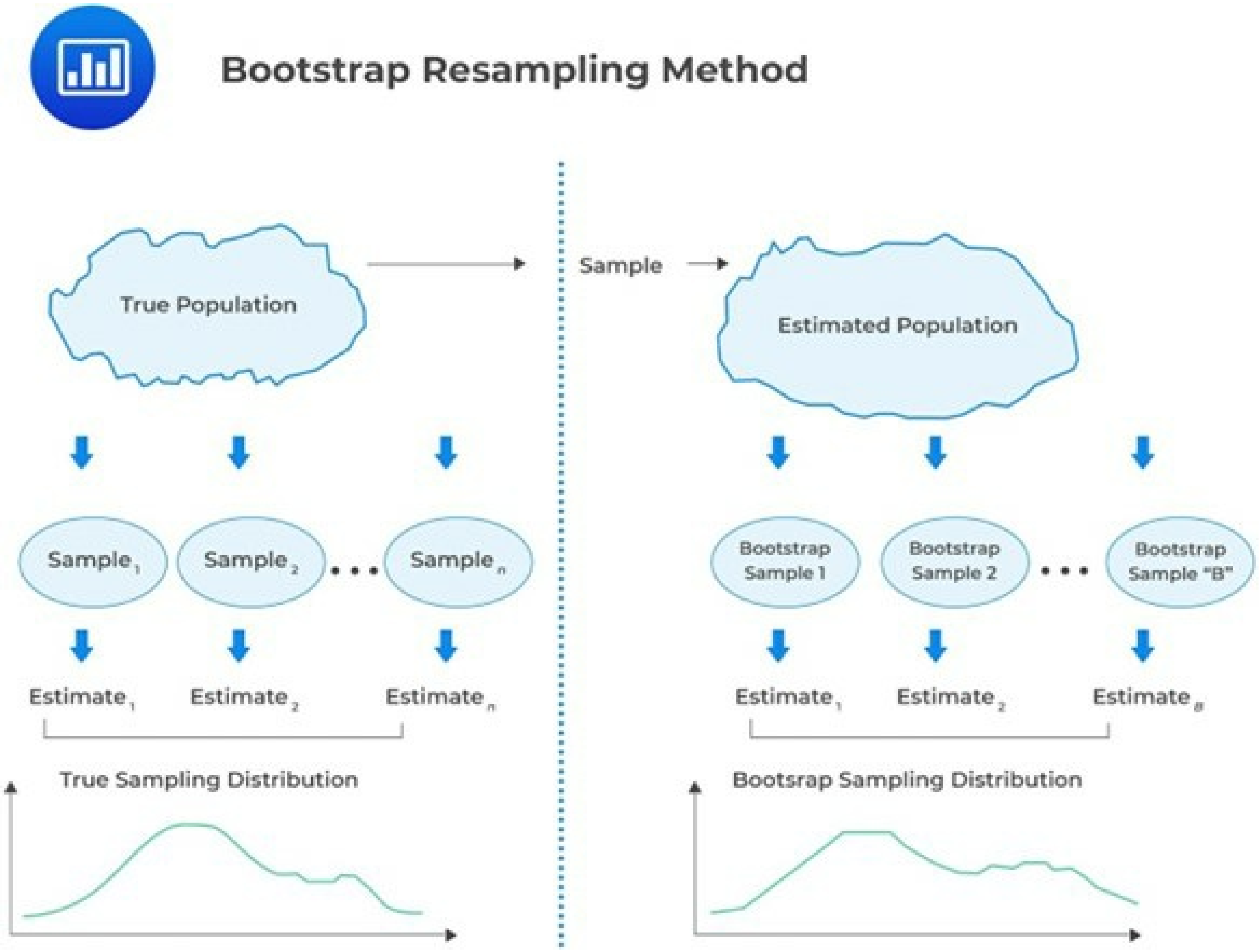
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$
$$\Rightarrow 0.02 = \frac{0.08}{\sqrt{n}}$$
$$\therefore n = 16$$

LOS 7c: describe the use of resampling (bootstrap, jackknife) to estimate the sampling distribution of a statistic

Resampling refers to the act of repeatedly drawing samples from the original observed data sample for the statistical inference of population parameters. The two commonly used methods of resampling are bootstrap and jackknife.

Bootstrap

Using a computer, the bootstrap resampling method simulates drawing multiple random samples from the original sample. Each resample is the same size as the original sample. These resamples are used to create a sampling distribution.



In the bootstrap method, the number of repeated samples to be drawn is at the the researcher's discretion. Note that bootstrap resampling is done with replacement.

Furthermore, we can calculate the standard error of the sample mean. This is done by resampling and calculating the mean of each sample. The following formula is used to estimate the standard error.

$$s_{\bar{X}} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b - \bar{\theta})^2}$$

Where:

$s_{\bar{X}}$ = Estimate of the standard error of the sample mean.

B = Number of resamples drawn from the original sample.

$\hat{\theta}_b$ = Mean of a resample.

$\bar{\theta}$= Mean across all the resample means.

The bootstrap resampling method can also be applied in estimating the confidence intervals for the statistic of other population parameters such as median.

Advantages of Bootstrap Resampling

1. **No Reliance on Analytical Formulas**: Bootstrap differs from traditional statistics because it doesn't rely on an analytical formula for estimating distributions. This makes it versatile for complex estimators and especially useful when analytical formulas are unavailable.
2. **Applicability to Complicated Estimators**: Bootstrap is a simple yet powerful method that can handle complicated estimators effectively. It can handle a wide range of statistical models, making it suitable for various applications in finance where complex estimations are common.
3. **Increased Accuracy**: Bootstrap can enhance accuracy by creating multiple resampled datasets and estimating population parameters on each. This helps understand estimator

variability and robustness, ultimately improving result accuracy.

Jackknife

Jackknife is a resampling method in which samples are drawn by omitting one observation at a time from the original data sample. This process involves drawing samples without replacement. For a sample size of n, we need n repeated samples. This method can be used to reduce the bias of an estimator or to estimate the standard error and the confidence interval of an estimator.

Question

Assume that you are studying the median height of 100 students in a university. You draw a sample of 1000 students and obtain 1000 median heights. The mean across all resample means is 5.8. The sum of squares of the differences between each sample mean, and the mean across all resample means $\sum_{b=1}^{B}(\hat{\theta}_b - \bar{\theta})^2$ is 2.3.

The Estimate of the standard error of the sample mean is *closest to*:

A. 0.05.

B. 0.08.

C. 0.10.

Solution

The correct answer is A.

$$s_{\bar{X}} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b - \bar{\theta})^2}$$

$$= \sqrt{\frac{1}{1000-1} \times 2.3} = 0.04798 \approx 0.05$$