LM07 · QUANTITATIVE METHODS · CFA LEVEL I

# Estimation and Inference

Sampling methods, Central Limit Theorem, standard error, resampling

### Statistical Inference

A **sample** is a method of obtaining information about a **population's parameters** (μ and σ) through **sample statistics** ($\bar{X}$ and S). Since we can rarely observe an entire population, we infer parameters from samples.

## Learning Outcome Statements

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

**LOS b** Explain the central limit theorem and its importance for the distribution and standard error of the sample mean

**LOS c** Describe the use of resampling (bootstrap, jackknife) to estimate the sampling distribution of a statistic

LOS A · PROBABILITY SAMPLING

# Probability Sampling Methods

**Probability sampling:** every member of a population has an *equal chance* of being selected. Samples will be more representative of the population.

| METHOD | HOW IT WORKS | WHEN TO USE |
|---|---|---|
| **Simple Random** | Each element has equal probability of selection. Example: random-number generator selects 50 from a population of 500. | Data are homogeneous |
| **Systematic** | Select every $K^{th}$ element until the desired sample size is reached. | When the population is too large to code |
| **Stratified Random** | Population sub-divided into sub-populations based on one or more classifications; simple random samples drawn from each sub-pop and pooled. Each sub-sample *proportionate* to its sub-population. | Guarantees sub-divisions are represented; statistics more precise |
| **Cluster (1-stage)** | Population divided into clusters (each a mini-representation of the whole). Random sample of clusters is taken as a whole. | Cost and time efficient |
| **Cluster (2-stage)** | Sub-samples are selected from each chosen cluster. | Lowest precision of the probability methods |

### Key Trade-off

Stratified sampling has **smaller sampling error** than cluster sampling when subgroups differ; two-stage cluster has **greater sampling error** than one-stage cluster.

LOS A · NON-PROBABILITY SAMPLING

# Non-Probability Sampling Methods

**Non-probability sampling:** selection depends on factors such as *judgment* or *convenience* (access to data). Risk: samples may be non-representative.

### Convenience Sampling

Observations are selected because they are **easy to obtain or accessible**. Not necessarily representative, but low cost.

### Judgmental Sampling

Selects observations based on the researcher's **experience and knowledge**. Useful when there is a time constraint or when the specialty of the researcher would result in better representation.

**Exam Tip:** All else equal, **probability sampling is more accurate and reliable** than non-probability sampling. Less randomness = greater sampling error.

### Sampling Error

The **difference between the observed value of a statistic and the population parameter**, arising from using only a subset of the population. As **n** ↑, sampling error ↓.

LOS B · CENTRAL LIMIT THEOREM

# The Central Limit Theorem (CLT)

If we draw simple random samples of size n from a population with mean **μ** and *finite* variance **$\sigma^2$**, the **sampling distribution of the sample mean** $\bar{X}$:

- Will be **normally distributed** (irrespective of the original population's distribution)
- Will have a mean of **μ**
- Will have variance **$\sigma^2/n$**

Applies when **n ≥ 30**.

WORKED EXAMPLE

## CLT Sampling Distribution

**GIVEN**

Universe of 10,000 stocks, population mean return μ = 12%, population SD σ = 10%. Keep drawing samples of n = 100 stocks and plot their average returns.

**SOLUTION**

By the CLT, the sampling distribution of the sample mean is **normal** with:

Mean = μ = **12%**

Variance = $\sigma^2/n = 10^2/100$ = **1%**

SD of sampling distribution (standard error) = **1%**.

### Why CLT Matters

The CLT lets us use the **normal distribution** for inference about $\bar{X}$ even when the underlying population is non-normal — provided n is large enough.

CHECK YOUR UNDERSTANDING · QUIZ 1

## Quiz 1

Select the best answer. Feedback appears after each click.

**Q1. Which sampling method divides the population into subgroups based on a characteristic and draws samples from each in proportion to subgroup size?**

**Q2. Which method is MOST likely to produce samples that are non-representative of the population?**

**Q3. Per the Central Limit Theorem, if σ = 10% and n = 100, the standard deviation of the sample-mean distribution is:**

**Q4. Two-stage cluster sampling is expected to have:**

---

LOS B · STANDARD ERROR OF THE SAMPLE MEAN

## Standard Error

The **standard error** of the sample mean is the standard deviation of the sampling distribution of the sample means.

$$When \sigma \; is known: \quad \sigma_{\bar{X}} = \sigma \; / \; \sqrt{n}$$

$$When \sigma \; is unknown: \quad s_{\bar{X}} = s / \sqrt{n}$$

WORKED EXAMPLE **Standard Error Calculation**

GIVEN

Average returns of all large-cap stocks = 10%, σ = 6%. Sample size n = 100.

**SOLUTION**

$\sigma_{\bar{X}} = \sigma / \sqrt{n} = 6 / \sqrt{100} = 6/10 =$ **0.6**

### SD vs SE

| | STANDARD DEVIATION (SD) | STANDARD ERROR (SE) |
|---|---|---|
| Measures | Dispersion from the mean | Sampling error |
| Used for | Data description | Data inference |

**Exam Tip:** $\bar{X}$ is the **best estimate** of μ. The smaller the SE, the more precise the estimate.

LOS C · RESAMPLING METHODS

# Resampling: Bootstrap & Jackknife

**Resampling**: repeatedly draw samples from an original data sample to estimate population parameters or the sampling distribution of a statistic.

### 1. Bootstrap Method

Uses computer simulation. From an observed sample of n = 30 (say), draw many bootstrap samples $B_1$, $B_2$, … $B_{1000}$, each of size n, **with replacement** (draw 1 observation, record, replace; repeat n times).

- Rather than estimate the distribution, this method **creates** the distribution.
- Can find the SE of an estimator even when no analytical formula is available (e.g., the **median**).

```
s_{X̄} = \sqrt{}[ (1/(B-1)) \; \cdot \; \Sigma _{i=1}^{B} (\theta
                    ^_{i} - \theta ¯)^{2} ]
```

### 2. Jackknife Method

Omit **one observation at a time** from the sample (without replacement).

Example: n = 30. $J_1$: n = 29, omit $X_1$. $J_2$: n = 29, omit $X_2$. … $J_{30}$: n = 29, omit $X_{30}$.

Will produce similar results from sample to sample (bootstrap may not).

| | BOOTSTRAP | JACKKNIFE |
|---|---|---|
| Draws | Repeated samples of size n *with replacement* | Leaves out one observation at a time, *without replacement* |
| Use | Complex statistics, no closed form | SE of sample mean; deterministic |

CHECK YOUR UNDERSTANDING · QUIZ 2

## Quiz 2

Select the best answer. Feedback appears after each click.

**Q1. The bootstrap method draws samples from the observed data:**

**Q2. A sample has σ = 8%, n = 64. The standard error of the sample mean is:**

**Q3. The jackknife method constructs resamples by:**

**Q4. According to the CLT, for a sufficiently large n the sampling distribution of $\bar{X}$ is approximately normal:**

**Q5. Bootstrap is PARTICULARLY useful when:**

CHECK YOUR UNDERSTANDING · QUIZ 3

## Quiz 3

Select the best answer. Feedback appears after each click.

**Q1. Which method divides the population into clusters and then randomly selects some clusters in their entirety?**

**Q2. Systematic sampling selects:**

**Q3. The standard error of the sample mean DECREASES when:**

MODULE COMPLETE

# LM07 · Estimation and Inference Summary

**Final Score**

**0 / 0**

## Cheat Sheet

$$SE: \sigma_{\bar{X}} = \sigma/\quad n(\sigma\ known); s_{\bar{X}} = s/\quad n(\sigma\ unknown)$$

$$CLT: \bar{X}\ N(\mu, \sigma^2/n), n\ \geq\ 30$$

$$Stratified: proportional; lowererrorthancluster$$

$$Cluster2 - stage > cluster1 - stagesamplingerror$$

$$Bootstrap: withreplacement, Bsamplesofsizen$$

$$Jackknife: leave - one - out, noreplacement$$

### Last-Minute Exam Checklist

**SE example:** σ=6%, n=100 → SE = 0.6

**CLT example:** μ=12%, σ=10%, n=100 → $\bar{X}$ ~ N(12%, 1%)

**Probability vs non-probability:** equal chance vs judgment/convenience

**Stratified advantage:** represents sub-populations

**Bootstrap:** creates the distribution when no formula exists

**SD ≠ SE:** SD describes data, SE describes inference

Step 1 of 10