

# Sampling and Estimation

Prepared by Dr. Mohamed Sleem

Subject code : FNT 631  
Data Analysis

## Lecture 6



# Sampling and Estimation

## Lecture 6

Subject code : FNT 631  
Data Analysis

Prepared by  
**Dr. Mohamed Sleem**

[www.midocean.ae](http://www.midocean.ae)

## Chapter 4 Review Questions

**Q1: What is the difference between discrete and continuous random variables?**

**A1:** Discrete random variables have countable outcomes, like the number of students in a class. Continuous random variables have outcomes over intervals, like the height of students.

**Q2: How is the expected value of a random variable calculated?**

**A2:** The expected value is the sum of all possible values each multiplied by its probability, given by  $E[X] = \sum x_i P(x_i)$ .

**Q3: What is the significance of the normal distribution in finance?**

**A3:** The normal distribution, characterized by the bell curve, is crucial in finance for modeling stock returns, assessing risks, and making investment decisions.

Chapter  
**05****Subject's Objectives**

- **Understanding Estimation:** Assess the value of unknown population parameters using sample data.
- **Point Estimates:** Learn the use of single numbers to estimate population parameters.
- **Confidence Interval Estimates:** Comprehend the range of values believed to contain the population parameter with a specified probability.
- **Sampling Methods:** Understand various sampling methods and their applications in data collection.
- **Unbiased Estimators:** Grasp the concept of unbiased estimators and their importance in statistical inference.
- **Central Limit Theorem:** Learn how the distribution of sample means approximates normal distribution as sample size increases.

# Need for Sampling

- Very large populations
- Destructive testing
- Continuous production process

The objective of sampling is to draw a valid inference about a population.

# History of Sampling *(Contd)*

- Dates back to 1920 and started by Literary Digest, a news magazine published in the U.S. between 1890 and 1938.
- Digest successfully predicted the presidential elections in 1920, 1924, 1928, 1932 but;
- Failed in 1936...
- The Literary Digest poll in 1936 used a sample of 10 million, drawn from government lists of automobile and telephone owners. Predicted Alf Landon would beat Franklin Roosevelt by a wide margin. But instead Roosevelt won by a landslide. The reason was that the sampling frame did not match the population. Only the rich owned automobiles and telephones, and they were the ones who favored Landon.

## *What is sampling*

- A sample is some part of a larger body specially selected to represent the whole
- Sampling is then is taking any portion of a population or universe as representative of that population or universe
- Sampling is the process by which this part is chosen

# Reasons for Drawing a Sample

- Less time consuming than a census
- Less costly to administer than a census
- Less cumbersome and more practical to administer than a census of the targeted population
- Enables More Detailed and In-Depth Analysis



# Key Definitions

- A **population** (universe) is the collection of things under consideration
- A **sample** is a portion of the population selected for analysis
- A **parameter** is a summary measure computed to describe a characteristic of the population
- A **statistic** is a summary measure computed to describe a characteristic of the sample

# *A Census*

- A survey in which information is gathered about all members of a population
- Gallup poll is able to develop representative samples of any adult population with interviews of approximately 1500 respondents
- That sample size allows them to be 95% confident that the results they obtain are accurate within + or – 3% points

---

## *Sampling concepts and terminologies*

- Population/Target population
- Sampling unit
- Sampling frame

# Sampling unit/Element/ Unit of analysis

- Sampling unit is the unit about which information is collected.
- Unit of analysis is the unit that provides the basis of analysis.
- Each member of a population is an element. (e.g. a child under 5)
- Sometimes it is household, e.g. any injury in the household in the last three months.

# Sampling Frame

- The actual list of sampling units from which the sample, or some stage of the sample, is collected. It is simply a list of the study population.
- Importance in Research: A well-defined sampling frame helps in reducing sampling errors and biases, leading to more reliable and valid results.
- Example in Finance: In a survey to understand customer satisfaction with a bank, the sampling frame could be the list of all customers who have an active account with the bank.
- Challenges: Maintaining an up-to-date and comprehensive sampling frame can be challenging, especially in dynamic populations.
- Types of Sampling Frames: Can include simple lists, stratified lists, or hierarchical structures depending on the complexity of the population and the research needs.
- Sampling Frame vs. Population: The sampling frame is not always equivalent to the population; it is a practical subset from which samples can be drawn.
- Implications of Errors: Errors in the sampling frame can lead to coverage errors, where some members of the population are inadequately represented or entirely omitted.

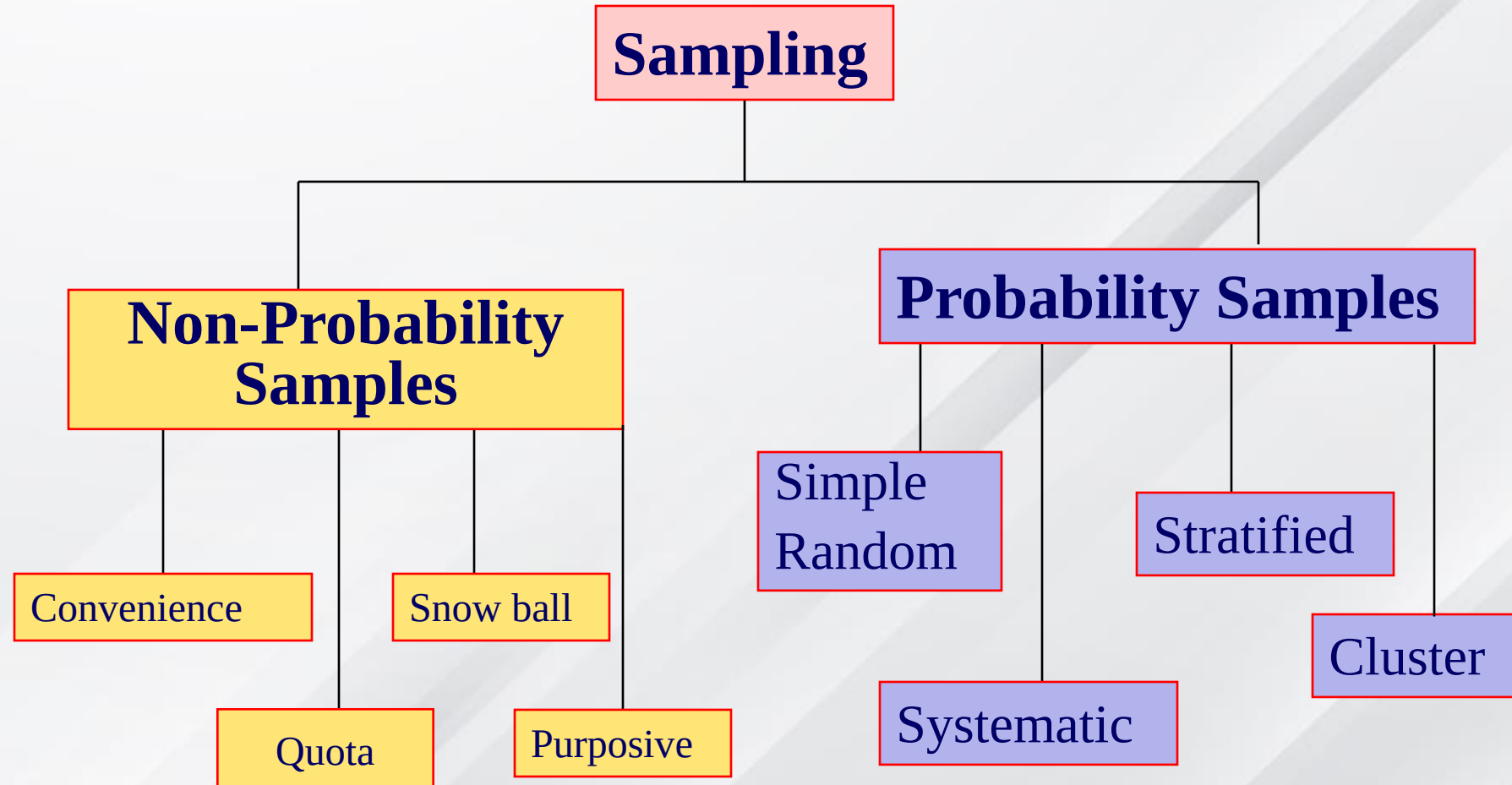
# Sample Design

- **Sampling Plan** – a description of the approach that will be used to obtain samples from a population
  - Objectives
  - Target population
  - Population frame
  - Method of sampling
  - Operational procedures for data collection
  - Statistical tools for analysis

# Sampling Methods

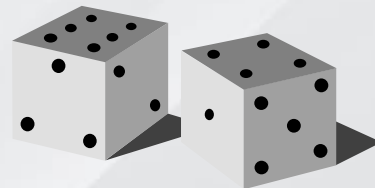
- Subjective
  - Judgment sampling
  - Convenience sampling
- Probabilistic
  - Simple random sampling – every subset of a given size has an equal chance of being selected

# Types of Sampling Methods



# Simple Random Samples

- Every individual or item from the frame has an equal chance of being selected
- Selection may be with replacement or, without replacement
- Samples obtained from table of random numbers or computer random number generators
- Random samples are unbiased and, on average, representative of the population



# Systematic sample

- This method is referred to as a systematic sample with a random start.
- This is done by picking every 5<sup>th</sup> or 10<sup>th</sup> unit at regular intervals.
- For example to carry out a filarial survey in a town, we take 10% sample. If the total population of the town is about 5000. The sample comes to 500.

# Stratified Random sample

- This involves dividing the population into distinct subgroups according to some important characteristics, such as age, or socioeconomic status, religion and selecting a random number from each subgroup. (e.g. African voodoo healers)
- Especially important when one group is so small (say, 3% of the population) that a random sample might miss them entirely.
- Population divided into two or more groups according to some common characteristic
- Simple random sample selected from each group
- The two or more samples are combined into one

# Cluster sample

- A sampling method in which each unit selected is a group of persons (all persons in a city block, a family, etc.) rather than an individual.
- Used when (a) sampling frame not available or too expensive, and (b) cost of reaching an individual element is too high
  - E.g., there is no list of automobile mechanics in the Myanmar. Even if you could construct it, it would cost too much money to reach randomly selected mechanics across the entire Myanmar : would have to have unbelievable travel budget
- In cluster sampling, first define large clusters of people. Fairly similar to other clusters. For example, cities make good clusters.
- Once you've chosen the cities, might be able to get a reasonably accurate list of all the mechanics in each of those cities. Is also much less expensive to fly to just 10 cities instead of 200 cities.
- Cluster sampling is less expensive than other methods, but less accurate.

## Non- Probability Sampling /(Non-Random)

This is where the probability of inclusion in the sample is unknown.

- Convenience sampling
- Purposive sampling
- Quota sampling
- Snow ball sampling

# Errors in Sampling

- Nonsampling error
  - Poor sample design
- Sampling (statistical) error
  - Depends on sample size
  - Tradeoff between cost of sampling and accuracy of estimates obtained by sampling

# Non-Sampling Error

- Coverage error
  - Wrong population definition
  - Flawed sampling frame
  - Interviewer or management error in following sampling frame
- Response error
  - Badly worded question results in invalid or incorrect response
  - Interviewer bias changes response
- Non-response error
  - Respondent refuses to take survey or is away
  - Respondent refuses to answer certain questions
- Processing errors
  - Error in data entry or recording of responses
- Analysis errors
  - Inappropriate analytical techniques, weighting or imputation are applied

# Sampling Error

- Sampling error is known after the data are collected by calculating the Margin of Error and confidence intervals
- Surveys don't have a Margin of Error, questions do
- Power analyses use estimates of the parameters involved in calculating the margin of error
- It is common to see sample sizes of 400 and 1000 for surveys (these are associated with 5% and 3% margins of error)
- In most cases the size of the population being sampled from is irrelevant
- The margin of error should be calculated using the size of the subgroups sampled

# Margin of Error Formula

$$H = \frac{zS}{\sqrt{n}}$$

- H = Half interval **expressed in units of standard deviation**
- z = z score associated with level of confidence (typically 95%)
- s = standard deviation
- n = sample size

# The **z** score

- The z value is the z score associated with a level of confidence
- Typically (almost exclusively) surveys use 95%
- This means that if the survey were replicated 100 times, 95 times out of 100 the estimate would be within the margin of error
- The z score associated with 95% is 1.96

# Statistical Sampling Experiment in Finance

- In finance, one way of evaluating capital budgeting projects is to compute a profitability index ( $PI$ ), which is defined as the ratio of the present value of future cash flows ( $PV$ ) to the initial investment ( $I$ ):

$$PI = PV/I$$

- Suppose that  $PV$  is estimated to be normally distributed with a mean of \$12 million and a standard deviation of \$2.5 million, and the initial investment is also estimated to be normal with a mean of \$3.0 million and standard deviation of \$0.8 million. What is the distribution of  $PI$ ?

# Central Limit Theorem

- If the sample size is large enough (generally at least 30, but depends on the actual distribution), the sampling distribution of the mean is approximately normal, *regardless* of the distribution of the population.
- If the population is normal, then the sampling distribution of the mean is exactly normal for any sample size.

# Applying Sampling Distributions – Example 1

- Suppose that the size of individual customer orders (in dollars),  $X$ , from a major discount book publisher Web site is normally distributed with a mean of \$36 and standard deviation of \$8. What is the probability that the next individual who places an order at the Web site will purchase more than \$40?
- $P(X > 40) = 1 - \text{NORMDIST}(40, 36, 8, \text{TRUE})$   
 $= 1 - 0.6915 = 0.3085$
- The calculation uses the standard deviation of the individual customer orders.

# Applying Sampling Distributions – Example 2

- Suppose that a sample of 16 customers is chosen. What is the probability that the *mean purchase* for these 16 customers will exceed \$40?
- The sampling distribution of the mean will have a mean of \$36, but a standard error of  $\$8/\sqrt{16} = \$2$ . —
- $P(\bar{X} > 40) = 1 - \text{NORMDIST}(40, 36, 2, \text{TRUE})$   
 $= 1 - .9772 = 0.0228$
- The calculation uses the standard error of the sampling distribution.

# Sampling and Estimation

- **Estimation** – assessing the value of an unknown population parameter using sample data.
- **Point estimate** – a single number used to estimate a population parameter
- **Confidence interval estimate** – a range of values between which a population parameter is believed to be along with the probability that the interval correctly estimates the true population parameter

# Common Point Estimates

**TABLE 4.4** Common Point Estimates

| Point Estimate                 | Population Parameter                    |
|--------------------------------|-----------------------------------------|
| Sample mean, $\bar{x}$         | Population mean, $\mu$                  |
| Sample variance, $s^2$         | Population variance, $\sigma^2$         |
| Sample standard deviation, $s$ | Population standard deviation, $\sigma$ |
| Sample proportion, $\hat{p}$   | Population proportion, $\pi$            |

# Example

|    | A                             | B               | C             | D            | E            | F             |
|----|-------------------------------|-----------------|---------------|--------------|--------------|---------------|
| 1  | <b>Cereal Name</b>            | <b>Calories</b> | <b>Sodium</b> | <b>Fiber</b> | <b>Carbs</b> | <b>Sugars</b> |
| 2  | Lucky Charms                  | 110             | 180           | 0            | 12           | 12            |
| 3  | Apple Jacks                   | 110             | 125           | 1            | 11           | 14            |
| 4  | Apple Cinn Cheerios           | 110             | 180           | 1.50         | 10.50        | 10            |
| 5  | Crispy Wheat & Raisins        | 100             | 140           | 2            | 11           | 10            |
| 6  | Honey Graham Ohs              | 120             | 220           | 1            | 12           | 11            |
| 7  | Puffed Wheat                  | 50              | 0             | 1            | 10           | 0             |
| 8  | Honey Nut Cheerios            | 110             | 250           | 1.50         | 11.50        | 10            |
| 9  | Trix                          | 110             | 140           | 0            | 13           | 12            |
| 10 | Post Nat. Raisin Bran         | 120             | 200           | 6            | 11           | 14            |
| 11 | Almond Delight                | 110             | 200           | 1            | 14           | 8             |
| 12 | Sample Mean                   | 105             | 163.5         | 1.5          | 11.6         | 10.1          |
| 13 | Sample Standard Deviation     | 20.1384         | 69.284        | 1.7          | 1.197        | 4.0125        |
| 14 |                               |                 |               |              |              |               |
| 15 | Population Mean               | 105.522         | 167.313       | 2.187        | 14.77        | 6.9552        |
| 16 | Population Standard Deviation | 18.631          | 79.9782       | 2.487        | 3.831        | 4.3758        |

# Theoretical Issues

- **Unbiased estimator** – one for which the expected value equals the population parameter it is intended to estimate
- The sample variance is an unbiased estimator for the population variance

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{N}$$

# Various Levels of Confidence

- When population standard deviation is known use Z table values:
  - For 95% CI: mean  $\pm$  1.96 s.e. of mean
  - For 99% CI: mean  $\pm$  2.58 s.e. of mean
- When population standard deviation is not known use “Critical Value of “t” table
  - For 95%CI: mean  $\pm$  2.04 s.e. of mean
  - For 99% CI: mean  $\pm$  2.75 s.e. of mean

# Interval Estimates

- Range within which we believe the true population parameter falls
  - Example: Gallup poll – percentage of voters favoring a candidate is 56% with a 3% margin of error.
  - Interval estimate is [53%, 59%]
- A  $100(1 - \alpha)\%$  probability interval is any interval  $[A, B]$  such that the probability of falling between  $A$  and  $B$  is  $1 - \alpha$

# Confidence Intervals

- **Confidence interval (CI)** – an interval estimated that specifies the likelihood that the interval contains the true population parameter
  - A defined interval of values that includes the statistic of interest, by adding and subtracting a specific amount from the computed statistic
  - A CI is the probability that the interval computed from the sample data includes the population parameter of interest
- **Level of confidence** ( $1 - \alpha$ ) – the likelihood that the CI contains the true population parameter, usually expressed as a percentage (90%, 95%, 99% are most common).

# Sampling From Finite Populations

- When  $n > 0.05N$ , use a correction factor in computing the standard error:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

# Key Observations

$$\bar{x} \pm z_{\alpha/2}(\sigma / \sqrt{n})$$

- As the confidence level  $(1 - \alpha)$  increases, the width of the confidence interval also increases.
- As the sample size increases, the width of the confidence interval decreases.

# Conclusion

- **Estimation:** Using sample data to assess unknown population parameters.
- **Point Estimates:** Single value estimates of population parameters, such as sample mean, sample variance, sample standard deviation, and sample proportion.
- **Confidence Interval:** Range within which the true population parameter is believed to fall, with a certain level of confidence.
- **Unbiased Estimator:** An estimator whose expected value equals the population parameter.
- **Central Limit Theorem:** As sample size increases, the sampling distribution of the mean approaches a normal distribution, regardless of the population's distribution.

# Review Questions

- Explain the importance of using the finite population correction factor when sampling from a finite population. Provide an example where this correction significantly impacts the standard error.
- Discuss the difference between sample variance and population variance, and explain why the sample variance is considered an unbiased estimator of the population variance.
- Describe the role of the z-score in estimating population parameters from a sample. How does the z-score facilitate the calculation of confidence intervals? Provide an example using a 95% confidence level.



# Thank You