Statistics and Probabilities for Computing (CEN1602)

Unit-6

Sampling Distributions and the Central Limit Theorem

Prepared by

Dr. Ahmad Alzahrani

Outline

1. **From population to sample**
   Why we use samples instead of studying the whole population
2. **Parameters and statistics**
   Population values versus sample values
3. **Point estimation**
   Using sample statistics to estimate population parameters
4. **Random samples**
   What makes a sample suitable for statistical analysis
5. **Sampling distributions**
   How statistics behave over repeated random sampling
6. **Sampling distribution of the sample mean**
   Mean, standard error, and shape
7. **Central Limit Theorem**
   Why $\bar{X}$ becomes approximately normal for large samples
8. **Using the sampling distribution**
   Standardizing $\bar{X}$ and computing probabilities

Learning Objectives for Unit 6

By the end of this unit, students should be able to:

- Explain how **samples** are used to make conclusions about **populations**.
- Distinguish between a population **parameter** and a sample **statistic**.
- Explain the idea of a **point estimator** and **point estimate**.
- Define a random sample and explain why **random sampling** matters.
- Describe the idea of a **sampling distribution**.
- Explain why a statistic can change from one sample to another.
- Identify the sampling distribution of the sample mean $\bar{X}$.
- Compute the **mean** and **standard error** of $\bar{X}$.
- Explain the main idea of the **Central Limit Theorem**.
- Use the normal distribution to compute probabilities involving $\bar{X}$.
- Explain how sampling distributions support **statistical inference**.

Main Idea

- In Unit 1, we learned that inferential statistics uses samples to make conclusions about populations.
- In this unit, we study **how sample results can be used to estimate unknown population values**.
- Usually, the population value we want to know is unknown.
- Examples:

$$\mu, \sigma, \sigma^2, p$$

- These population values are called **parameters**.
- Since studying the whole population may be difficult, expensive, or impossible, we collect a sample.
- From the sample, we calculate values such as:

$$\bar{x}, s, s^2, \hat{p}$$

- These sample values are called **statistics**.
- The main idea of this unit is:

Use *a sample statistic to estimate a population parameter*

- In this unit, we focus mainly on using the sample mean $\bar{X}$ to estimate the population mean μ.

Parameters and Statistics

- A **parameter** is a numerical value that describes a population.
- A **statistic** is a numerical value calculated from a sample.

Quantity	Population Parameter	Sample Statistic
Mean	μ	$\bar{x}$
Standard deviation	σ	s
Variance	σ^2	s^2
Proportion	p	$\hat{p}$

- The proportion p is used for success/failure trials, such as binomial, geometric, and negative binomial cases.
- Parameters are usually unknown because they require information from the whole population.
- Statistics are computed from sample data and are used to estimate population parameters.
- Different samples may produce different values of the same statistic.
- Because statistics can change from sample to sample, we need to understand how they behave over repeated sampling.

Point Estimation

- A **point estimator** is a statistic used to estimate a population parameter.
- A **point estimate** is the numerical value obtained from one sample.
- For example, the sample mean $\bar{x}$ is a point estimator of the population mean μ.

$$\bar{X} \to \mu$$

- If one sample gives:

$$\bar{x} = 11.8$$

- then 11.8 is the point estimate of μ.
- Similarly:

Parameter	Point Estimator
μ	$\bar{X}$
σ^2	S^2
p	$\hat{p}$

- A good point estimator should usually give values close to the true population parameter.
- Different samples may give different point estimates, even when they are estimating the same parameter.
- In this unit, we mainly focus on $\bar{x}$ as an estimator of μ.

Common Estimation Techniques

- After choosing a population parameter to estimate, we need a method to choose a suitable estimator.
- Some common estimation techniques are:

Technique	Main Idea
Sample analogue estimation	Use the sample version of the population quantity
Method of moments (MoM)	Match sample summaries with population formulas
Maximum likelihood estimation	Choose the parameter value that makes the observed sample most likely
Bayesian estimation	Combine sample data with prior information

- For this course, we only need the main idea of these techniques.
- The rest of the unit focuses on the sample mean $\bar{X}$,its sampling distribution, and the CLT.

Estimation Techniques: Simple Examples

1. Sample Analogue Estimation

- Sample analogue estimation uses the sample version of the population quantity.
- For example, if the unknown population mean is μ we estimate it directly using the observed sample mean $\bar{x}$ ($\mu \approx \bar{x}$).
- Similarly, if the unknown population variance is σ^2 ,we estimate it using the sample variance s^2 , so $\sigma^2 \approx s^2$.

2. Method of Moments (MoM)

- A moment is an expected value of a power of X.

$$\textit{k-th moment} = E(X^k)$$

- So, the first moment is $E(X)$ which is the mean. The second moment $E(X^2)$ is related to the spread of the distribution ($\sigma^2 = E(X^2) - [E(X)]^2$)
- For example, for a normal distribution with μ and σ, we know:

$$E(X) = \mu \quad \text{and} \quad E(X^2) = \mu^2 + \sigma^2$$

- MoM matches population moments with sample moments:

$$E(X) \approx \bar{x}$$

$$E(X^2) \approx \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

- Then these equations can be solved to estimate the unknown parameters μ and σ^2.

Estimation Techniques: Simple Examples

3. Maximum Likelihood Estimation (MLE)

- Suppose a random variable X has a PDF that depends on an unknown parameter θ we want to estimate:

$$f(x, \theta)$$

- If the observed sample values are:

$$x_1,\ x_2, \ldots, x_n$$

- then MLE asks: Which value of θ makes the observed sample most likely?
- The likelihood function is:

$$L(\theta) = f(x_1, \theta) f(x_2, \theta) \cdots f(x_n, \theta)$$

- The MLE is the value of θ that gives the largest likelihood

$$\theta_{MLE} = value\ of\ \theta\ that\ maximizes\ L(\theta)$$

- Example: Suppose response time is normally distributed with known variance (σ^2), but unknown mean (μ). IF observed sample is $70,\ 72,\ 74$, then data is centered around $\bar{x} = 72$. The MLE chooses: $\mu_{MLE} = 72.$
- The MLE estimator in this example is the same as sample analogue estimation. The difference is that MLE chooses μ **by maximizing likelihood**, not simply because $\bar{x}$ is the sample version of μ.

Estimation Techniques: Simple Examples

4. Bayesian Estimation

- In Unit 3, Bayes' theorem updated the probability of an event after new evidence:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

- The same idea is used in Bayesian estimation, but now we update our belief about an unknown parameter, such as μ.

$$p(\mu \mid data) \propto p(data \mid \mu)p(\mu)$$

Bayes' Theorem	Bayesian Estimation
$P(A)$: prior probability	$p(\mu)$:prior belief about (μ)
B: new evidence	$data$: new sample
$P(B \mid A)$: likelihood	$p(\text{data} \mid \mu)$: likelihood of the data for a given (μ)
$P(A \mid B)$: posterior probability	$p(\mu \mid \text{data})$: updated belief about (μ)

- Example: Suppose the response time is a random variable X, and we want to estimate its population mean μ.
  Previous system logs suggest $\mu \approx 70\ ms$. A new sample gives $\bar{x} = 72\ ms$. Bayesian estimation updates the estimate using both sources of information. A simplified weighted update could be $0.7(70) + 0.3(72) = 70.6\ ms$.
- This weighted average shows the intuition of Bayesian updating here. In formal Bayesian estimation, these weights are calculated from the prior and the observed data.

From Population to Sample

- To learn about a population, we select a smaller group called a sample.
- The sample should be selected randomly in a way that gives a fair picture of the population.
- For each item in the sample, we measure the random variable X.
- If the sample contains n selected items, then the observed values are:

$$X_1, X_2, X_3, \dots, X_n$$

- These values may be different from one item to another.
- To summarize the sample, we calculate the sample mean:

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

- The sample mean gives one average value for the selected sample.
- If a different sample is selected, the sample mean $\bar{X}$ can also be different.
- Therefore, $\bar{X}$ is considered a random variable, and its probability distribution is called the **sampling distribution**.
- This moves us from studying one value X to studying a sample average $\bar{X}$.

Applied Example: UQU WiFi Coverage

- Suppose UQU wants to answer the following practical question:

  Is the WiFi download speed in UQU study buildings sufficient for video-based learning during peak hours?
- We use 10 Mbps as a practical reference point for video-based learning.
- During peak hours, WiFi speed may decrease because many students are using the network at the same time.
- To study this, we define:

  Population: All WiFi coverage locations in UQU study buildings during peak hours

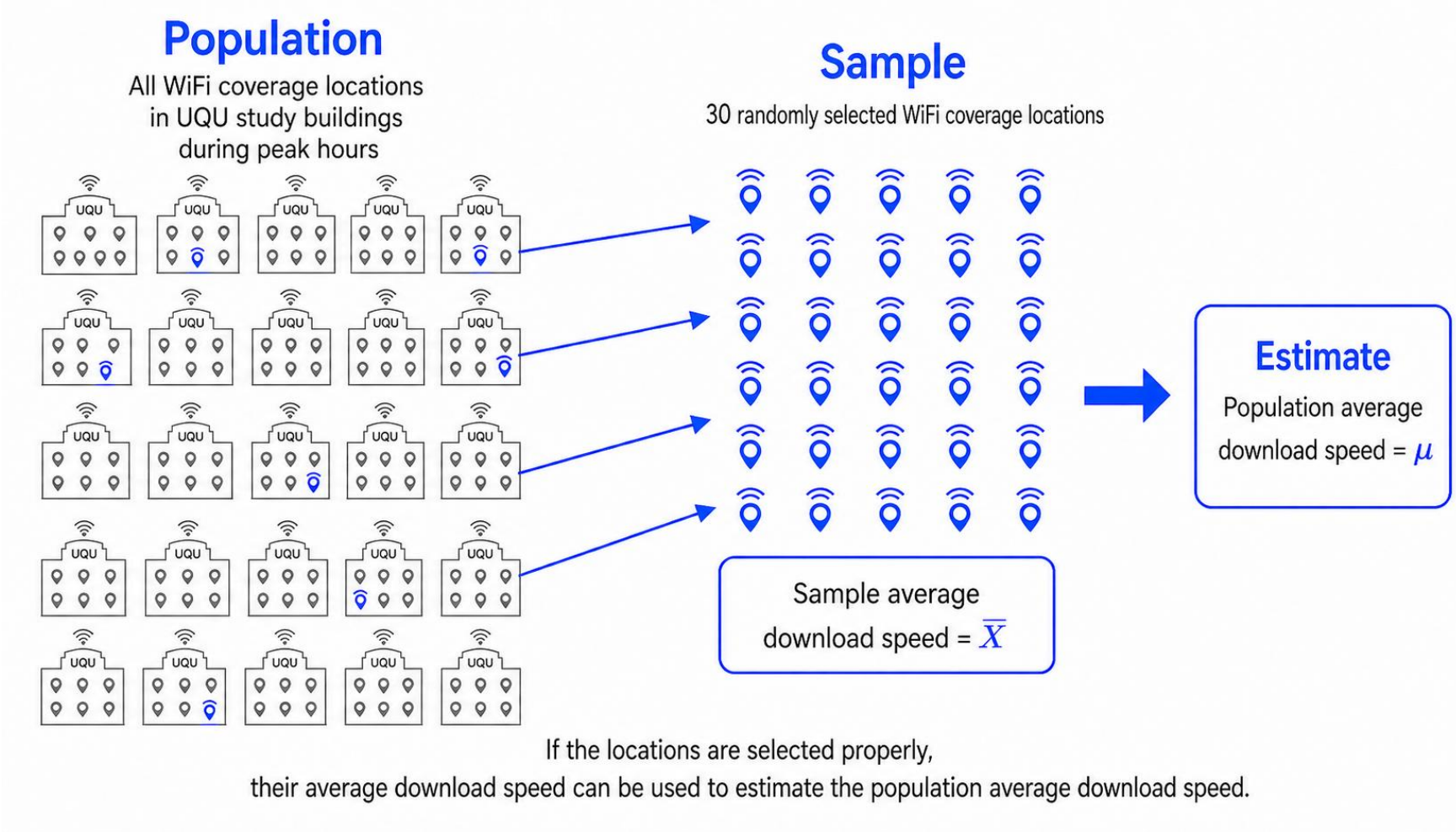
  Sample: 30 selected WiFi coverage locations

  Variable: $X = \textit{download speed at one selected location in Mbps}$
- From the sample, we calculate:

  $$\bar{X} = \textit{sample average download speed}$$
- The goal is to estimate:

  $$\mu = \textit{population average download speed}$$
- This can help decide whether WiFi quality is sufficient and where improvements may be needed.

Applied Example: UQU WiFi Coverage

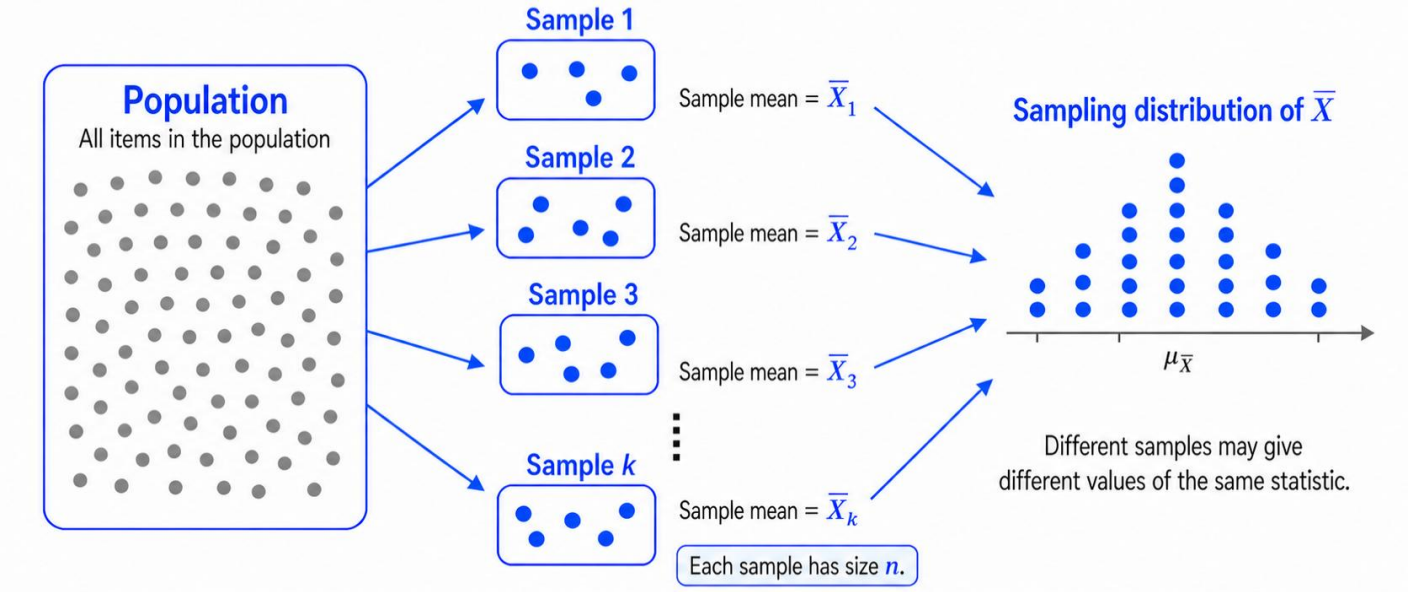


Random Samples

- A sample should be selected in a way that gives useful information about the population.
- A **simple random sample** is selected so that each possible sample of size n has the same chance of being selected.
- Random sampling helps reduce **bias**.
- Bias can happen when some parts of the population are more likely to be selected than others.
- For example, if WiFi speed is measured only near routers, the sample may give an unfair picture of WiFi quality.
- A better sample should include selected locations from different parts of the study buildings.
- **Undercoverage** happens when some parts of the population are left out of the sample.
- For example, if basement study areas or crowded floors are ignored, the sample may not represent the whole WiFi coverage area.
- A suitable random sample helps us make more reliable conclusions about the population.

Sampling Distribution

- A **sampling distribution** describes the behavior of a statistic over many possible random samples.
- In this unit, the main statistic we study is the sample mean $\bar{X}$.
- Imagine selecting k random samples of the same size n. Each sample produces one sample mean. $\bar{X}_1, \bar{X}_2, \bar{X}_3, .., \bar{X}_k$
- These sample means form a distribution called the **sampling distribution** of $\bar{X}$.
- It helps us understand how much sample means can vary when estimating the population mean.
- It is the bridge between sample data and statistical inference.



Mean and Standard Error of $\bar{X}$

- Suppose a population has mean μ and standard deviation σ.
- If we take random samples of size n, each sample gives one sample mean $\bar{X}$.
- The sampling distribution of $\bar{X}$ has:

$$\boldsymbol{\mu_{\bar{X}} = \mu}$$

and

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

- $\sigma_{\bar{X}}$ is called the **standard error (SE)** of the sample mean.
- The mean $\mu_{\bar{X}}$ tells us where the sample means are centered.
- The standard error tells us how much the sample means vary from one sample to another.
- As n increases, the standard error decreases, so $\bar{X}$ becomes more stable.

Central Limit Theorem

- The **Central Limit Theorem (CLT)** states that:

  If we take many random samples of size n from a population with mean μ and standard deviation σ, then the sampling distribution of the sample mean $\bar{X}$ becomes approximately normal when n is large enough.

- For many practical cases, $\boldsymbol{n \geq 30}$ is used as a common rule of thumb. Therefore, for large n:

$$\bar{X} \approx N(\mu, \mathrm{SE}^2)$$

- This approximation allows us to use the normal distribution to answer probability questions about $\bar{X}$.
- For example, we can find probabilities such as:

$$P(\bar{X} < a) \quad \text{or} \quad P(a < \bar{X} < b)$$

- So, instead of studying individual values of X, we can study the behavior of sample averages.
- This is the main reason the CLT is important in **statistical inference.**

Standardizing the Sample Mean

- When the sampling distribution of $\bar{X}$ is normal or approximately normal, we can convert $\bar{X}$ to a standard normal variable.

$$Z = \frac{\bar{X} - \mu}{\text{SE}}$$

- The numerator $(\bar{X} - \mu)$ measures how far the sample mean is from the population mean.
- The denominator is the standard error of the sample mean.

$$\text{SE} = \frac{\sigma}{\sqrt{n}}$$

- So, Z tells us how many standard errors $\bar{X}$ is away from μ.
- After standardizing, probabilities involving $\bar{X}$ can be found using the standard normal table.
- Example forms:

$$P(\bar{X} < a) \rightarrow P\left(Z < \frac{a - \mu}{\sigma/\sqrt{n}}\right)$$

$$P(a < \bar{X} < b) \rightarrow P\left(\frac{a - \mu}{\sigma/\sqrt{n}} < Z < \frac{b - \mu}{\sigma/\sqrt{n}}\right)$$

Example 1

Suppose the WiFi download speed in UQU study buildings has:

$$\mu = 12\ Mbps \quad , \quad \sigma = 4\ Mbps$$

A random sample of $n = 36$ locations is selected.

Find: $P(\bar{X} < 11)$.

Since $n = 36$, we use the sampling distribution of $\bar{X}$:

$$\mu_{\bar{X}} = \mu = 12$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{36}} = \frac{4}{6} = 0.667$$

Standardize:

$$Z = \frac{11 - 12}{0.667} = -1.50$$

Therefore:

$$P(\bar{X} < 11) = P(Z < -1.50)$$

From the standard normal table:

$$P(Z < -1.50) = 0.0668$$

So, the probability is:

$$0.0668 \text{ or } 6.68\%$$

Example 2

Using the same population $\mu = 12$ *Mbps* , $\sigma = 4$ *Mbps*
and the same sample size $n = 36$
Find: $P(11 < \bar{X} < 13)$

The standard error is $\sigma_{\bar{X}} = \frac{4}{\sqrt{36}} = 0.667$

Standardize both values:

$$Z_1 = \frac{11 - 12}{0.667} = -1.50$$

$$Z_2 = \frac{13 - 12}{0.667} = 1.50$$

Therefore:

$$P(11 < \bar{X} < 13) = P(-1.50 < Z < 1.50)$$

From the standard normal table:

$$P(Z < 1.50) = 0.9332$$

$$P(Z < -1.50) = 0.0668$$

So:

$$P(-1.50 < Z < 1.50) = 0.9332 - 0.0668 = 0.8664$$

The probability is:

$$0.8664 \text{ or } 86.64\%$$

Unit 6 Summary

- We use samples to learn about unknown population parameters.
- A statistic is calculated from a sample and can be used as a point estimator.
- Different samples may give different values of the same statistic.
- A sampling distribution describes how a statistic behaves over repeated random sampling.
- For the sample mean $\bar{X}$:

$$\mu_{\bar{X}} = \mu$$

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

- The Central Limit Theorem explains why $\bar{X}$ becomes approximately normal for large samples.
- When $\bar{X}$ is normal or approximately normal, we can standardize:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

- This allows us to use the standard normal table to compute probabilities involving $\bar{X}$.