Chapter 7

Sampling and Sampling Distributions

LO 7.1: Differentiate between a population parameter and a sample statistic.

LO 7.2: Explain common sample biases.

LO 7.3: Describe simple random sampling.

LO 7.4: Distinguish between stratified random sampling and cluster sampling.

LO 7.5: Describe the properties of the sampling distribution of the same mean.

LO 7.6: Explain the importance of the central limit theorem.

LO 7.7: Describe the properties of the sample distribution of the sample proportion.

LO 7.8: Use a finite population correction factor.

LO 7.9: Construct and interpret control charts from quantitative and qualitative data.

7.1 Sampling

LO 7.1 **Differentiate between a population parameter and a sample statistic.**

- **Population** – consists of all items of interest in a statistical problem.
  - **Population Parameter** is unknown.
- **Sample** – a subset of the population.
  - **Sample statistic** is calculated from sample and used to make inferences about the population.
- **Bias** – the tendency of a sample statistic to systematically over- or under-estimate a population parameter.

LO 7.2 **Explain common sample biases.**

- Classic Case of a "Bad" Sample: The *Literary Digest* Debacle of 1936
  - During the 1936 presidential election, the *Literary Digest* predicted a landslide victory of Alf Landon over Franklin D. Roosevelt (FDR) with only a 1% margin or error.
  - They were wrong! FDR won in a landslide election.
- **Selection bias** – a systematic exclusion of certain groups from consideration for the sample.
  - The *Literary Digest* committed selection bias by excluding a large portion of the population (e.g., lower income voters).
- **Nonresponse bias** – a systematic difference in preferences between respondents and non-respondents to a survey or a poll.
  - The *Literary Digest* had only a 24% response rate. This indicates that only those who cared a great deal about the election took the time to respond to the survey. These respondents may be atypical of the population as a whole.

LO 7.3 **Describe simple random sampling.**

7.1.1 Sampling Methods

- Simple random sample is a sample of n observations which have the sample probability of being selected from the population as any other sample of n observations.
  - Most statistical methods presume simple random samples.
  - However, in some situations, other sampling methods have an advantage over simple random samples.

LO 7.4 **Distinguish between stratified random sampling and cluster sampling.**

7.1.2 Stratified Random Sampling

- Divide the population into mutually exclusive and collectively exhaustive groups, called **strata**.
- Randomly select observations from each stratum, which are proportional to the stratum's size.

- Advantages:
  - Guarantees that each population's subdivision is represented in the sample.
  - Parameter estimates have greater precision than those estimated from simple random sampling.

7.1.3 Cluster Sampling

- Divide population into mutually exclusive and collectively exhaustive groups, called clusters.
- Random select clusters.
- Sample every observation in those randomly selected clusters.
- Advantages and disadvantages:
  - Less expensive than other sampling methods.
  - Less precision than simple random sampling or stratified sapling.
  - Useful when clusters occur naturally in the population.

Table 7.1: Stratified vs. Cluster Sampling

Stratified Sampling	**Cluster Sampling**
Sample consists of elements from each group.	Sample consists of elements from the selected groups.
Preferred when the objective is to increase precision.	Preferred when the objective is to reduce costs.

7.2 The Sampling Distribution of the Means

LO 7.5 **Describe the properties of the sampling distribution of the same mean.**

- Population is described by parameters.
  - A *parameter* is a constant, whose value may be unknown.
  - Only one population.
- Sample is described by statistics.
  - A **statistic** is a random variable whose value depends on the chosen random sample.
  - Statistics are used to make **inferences** about the population parameters.
  - Can draw multiple random samples of size n.

7.2.1 Estimator

- A statistic that is used to estimate a population parameter.
- For example, $\bar{X}$, the mean of the sample, is an estimate of μ, the mean of the population.

7.2.2 Estimate

- A particular value of the estimator.
- For example, the mean of the sample $\bar{x}$ is an estimate of μ, the mean of the population.

7.2.3 Sampling Distribution of the Mean $\bar{x}$

- Each random sample size n drawn from the population provides an estimate of μ–the sample mean.
- Drawing many samples of size n results in many different sample means, one for each sample.
- The sampling distribution of the mean is the frequency or probability distribution of these sample means.

7.2.4 The Expected Value and Standard Deviation of the Sample Mean

- The expected value of X,
$$E(X) = \mu \tag{7.1}$$
- The expected value of the mean,
$$E(\bar{X}) = E(X) = \mu \tag{7.2}$$
- Variance of X
$$\text{Var}(X) = \sigma^2 \tag{7.3}$$
- Standard Deviation
  - of X
$$SD(X) = \sqrt{\sigma^2} = \sigma \tag{7.4}$$
  - of $\bar{X}$
$$SD(\bar{X}) = \frac{\sigma}{\sqrt{n}} \tag{7.5}$$

    where n is the sample size. Also known as the **standard error of the mean**.

7.2.5 Sampling from a Normal Distribution

- For any sample size n, the sampling distribution of $\bar{X}$ is **normal** if the population X from which the sample is drawn is normally distributed.
- If X is normal, then we can transform it into the **standard normal random variable** as:
  - For a sampling distribution:

$$\begin{aligned} Z &= \frac{\bar{X} - E(\bar{X})}{\mathrm{SD}(\bar{X})} \\ &= \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \end{aligned} \tag{7.6}$$

  - For a distribution of the values of X.

$$\begin{aligned} Z &= \frac{X - E(X)}{\mathrm{SD}(X)} \\ &= \frac{X - \mu}{\sigma} \end{aligned} \tag{7.7}$$

7.2.6 The Central Limit Theorem

LO 7.6 **Explain the importance of the central limit theorem.**

- For any population X with expected value μ and standard deviation σ, the sampling distribution of $\bar{X}$ will be approximately normal if the sample size n is sufficiently large.
- As a general guideline, the normal distribution approximation is justified when $n \geq 30$.
- As before, if $\bar{X}$ is approximately normal, then we can transform it using (7.7).

LO 7.7 **Describe the properties of the sample distribution of the sample proportion.**

- **Estimator** – Sample proportion $\bar{P}$ is used to estimate the population parameter p.
- **Estimate** – a particular value of the estimator $\bar{p}$.

LO 7.8 **Use a finite population correction factor.**

LO 7.9 **Construct and interpret control charts from quantitative and qualitative data.**

Chapter 8

Estimation

LO 8.1: Discuss point estimators and their desirable properties.

LO 8.2: Explain an interval estimator.

LO 8.3: Calculate a confidence interval for the population mean when the population standard deviation is known.

LO 8.4: Describe the factors that influence the width of a confidence interval.

LO 8.5: **Discuss features of the t distribution.**

LO 8.6: **Calculate a confidence interval for the population mean when the population standard deviation is not known.**

LO 8.7: **Calculate a confidence interval for the population proportion.**

LO 8.8: **Select a sample size to estimate the population mean and the population proportion.**

8.1 Point Estimators and Their Properties

LO 8.1 **Discuss point estimators and their desirable properties.**

8.1.1 Point Estimator

- A function of the random sample used to make inferences about the value of an unknown population parameter.
- For example, $\bar{X}$ is a point estimator for μ and $\bar{P}$ is a point estimator for p.

8.1.2 Point Estimate

- The value of the point estimator derived from a given sample.
- For example, $\bar{X} = 96.5$ is a point estimate of the mpg for all ultra-green cars.

8.1.3 Properties of Point Estimators

- **Unbiased** – an estimator is unbiased if its expected value equals the unknown population parameter being estimated.
- **Efficient** – an unbiased estimator is efficient if its standaed error is lower than that of other unbiased estimators.
- **Consistent** – an estimator is consistent if it approaches the unknown population parameter being estimated as the sample size grows larger.

8.2 Confidence Interval of the Population Mean When σ Is Known

LO 8.2 **Explain an interval estimator.**

- **Confidence Interval** – provides a range of values that, with a certain level of confidence, contains the population parameter of interest.
  - Also referred to as an **interval estimate**.
- Construct a confidence interval as: Point estimate ± Margin of error.
  - **Margin of error** accounts for the variability of the estimator and the desired confidence level of the interval.

8.2.1 Constructing a Confidence Interval for μ When σ is Known

LO 8.3 **Calculate a confidence interval for the population mean when the population standard deviation is known.**

- Consider a standard normal random variable:

$$P(-1.96 \leq Z \leq 1.96) = 0.95$$

- Because of (7.7), we get:

$$P\left(-1.96 \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq 1.96\right) = 0.95$$

- Which, after algebraically manipulating, is equal to:

$$P\left(\mu - 1.96\frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95 \tag{8.1}$$

- Note that (8.1) implies there is a 95% probability that the sample mean $\bar{X}$ will fall within the interval $\mu \pm 1.96\frac{\sigma}{\sqrt{n}}$.
  - Thus, if samples of size n are drawn repeatedly from a given population, 95% of the computed sample means, ___, will fall within the interval and the remaining 5% will fall outside the interval.
- Since we do not know μ, we cannot determine if a particular $\bar{X}$ falls within the interval of not.
  - However, we do know that $\bar{X}$ will fall within the interval $\mu \pm 1.96\frac{\sigma}{\sqrt{n}}$ iff μ falls within the interval $\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$.
- This will happen 95% of the time given the interval construction. Thus, this is a 95% confidence interval for the population mean.
- Level of significance (i.e., probability of error) $= \alpha$.
- Confidence coefficient $= 1 - \alpha \Rightarrow \alpha = 1-$ confidence coefficient.
- A $100(1-\alpha)\%$ confidence interval of the population mean μ when the standard deviation σ is known is computed as

$$\bar{x} \pm z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \tag{8.2}$$

  or equivalently

$$\left[\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right] \tag{8.3}$$

- $z_{\frac{\alpha}{2}}$ is the z-value associated with the probability of $\frac{\alpha}{2}$ being in the upper-tail.

8.2.2 The Width of a Confidence Interval

LO 8.4 Describe the factors that influence the width of a confidence interval.

- Margin of Error: $z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$
- Confidence Interval Width: $2\left(z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$
- The width of the confidence interval is influenced by the:
  - Sample size n,
  - Standard deviation α, and
  - Confidence level $100(1-\alpha)\%$.

8.3 Confidence Interval of the Population Mean When σ Is Unknown

8.3.1 The t-Distribution

LO 8.5 Discuss features of the t distribution.

- If repeated samples of size n are taken from a normal population with a finite variance, then the statistic T follows the t-distribution __ with $n-1$ degrees of freedom, __.
- **Degrees of freedom** – determines the extend of the broadness of the tails of the distribution; the fewer the degrees of freedom, the broader the tails.

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \tag{8.4}$$

LO 8.6 Calculate a confidence interval for the population mean when the population standard deviation is not known.

8.4 Confidence Interval of the Population Proportion

LO 8.7 Calculate a confidence interval for the population proportion.

8.5 Selecting a Useful Sample Size

LO 8.8 Select a sample size to estimate the population mean and the population proportion.

8.5.1 Selecting n to Estimate μ

- Consider a confidence interval for μ with a known σ and let D denote the desired margin or error.
- Since

$$D = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \tag{8.5}$$

  we may rearrange to get

$$n = \left(\frac{z_{\frac{\alpha}{2}} \sigma}{D} \right)^2. \tag{8.6}$$

- If σ is unknown, estimate it with $\hat{\sigma}$.

- For a desired margin of error D, the minimum sample size n required to estimate a $100(1-\alpha)\%$ confidence interval of the population mean μ is

$$n = \left(\frac{z_{\frac{\alpha}{2}}\hat{\sigma}}{D}\right)^2. \tag{8.7}$$

  where $\hat{\sigma}$ is a reasonable estimate of σ in the planning stage.

8.5.2 Selecting n to Estimate p

- Consider a confidence interval for p and let D denote the desired margin of error.
- Since

$$D = z_{\frac{\alpha}{2}}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \tag{8.8}$$

  (where $\bar{p}$ is the sample proportion), we may rearrange to get

$$n = \left(\frac{z_{\alpha/2}}{D}\right)\bar{p}(1-\bar{p}) \tag{8.9}$$

- Since $\bar{p}$ comes from a sample, we must use a reasonable estimate of p, that is, $\hat{p}$.
- For a desired margin of error D, the minimum sample size n required to estimate a $100(1-\alpha)\%$ confidence interval of the population proportion is

$$n = \left(\frac{z_{\alpha/2}}{D}\right)p(1-p) \tag{8.10}$$

  where $\hat{p}$ is a reasonable estimate of p in the planning stage.