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 sample 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.
  - The *Literary Digest* had committed selection bias by randomly sampling from their own subscriber/membership lists, etc.
  - In addition, with only a 24% response rate, the *Literary Digest* had a great deal of non-response bias.
- 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 $\bar{x}$.
- 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
$$\text{Var}(X) = \sigma^2 = \sum \frac{(X_i - \bar{X})^2}{n-1} \tag{7.3}$$

- 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:

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

  - For a distribution of the values of X.

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

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 (**??**).

7.3 The Sampling Distribution of the Sample Proportion

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 $\bar{p}$.

7.3.1 The Expected Value and Standard Deviation of the Sample Proportion

- The expected value of $\bar{P}$ is

$$E(\bar{P}) = p \tag{7.8}$$

- The standard deviation of $\bar{P}$ is

$$\text{SD}(\bar{P}) = \sqrt{\frac{p(1-p)}{n}} \tag{7.9}$$

7.3.2 The Central Limit Theorem for the Sample Proportion

- For any population proportion p, the sampling distribution of $\bar{P}$ is approximately normal if the sample size n is sufficiently large.
- As a general guideline, the normal distribution approximation is justified when $np \geq 5$ and $n(1-p) \geq 5$.
- If $\bar{P}$ is normal, we can transform it into the standard normal random variable as

$$\begin{aligned} Z &= \frac{\bar{P} - E(\bar{P})}{SD(\bar{P})} \\ &= \frac{\bar{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \end{aligned} \tag{7.10}$$

- Therefore, any value $\bar{p}$ on $\bar{P}$ has a corresponding value z on Z given by

$$z = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \tag{7.11}$$

7.4 The Finite Population Correction Factor

LO 7.8 Use a finite population correction factor.

- Used to reduce the sampling variation of $\bar{X}$.
- The resulting standard deviation is

$$SD(\bar{X}) = \frac{\sigma}{\sqrt{n}} \left(\sqrt{\frac{N-n}{N-1}} \right) \tag{7.12}$$

- The transformation of $\bar{x}$ to Z is made accordingly.
- Apparently, only used when $\frac{n}{N} > 5\%$.

7.4.1 The Finite Population Correction Factor for the Sample Proportion

- Used to reduce the sampling variation of the sample proportion $\bar{P}$.
- The resulting standard deviation is:

$$SD(\bar{P}) = \sqrt{\frac{p(1-p)}{n}} \left(\sqrt{\frac{N-n}{N-1}} \right) \tag{7.13}$$

- The transformation of $\bar{P}$ to Z is made accordingly.

7.5 Statistical Quality Control

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

- Involves statistical techniques used to develop and maintain a firm's ability to produce high-quality goods and services.
- Two Approaches for Statistical Quality Control
  - Acceptance Sampling
  - Detection Approach

7.5.1 Acceptance Sampling

- Used at the completion of a production process or service.
- If a particular product does not conform to certain specifications, then it is either discarded or repaired.
- Disadvantages
  - It is costly to discard or repair a product.
  - The detection of all defective products is not guaranteed.

7.5.2 Detection Approach

- Inspection occurs during the production process in order to detect any nonconformance to specifications.
- Goal is to determine whether the production process should be continued or adjusted before producing a large number of defects.
- Types of variation.
  - Chance variation.
  - Assignable variation.

Chance Variation (Common Variation)

- Caused by a number of randomly occurring events that are part of the production process.
- Not controllable by the individual worker or machine.
- Expected, so not a source of alarm as long as its magnitude is tolerable and the end product meets specifications.

Assignable variation

- Caused by specific events or factors that can usually be identified and eliminated.
- Identified and corrected or removed.

7.5.3 Control Charts

- Developed by Walter A. Shewhart.
- A plot of calculated statistics of the production process over time.
- Production process is "in control" if the calculated statistics fall in an expected range.

- Production process is "out of control" if calculated statistics reveal an undesirable trend.
  - For quantitative data–$\bar{x}$ chart.
  - For qualitative data–$\bar{p}$ chart.

Control Charts for Quantitative Data

- Centerline–the mean when the process is under control.
- Upper control limit (UCL)–set at $+3\sigma$ from the mean.

$$\mu + 3\frac{\sigma}{\sqrt{n}} \tag{7.14}$$

  - Points falling above the upper control limit are considered to be out of control.
- Lower control limit (LCL)–set at -3σ from the mean.

$$\mu - 3\frac{\sigma}{\sqrt{n}} \tag{7.15}$$

  - Points falling below the lower control limit are considered to be out of control.
- Process is in control–all points fall within the control limits.

Control Charts for Qualitative Data

- $\bar{p}$ chart (fraction defective or percent defective chart).
- Tracks proportion of defects in a production process.
- Relies on central limit theorem for normal approximation for the sampling distribution of the sample proportion.
- Centerline–the mean when the process is under control.
- Upper control limit (UCL)–set at $+3\sigma$ from the mean.

$$p + 3\sqrt{\frac{p(1-p)}{n}} \tag{7.16}$$

  - Points falling above the upper control limit are considered to be out of control.
- Lower control limit (LCL)–set at -3σ from the mean.

$$p - 3\sqrt{\frac{p(1-p)}{n}} \tag{7.17}$$

  - Points falling below the lower control limit are considered to be out of control.
- Process is out of control–some points fall above the UCL.