Contents

Chapter 2

Chapter 2 Name Here

2.1 Tabular . . .

2.2 Summarizing Qualitative Data

- A bar chart depicts the frequency or the relative frequency for each category of the qualitative data as a bar rising vertically from the horizontal axis.
- For example, Adidas' sales may be proportionally compared for each Region over these two periods.
- A frequency distribution for quantitative data groups data into intervals called classes, and records the number of observations that fall into each class.
- Guidelines when constructing frequency distribution:
  - Classes are *mutually exclusive*.
  - Classes are *exhaustive*.
- The number of classes usually ranges from 5 to 20.
- Approximating the class width:

$$\frac{\text{Largest Value - Smallest Value}}{\text{\# of classes}}$$

- The
- A cumulative frequency distribution specifies how many observations fall below the upper limit of a particular class.

Chapter 3

Numerical Descriptive Measures

3.1 Investment Decision

Table 3.1: Investment Decision

Year	**Metals**	**Income**	**Year**	**Metals**	**Income**
2000	-7.34	4.07	2005	43.79	3.12
2001	18.33	6.52	2006	34.30	8.15
2002	33.35	9.38	2007	36.13	5.44
2003	59.45	18.62	2008	-56.02	-11.37
2004	8.09	9.44	2009	76.46	31.77

- Rebecca would like to
  1. Determine the typical return of the mutual funds.
  2. Evaluate the investment risk of the mutual funds.
- As an investment counselor at a large bank, Rebecca Johnson was asked by an inexperienced investor to explain the differences between two top-performing mutual funds:
  - Vanguard's Precious Metals and Mining fund (Metals)
  - Fidelity's Strategic Income Fund (Income)
- The investor has collected sample returns for these two funds for years 2000 through 2009. These data are presented in the next slide.

3.2 Measures of Central Location

3.2.1 Mean

- The arithmetic mean is a primary measure of central location.

- Sample mean $\bar{x}$

$$\bar{x} = \frac{1}{n}\sum x_i \tag{3.1}$$

- Population mean μ

$$\mu = \frac{1}{N}\sum x_i \tag{3.2}$$

Metals fund metal return:

$$\frac{-7.34 + 18.33 + 33.35 + 59.45 + 8.09 + 43.79 + 34.30 + 36.13 - 56.02 + 76.46}{10} = \frac{246.54}{10} = 24.654\%$$

Income fund mean return:

$$\frac{4.07 + 3.12 + 6.52 + 8.15 + 9.38 + 5.44 + 18.62 - 11.37 + 9.44 + 31.77}{10} = \frac{85.14}{10} = 8.514\%$$

3.2.2 Median

3.2.3 Mode

3.2.4 Percentiles and Box Plots

- In general, the pth percentile divides a data set into two parts:
  - Approximately p% of the observations have values less than the pth percentile;
  - Approximately $(100 - p)$% of the observations have values greater than the pth percentile.
- Calculating the pth percentile:
  - First, arrange the data in ascending order.
  - Locate the position, L_p, of the pth percentile ...
- Consider the sorted data from the introductory case.

Position	1	2	3	4	5	6	7	8	9	10
Value	4.07	3.12	6.52	8.15	9.38	5.44	18.62	-11.37	9.44	31.77

- For the 25th percentile, we locate
- A box plot allows you to:
  - Graphically display the distribution of a data set.
  - Compare two or more distributions.
  - Identify outliers in a data set.
- Detecting outliers

  - Calculate IQR $= Q_3 - Q_1$
  - Calculate $1.5 \times IQR$
  - There are outliers if:
    * $Q_1 - \min > 1.5$IQR, or if
    * $\max - Q_3 > 1.5$IQR, or if

3.2.5 Geometric Mean

For multiperiod returns $R_1, R_2, \ldots, R_n$, the geometric mean return G_R is calculated as:

$$G_R = \sqrt[n]{(1+R_1)(1+R_2)\ldots(1+R_n)} - 1 \tag{3.3}$$

where n is the number of multiperiod returns.

3.3 Measures of Dispersion

- Measures of dispersion gauge the variability of a data set.
- Measures of dispersion include:
  - Range
  - Mean Absolute Deviation (MAD)
  - Variance and Standard Deviation
    * In finance, standard deviation of a return is known as volatility
  - Coefficient of Variation (CV)

3.3.1 Mean Absolute Deviation (MAD)

- MAD is an average of the absolute difference of each observation from the mean.

Sample MAD:

$$\text{Sample MAD} = \frac{\sum |x_i - \bar{x}|}{n} \tag{3.4}$$

$$\text{Population MAD} = \frac{\sum |x_i - \mu|}{N} \tag{3.5}$$

3.3.2 Variance and Standard Deviation

For a given sample,

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \text{ and } s = \sqrt{s^2} \tag{3.6}$$

where s is the sample standard deviation and s^2 is the sample variance.

For a given population:

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} \text{ and } \sigma = \sqrt{\sigma^2} \tag{3.7}$$

where σ is the standard deviation and σ^2 is the variance.

Table 3.2: Volatility Index (VIX)

Period	**Typical VIX Levels**	**What It Means**
Quiet Markets	$\sim 10 - -15$	Low fear, high confidence in equity returns
Normal Conditions	$\sim 15 - -25$	Modest uncertainty, moderate stability
Crisis Conditions	> 30	Elevated fear–uncertain markets
Peak Crises	$> 60 - -80$	Extreme panic or sharp dislocations

3.3.3 Coefficient of Variation (CV)

- CV adjusts for differences in the magnitudes of the means.
- CV is unitless, allowing easy comparison of mean-adjusted dispersion across different data sets.

$$\text{Sample CV} = \frac{s}{\bar{x}} \tag{3.8}$$

$$\text{Population CV} = \frac{\sigma}{\mu} \tag{3.9}$$

3.4 Sharpe Ratio

- Measures the extra reward per unit of risk.
- For an investment I, the Sharpe ratio is computed as

$$\text{Sharpe Ratio} = \frac{\bar{x}_I - \bar{R}_f}{s_I} \tag{3.10}$$

where $\bar{x}_I$ is the mean return for the investment, $\bar{R}_f$ is the mean return for a risk-free asset, and s_I is the standard deviation for the investment.

3.5 Chebyshev's Theorem and the Empirical Rule

- Chebyshev's Theorem – For any data set, the proportion of observation that lie within k standard deviations from the mean is at least $1 - \frac{1}{k^2}$, where k is any number greater than 1.
- Consider a large lecture class with 280 students. The mean score on an exam is 74 with a standard deviation of 8. At least how many students scored within 85 and 90?
- With $k = 2$, we have $1 - \frac{1}{2^2} = 0.75 \ldots$

3.5.1 The Empirical Rule

- Approximately 68% of all observations fall in the interval $\bar{x} \pm s$.
- Approximately 95% of all observations fall in the interval $\bar{x} \pm 2s$.
- Approximately 99.7% of all observations fall in the interval $\bar{x} \pm 3s$.

Chapter 4

Introduction to Probability

4.1 Sportsware Brands

- Annabel Gonzalez, chief retail analyst at marketing firm Longmeadow Consultants is tracking the sales of compression-gear produced by Under Armour, Inc., Nike, Inc, and Adidas Group.
- After collecting data from 600 recent purchases, Annabel wants to determine whether age influences brand choice.

	Brand Name		
Age Group	**Under Armour**	**Nike**	**Adidas**
Under 25 years	174	132	90
35 years and older	54	72	78

Table 4.1: Live Example for Chapter 4

4.2 Fundamental Probability Concepts

- A probability is a numerical value that . . .
- An experiment

4.2.1 Assigning Probabilities

Subjective Probabilities

- Draws on personal and subjective judgment.

Objective Probabilities

- Empirical probability: a relative frequency of occurrence
- a priori probability: a logical analysis

4.2.2 Probabilities expressed as odds

Percentages and **odds** are an alternative approach to expressing probabilities include.

4.2.3 Converting an odds ratio to a probability

Given **odds** for event A occurring of "a to b", the probability of A is:

$$\frac{a}{a+b}$$

Given **odds** against event A occurring of "a to b", the probability of A is:

$$\frac{b}{a+b}$$

4.2.4 Converting probability to an odds ratio

4.3 Rules of Probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{4.1}$$

4.3.1 Multiplication Rule

$$P(A \cap B) = P(A|B) \times P(B) = P(B|A) \times P(A) \tag{4.2}$$

4.4 Contingency Tables and Probabilities

4.4.1 Contingency Tables

- A contingency table generally shows frequencies for two qualitative ...

4.5 Bayes' Rule

$$\begin{aligned} P(B|A) &= \frac{P(A \cap B)}{P(A \cap B) + P(A \cap B^c)} \\ &= \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B)^c} \end{aligned} \tag{4.3}$$

Prior Probability	Conditional Probability	Joint Probability	Posterior Probability
$P(T) = 0.99$			
$P(T^c) = 0.01$			
$P(T) + P(T^c) = 1$			

Table 4.2: Bayes' Rule Example

We find:

$$\begin{aligned} P(T|D) &= \frac{(0.005)(0.99)}{(0.005)(0.99) + (0.95)(0.01)} \\ &= \frac{0.00495}{0.00495 + 0.0095} \\ &= \frac{0.00495}{0.01445} \\ &= 0.342560554 \end{aligned}$$

4.6 Counting Rules

$$_nC_x = \begin{pmatrix} n \\ x \end{pmatrix} = \frac{n!}{(n-x)!x!} \tag{4.4}$$

$$_nP_x = \dots \tag{4.5}$$

Chapter 5

Discrete Random Variables

LO 5.1: Distinguish between discrete and continuous random variables.

LO 5.2: Describe the probability distribution of a discrete random variable.

LO 5.3: Calculate and interpret summary measures for a discrete random variable.

LO 5.4: Differentiate among risk neutral, risk averse, and risk loving consumers.

LO 5.5: Compute summary measures to evaluate portfolio returns.

LO 5.6: Describe the binomial distribution and compute relevant probabilities.

LO 5.7: Describe the Poisson distribution and compute relevant probabilities.

5.1 Random Variables and Discrete Probability Distributions

LO 5.1 **Distinguish between discrete and continuous random variables.**

- Random Variable
  - A function that assigns numerical values to the outcomes of a random experiment.
  - Denoted by uppercase letters (e.g., X)
- Values of the random variable are denoted by corresponding lowercase letters.
  - Corresponding values of the random variable: $x_1, x_2, x_3, \ldots$
- Random variables may be classified as:

  Discrete The random variable assumes a countable number of distinct values.

  Continuous The random variable is characterized by (infinitely) uncountable values within any interval.

- Consider an experiment in which two shirts are selected from the production line and each can be defective (D) or non-defective (N).
  - Here is the sample space:
  - The random variable X is the number of defective shirts.
  - The possible number of defective shirts is the set $\{0, 1, 2\}$.
- Since these are the only possible outcomes, this is a discrete random variable.

LO 5.2 **Describe the probability distribution of a discrete random variable.**

- Every random variable is associated with a probability distribution that describes the variable completely.
  - A probability mass function is used to describe discrete random variables.
  - A probability density function is used to describe continuous random variables.
  - A cumulative distribution function may be used to describe both discrete and continuous random variables.
- The probability mass function of a discrete random variable X is a list of the values of X with the associated probabilities, that is, the list of all possible pairs:

$$(x, P(X = x)) \tag{5.1}$$

- The cumulative distribution function of X is defined as

$$P(X \leq x) \tag{5.2}$$

- Two key properties of discrete probability distributions:
  - The probability of each value x is a value between 0 and 1, or equivalently

    $$0 \leq P(X = x) \leq 1$$

  - The sum of the probabilities equals 1. In other words,

    $$\sum_i P(X = x_i) = 1$$

    where the sum extends over all values x_i of X.
- A discrete probability distribution may be viewed as a table, algebraically, or graphically.

- For example, consider the experiment of rolling a six-sided die. A tabular presentation is:

Table 5.1: Tabular representation of rolling a six-sided die.

x	1	2	3	4	5	6
$P(X=x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

- Each outcome has an associated probability of $\frac{1}{6}$. Thus, the pairs of values and their probabilities form the probability mass function for X.
- Another tabular view of a probability distribution is based on the cumulative probability distribution.
  - For example, consider the experiment of rolling a six-sided die. The cumulative probability distribution is

Table 5.2: Tabular cumulative probability distribution of rolling a six-sided die.

x	1	2	3	4	5	6
$P(X=x)$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{6}$	$\frac{4}{6}$	$\frac{5}{6}$	$\frac{6}{6}$

- The cumulative probability gives the probability of X being less than or equal to x. For example, $P(x \leq 4) = \frac{4}{6} = \frac{2}{3}$.
- A probability distribution may be expressed algebraically.
- For example, for the sid-sided die experiment, the probability distribution of the random variable X is:

$$P(X=x) = \begin{cases} \frac{1}{6} & \text{if } x = 1,2,3,4,5,6 \\ 0 & \text{otherwise.} \end{cases}$$

- Using this formula, we can find

$$P(X=5) = \frac{1}{6}$$
$$P(X=7) = 0$$

- A probability distribution may be expressed graphically.
  - The values x of X are placed on the horizontal axis and the associated probabilities on the vertical axis.
  - A line is drawn such that its height is associated with the probability of x.
  - ...
  - This is a uniform distribution since the bar heights are all the same.

5.2 Expected Value, Variance, and Standard Deviation

LO 5.3 **Calculate and interpret summary measures for a discrete random variable.**

- Summary measures for a random variable include the
  - Mean (Expected Value)
  - Variance
  - Standard Deviation

5.2.1 Expected Value

$$\text{Expected Value} \Leftrightarrow \text{Population Mean}$$
$$E(X) \Leftrightarrow \mu$$

- $E(X)$ is the long-run average value of the random variable over infinitely many independent repetitions of an experiment.
- For a discrete random variable X with values $x_1, x_2, x_3, \ldots$ that occur with probabilities $P(X = x_i)$, the expected value of X is

$$\begin{aligned} E(X) &= \mu \\ &= \sum_i x_i \times P(X = x_i) \end{aligned} \tag{5.3}$$

5.2.2 Variance and Standard Deviation

- For a discrete random variable X with values $x_1, x_2, x_3, \ldots$ that occur with probabilities $P(X = x)$,

$$\begin{aligned} \text{Var}(X) = \sigma^2 &= \sum_i (x_i - \mu)^2 P(X = x_i) \\ &= \sum_i x_i^2 P(X = x_i) - \mu_2 \end{aligned} \tag{5.4}$$

- The standard deviation is the square root of the variance.

$$\text{SD}(X) = \sigma = \sqrt{\sigma^2} \tag{5.5}$$

5.2.3 Risk Neutrality and Risk Aversion

LO 5.4 **Differentiate among risk neutral, risk averse, and risk loving consumers.**

- Risk average consumers:

  - Expect a reward for taking a risk.
  - May decline a risky prospect even if it offers a positive expected gain.
- Risk neutral consumers:
  - Completely ignore risk.
  - Always accept a prospect that offers a positive expected gain.
- Risk loving consumers:
  - May accept a risky prospect even if the expected gain is negative.

5.2.4 Application of Expected Value to Risk

- Suppose you have a choice of receiving \$1,000 in cash or receiving a beautiful painting from your grandmother.
- The actual value of the painting is uncertain. Here is a probability distribution of the possible worth of the painting. What should you do?

Table 5.3: Painting Value Probabilities

x	$P(X = x)$
\$2,000	0.20
\$1,000	0.50
\$500	0.30

5.3 Portfolio Returns

LO 5.5 **Compute summary measures to evaluate a portfolio's return.**

- Investment opportunities often use both:
  - Expected return as a measure of reward.
  - Variance or standard deviation of return as a measure of risk.
- Portfolio is defined as a collection of assets such as stocks and bonds.
  - Let X and Y two random variables of interest, denoting, say, the returns of two assets.
  - Since an investor may have invested in both assets, we would like to evaluate the portfolio return formed by a linear combination of X and Y.

5.3.1 Properties of random variables useful in evaluating portfolio returns

- Given two random variables X and Y,
  - The expected value of X and Y is

$$E(X+Y)=E(X)+E(Y) \tag{5.6}$$

  - The variance of X and Y is

$$\text{Var}(X+Y)=\text{Var}(X)+\text{Var}(Y)+2\text{Cov}(X,Y) \tag{5.7}$$

    where $\text{Cov}(X,Y)$ is the covariance between X and Y.
  - For constants a, b, the formulas extend to

$$\begin{aligned}E(aX+bY)&=aE(X)+bE(Y)\\ \text{Var}(aX+bY)&=a^2\text{Var}(X)+b^2\text{Var}(Y)+2ab\text{Cov}(X,Y)\end{aligned}$$

5.3.2 Expected return, variance, and standard deviation of portfolio returns

- Given a portfolio with two assets, Asset A and Asset B, the expected return of the portfolio $E(R_p)$ is computed as:

$$E(R_p)=w_AE(R_A)+w_BE(R_B) \tag{5.8}$$

  where w_A and w_B are the portfolio weights, $w_A+w_B=1$, and $E(R_A)$ and $E(R_B)$ are the expected returns on assets A and B, respectively.

- Using the covariance or the correlation coefficient of the two returns, the portfolio variance of return is:

$$\text{Var}(R_p)=w_A^2\sigma_A^2+w_B^2\sigma_B^2+2w_Aw_B\rho_{AB}\sigma_A\sigma_B \tag{5.9}$$

  where σ_A^2 and σ_B^2 are the variances of the returns for Asset A and Asset B, respectively, σ_{AB} is the covariance between the returns for Assets A and B, and ρ_{AB} is the correlation coefficient between the returns for Asset A and Asset B.

$$\rho_{AB}=\frac{\sigma_{AB}}{\sigma_A\sigma_B} \tag{5.10}$$

5.4 The Binomial Probability Distribution

LO 5.6 **Describe the binomial distribution and compute relevant probabilities.**

- A binomial random variable is defined as the number of successes achieved in the n trials of a Bernoulli process.
  - A Bernoulli process consists of a series of n independent and identical trials of an experiment such that on each trial:
    * There are only two possible outcomes:
      p probability of a success
      $1-p=q$ probability of a failure
    * Each time the trial is repeated, the probabilities of success and failure remain the same.
- A binomial random variable X is defined as the number of successes achieved in the n trials of a Bernoulli process.
- A binomial probability distribution shows the probabilities associated with the possible values of the binomial random variable (that is, 0, 1, . . . , n).
  - For a binomial random variable X, the probability of x successes in n Bernoulli trials is:

$$\begin{aligned} P(X=x) &= \binom{n}{x} p^x q^{n-x} \\ &= \frac{n!}{(n-x)!x!} p^x q^{n-x} \end{aligned} \tag{5.11}$$

    for $x = 0, 1, 2, \ldots, n$.
- For a binomial distribution:
  - The expected value $E(X)$ is:

$$E(X) = \mu = np \tag{5.12}$$

  - The variance $\mathrm{Var}(X)$ is:

$$\mathrm{Var}(X) = \sigma^2 = npq \tag{5.13}$$

  - The standard deviation $\mathrm{SD}(X)$ is:

$$\mathrm{SD}(X) = \sigma = \sqrt{npq} \tag{5.14}$$

5.5 The Poisson Probability Distribution

LO 5.7 **Descrive the Poisson distribution and compute relevant probabilities.**

- A binomial random variable counts the number of successes in a fixed number of Bernoulli trials.

- In contrast, a Poisson random variable counts the number of successes over a given interval of time or space.
  - Examples of a Poisson random variable include:

    With respect to time the number of cars that cross the Brooklyn Bridge between 9:00 am and 10:00 am on a Monday morning.

    With respect to space the number of defects in a 50-year roll of fabric.
- A random experiment satisfies a Poisson process if:
  - The number of successes within a specified time or space interval equals any integer between 0 and ∞.
  - The number of successes in non-overlapping intervals are independent.
  - The probability that successes occurs in any interval is the same for all intervals of equal size and is proportional to the size of the interval.
- For a Poisson random variable X, the probability of x successes over a given interval of time or space is:

$$P(X = x) = \frac{e^{-\mu}\mu^x}{x!} \text{ for } x = 0, 1, 2, \ldots \tag{5.15}$$

  where μ is the mean number of successes and $e \approx 2.718$ is the base of the natural logarithm.
- For a Poisson distribution:
  - The expected value $E(X)$ is:

$$E(X) = \mu \tag{5.16}$$

  - The variance $\mathrm{Var}(X)$ is:

$$\mathrm{Var}(X) = \sigma^2 = \mu \tag{5.17}$$

  - The standard deviation $\mathrm{SD}(X)$ is:

$$\mathrm{SD}(X) = \sigma = \sqrt{\mu} \tag{5.18}$$

Chapter 6

Continuous Random Variables

LO 6.1: Describe a continuous random variable.

LO 6.2: Describe a continuous uniform distribution and calculate associated probabilities.

LO 6.3: Explain the characteristics of the normal distribution.

LO 6.4: Use the standard normal table of the z-table.

LO 6.5: **Calculate and interpret probabilities or a random variable that follows the normal distribution.**

LO 6.6: **Calculate and interpret probabilities or a random variable that follows the exponential distribution.**

LO 6.7: **Calculate and interpret probabilities or a random variable that follows the lognormal distribution.**

6.1 Continuous Random Variables and the Uniform Probability Distribution

LO 6.1 **Describe a continuous random variable.**

- Remember that random variables may be classified as

  Discrete The random variable assumes a countable number of distinct values.

  Continuous The random variable is characterized by (infinitely) uncountable values within any interval.

- When computing probabilities for a continuous random variable, keep in mind that $P(X = x) = 0$.
  - We cannot assign a nonzero probability to each infinitely uncountable value and still have the probabilities sum to one.

- Thus, since $P(X = a)$ and $P(X = b)$ both equal zero, the following holds true for continuous random variables:

$$P(a \leq X \leq b) = P(a < X < b) = P(a \leq X < b) = P(a < X \leq b)$$

6.1.1 Probability Density Function $f(x)$ of a continuous random variable X

- Describes the relative likelihood that X assumes a value within a general interval (e.g., $P(a \leq X \leq b)$), where
  - $f(x) > 0$ for all possible values of X.
  - The area under $f(x)$ over all values of x equals 1.

6.1.2 Cumulative Density Function $F(x)$ of a continuous random variable X

- For any value x of the random variable X, the cumulative distribution function $F(x)$ is computes as:

$$F(X) = P(X \leq x)$$

- As a result:

$$P(a \leq X \leq b) = F(b) - F(a)$$

6.1.3 The Continuous Uniform Distribution

LO 6.2 **Describe a continuous uniform distribution and calculate associated probabilities.**

- Describe a random variable that has an equally likely chance of assuming a value within a specified range.
- Probability density function:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b, \text{ and} \\ 0 & \text{for } x < a \text{ or } x > b \end{cases} \tag{6.1}$$

  where a and b are the lower and upper limits, respectively.

- The expected value and standard deviation of X are:

$$E(X) = \mu = \frac{a+b}{2} \tag{6.2}$$

$$\begin{aligned} \mathrm{SD}(X) &= \sigma \\ &= \sqrt{\frac{(b-a)^2}{12}} \end{aligned} \tag{6.3}$$

6.1.4 Graph of the continuous uniform distribution

- The values of a and b on the horizontal axis represent the lower and upper limits, respectively.
- The height of the distribution does not directly represent a probability.
- It is the area under $f(x)$ that corresponds to probability.

Cumulative function:

$$\begin{aligned} P(X > x_1) &= \text{base} \times \text{height} \\ &= (b - x_1) \times \frac{1}{b-a} \end{aligned}$$

6.2 The Normal Distribution

- For a random variable X with mean μ and variance σ^2:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{6.4}$$

6.2.1 The Normal Distribution

LO 6.3 **Explain the characteristics of the normal distribution.**

- Symmetric
- Bell-shaped
- Closely approximates the probability distribution of a wide range of random variables, such as the
  - Heights and weights of newborn babies
  - Scores on SAT
  - Cumulative debt of college graduates
- Serves as the cornerstone of statistical inference.

6.2.2 Characteristics of the Normal Distribution

- Symmetric about its mean
  - Mean = Median = Mode
- Asymptotic–that is, the tail gets closer and closer to the horizontal axis but never touches it.
- The normal distribution is completely described by two parameters: μ and σ^2.

  μ is the population mean which describes the central location of the distribution.
  σ^2 is the population variance which describes the dispersion of the distribution.

6.2.3 Probability Density Function of the Normal Distribution

- For a random variable X with mean μ and variance σ^2:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{6.5}$$

6.2.4 The Standard Normal (Z) Distribution

LO 6.4 **Use the standard normal table of the z-table.**

- A special case of the normal distribution:
  - Mean μ is qual to zero ($E(X) = 0$).
  - Standard deviation σ is equal to 1 ($\text{SD}(Z) = 1$).

6.2.5 Standard Normal Table (Z-Table)

- Gives the cumulative probabilities $P(Z \leq z)$ for positive and negative values of z.
- Since the random variable Z is symmetric about its mean of 0,

$$P(Z < 0) = P(Z > 0) = 0.5$$

- To obtain the $P(Z < z)$, read down the z-column first, then across the top.

6.2.6 Finding the Probability for a Given z-Value

- Transform normally distributed random variables into standard normal random variables and use the z-table to compute the relevant probabilities.
- The z-table provides cumulative probabilities $P(Z \leq z)$ for a given z.

6.3 Solving Problems with the Normal Distribution

LO 6.5 **Calculate and interpret probabilities or a random variable that follows the normal distribution.**

6.3.1 The Normal Transformation

- Any normally distributed random variable X with mean μ and standard deviation σ can be transformed into the standard normal random variable Z as:

$$Z = \frac{X-\mu}{\sigma} \text{ with corresponding values } z = \frac{x-\mu}{\sigma} \tag{6.6}$$

- As constructed: $E(Z) = 0$ and $\mathrm{SD}(Z) = 1$.
- A z-value specifies by how many standard deviations the corresponding x value falls above ($z > 0$) or below ($z < 0$) the mean.
  - A positive z indicates by how many standard deviations the corresponding x lies above μ.
  - A zero z indicates that the corresponding x equals μ.
  - A negative z indicates by how many standard deviations the corresponding x lies below μ.

6.3.2 Use the Inverse Transformation to Compute Probabilities for Given x values

- A standard normal variable Z can be transformed to the normally distributed random variable X with mean μ and standard deviation σ as

$$X = \mu + Z\sigma \text{ with corresponding values } x = \mu + z\sigma \tag{6.7}$$

6.4 Other Continuous Probability Distributions

LO 6.6 **Calculate and interpret probabilities or a random variable that follows the exponential distribution.**

6.4.1 Exponential Distribution

- A random variable X follows the exponential distribution if its probability density function is:

$$f(x) = \lambda e^{-\lambda x} \text{ for } x \geq 0 \tag{6.8}$$

  where λ is the rate parameter and $E(X) = \mathrm{SD}(X) = \frac{1}{\lambda}$.
- The cumulative distribution function is:

$$P(X \leq x) = 1 - e^{-\lambda x} \tag{6.9}$$

6.4.2 The Lognormal Distribution

LO 6.7 **Calculate and interpret probabilities or a random variable that follows the lognormal distribution.**

- Defined for a positive random variable, the lognormal distribution is positively skewed.
- Useful for describing variables such as

  - Income
  - Real estate values
  - Asset prices

- Failure rate may increase or decrease over time.
- Let X be a normally distributed random variable with mean μ and standard deviation σ. The random variable $Y = e^X$ follows the lognormal distribution with a probability density function as

$$f(y) = \frac{1}{y\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln(y)-\mu)^2}{2\sigma^2}\right) \text{ for } y > 0 \tag{6.10}$$

- The lognormal distribution is clearly positively skewed for $\sigma > 1$. For $\sigma < 1$, the lognormal distribution somewhat resembles to normal distribution.

6.4.3 Expected values and standard deviations of the lognormal and normal distributions

- Let X be a normal random variable with mean μ and standard deviation σ and let $Y = e^X$ by the corresponding lognormal variable. The mean μ_y and standard deviation σ_Y or Y are derived as:

$$\mu_Y = \exp\left(\frac{2\mu + \sigma^2}{2}\right) \tag{6.11}$$

$$\sigma_Y = \sqrt{(\exp(\sigma^2) - 1)\exp\left(2\mu + \sigma^2\right)} \tag{6.12}$$

- Equivalently, the mean and standard deviation of the normal variable $X = \ln(Y)$ are derived as

$$\mu = \ln\left(\frac{\mu_Y^2}{\sqrt{\mu_Y^2 + \sigma_Y^2}}\right) \tag{6.13}$$

$$\sigma = \sqrt{\ln\left(1 + \frac{\sigma_Y^2}{\mu_Y^2}\right)} \tag{6.14}$$

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

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

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.

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 standard 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 $\pm$ 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 (7.6), 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.
- Confidence Intervals:
  - 90%, $\alpha = 0.10$, $\frac{\alpha}{2} = 0.05$, $z_{0.05} = 1.645$.
  - 95%, $\alpha = 0.05$, $\frac{\alpha}{2} = 0.025$, $z_{0.025} = 1.96$.
  - 99%, $\alpha = 0.01$, $\frac{\alpha}{2} = 0.005$, $z_{0.005} = 2.575$.

8.2.2 Interpreting a Confidence Interval

- Interpreting a confidence interval requires care.
- Incorrect: the probability that μ falls in the interval is 0.95.

- Correct: If numerous samples of size n are drawn from a given population, then 95% of the intervals formed by the ____ $\bar{x} \pm z_{\alpha/2}\frac{\sigma}{n}$ will contain μ.
  - Since there are many possible samples, we will be right 95% of the time, thus giving us 95% confidence.

8.2.3 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 σ, and
  - Confidence level $100(1-\alpha)\%$.

8.2.4 Summary of the t_{df} Distribution

- Bell-shaped and symmetric around 0 with asymptotic tails (the tails get closer and closer to the horizontal axis, but never touch it).
- Has slightly broader tails than the z distribution.
- Consists of a family of distributions where the actual shape of each one depends on the df. As df increases, the t_{df} distribution becomes similar to the z distribution; it is identical to the z distribution when $df \to \infty$.

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 extent of the broadness of the tails of the distribution; the fewer the degrees of freedom, the broader the tails.

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

8.3.2 Constructing a Confidence Interval for μ When σ Is Unknown

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

- A $100(1-\alpha)\%$ confidence interval of the population mean μ when the population standard deviation σ is not known is ____

$$\bar{x} \pm t_{\alpha/2,df}\frac{s}{\sqrt{n}} \tag{8.5}$$

or equivalently

$$\left[\bar{x} - t_{\alpha/2,df}\frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2,df}\frac{s}{\sqrt{n}}\right] \tag{8.6}$$

where s is the sample standard deviation.

8.4 Confidence Interval of the Population Proportion

LO 8.7 Calculate a confidence interval for the population proportion.

- Let the parameter p represent the proportion of successes in the population, where success is defined by a particular output.
  - p is the point estimator of the population proportion p.
- By the central limit theorem, $\bar{P}$ can be approximated by a normal distribution for large samples (i.e., $np \geq 5$ and $n(1-p) \geq 5$).
- Thus, a $100(1-\alpha)\%$ confidence interval of the population proportion is

$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \text{ or } \left[\bar{p} - z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}, \bar{p} + z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}\right] \tag{8.7}$$

where $\bar{p}$ is used to estimate the population parameter p.

8.5 Selecting a Useful Sample Size

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

- **Precision** in interval estimates is implied by a low margin of error.
- The larger n reduces the margin of error for the interval estimates.
- How large should the sample size by for a given margin of error?

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
$$D = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \tag{8.8}$$
we may rearrange to get
$$n = \left(\frac{z_{\frac{\alpha}{2}} \sigma}{D} \right)^2. \tag{8.9}$$
- 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
$$n = \left(\frac{z_{\frac{\alpha}{2}} \hat{\sigma}}{D} \right)^2. \tag{8.10}$$
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
$$D = z_{\frac{\alpha}{2}} \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \tag{8.11}$$
(where $\bar{p}$ is the sample proportion), we may rearrange to get
$$n = \left(\frac{z_{\alpha/2}}{D} \right)^2 \bar{p}(1-\bar{p}) \tag{8.12}$$
- 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 p is
$$n = \left(\frac{z_{\alpha/2}}{D} \right)^2 p(1-p) \tag{8.13}$$
where $\hat{p}$ is a reasonable estimate of p in the planning stage.

Chapter 9

Hypothesis Testing

LO 9.1: Define the null hypothesis and the alternative hypothesis.

LO 9.2: Distinguish between Type I and Type II errors.

LO 9.3: Explain the steps of a hypothesis test using the p-value approach.

LO 9.4: Explain the steps of a hypothesis test using the critical value approach.

LO 9.5: Differentiate between the test statistics for the population mean.

LO 9.6: Specify the test statistic for the population proportion.

9.1 Point Estimators and Their Properties

LO 9.1 **Define the null hypothesis and the alternative hypothesis.**

- Hypothesis tests resolve conflicts between two competing opinions (hypotheses).
- In a hypothesis test, define

  H_0 the null hypothesis, the presumed default state of nature or status quo.

  H_A the alternative hypothesis, a contradiction of the default state of nature or status quo.

- In statistics, we use sample information to make inferences regarding the unknown population parameters of interest.
- We conduct hypothesis tests to determine if sample evidence contradicts H_0.
- On the basis of sample information, we either
  - "Reject the null hypothesis"
    - Sample evidence **is** inconsistent with H_0.

- "Do not reject the null hypothesis"
  - Sample evidence **is not** inconsistent with H_0.
  - We do not have enough evidence to "accept" H_0.

9.1.1 Defining the Null Hypothesis and Alternative Hypothesis

General guidelines:

- Null hypothesis, H_0, states the status quo.
- Alternative hypothesis, H_A, states whatever we wish to establish (i.e., contests the status quo)
- Note that H_0 always contains the "equality".

9.1.2 One-Tailed vs Two-Tailed Hypothesis Tests

Two-Tailed Test

- Reject H_0 on either side of the hypothesized value of the population parameter.
- For example:
  - H_0: $\mu = \mu_0$ versus H_A: $\mu \neq \mu_0$
  - H_0: $p = p_0$ versus H_A: $p \neq p_0$
- The $\neq$ symbol in H_A indicates that both tail areas of the distribution will be used to make the decision regarding the rejection of H_0.

One-Tailed Test

- Reject H_0 only on one side of the hypothesized value of the population parameter.
- For example:
  - H_0: $\mu \leq \mu_0$ versus H_A: $\mu > \mu_0$ (right-tail test)
  - H_0: $\mu \geq \mu_0$ versus H_A: $\mu < \mu_0$ (left-tail test)
- Note that the inequality in H_A determines which tail area will be used to make the decision regarding the rejection of H_0.

9.1.3 Three Steps to Formulate Hypotheses

1. Identify the relevant population parameter of interest (e.g., μ or p).
2. Determine whether it is a one- or a two-tailed test.
3. Include some form of the equality sign in H_0 and use H_A to establish a claim.

H_0	H_A	Test Type
$=$	$\neq$	Two-tail
$\geq$	$<$	One-tail, Left-tail
$\leq$	$>$	One-tail, Right-tail

9.1.4 Type I and Type II Errors

LO 9.2 **Distinguish between Type I and Type II errors.**

- **Type I Error** – Committed when we reject H_0 when H_0 is actually true.
  - Occurs with probability α. α is chosen **a priori**.
- **Type II Error** – Committed when we do not reject H_0 when H_0 is actually false.
  - Occurs with probability β. Power of the test $= 1 - \beta$
- For a given sample size n, a decrease in α will increase β and vice versa.
- Both α and β decreases as n increases.

Decision	**Null hypothesis is true**	**Null hypothesis is false**
Reject the null hypothesis	Type I error	Correct decision
Do not reject the null hypothesis	Correct decision	Type I error

9.2 Hypothesis Test of the Population Mean When σ Is Known

LO 9.3 **Explain the steps of a hypothesis test using the p-value approach.**

- Hypothesis testing enables us to determine whether the sample evidence is inconsistent with what is hypothesized under the null hypothesis (H_0).
- Basic principle: First assume that H_0 is true and then determine if sample evidence contradicts this assumption.
- Two approaches to hypothesis testing:
  - The p-value approach.
  - The critical value approach.

9.2.1 The p-value Approach

- The value of the test statistic for the hypothesis test of the population mean μ when the population standard deviation σ is known is computed as

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \tag{9.1}$$

where μ_0 is the hypothesized mean value.

- p-value: the likelihood of obtaining a sample mean that is at least as extreme as the one derived from the given sample, under the assumption that the null hypothesis is true.
- Under the assumption that $\mu = \mu_0$, the p-value is the likelihood of observing a sample mean that is at least as extreme as the one derived from the given sample.
- The calculation of the p-value depends on the ______.

Alternative hypothesis	p-value
$H_A : \mu > \mu_0$	Right-tail probability: $P(Z \geq z)$
$H_A : \mu < \mu_0$	Left-tail probability: $P(Z \leq z)$
$H_A : \mu \neq \mu_0$	Two-tail probability: $2P(Z \geq z)$ if $z > 0$ or $2P(Z \leq z)$ if $z < 0$

- Decision rule: Reject H_0 if p-value $< \alpha$.

9.2.2 Four Step Procedure Using the p-value Approach

Step 1. Specify the null and the alternative hypotheses.

Step 2. Specify the test statistic and compute its value.

Step 3. Calculate the p-value.

Step 4. State the conclusion and interpret the results.

LO 9.4 **Explain the steps of a hypothesis test using the critical value approach.**

9.2.3 The Critical Value Approach

- **Rejection region** – a region of values such that if the test statistic falls into this region, then we reject H_0.
  - The location of this region is determined by H_A.

- **Critical value** – a point that separates the rejection region from the nonrejection region.
- The critical value approach specifies a region such that if the value of the test statistic falls into the region, the null hypothesis is rejected.
- The critical value depends on the alternative.

Alternative hypothesis	**Critical Value**
$H_A : \mu > \mu_0$	Right-tail critical value is z_α, where $P(Z \geq z_\alpha) = \alpha$
$H_A : \mu < \mu_0$	Left-tail critical value is $-z_\alpha$, where $P(Z \leq -z_\alpha) = \alpha$
$H_A : \mu \neq \mu_0$	Two-tail critical value $-z_{\alpha/2}$ and $z_{\alpha/2}$, where $P(Z \geq z_{\alpha/2}) = \frac{\alpha}{2}$

- Decision Rule: Reject H_0 if:
  - $z > z_\alpha$ for a right-tailed test
  - $z < -z_\alpha$ for a left-tailed test
  - $z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$ for a two-tailed test

9.2.4 Four Step Procedure Using the Critical Value Approach

Step 1. Specify the null and the alternative hypotheses.

Step 2. Specify the test statistic and compute its value.

Step 3. Find the critical value **or** values.

Step 4. State the conclusion and interpret the results.

9.2.5 Confidence Intervals and Two-Tailed Hypothesis Tests

- Given the significance level α, we can use the sample data to construct a $100(1 - \alpha)\%$ confidence interval for the population mean μ.
- Decision Rule
  - Reject H_0 if the confidence interval **does not** contain the value of the hypothesized mean μ_0.
  - Do not reject H_0 if the confidence interval **does** contain the value of the hypothesized mean μ_0.

9.2.6 Implementing a Two-Tailed Test Using a Confidence Interval

- The general specification for a $100(1-\alpha)\%$ confidence interval of the population mean μ when the population standard deviation σ is known as

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \text{or} \quad \left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \tag{9.2}$$

- Decision Rule: Reject H_0 if $\mu_0 < \bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ or if $\mu_0 > \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

9.3 Hypothesis Test of the Population Mean When σ Is Unknown

9.3.1 Test Statistic for μ When σ is Unknown

LO 9.5 **Differentiate between the test statistics for the population mean.**

- When the population standard deviation σ is unknown, the test statistic for testing the population mean μ is assumed to follow the t_{df} distribution with $(n-1)$ degrees of freedom (df).

- The value of t_{df} is computed as

$$t_{df} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \tag{9.3}$$

9.4 Hypothesis Test of the Population Proportion

LO 9.6 **Specify the test statistic for the population proportion.**

- $\bar{P}$ can be approximated by a normal distribution if $np \geq 5$ and $n(1-p) \geq 5$.

- Test statistic for the hypothesis test of the population proportion p is assumed to follow the z distribution:

$$z = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \tag{9.4}$$

where $\bar{p} = \frac{x}{n}$ and p_0 is the hypothesized value of the population proportion.