Student Workbook with Solutions

Heecheon You
Pohang University of Science and Technology

Applied Statistics and Probability for Engineers

Third Edition

Douglas C. Montgomery
Arizona State University

George C. Runger
Arizona State University

WILEY

JOHN WILEY & SONS, INC.

Cover Image: Norm Christiansen

To order books or for customer service call 1-800-CALL-WILEY (225-5945).

ISBN 0-471-42682-2

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

Printed and bound by Courier Kendallville, Inc.

This book is dedicated to:

Jiyeon, *Sophia*, *Jaechang*, *Sookhyun*, *Young*, and *Isik*

Preface

Objectives of the Book

As a supplement of *Applied Statistics and Probability for Engineers* (3^{rd} edition) authored by Drs. Douglas C. Montgomery and George C. Runger (abbreviated by MR hereinafter), this book is developed to serve as an effective, learner-friendly guide for students who wish to establish a solid understanding of probability and statistics. First, this study guide outlines learning goals and presents detailed information accordingly; this goal-oriented presentation is intended to help the reader see the destination before learning and check his/her achievement after learning. Second, this book helps the reader build both scientific knowledge and computer application skill in probability and statistics. Third, this book provides concise explanations, step-by-step algorithms, comparison tables, graphic charts, daily-life examples, and real-world engineering problems, which are designed to help the reader to efficiently learn concepts and applications of probability and statistics. Lastly, this book covers essential concepts and techniques of probability and statistics; thus, the reader who has interest in the advanced topics indicated as CD only in MR should refer to the MR CD.

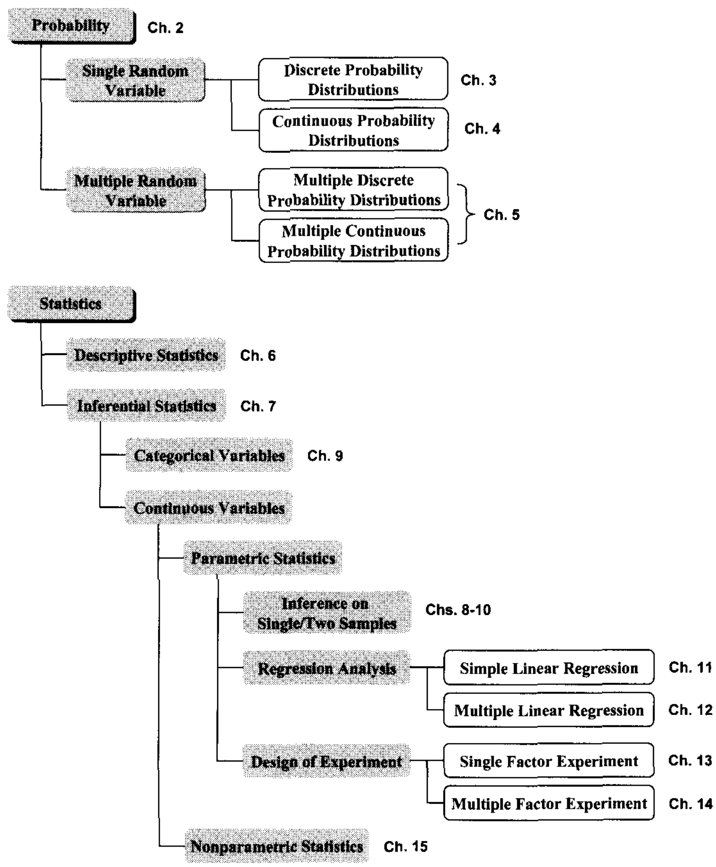
Features of the Book

The reader will enjoy the following learner-friendly features of the book:

- *Learning Goals*: Each section starts with a list of learning goals which states topics covered and competency levels targeted.
- *Organized, Concise Explanation*: According to the learning goals defined, topics are presented with concise, clear explanations.
- *Learning Aids*: Tables, charts, and algorithms are provided wherever applicable to help the reader form integrated mental pictures by contrasting or visualizing concepts and techniques.
- *Easy-to Understand Examples*: For topics requiring problem-solving practices, daily-life examples (which are readily understandable) are used to help the reader better see the application process of statistical techniques.
- *Engineering-Context Exercises*: Together with daily-life examples, engineering problems (selected from exercises in the MR textbook; the original exercise numbers are indicated) are used to have the reader explore the application of statistical techniques to the real-world engineering domain.
- *Complete Solutions of Exercises*: At the end of each chapter, complete solutions of exercises are provided so that the reader checks his/her own answers with ease.
- *Software Demonstration*: To help students learn use of Excel and Minitab (student version) in solving probability and statistics problems, respectively, step-by-step directions and screen snapshots are provided for examples as applicable.

Organization of the Book

This book presents various topics of probability and statistics with the structure displayed in the figure on next page. Chapters 2 through 5 describe probability and probability distributions. After explaining fundamental concepts, rules, and theorems of probability, this book presents various discrete and continuous probability distributions of single and multiple random variables. Then, chapters 6 through 15 explain descriptive and inferential statistics. Descriptive statistics deals with measures and charts to describe collected data, whereas inferential statistics covers categorical data analysis, statistical inference on single/two samples, regression analysis, design of experiment, and nonparametric analysis.



Acknowledgements

I would like to express my considerable gratitude to Jenny Welter for her sincere support, trust, and encouragement in developing this book. I also wish to thank Drs. Douglas C. Montgomery and George C. Runger at Arizona State University for their pleasant partnership and helpful advice in producing this book. Finally, I greatly appreciate the insightful feedback and useful suggestions from the professors and students who reviewed the various parts of the manuscript: Dr. Yasser M. Dessouky at San Jose State University; Major Brian M. Layton at U.S. Military Academy; Dr. Arunkumar Pennathur at University of Texas at El Paso; Dr. Manuel D. Rossetti at University of Arkansas; Dr. Bruce Schmeiser at Purdue University; Dr. Lora Zimmer at Arizona State University; Christi Barb, Shea Kupper, Maria Fernanda Funez-Mejia, Tram Huynh, Usman Javaid, Thuan Mai, Prabaharan Veluswamy, Cheewei Yip, and Ronda Young at Wichita State University. Without their invaluable contributions, this book may not have been completed. Peace be with them all.

Heecheon You

Contents

Chapter 7 Point Estimation of Parameters 7-1

Chapter 8 Statistical Intervals for a Single Sample 8-1

Chapter 9 Tests of Hypotheses for a Single Sample 9-1

Chapter 10 Statistical Inference for Two Samples 10-1

Chapter 11 Simple Linear Regression and Correlation 11-1

Chapter 12 Multiple Linear Regression 12-1

Chapter 13 Design and Analysis of Single-Factor Experiments: The Analysis of Variance 13-1

Chapter 14 Design of Experiment with Several Factors 14-1

Chapter 15 Nonparametric Statistics 15-1

1 The Role of Statistics in Engineering

OUTLINE

1-1 The Engineering Method and Statistical Thinking

Learning Goals

- ❑ Explain engineering problem solving process.
- ❑ Describe the application of statistics in the engineering problem solving process.
- ❑ Distinguish between enumerative and analytic studies.

Engineering Problem Solving Process

Engineers develop practical solutions and techniques for engineering problems by applying scientific principles and methodologies. Existing systems are improved and/or new systems are introduced by the engineering approach for better harmony between humans, systems, and environments.

In general, the engineering problem solving process includes the following steps:

1. **Problem definition**: Describe the problem to solve.
2. **Factor identification**: Identify primary factors which cause the problem.
3. **Model (hypothesis) suggestion**: Propose a model (hypothesis) that explains the relationship between the problem and factors.
4. **Experiment**: Design and run an experiment to test the tentative model.
5. **Analysis**: Analyze data collected in the experiment.
6. **Model modification**: Refine the tentative model.
7. **Model validation**: Validate the engineering model by a follow-up experiment.
8. **Conclusion (recommendation)**: Draw conclusions or make recommendations based on the analysis results.

(Note) Some of these steps may be iterated as necessary.

Application of Statistics

Statistical methods are applied to interpret data with variability. Throughout the engineering problem solving process, engineers often encounter data showing variability. Statistics provides essential tools to deal with the observed variability. Examples of the application of statistics in the engineering problem solving process include, but are not limited to, the following:

1. Summarizing and presenting data: numerical summary and visualization in descriptive statistics (Chapter 6).
2. Inferring the characteristics (mean, median, proportion, and variance) of single/two populations: z, t, χ^2, and F tests in parametric statistics (Chapters 7-10); signed, signed-rank, and rank-sum tests in nonparametric statistics (Chapter 15).
3. Testing the relationship between variables: correlation analysis (Chapter 11); categorical data analysis (Chapter 8).
4. Modeling the causal relationship between the response and independent variables: regression analysis (Chapters 11-12); analysis of variance (Chapters 13-14).
5. Identifying the sources of variability in response: analysis of variance (Chapters 13-14).
6. Evaluating the relative importance of factors for the response variable: regression analysis (Chapter 12); analysis of variance (Chapter 14).
7. Designing an efficient, effective experiment: design of experiment (Chapters 13-14).

These applications would lead to development of general laws and principles such as Ohm's law and design guidelines.

Enumerative vs. Analytic Studies

Two types of studies are defined depending on the use of a sample in statistical inference:

1. **Enumerative study**: Makes an inference to the well-defined population from which the sample is selected. (e.g.) defective rate of products in a lot
2. **Analytic study**: Makes an inference to a future (conceptual) population. (e.g.) defective rate of products at a production line

1-2 Collecting Engineering Data

Learning Goals

- ❑ Explain three methods of data collection: retrospective study, observational study, and designed experiment.

Data Collection Methods

Three methods are available for data collection:

1. **Retrospective study**: Use existing records of the population. Some crucial information may be unavailable and the validity of data be questioned.
2. **Observational study**: Collect data by observing the population with as minimal interference as possible. Information of the population for some conditions of interest may be unavailable and some observations be contaminated by extraneous variables.
3. **Designed experiment**: Collect data by observing the population while controlling conditions on the experiment plan. The findings would obtain scientific rigorousness through deliberate control of extraneous variables.

1-3 Mechanistic and Empirical Models

Learning Goals

- ☐ Explain the difference between mechanistic and empirical models.

Mechanistic vs. Empirical Models

Models (explaining the relationship between variables) can be divided into two categories:

1. **Mechanistic model**: Established based on the underlying theory, principle, or law of a physical mechanism.

   (e.g.) $I = \frac{E}{R} + \varepsilon$ (Ohm's law)

   where: I = current, E = voltage, R= resistance, and ε = random error

2. **Empirical model**: Established based on the experience, observation, or experiment of a system (population) under study.

   (e.g.) $y = \beta_0 + \beta_1 x + \varepsilon$

   where: y = sitting height, x = stature, and ε = random error

1-4 Probability and Probability Models

Learning Goals

- ☐ Describe the application of probability in the engineering problem solving process.

Application of Probability

Along with statistics, the concepts and models of probability are applied in the engineering problem solving process for the following:

1. Modeling the stochastic behavior of the system: discrete and continuous probability distributions.
2. Quantifying the risks involved in statistical inference: error probabilities in hypothesis testing.
3. Determining the sample size of an experiment for a designated test condition: sample size selection

2 Probability

OUTLINE

2-1 Sample Spaces and Events

Learning Goals

- ❑ Explain the terms *random experiment*, *sample space*, and *event*.
- ❑ Define the sample space and events of a random experiment.
- ❑ Define a new joint event from existing events by using set operations.
- ❑ Assess if events are mutually exclusive and/or exhaustive.
- ❑ Use a Venn diagram and/or tree diagram to display a sample space and events.

Random Experiment

When different outcomes are obtained in repeated trials, the experiment is called a random experiment. Some sources of the variation in outcomes are controllable and some are not in the random experiment.

(e.g.) Sources of variability in testing the life length of an INFINITY light bulb
(1) Material
(2) Manufacturing process
(3) Production environment (temperature, humidity, etc.)
(4) Measurement instrument
(5) Drift of current
(6) Observer

Sample Space (S)

A sample space is the set of possible outcomes of a random experiment (denoted as S). Two types of sample space are defined:

1. **Discrete Sample Space**: Consists of finite (or countably infinite) outcomes.
   (e.g.) Tossing a coin: $S = \{\text{head, tail}\}$
2. **Continuous Sample Space**: Consists of infinite and innumerable outcomes.
   (e.g.) The life length of the INFINITY light bulb (X): $S = \{x; x \geq 0\}$

Event (E)

An event is a subset of the sample space of a random experiment (denoted by E).

Set Operations

Three set operations can be applied to define a new joint event from existing events:

1. **Union** ($E_1 \cup E_2$): Combines all outcomes of E_1 and E_2.
2. **Intersection** ($E_1 \cap E_2$): Includes outcomes that are common in E_1 and E_2.
3. **Complement** (E' or $\overline{E}$): Contains outcomes that are not in E. Note that $(E')' = E$.

Laws for Set Operations

The following laws are applicable to the set operations:

1. **Commutative Law**
   - $E_1 \cap E_2 = E_2 \cap E_1$
   - $E_1 \cup E_2 = E_2 \cup E_1$
2. **Distributive Law**
   - $(E_1 \cap E_2) \cup E_3 = (E_1 \cup E_3) \cap (E_2 \cup E_3)$
   - $(E_1 \cup E_2) \cap E_3 = (E_1 \cap E_3) \cup (E_2 \cap E_3)$
3. **DeMorgan's Law**
   - $(E_1 \cap E_2)' = E_1' \cup E_2'$
   - $(E_1 \cup E_2)' = E_1' \cap E_2'$

Mutually Exclusive/ Exhaustive Events

A collection of events $E_1, E_2, \ldots, E_k$ is said to be mutually exclusive (see Figure 2-1), if the events do not have any outcomes in common, i.e.,

$E_i \cap E_j = \varnothing$ for all pairs (i, j), $i \neq j$

Next, the collection of events $E_1, E_2, \ldots, E_k$ is said to be exhaustive (see Figure 2-1), if the union of the events is equal to S, i.e.,

$E_1 \cup E_2 \cup \cdots \cup E_k = S$

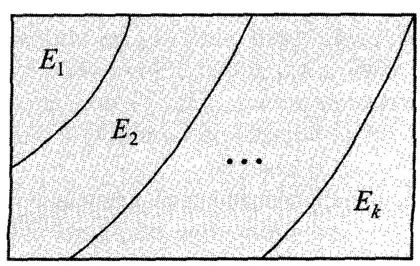


Figure 2-1 Mutually exclusive and exhaustive events.

Example 2.1

Two dice are cast at the same time in an experiment.

1. (**Sample Space**) Define the sample space of the experiment.

   The sample space is $S = \{(x, y); x = 1, \ldots, 6, y = 1, \ldots, 6\}$, where x and y represent the outcome of each die. Since there are 6 possible outcomes with the first and second dice each, the sample space consists of a total of $6 \times 6 = 36$ outcomes.
2. (**Event**) Of the sample space, find the pairs whose sum is 5 (E_1) and the pairs whose first die is odd (E_2).

Example 2.1 (*cont.*)

$E_1 = \{(1, 4), (2, 3), (3, 2), (4, 1)\}$
$E_2 = \{(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6), (5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6)\}$

3. (**Set Operation**) Find $E_1 \cup E_2$, $E_1 \cap E_2$, and E_2'.

$$E_1 \cup E_2 = \{(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), (2, 3), (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6), (4, 1), (5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6)\}$$
$$E_1 \cap E_2 = \{(1, 4), (3, 2)\}$$
$$E_2' = \{(2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6), (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)\}$$

4. (**Mutually Exclusive Events**) Are E_1 and E_2 mutually exclusive?

No, because $E_1 \cap E_2 \neq \varnothing$.

5. (**Exhaustive Events**) Are E_1 and E_2 exhaustive?

No, because $E_1 \cup E_2 \neq S$.

Exercise 2.1 (MR 2-33)

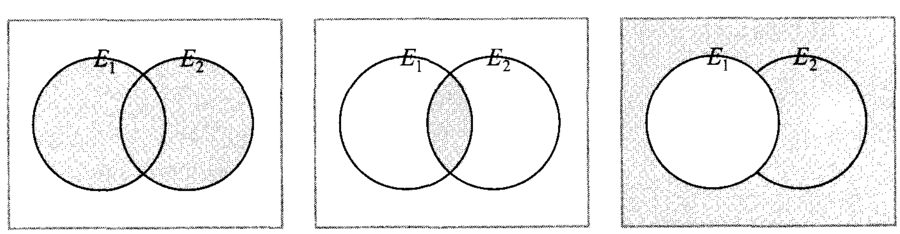
The rise times (unit: min.) of a reactor for two batches are measured in an experiment.

1. Define the sample space of the experiment.
2. Define E_1 where the reactor rise time of the first batch is <u>less</u> than 55 min. and E_2 where the reactor rise time of the second batch is <u>greater</u> than 70 min.
3. Find $E_1 \cup E_2$, $E_1 \cap E_2$, and E_1'.
4. Are E_1 and E_2 mutually exclusive?
5. Are E_1 and E_2 exhaustive?

Diagrams

Diagrams are often used to display a sample space and events in an experiment:

1. **Venn Diagram**: A rectangle represents the sample space and circles indicate individual events, as illustrated in Figure 2-2.



(a) (b) (c)

Figure 2-2 Venn diagrams: (a) $E_1 \cup E_2$; (b) $E_1 \cap E_2$; (c) E_1'. The areas shaded indicate the results of union, intersection, and complement, respectively.

Diagrams (*cont.*)

2. **Tree Diagram**: Branches represent possible outcomes (see Figure 2-3). The tree diagram method is useful when the sample space is established through a series of steps or stages.

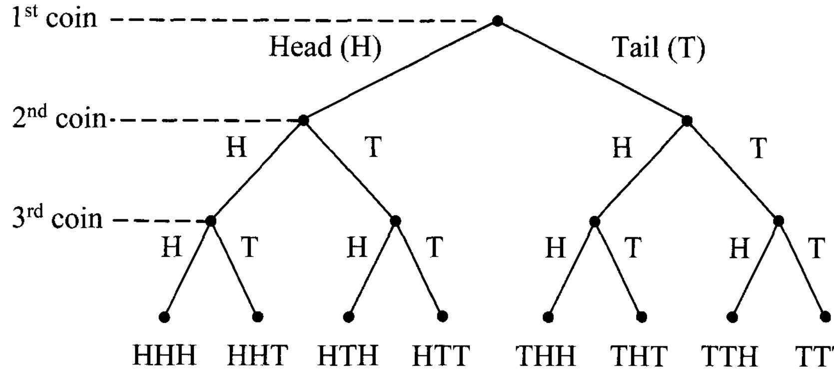


Figure 2-3 Tree diagram for outcomes of tossing three coins at the same time.

2-2 Interpretations of Probability

Learning Goals

- ❑ Explain the term *probability*.
- ❑ Determine the probability of an event.

Probability

The probability of an event means the likelihood of the event occurring in a random experiment. If S denotes the sample space and E denotes an event, the following conditions should be met:

(1) $P(S) = 1$

(2) $0 \le P(E) \le 1$

(3) $P(E_1 \cup E_2 \cup E_3 \cup \cdots) = P(E_1) + P(E_2) + P(E_3) + \cdots$, where the events are mutually exclusive.

Probability of Event

If a sample space includes n outcomes that are equally likely, the probability of each outcome is $1/n$. Thus, the probability of an event E consisting of k equally likely outcomes is

$$P(E) = k \times \frac{1}{n} = \frac{k}{n}$$

where: n = number of possible outcomes in S
k = number of equally likely elements in E

(Note) For any event E,

$$P(E') = 1 - P(E)$$

2-3 Addition Rules

Learning Goals

- ☐ Find the probability of a joint event by using the probabilities of individual events.

Probability of Joint Event

The probability of a joint event can often be calculated by using the probabilities of the individual events involved. The following addition rules can be applied to determine the probability of a joint event when the probabilities of existing events are known:

- $$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$
  $$= P(E_1) + P(E_2) \quad \text{if } E_1 \cap E_2 \neq \varnothing$$
- $$P(E_1 \cup E_2 \cup E_3) = P(E_1) + P(E_2) + P(E_3)$$
  $$-P(E_1 \cap E_2) - P(E_1 \cap E_3) - P(E_2 \cap E_3) + P(E_1 \cap E_2 \cap E_3)$$

Example 2.2

(Probability of Joint Event) An instructor of a statistics class tells students that the probabilities of earning an A, B, C, and D or below are $\frac{1}{5}$, $\frac{2}{5}$, $\frac{3}{10}$, and $\frac{1}{10}$, respectively. Find the probabilities of (1) earning an A or B and (2) earning a B or below.

Let E_1, E_2, E_3, and E_4 denote the events of earning an A, B, C, and D or below, respectively. These individual events are mutually exclusive and exhaustive because

$$P(E_1 \cup E_2 \cup E_3 \cup E_4) = P(E_1) + P(E_2) + P(E_3) + P(E_4)$$
$$= \frac{1}{5} + \frac{2}{5} + \frac{3}{10} + \frac{1}{10} = 1$$

(1) The event of earning an A or B is $E_1 \cup E_2$. Therefore,

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2) = \frac{1}{5} + \frac{2}{5} + 0 = \frac{3}{5}$$

(2) The event of earning a B or below is $E_2 \cup E_3 \cup E_4$, which is equal to E_1'. Therefore,

$$P(E_2 \cup E_3 \cup E_4) = P(E_1') = 1 - P(E_1) = 1 - \frac{1}{5} = \frac{4}{5}$$

Exercise 2.2 (MR 2-45)

Test results of scratch resistance and shock resistance for 100 disks of polycarbonate plastic are as follows:

		Shock Resistance	
		High (B)	Low (B')
Scratch	High (A)	80	9
Resistance	Low (A')	6	5

Let A and A' denote the event that a disk has high scratch resistance and the event that a disk has low scratch resistance, respectively. Let B and B' denote the event that a disk has high shock resistance and the event that a disk has low shock resistance, respectively.

1. When a disk is selected at random, find the probability that both the scratch <u>and</u> shock resistances of the disk are high.

Exercise 2.2 (*cont.*)

2. When a disk is selected at random, find the probability that the scratch <u>or</u> shock resistance of the disk is high.
3. Consider the event that a disk has high scratch resistance and the event that a disk has high shock resistance. Are these two events mutually exclusive?

2-4 Conditional Probability

Learning Goals

- ❑ Explain the term *conditional probability* between events.
- ❑ Calculate the conditional probability of events.

Conditional Probability

The conditional probability $P(B \mid A)$ is the probability of an event B, given an event A. The following formula is used to calculate the conditional probability:

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}, \text{ where } P(A) > 0$$

Example 2.3

(**Conditional Probability**) A novel method to screen carpal tunnel syndrome (CTS) at the workplace is tested with two groups of people: 50 workers having CTS and 50 healthy workers without CTS. Let A and A' denote the event that a worker has CTS and the event that a worker does not have CTS, respectively. Let B and B' denote the event that a CTS test is positive and the event that a CTS test is negative, respectively. The summary of CTS test results is as follows:

		Test Result	
		Negative (B')	Positive (B)
Group	CTS (A)	10	40
	Healthy (A')	45	5

Find the probability that a CTS test is positive (B) when a worker has CTS (A).

$$P(A) = \frac{50}{100},\ P(B) = \frac{45}{100},\ \text{and } P(A \cap B) = \frac{40}{100}$$

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{40/100}{50/100} = \frac{4}{5}$$

Exercise 2.3

In Exercise 2.2, find the probability that a disk has high scratch resistance (A) when it has high shock resistance (B).

2-5 Multiplication and Total Probability Rules

Learning Goals

- ❑ Apply a total probability rule to find the probability of an event when the event is partitioned into several mutually exclusive and exhaustive subsets.

Multiplication Rule

From the definition of conditional probability,

$$P(B \cap A) = P(B \mid A)P(A) = P(A \mid B)P(B)$$

Total Probability Rule

When event B is partitioned into two mutually exclusive subsets $B \cap A$ and $B \cap A'$ (see Figure 2-4),

$$\begin{aligned} P(B) &= P(B \cap A) + P(B \cap A') \\ &= P(B \mid A)P(A) + P(B \mid A')P(A') \end{aligned}$$

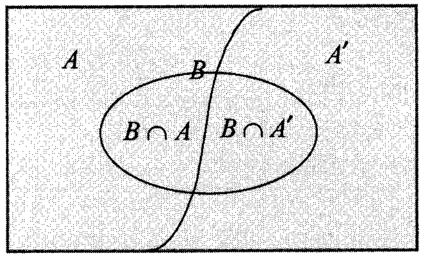


Figure 2-4 Partitioning event B into two mutually exclusive subsets: $B \cap A$ and $B \cap A'$.

As an extension, when event B is partitioned by a collection of k mutually exclusive and exhaustive events $E_1, E_2, \ldots, E_k$,

$$\begin{aligned} P(B) &= P(B \cap E_1) + P(B \cap E_2) + \cdots + P(B \cap E_k) \\ &= P(B \mid E_1)P(E_1) + P(B \mid E_2)P(E_2) + \cdots + P(B \mid E_k)P(E_k) \end{aligned}$$

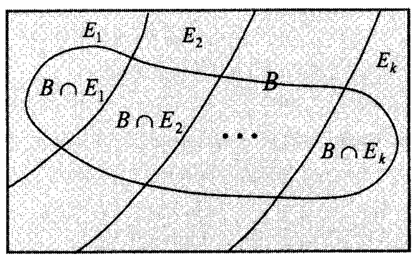


Figure 2-5 Partitioning event B into k mutually exclusive subsets: $B \cap E_1$, $B \cap E_2, \cdots, B \cap E_k$.

Example 2.4

(**Total Probability Rule**) In Example 2.3, the CTS screening method experiment indicates that the probability of screening a worker having CTS (A) as positive (B) is 0.8 and the probability of screening a worker without CTS (A') as positive (B) is 0.1, i.e.,

$$P(B \mid A) = 0.8 \quad \text{and} \quad P(B \mid A') = 0.1$$

Suppose that the industry-wide incidence rate of CTS is $P(A) = 0.0017$. Find the probability that a worker shows a positive test (B) for CTS at the workplace.

$P(A) = 0.0017$ and $P(A') = 1 - P(A) = 0.9983$

By applying a total probability rule,

$$\begin{aligned} P(B) &= P(B \mid A)P(A) + P(B \mid A')P(A') \\ &= 0.8 \times 0.0017 + 0.1 \times 0.9983 = 0.101 \end{aligned}$$

Exercise 2.4 (MR 2-97)

Customers are used to evaluate preliminary product designs. In the past, 95% of highly successful products, 60% of moderately successful products, and 10% of poor products received good reviews. In addition, 40% of product designs have been highly successful, 35% have been moderately successful, and 25% have been poor products. Find the probability that a product receives a good review.

2-6 Independence

Learning Goals

- ☐ Explain the term *independence* between events.
- ☐ Assess the independence of two events.

Independence of Events

Two events A and B are stochastically independent if the occurrence of A does not affect the probability of B and vice versa. In other words, two events A and B. are independent if and only if

(1) $P(A \mid B) = P(A)$

(2) $P(B \mid A) = P(B)$

(3) $P(A \cap B) = P(A)P(B)$

(Derivation) $P(A \cap B) = P(A)P(B)$

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = P(B) \text{ if } A \text{ and } B \text{ are independent.}$$

$$\Rightarrow P(A \cap B) = P(A)P(B)$$

Example 2.5

(Independence) For the CTS test results in Example 2.3, the following probabilities have been calculated:

$$P(B) = \frac{45}{100} \text{ and } P(B \mid A) = \frac{4}{5}$$

Check if events A and B are independent.

Since $P(B \mid A) = \frac{4}{5} \neq P(B) = \frac{45}{100}$, A and B are not independent. This indicates that information from the CTS test is useful to screen workers having CTS at the workplace.

Exercise 2.5

For the resistance test results in Exercise 2.2, the following probabilities have been calculated:

$$P(A) = \frac{89}{100} \text{ and } P(A \mid B) = \frac{40}{43}$$

Assess if events A and B are independent.

2-7 Bayes' Theorem

Learning Goals

- ☐ Apply Bayes' theorem to find the conditional probability of an event when the event is partitioned into several mutually exclusive and exhaustive subsets.

Bayes' Theorem

From the definition of conditional probability,

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

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

The multiplication rule for a collection of k mutually exclusive and exhaustive events $E_1, E_2, \ldots, E_k$ and any event B is

$$P(B) = P(B \mid E_1)P(E_1) + P(B \mid E_2)P(E_2) + \cdots + P(B \mid E_k)P(E_k)$$

From the two expressions above, the following general result (known as Bayes' theorem) is derived:

$$P(E_i \mid B) = \frac{P(B \mid E_i)P(E_i)}{P(B)} = \frac{P(B \mid E_i)P(E_i)}{P(B \cap E_1) + P(B \cap E_2) + \cdots + P(B \cap E_k)}$$

$$= \frac{P(B \mid E_i)P(E_i)}{P(B \mid E_1)P(E_1) + P(B \mid E_2)P(E_2) + \cdots + P(B \mid E_k)P(E_k)}$$

Example 2.6

(Bayes' Theorem) In Examples 2.3 and 2.4, the following probabilities have been identified:

$$P(B \mid A) = 0.8, \ P(B \mid A') = 0.1, \ P(A) = 0.0017, \text{ and } P(B) = 0.101$$

Find the probability that a worker has CTS (A) when the test is positive (B).

By applying Bayes' theorem,

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

$$= \frac{0.8 \times 0.0017}{0.101} = 0.013$$

Since the incidence rate of CTS in industry is low, the probability that a worker has CTS is quite small even if the test is positive.

Exercise 2.6

In Exercise 2.4, what is the probability that a new design will be highly successful if it receives a good review?

2-8 Random Variables

Learning Goals

- ❑ Describe the terms *random variable X* and *range of X*.
- ❑ Distinguish between discrete and continuous random variables.

Random Variable

A random variable, denoted by an uppercase such as X, associates real numbers with individual outcomes of a random experiment. Note that a measured value of X is denoted by a lowercase such as $x = 10$.

The set of possible numbers of X is referred to as the **range** of X. Depending on the type of range, two categories of random variables are defined:

1. **Discrete Random Variable**: Has a finite (or countably infinite) range.
   (e.g.) Tossing a coin: $X = 0$ for head and $X = 1$ for tail
2. **Continuous Random Variable**: Has an interval of real numbers for its infinite range.
   (e.g.) The life length of an INFINITY light bulb: $X \geq 0$

Answers to Exercises

Exercise 2.1

1. **(Sample Space)**

   $S = \{x; x > 0\}$, where x represents the rise time of the reactor for a certain batch.

2. **(Event)**

   $E_1 = \{x; 0 < x < 55\}$
   $E_2 = \{x; x > 70\}$

3. **(Set Operation)**

   $E_1 \cup E_2 = \{x; 0 < x < 55 \text{ or } x > 70\}$
   $E_1 \cap E_2 = \varnothing$
   $E_1' = \{x; x \geq 55\}$

4. **(Mutually Exclusive Events)**

   Yes, because $E_1 \cap E_2 = \varnothing$.

5. **(Exhaustive Events)**

   No, because $E_1 \cup E_2 \neq S$.

Exercise 2.2

(Probability of Joint Event)

1. $P(A \cap B) = \dfrac{80}{100}$

2. $P(A) = \dfrac{89}{100}$, $P(B) = \dfrac{89}{100}$, and $P(A \cap B) = \dfrac{80}{100}$

   $$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{89}{100} + \frac{86}{100} - \frac{80}{100} = \frac{95}{100}$$

3. Since $P(A \cap B) = \dfrac{80}{100} \neq 0$, A and B are not mutually exclusive.

Exercise 2.3

(Conditional Probability)

$P(A) = \dfrac{89}{100}$, $P(B) = \dfrac{86}{100}$, and $P(A \cap B) = \dfrac{80}{100}$

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{80/100}{86/100} = \frac{80}{86} = \frac{40}{43}$$

Exercise 2.4

(Total Probability Rule)

Let E_1, E_2, and E_3 represent the event of being a highly successful product, the event of being a moderately successful product, and the event of being a poor product, respectively. Also, let G denote the event of receiving good reviews from customers.

Exercise 2.5 (*cont.*)

Then,

$$P(G \mid E_1) = 0.95\,,\ P(G \mid E_2) = 0.60\,,\ P(G \mid E_3) = 0.10\,,$$
$$P(E_1) = 0.40\,,\ P(E_2) = 0.35\text{, and } P(E_3) = 0.25$$

The events E_1, E_2, and E_3 are mutually exclusive and exhaustive because

$$P(E_1 \cup E_2 \cup E_3) = P(E_1) + P(E_2) + P(E_3) = 1 = P(S)$$

By applying a total probability rule,

$$\begin{aligned} P(G) &= P(G \mid E_1)P(E_1) + P(G \mid E_2)P(E_2) + P(G \mid E_3)P(E_3) \\ &= 0.95 \times 0.40 + 0.60 \times 0.35 + 0.10 \times 0.25 = 0.62 \end{aligned}$$

Exercise 2.5

(Independence)

Since $P(A \mid B) = \dfrac{40}{43} \neq P(A) = \dfrac{89}{100}$, A and B are not independent. This indicates that the scratch resistance and shock resistance of disk are stochastically related.

Exercise 2.6

(Bayes' Theorem)

By applying Bayes' theorem,

$$\begin{aligned} P(E_1 \mid G) &= \frac{P(G \mid E_1)P(E_1)}{P(G \mid E_1)P(E_1) + P(G \mid E_2)P(E_2) + P(G \mid E_3)P(E_3)} \\ &= \frac{0.95 \times 0.40}{0.95 \times 0.40 + 0.60 \times 0.35 + 0.10 \times 0.25} = 0.62 \end{aligned}$$

3 Discrete Random Variables and Probability Distributions

OUTLINE

3-1 Discrete Random Variables

See Section 2-8. Random Variables to review the notion of discrete random variable.

3-2 Probability Distributions and Probability Mass Functions

Learning Goals

- ❑ Distinguish between probability mass function and cumulative distribution function.
- ❑ Determine the probability mass function of a discrete random variable.

Probability Distribution

A probability distribution indicates how probabilities are distributed over possible values of X.

Two types of functions are used to express the probability distribution of a discrete random variable X:

1. **Probability Mass Function (p.m.f.)**: Describes the probability of a value of X, i.e., $P(X = x_i)$.

Probability Distribution (*cont.*)

2. **Cumulative Distribution Function (c.d.f.)**: Describes the sum of the probabilities of values of X that are less than or equal to a specified value, i.e., $P(X \le x_i)$.

Probability Mass Function (p.m.f.)

The probability mass function of a discrete random variable X, denoted as $f(x_i)$, is

$$f(x_i) = P(X = x_i),\ x_i = x_1, x_2, \dots, x_n$$

and satisfies the following properties (see Figure 3-1):

(1) $f(x_i) \ge 0$ for all x_i

(2) $\sum_{i=1}^{n} f(x_i) = 1$

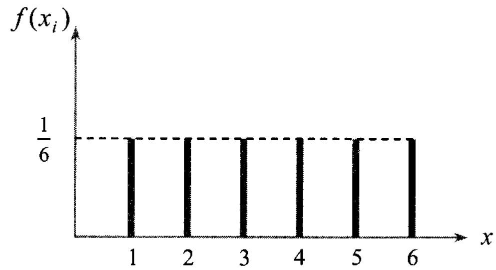


Figure 3-1 Probability mass function of casting a die.

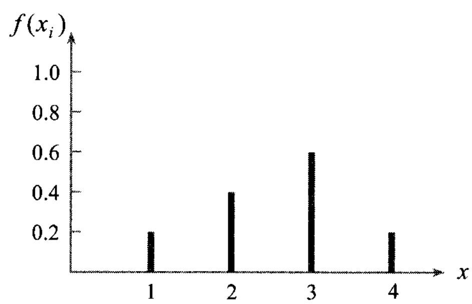
Example 3.1

(Probability Mass Function) The grades of n = 50 students in a statistics class are summarized as follows:

	Grade (X)			
	A (x = 1)	B (x = 2)	C (x = 3)	D or below (x = 4)
No. of students	10	20	15	5

Let X denote a grade in statistics. Let the values 1, 2, 3, and 4 represent an A, B, C, and D or below, respectively. Determine the probability mass function of X and plot $f(x_i)$.

$P(X=1)=\frac{1}{5}$, $P(X=2)=\frac{2}{5}$, $P(X=3)=\frac{3}{10}$, and $P(X=4)=\frac{1}{10}$



Probability mass function of X

Exercise 3.1 (MR 3-10)

The orders from $n = 100$ customers for wooden panels of various thickness (X) are summarized as follows:

	Wooden Panel Thickness (X; unit: inch)		
	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{3}{8}$
No. of customer orders	20	70	10

Determine the probability mass function of X and plot $f(x_i)$.

3-3 Cumulative Distribution Functions

Learning Goals

- ❑ Determine the cumulative distribution function of a discrete random variable.

Cumulative Distribution Function (c.d.f.)

The cumulative distribution function of a discrete random variable X, denoted as $F(x)$, is

$$F(x) = P(X \le x) = \sum_{x_i \le x} f(x_i)$$

and satisfies the following properties (see Figure 3-2):

(1) $0 \le F(x) \le 1$

(2) $F(x) \le F(y)$ if $x \le y$

(3) $f(x_i) = F(x_i) - F(x_{i-1})$

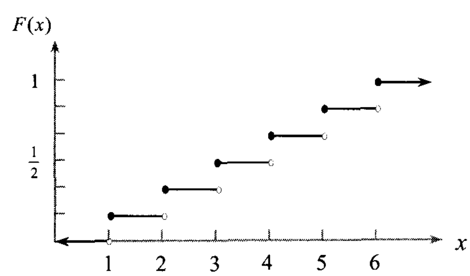


Figure 3-2 Cumulative distribution function of casting a die.

Example 3.2

(**Cumulative Distribution Function**) In Example 3-1, the following probabilities have been calculated:

$$P(X = 1) = \frac{1}{5},\ P(X = 2) = \frac{2}{5},\ P(X = 3) = \frac{3}{10},\text{ and } P(X = 4) = \frac{1}{10}$$

Determine the cumulative distribution function of X and plot $F(x)$.

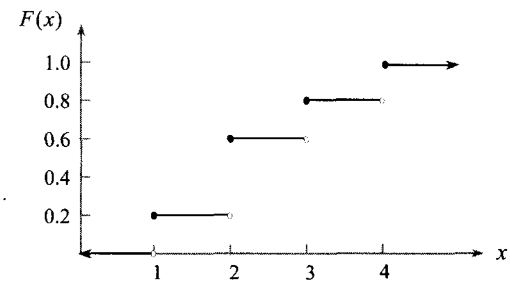
By using the probability mass function of X,

$$F(1) = P(X \le 1) = \frac{1}{5} \qquad F(2) = P(X \le 2) = \frac{1}{5} + \frac{2}{5} = \frac{3}{5}$$

$$F(3) = P(X \le 3) = \frac{3}{5} + \frac{3}{10} = \frac{9}{10} \qquad F(4) = P(X \le 4) = \frac{9}{10} + \frac{1}{10} = 1$$

Example 3.2 (*cont.*)

$$F(x) = \begin{cases} 0, & x < 1 \\ 0.2, & 1 \le x < 2 \\ 0.6, & 2 \le x < 3 \\ 0.9, & 3 \le x < 4 \\ 1, & 4 \le x \end{cases}$$



Cumulative distribution function of X

Exercise 3.2

In Exercise 3-1, the following probabilities have been calculated:

$P(X = \frac{1}{8}) = 0.2$, $P(X = \frac{1}{4}) = 0.7$, and $P(X = \frac{3}{8}) = 0.1$

Determine the cumulative distribution function of X and plot $F(x)$.

3-4 Mean and Variance of a Discrete Random Variable

Learning Goals

- ☐ Calculate the mean, variance, and standard deviation of a discrete random variable.

Mean of X (μ)

The mean of X, denoted as μ or $E(X)$, means the expected value of X:

$$\mu = E(X) = \sum_x x f(x)$$

Variance of X (σ^2)

The variance of X, denoted as σ^2 or $V(X)$, indicates the dispersion of X about μ:

$$\sigma^2 = V(X) = \sum_x (x-\mu)^2 f(x) = \sum_x x^2 f(x) - \mu^2$$

The **standard deviation** of X is $\sigma = \sqrt{V(X)}$.

(Derivation) $V(X) = \sum_x x^2 f(x) - \mu^2$ (see page 67 of MR)

Example 3.3

(**Mean, Variance, and Standard Deviation**) Suppose that, in an experiment of casting a single die, six outcomes (X) are equally likely. Determine the mean, variance, and standard deviation of X.

Example 3.3 (*cont.*)

The probability mass function of X is

$$f(x) = \frac{1}{6}, \quad x = 1, 2, \ldots, 6$$

Therefore,

$$\mu = \sum_x x f(x) = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + \cdots + 6 \times \frac{1}{6} = 3.5$$

$$\sigma^2 = \sum_x x^2 f(x) - \mu^2 = 1^2 \times \frac{1}{6} + 2^2 \times \frac{1}{6} + \cdots + 6^2 \times \frac{1}{6} - 3.5^2 = 2.9$$

$$\sigma = \sqrt{V(X)} = \sqrt{2.9} = 1.7$$

Exercise 3.3

In Exercise 3-1, the following probabilities have been calculated:

$P(X = \tfrac{1}{8}) = 0.2$, $P(X = \tfrac{1}{4}) = 0.7$, and $P(X = \tfrac{3}{8}) = 0.1$

Determine the mean, variance, and standard deviation of X.

3-5 Discrete Uniform Distribution

Learning Goals

- ❑ Describe the probability distribution of a discrete uniform random variable.
- ❑ Determine the probability mass function, mean, and variance of a discrete uniform random variable.

Discrete Uniform Random Variable

A discrete uniform random variable X has an equal probability for each value in the range of $X = [a, b]$, $a < b$ (see Figure 3-3). Thus, the **probability mass function** of X is

$$f(x) = \frac{1}{b - a + 1}, \quad \text{where } x = a, a + 1, \ldots, b$$

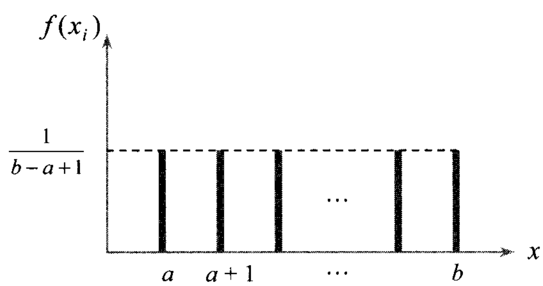


Figure 3-3 A discrete uniform distribution.

The **mean** and **variance** of X are

$$\mu = \frac{b + a}{2} \quad \text{and} \quad \sigma^2 = \frac{(b - a + 1)^2 - 1}{12}$$

Example 3.4

Suppose that six outcomes are equally likely in the experiment of casting a single die.

1. (**Probability Mass Function; Discrete Uniform Distribution**) Determine the probability mass function of the number (X) of the die.

Since $x = 1, 2, \ldots, 6$, $a = 1$ and $b = 6$. Therefore, the probability mass function of X is

$$f(x) = \frac{1}{b-a+1} = \frac{1}{6-1+1} = \frac{1}{6}, \quad x = 1 \text{ to } 6$$

2. (**Probability**) Find the probability that the number (X) of the die in the experiment is <u>greater</u> than three.

$$P(X > 3) = 1 - P(X \le 3) = 1 - \sum_{x=1}^{3} \frac{1}{6} = 1 - \frac{1}{2} = \frac{1}{2}$$

3. (**Mean and Variance**) Calculate the mean and variance of X.

Since $a = 1$ and $b = 6$, the mean and variance of X are

$$\mu = \frac{b+a}{2} = \frac{6+1}{2} = 3.5$$

$$\sigma^2 = \frac{(b-a+1)^2 - 1}{12} = \frac{(6-1+1)^2 - 1}{12} = 2.9 = 1.7^2$$

(Note) Check that the values of μ and σ^2 above are the same with the answers in Example 3.2.

Exercise 3.4 (MR 3-50)

Suppose that product codes of 2, 3, or 4 letters are equally likely.

1. Determine the probability mass function of the number of letters (X) in a product code.
2. Calculate the mean and variance of X.
3. Calculate the mean and variance of the number of letters in 100 product codes ($Y = 100X$). Note that $E(cX) = cE(X)$ and $V(cX) = c^2V(X)$ (refer to Section 5-7. Linear Combinations of Random Variables for further detail).

3-6 Binomial Distribution

Learning Goals

- ❑ Describe the terms *Bernoulli trial* and *binomial experiment.*
- ❑ Describe the probability distribution of a binomial random variable.
- ❑ Determine the probability mass function, mean, and variance of a binomial random variable.

Binomial Experiment

A binomial experiment refers to a random experiment consisting of n repeated trials which satisfy the following conditions:

(1) The trials are independent, i.e., the outcome of a trial does not affect the outcomes of other trials,

(2) Each trial has only two outcomes, labeled as 'success' and 'failure,' and

(3) The probability of a success (p) in each trial is constant.

Binomial Experiment (*cont.*)

In other words, a binomial experiment consists of a series of n independent Bernoulli trials (see the definition of Bernoulli trial below) with a constant probability of success (p) in each trial.

Bernoulli Trial

A Bernoulli refers to a trial that has only two possible outcomes.

(e.g.) Bernoulli trials

(1) Flipping a coin: $S = \{\text{head, tail}\}$
(2) Truth of an answer: $S = \{\text{right, wrong}\}$
(3) Status of a machine: $S = \{\text{working, broken}\}$
(4) Quality of a product: $S = \{\text{good, defective}\}$
(5) Accomplishment of a task: $S = \{\text{success, failure}\}$

The **probability mass function** of a Bernoulli random variable X is

$$f(x) = \begin{cases} 1-p, & x = 0 \\ p, & x = 1 \end{cases}$$

The **mean** and **variance** of a Bernoulli random variable are

$$\mu = p \quad \text{and} \quad \sigma^2 = p(1-p)$$

(Derivation) μ and σ^2 of a Bernoulli random variable

$$\mu = \sum_x xf(x) = 0 \times (1-p) + 1 \times p = p$$

$$\sigma^2 = \sum_x x^2 f(x) - \mu^2 = 0^2 \times (1-p) + 1^2 \times p - p^2 = p(1-p)$$

Binomial Random Variable

A binomial random variable X represents the number of trials whose outcome is a success out of n trials in a binomial experiment with a probability of success p (see Table 3-1).

Table 3-1 The Characteristics of a Binomial Distribution

Distribution	Population*	Parameters		
		Probability of Success (p)†	No. of Trials	No. of Successes
Binomial	Infinite	Constant	Constant (n)	Variable (X)

* If an item selected from a population is replaced before the next trial, the size of the population is considered infinite even if it may be finite.

† If the probability of success p is constant, the trials are considered independent; otherwise, the trials are dependent.

The **probability mass function** of X, denoted as $B(n, p)$, is

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n$$

(Note) The number of combinations of x from n: $C_x^n = \binom{n}{x} = \dfrac{n!}{x!(n-x)!}$

The **mean** and **variance** of X are

$$\mu = np \quad \text{and} \quad \sigma^2 = np(1-p)$$

Example 3.5

A test has $n = 50$ multiple-choice questions; each question has four choices but only one answer is right. Suppose that a student gives his/her answers by simple guess.

1. **(Probability Mass Function; Binomial Distribution)** Determine the probability mass function of the number of right answers (X) that the student gives in the test.

Since a 'right answer' is a success, the probability of a success for each question is $p = \frac{1}{4}$. Thus, the probability mass function of X is

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{50}{x} \frac{1}{4}^{x} \frac{3}{4}^{50-x}, \quad x = 0, 1, \cdots, 50$$

2. **(Probability)** Find the probability that the student answers <u>at least</u> 30 questions correctly.

$$P(X \geq 30) = 1 - P(X < 30) = 1 - \sum_{x=0}^{29} \binom{50}{x} \frac{1}{4}^{x} \frac{3}{4}^{50-x} = 1.6 \times 10^{-7}$$

3. **(Mean and Variance)** Calculate the mean and variance of X.

$\mu = np = 50 \times 0.25 = 12.5$ right answers

$\sigma^2 = np(1-p) = 50 \times 0.25 \times 0.75 = 9.4 = 3.1^2$

Exercise 3.5 (MR 3-67)

Because all airline passengers do not show up for their reserved seat, an airline sells $n = 125$ tickets for a flight that holds only 120 passengers. The probability that a passenger does not show up is 0.10. Assume that the passengers behave independently.

1. Determine the probability mass function of the number of passengers (X) who show up for the flight.
2. Find the probability that the flight departs with empty seats.
3. Calculate the mean and variance of X.

3-7 Geometric and Negative Binomial Distributions

3-7.1 Geometric Distribution

Learning Goals

- ❑ Describe the probability distribution of a geometric random variable.
- ❑ Compare the geometric distribution with the binomial distribution.
- ❑ Explain the lack of memory property of the geometric distribution.
- ❑ Determine the probability mass function, mean, and variance of a geometric random variable.

Geometric Random Variable

A geometric random variable X represents the number of trials conducted until the first success in a binomial experiment with a probability of success p:

$$\underbrace{\textit{Fail} \quad \textit{Fail} \quad \cdots \quad \textit{Fail}}_{\text{Fail } (x - 1) \text{ times}} \qquad \underset{\text{Succeed at the } x^{\text{th}} \text{ time}}{\underset{\uparrow}{\textit{Success}}}$$

$$P(X = x) \quad = \quad (1-p)^{x-1} \quad \times \quad p$$

The **probability mass function** of X is

$$f(x) = (1-p)^{x-1} p, \quad x = 1, 2, \ldots$$

The **mean** and **variance** of X are

$$\mu = \frac{1}{p} \quad \text{and} \quad \sigma^2 = \frac{1-p}{p^2}$$

The probability of X decreases geometrically as illustrated in Figure 3-4.

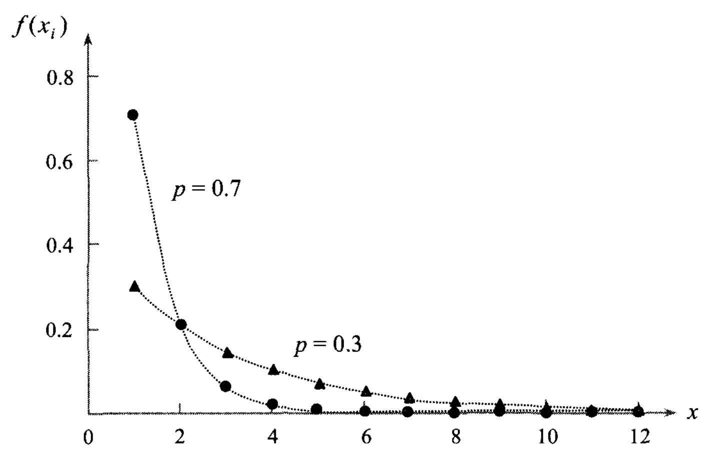


Figure 3-4 Geometric distributions with selected values of p.

Geometric vs. Binomial Distributions

In the geometric distribution the number of trials is a random variable and the number of successes is one, whereas in the binomial distribution the number of trials is constant and the number of successes is a random variable (see Table 3-2).

Table 3-2 The Characteristics of Binomial and Geometric Distributions

Distribution	Population*	Parameters		
		Probability of Success (p)†	No. of Trials	No. of Successes
Binomial	Infinite	Constant	Constant (n)	Variable (X)
Geometric	Infinite	Constant	Variable (X)	1

* If an item selected from a population is replaced before the next trial, the size of the population is considered infinite even if it may be finite.

† If the probability of success p is constant, the trials are considered independent; otherwise, the trials are dependent.

Lack of Memory Property

The lack of memory property means that the system does not remember the history of previous outcomes so that the timer of the system can be reset at any time, i.e.,

$$P(X < t + \Delta t \mid X > t) = \frac{P(t < X < t + \Delta t)}{P(X > t)} = P(X < \Delta t)$$

The lack of memory property of the geometric distribution implies that the value of p does not change with trial. Due to the memoryless property of the geometric distribution, the number of trials (X) until the next success can be counted from any trial. The geometric distribution is the only discrete distribution having the lack of memory property.

(Proof) The lack of memory property
The conditional probability $P(X < t + \Delta t \mid X > t)$ is

$$P(X < t + \Delta t \mid X > t) = \frac{P(t < X < t + \Delta t)}{P(X > t)} = \frac{P(X < t + \Delta t) - P(X \le t)}{1 - P(X \le t)}$$

$$= \frac{\sum_{x=1}^{t+\Delta t-1}(1-p)^{x-1}p - \sum_{x=1}^{t}(1-p)^{x-1}p}{1 - \sum_{x=1}^{t}(1-p)^{x-1}p}$$

Note that the sum of the first n terms of the geometric sequence $a, ar, ar^2, \ldots, ar^{n-1}$ is

$$S_n = \sum_{k=0}^{n-1} ar^k = \frac{a - ar^n}{1 - r}, \quad \text{where } r \ne 1$$

Accordingly,

$$P(X < t + \Delta t \mid X > t) = \frac{P(t < X < t + \Delta t)}{P(X > t)} = \frac{P(X < t + \Delta t) - P(X \le t)}{P(X > t)}$$

$$= \frac{P(X \le t + \Delta t - 1) - P(X \le t)}{1 - P(X \le t)}$$

$$= \frac{\sum_{x=1}^{t+\Delta t-1} p(1-p)^{x-1} - \sum_{x=1}^{t} p(1-p)^{x-1}}{1 - \sum_{x=1}^{t} p(1-p)^{x-1}}$$

$$= \frac{\sum_{k=0}^{(t+\Delta t-1)-1} p(1-p)^{k} - \sum_{k=0}^{t-1} p(1-p)^{k}}{1 - \sum_{k=0}^{t-1} p(1-p)^{k}}, \quad \text{where } k = x - 1$$

$$= \frac{\dfrac{p - p(1-p)^{t+\Delta t-1}}{1-(1-p)} - \dfrac{p - p(1-p)^{t}}{1-(1-p)}}{1 - \dfrac{p - p(1-p)^{t}}{1-(1-p)}}$$

Lack of Memory Property (*cont.*)

$$= \frac{(1-p)^t - (1-p)^{t+\Delta t-1}}{(1-p)^t} = 1-(1-p)^{\Delta t-1}$$

The probability $P(X < \Delta t)$ is

$$P(X < \Delta t) = P(X \le \Delta t - 1) = \sum_{x=1}^{\Delta t-1} (1-p)^{x-1} p$$

$$= \sum_{k=0}^{(\Delta t-1)-1} (1-p)^{x-1} p, \text{ where } k = x-1$$

$$= \frac{p-(1-p)^{\Delta t-1}}{1-(1-p)} = 1-(1-p)^{\Delta t-1}$$

Therefore,

$$P(X < t+\Delta t \mid X > t) = P(X < \Delta t)$$

Example 3.6

Suppose that the probability of meeting a 'right' person for marriage is 0.1 at a blind date and blind dates are independent (i.e., the probability of meeting a right person for marriage is constant).

1. (**Probability Mass Function; Geometric Distribution**) Determine the probability mass function of the number of blind dates (X) to make until succeeding at meeting a right person for marriage.

Since meeting a right person for marriage is a success, the probability of success is $p = 0.1$. Thus, the probability mass function of X is

$$f(x) = (1-p)^{x-1} p = 0.9^{x-1} 0.1, \quad x = 1, 2, \ldots$$

2. (**Probability**) Find the probability of meeting a right person for marriage within 10 blind dates.

$$P(X \le 10) = \sum_{x=1}^{10} 0.9^{x-1} 0.1 = 0.65$$

3. (**Mean and Variance**) Calculate the mean and variance of X.

$$\mu = \frac{1}{p} = \frac{1}{0.1} = 10 \text{ blind dates}$$

$$\sigma^2 = \frac{1-p}{p^2} = \frac{1-0.1}{0.1^2} = 90 = 9.5^2$$

Exercise 3.6 (MR 3-73)

Suppose that the probability of a successful optical alignment in the assembly of an optical data storage product is 0.8. Assume that the trials are independent.

1. Determine the probability mass function of the number of trials (X) required until the first successful alignment.
2. Find the probability that the first successful alignment requires four trials at the maximum.
3. Calculate the mean and variance of X.

3-7.2 Negative Binomial Distribution

Learning Goals

- ☐ Describe the probability distribution of a negative binomial random variable.
- ☐ Compare the negative binomial distribution with the binomial distribution.
- ☐ Determine the probability mass function, mean, and variance of a negative binomial random variable.

Negative Binomial Random Variable

A negative binomial random variable X represents the number of trials conducted until r successes in a binomial experiment with a probability of success p:

$$\underbrace{\textit{Fail}\ \ \textit{Fail}\ \ \textit{Success}\ \ \cdots\ \ \textit{Success}\ \ \textit{Fail}}_{\substack{\text{Succeed } (r-1) \text{ times during } (x-1) \text{ times} \\ (= \text{Fail } (x-r) \text{ times during } (x-1) \text{ times})}} \qquad \underset{\substack{\uparrow \\ \text{Make the } r^{\text{th}} \text{ success} \\ \text{at the } x^{\text{th}} \text{ time}}}{\textit{Success}}$$

The **probability mass function** of X is

$$f(x) = \binom{x-1}{r-1}(1-p)^{x-r}p^r, \quad x = r, r+1, \ldots$$

The **mean** and **variance** of X are

$$\mu = \frac{r}{p} \quad \text{and} \quad \sigma^2 = \frac{r(1-p)}{p^2}$$

Negative Binomial vs. Binomial Distributions

In the binomial distribution the number of trials is constant and the number of successes is a random variable, whereas the opposite becomes true in the negative binomial distribution (this is how the 'negative binomial' is named) (see Table 3-3).

Table 3-3 The Characteristics of Binomial and Negative Binomial Distributions

Distribution	Population*	Probability of Success (p)†	Parameters: No. of Trials	Parameters: No. of Successes
Binomial	Infinite	Constant	Constant (n)	Variable (X)
Negative binomial	Infinite	Constant	Variable (X)	Constant (r)

* If an item selected from a population is replaced before the next trial, the size of the population is considered infinite even if it may be finite.

† If the probability of success p is constant, the trials are considered independent; otherwise, the trials are dependent.

Example 3.7

A travel company is recruiting new members by making calls for the membership program ECONOTRAVEL. A cost analysis indicates that at least 20 new members per day be recruited to maintain the program. Suppose that the probability of recruiting a new member per call is 0.2.

1. (**Probability Mass Function; Negative Binomial Distribution**) Determine the probability mass function of the number of calls (X) to make to recruit 20 new members per day.

Example 3.7 (*cont.*)

Since recruiting a new member is a success, the probability of success is p = 0.2 and the number of successes is r = 20. Thus, the probability mass function of X is

$$f(x) = \binom{x-1}{r-1}(1-p)^{x-r}p^r = \binom{x-1}{19}0.8^{x-20}0.2^{20}, \quad x = 20, 21, \ldots$$

2. (**Probability**) Find the probability of recruiting 20 new members per day by making 100 calls.

$$P(X=100) = \binom{x-1}{19}0.8^{x-20}0.2^{20} = \binom{100-1}{19}0.8^{100-20}0.2^{20} = 0.02$$

3. (**Mean and Variance**) Calculate the mean and variance of X.

$$\mu = \frac{r}{p} = \frac{20}{0.2} = 100 \text{ calls per day}$$

$$\sigma^2 = \frac{r(1-p)}{p^2} = \frac{20(1-0.2)}{0.2^2} = 400 = 20^2$$

Exercise 3.7 (MR 3-83)

An electronic scale at an automated filling operation stops the manufacturing line after three underweight packages are detected. Suppose that the probability of an underweight package is 0.01 and fills are independent.

1. Determine the probability mass function of the number of fills (X) before the manufacturing line is stopped.
2. Find the probability that at least 100 fills can be completed before the manufacturing line is stopped.
3. Calculate the mean and variance of X.

3-8 Hypergeometric Distribution

Learning Goals

- ❑ Describe the probability distribution of a hypergeometric random variable.
- ❑ Compare the hypergeometric distribution with the binomial distribution.
- ❑ Determine the probability mass function, mean, and variance of a hypergeometric random variable.
- ❑ Apply the binomial approximation to a hypergeometric random variable.

Hypergeometric Random Variable

A hypergeometric random variable X represents the number of successes in a sample of size n that is selected at random without replacement from a finite population of size N consisting of K successes and $N - K$ failures. Since each item selected from the population is not replaced, the outcome of a trial depends on the outcome(s) of the previous trial(s); therefore, the probability of success p at each trial is not constant.

Hyper-geometric Random Variable (*cont.*)

The **probability mass function** of X is

$$f(x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}, \quad x = \max\{0, n-(N-K)\} \text{ to } \min\{K, n\}$$

The **mean** and **variance** of X are

$$\mu = np \quad \text{and} \quad \sigma^2 = np(1-p)\left(\frac{N-n}{N-1}\right) \quad \text{where } p = \frac{K}{N}$$

(Note) The variance of a hypergeometric random variable is different from the variance of a binomial random variable by $(N - n)/(N - 1)$, which is called **finite population correction factor**.

Hyper-geometric vs. Binomial Distributions

In the hypergeometric distribution the population is finite and the probability of success is changing, whereas in the binomial distribution the population is infinite and the probability of success is constant (see Table 3-4).

Table 3-4 The Characteristics of Binomial and Hypergeometric Distributions

Distribution	Population*	Probability of Success (p)[†]	Parameters: No. of Trials	Parameters: No. of Successes
Binomial	Infinite	Constant	Constant (n)	Variable (X)
Hyper-geometric	Finite (K successes; N - K failures)	Changing	Constant (n)	Variable (X)

* If an item selected from a population is replaced before the next trial, the size of the population is considered infinite even if it may be finite.

[†] If the probability of success p is constant, the trials are considered independent; otherwise, the trials are dependent.

Example 3.8

An instructor of a statistics class is planning to interview a sample of $n = 10$ students who are randomly selected from the class. The class has a total of 30 students, consisting of 20 traditional and 10 non-traditional students.

1. (**Probability Mass Function; Hypergeometric Distribution**) Determine the probability mass function of the number of non-traditional students (X) in the sample.

The size of the population is $N = 30$. Since a non-traditional student is a success, the number of successes in the population is $K = 10$.

To determine the range of the number of non-traditional students (X) in the sample, calculate the following:

$$\max\{0, n-(N-K)\} = \max\{0, 10-(30-10)\} = \max\{0, -10\} = 0$$
$$\min\{K, n\} = \min\{10, 10\} = 10$$

Therefore, the probability mass function of X is

Example 3.8 (*cont.*)

$$f(x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}} = \frac{\binom{10}{x}\binom{30-10}{10-x}}{\binom{30}{10}} = \frac{\binom{10}{x}\binom{20}{10-x}}{\binom{30}{10}}, \quad x = 0 \text{ to } 10$$

2. (**Probability**) Find the probability that <u>at least</u> one non-traditional student is in the sample.

$$P(X \ge 1) = 1 - P(X = 0) = 1 - \binom{10}{x}\binom{20}{10-x}\Big/\binom{30}{10} = 1 - \binom{10}{0}\binom{20}{10-0}\Big/\binom{30}{10}$$
$$= 1 - 0.006 = 0.994$$

3. (**Mean and Variance**) Calculate the mean and variance of X.

$$p = \frac{K}{N} = \frac{10}{30} = \frac{1}{3}$$

$\mu = np = 10 \times \frac{1}{3} = 3.3$ non-traditional students in the sample of 10 students

$$\sigma^2 = np(1-p)\left(\frac{N-n}{N-1}\right) = 10 \times \frac{1}{3} \times \frac{2}{3} \times \frac{30-10}{30-1} = 1.5 = 1.2^2$$

Exercise 3.8 (MR 3-90)

A lot of 75 washers contains 5 defectives whose variability in thickness is unacceptably large. A sample of 10 washers is selected at random <u>without</u> replacement.

1. Determine the probability mass function of the number of defective washers (X) in the sample.
2. Find the probability that <u>at least</u> one unacceptable washer is in the sample.
3. Calculate the mean and variance of X.

Binomial Approximation

The hypergeometric distribution approximates to the binomial distribution with $p = K/N$, if the sample size n is relatively small to the population size N. A rule of thumb is that the binomial approximation of a hypergeometric distribution is satisfactory if $n/N < 0.1$.

Example 3.9

(**Binomial Approximation; Hypergeometric Distribution**) Suppose that X has a hypergeometric distribution with $N = 200$, $n = 10$, and $K = 20$. Find $P(X = 2)$ by using the hypergeometric distribution and approximate binomial distribution. Is the binomial approximation reasonable?

The probability mass function of X is

$$f(x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}} = \frac{\binom{20}{x}\binom{200-20}{10-x}}{\binom{200}{10}} = \frac{\binom{20}{x}\binom{180}{10-x}}{\binom{200}{10}}, \quad x = 0 \text{ to } 10$$

(Note) $\max\{0, n-(N-K)\} = \max\{0,\ 10-(200-20)\}$
$= \max\{0, -170\} = 0$
$\min\{K, n\} = \min\{20, 10\} = 10$

Example 3.9 (*cont.*)

Thus,

$$P(X=2)=f(2)=\frac{\binom{20}{x}\binom{180}{10-x}}{\binom{200}{10}}=\frac{\binom{20}{2}\binom{180}{10-2}}{\binom{200}{10}}=0.198$$

The approximate binomial distribution of the hypergeometric random variable X is

$$f(x)=\binom{n}{x}p^x(1-p)^{n-x}=\binom{10}{x}0.1^x 0.9^{10-x},\quad x=0 \text{ to } 10$$

$$\text{(Note)}\ p=\frac{K}{N}=\frac{20}{200}=0.1$$

Based on the binomial approximation,

$$P(X=2)=f(2)=\binom{10}{x}0.1^x 0.9^{10-x}=\binom{10}{2}0.1^2 0.9^{10-2}=0.194$$

Since n is relatively small to N ($n/N = 10/200 = 0.05$), the binomial approximation of the hypergeometric random variable X is reasonable.

Exercise 3.9

Suppose that X has a hypergeometric distribution with $N = 500$, $n = 50$, and $K = 100$. Find $P(X = 10)$ by using the hypergeometric distribution and approximate binomial distribution. Is the binomial approximation reasonable?

3-9 Poisson Distribution

Learning Goals

- ❑ Explain the term *Poisson process.*
- ❑ Describe the probability distribution of a Poisson random variable.
- ❑ Compare the Poisson distribution with the binomial distribution.
- ❑ Determine the probability mass function, mean, and variance of a Poisson random variable.

Poisson Process

Suppose that the occurrence of an event over an interval (of time, length, area, space, etc.) is countable and the interval can be partitioned into subintervals. Then, a random experiment is defined as a Poisson process (see Figure 3-5) if

(1) The probability of more than one occurrence in a subinterval is infinitesimal (approximately zero),

(2) The occurrences of the event in non-overlapping subintervals are stochastically independent, and

(3) The probability of one occurrence of the event in a subinterval is the <u>same</u> throughout all subintervals and <u>proportional</u> to the length of the subinterval.

Poisson Process (*cont.*)

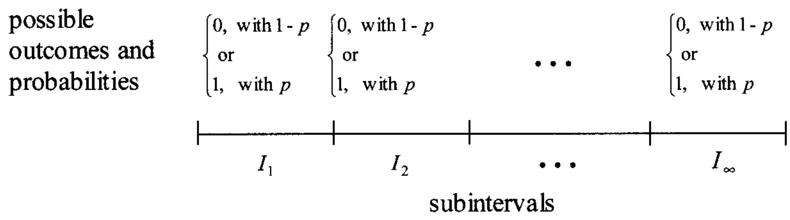


Figure 3-5 Poisson process.

In other words, the Poisson process is a binomial experiment with infinite n trials.

(e.g.) Poisson Process

(1) The number of defects of a product
(2) The number of customers in a store
(3) The number of automobile accidents
(4) The number of e-mails received

Poisson Random Variable

A Poisson random variable X represents the number of occurrences of an event of interest in a unit interval (of time, space, etc.) specified.

The **probability mass function** of X is

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots$$

The **mean** and **variance** of X are

$$\mu = \lambda \quad \text{and} \quad \sigma^2 = \lambda$$

(Caution) Use consistent units to define a Poisson random variable X and the corresponding parameter λ. For example, the following pairs of X and λ are equivalent to each other:

X (counts/unit interval)	λ (average no. of counts/unit interval)
No. of flaws a disk	1
No. of flaws every 10 disks	10
No. of flaws every 100 disks	100

Poisson vs. Binomial Distributions

In the Poisson distribution, the number of trials is infinite, whereas in the binomial distribution the number of trials is finite (see Table 3-5). In other words, the Poisson distribution with $E(X) = \lambda$ is the limiting form of the binomial distribution with $E(X) = np$:

$$\lim_{n\to\infty} B(n,p) = \lim_{n\to\infty}\binom{n}{x}p^x(1-p)^{n-x} = \lim_{n\to\infty}\binom{n}{x}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x} = \frac{e^{-\lambda}\lambda^x}{x!}$$

(Proof) Poisson vs. binomial distributions

Suppose that X is a binomial random variable with parameters n and p, and let $\lambda = np$.

Then,

$$\lim_{n\to\infty}\binom{n}{x}p^x(1-p)^{n-x} = \lim_{n\to\infty}\frac{n!}{x!(n-x)!}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x}$$

Poisson vs. Binomial Distributions (*cont.*)

$$= \frac{\lambda^x}{x!} \lim_{n\to\infty} \frac{n(n-1)\cdots(n-x+1)}{n^x}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x}$$

As n becomes large,

$$\lim_{n\to\infty} \frac{n(n-1)\cdots(n-x+1)}{n^x} = 1,$$

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n = \lim_{n\to\infty}\left[\left(1+\frac{1}{(-n/\lambda)}\right)^{-n/\lambda}\right]^{-\lambda} = e^{-\lambda}, \text{ and}$$

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-x} = 1$$

Therefore,

$$\lim_{n\to\infty}\binom{n}{x}p^x(1-p)^{n-x} = \frac{e^{-\lambda}\lambda^x}{x!}$$

Table 3-5 The Characteristics of Binomial and Poisson Distributions

Distribution	Population*	Parameters: Probability of Success (p)†	Parameters: No. of Trials	Parameters: No. of Successes
Binomial	Infinite	Constant	Constant (n)	Variable (X)
Poisson	Infinite	Constant ($p = \lambda/n$)	Infinite	Variable (X)

* If an item selected from a population is replaced before the next trial, the size of the population is considered <u>infinite</u> even if it may be finite.

† If the probability of success p is constant, the trials are considered <u>independent</u>; otherwise, the trials are dependent.


Example 3.10

The number of customers who come to the YUMMY donut store follows a Poisson distribution with a mean of 5 customers every 10 minutes.

1. (**Probability Mass Function; Poisson Distribution**) Determine the probability mass function of the number of customers (X) per <u>hour</u> coming to the donut store.

For a consistent unit for X and λ,

$$\lambda = E(X) = 5 \text{ customers/10 min.} \times 60 \text{ min.} = 30 \text{ customers per hour}$$

Therefore, the probability mass function of X is

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-30}30^x}{x!}, \quad x = 0, 1, 2, \ldots$$

2. (**Probability**) Find the probability that 40 customers come to the donut store in an hour.

$$P(X = 40) = \frac{e^{-30}30^x}{x!} = \frac{e^{-30}30^{40}}{40!} = 0.014$$

Example 3.10
(cont.)

3. (**Mean and Variance**) Calculate the mean and variance of X.

$\mu = \lambda = 30$ customers per hour

$\sigma^2 = \lambda = 30 = 5.5^2$

Exercise 3.10
(MR 3-103)

The number of cracks that need repair in a section of interstate highway follows a Poisson distribution with a mean of two cracks per mile.

1. Determine the probability mass function of the number of cracks (X) in <u>5 miles</u> of highway.
2. Find the probability that there are <u>at least</u> five cracks in 5 miles of highway that require repair.
3. Calculate the mean and variance of X.

Summary of Discrete Probability Distributions

A. Probability Distributions

Distribution	Probability Mass Function	Mean	Variance	Section
Uniform	$\dfrac{1}{b-a+1}, \quad x = a \text{ to } b, a \le b$	$\dfrac{b+a}{2}$	$\dfrac{(b-a+1)^2-1}{12}$	3-5
Binomial	$\binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0 \text{ to } n$	np	$np(1-p)$	3-6
Geometric	$(1-p)^{x-1} p, \quad x = 1, 2, \ldots$	$\dfrac{1}{p}$	$\dfrac{1-p}{p^2}$	3-7.1
Negative Binomial	$\binom{x-1}{r-1} (1-p)^{x-r} p^r, \quad x = r, r+1, \ldots$	$\dfrac{r}{p}$	$\dfrac{r(1-p)}{p^2}$	3-7.2
Hypergeometric	$\dfrac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}, \quad K \le N, n \le N,$ $x = \max\{0, n-(N-K)\} \text{ to } \min\{K, n\}$	np, $p = \dfrac{K}{N}$	$np(1-p)\left(\dfrac{N-n}{N-1}\right)$	3-8
Poisson	$\dfrac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots$	λ	λ	3-9

B. Comparison

Distribution	Population*	Parameters		
		Probability of Success (p)†	No. of Trials	No. of Successes
Binomial	Infinite	Constant	Constant (n)	Variable (X)
Geometric	Infinite	Constant	Variable (X)	1
Negative Binomial	Infinite	Constant	Variable (X)	Constant (r)
Hypergeometric	Finite (K successes; N - K failures)	Changing	Constant (n)	Variable (X)
Poisson	Infinite	Constant ($p = \lambda/n$)	Infinite	Variable (X)

* If an item selected from a population is replaced before the next trial, the size of the population is considered <u>infinite</u> even if it may be finite.

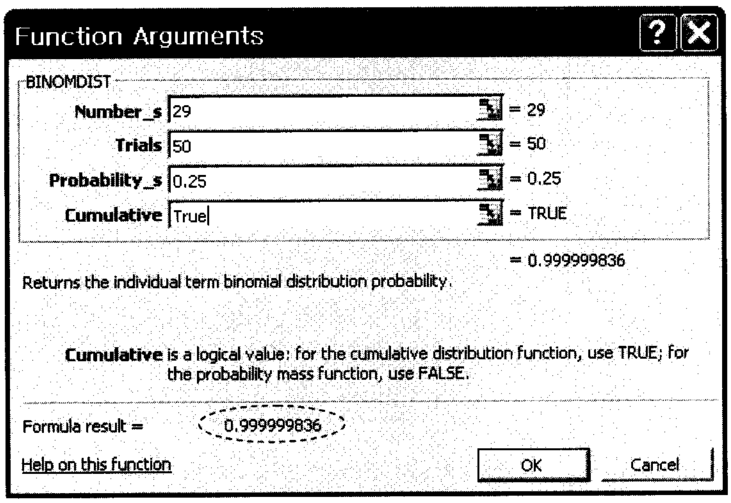
† If the probability of success p is constant, the trials are considered <u>independent</u>; otherwise, the trials are dependent.

EXCEL Applications

Example 3.5

2. **(Probability; Binomial Distribution)**

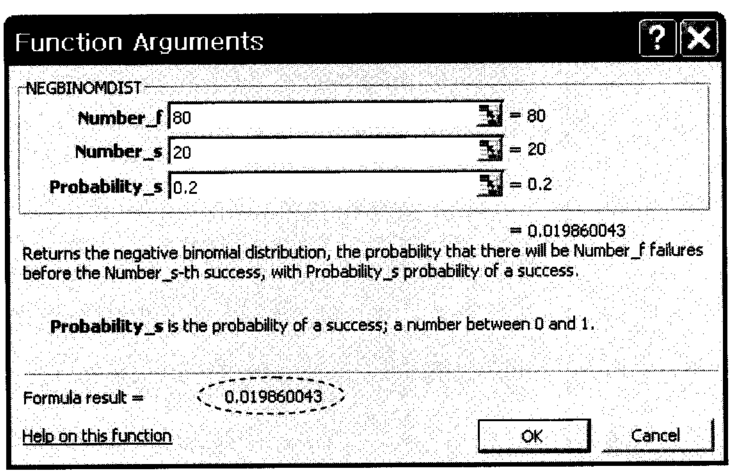
   (1) Choose **Insert ➤ Function**.
   (2) Select the category **Statistical** and the function **BINOMDIST**.
   (3) Enter the number of successes (**Number_s**; X), number of trials (**Trials**; n), probability of a success of interest (**Probability_s**; p), and logical value (**Cumulative**; true for the cumulative distribution function and false for the probability mass function).
   (4) Find the result circled with a dotted line below.



Example 3.7

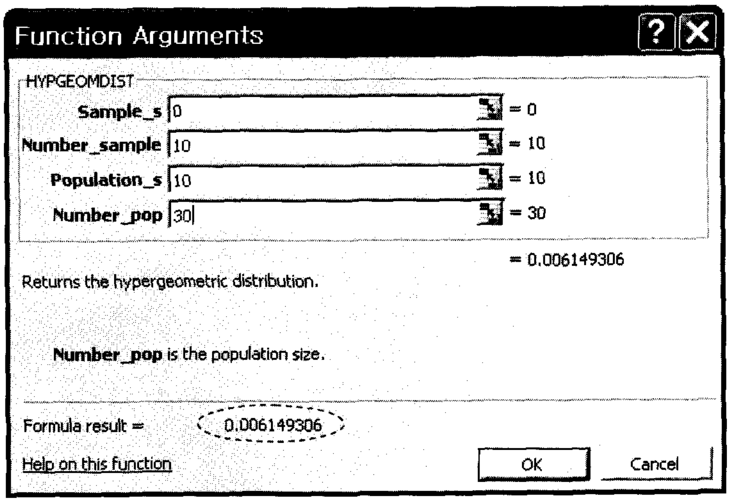
2. **(Probability; Negative Binomial Distribution)**

   (1) Choose **Insert ➤ Function**.
   (2) Select the category **Statistical** and the function **NEGBINOMDIST**.
   (3) Enter the number of failures (**Number_f**; $x - r$), number of successes (**Number_s**; r), and probability of a success of interest (**Probability_s**; p)
   (4) Find the result circled with a dotted line below.



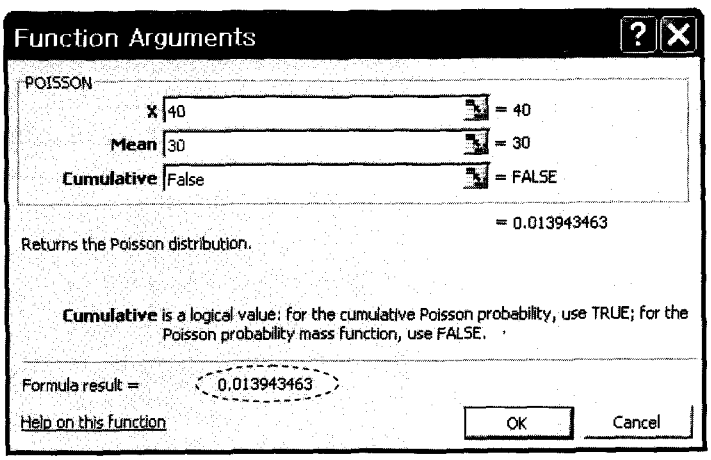
Example 3.8

2. **(Probability; Hypergeometric Distribution)**
   (1) Choose **Insert ➤ Function**.
   (2) Select the category **Statistical** and the function **HYPGEOMDIST**.
   (3) Enter the number of successes in the sample (**Sample_s**; x), size of the sample (**Number_sample**; n), number of successes in the population (**Population_s**; K), and size of the population (**Number_pop**; N)
   (4) Find the result circled with a dotted line below.



Example 3.10

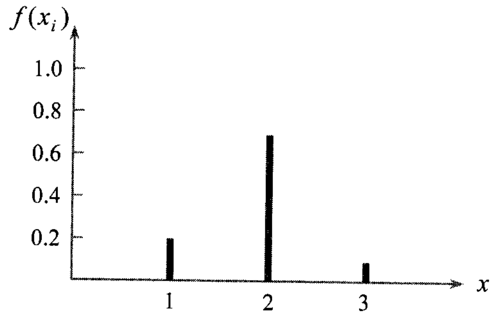
2. **(Probability; Poisson Distribution)**
   (1) Choose **Insert ➤ Function**.
   (2) Select the category **Statistical** and the function **POISSON**.
   (3) Enter the number of events (**X**), mean (**Mean**; λ), number of successes in the population (**Population_s**; K), and logical value (**Cumulative**; true for the cumulative distribution function and false for the probability mass function).
   (4) Find the result circled with a dotted line below.



Answers to Exercises

Exercise 3.1 **(Probability Mass Function)**

$P(X=\frac{1}{8})=0.2$, $P(X=\frac{1}{4})=0.7$, and $P(X=\frac{3}{8})=0.1$



Probability mass function of X

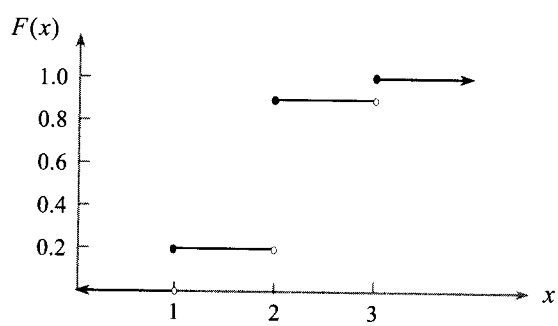
Exercise 3.2 **(Cumulative Distribution Function)**

By using the probability mass function of X,

$F(\frac{1}{8}) = P(X \le \frac{1}{8}) = 0.2 \qquad F(\frac{1}{4}) = P(X \le \frac{1}{4}) = 0.2 + 0.7 = 0.9$

$F(\frac{3}{4}) = P(X \le \frac{3}{4}) = 0.9 + 0.1 = 1$

$$F(x) = \begin{cases} 0, & x < \frac{1}{8} \\ 0.2, & \frac{1}{8} \le x < \frac{1}{4} \\ 0.9, & \frac{1}{4} \le x < \frac{3}{8} \\ 1, & \frac{3}{8} \le x \end{cases}$$



Cumulative distribution function of X

Exercise 3.3 **(Mean, Variance, and Standard Deviation)**

$$\mu = \sum_x x f(x) = \tfrac{1}{8}\times 0.2 + \tfrac{1}{4}\times 0.7 + \tfrac{3}{8}\times 0.1 = 0.2375 \text{ inch}$$

$$\sigma^2 = \sum_x x^2 f(x) - \mu^2 = (\tfrac{1}{8})^2 \times 0.2 + (\tfrac{1}{4})^2 \times 0.7 + (\tfrac{3}{8})^2 \times 0.1 - 0.2375^2$$

$$= 0.0045$$

$$\sigma = \sqrt{V(X)} = \sqrt{0.0045} = 0.067 \text{ inch}$$

Exercise 3.4

1. **(Probability Mass Function; Discrete Uniform Distribution)**

   Since $x = 2, 3, 4$, $a = 2$ and $b = 4$. Therefore, the probability mass function of X is

   $$f(x) = \frac{1}{b-a+1} = \frac{1}{4-2+1} = \frac{1}{3}, \quad x = 2, 3, 4$$

2. **(Mean and Variance)**

   Since $a = 2$ and $b = 4$, the mean and variance of X are

   $$E(X) = \frac{b+a}{2} = \frac{2+4}{2} = 3$$

   $$V(X) = \frac{(b-a+1)^2 - 1}{12} = \frac{(4-2+1)^2 - 1}{12} = \frac{2}{3}$$

3. **(Mean and Variance of a Linearly Combination)**

   The mean and variance of $Y = 100X$ are

   $$E(Y) = E(100X) = 100E(X) = 100 \times 3 = 300$$

   $$V(Y) = V(100X) = 100^2 V(X) = 100^2 \times \frac{2}{3} = 6{,}666.7 = 81.6^2$$

Exercise 3.5

1. **(Probability Mass Function; Binomial Distribution)**

   Since "a passenger showing up" is a success and the probability of a success is $p = 1 - 0.1 = 0.9$. Thus, the probability mass function of X is

   $$f(x) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{125}{x} 0.9^x 0.1^{125-x}, \quad x = 0, 1, \cdots, 125$$

2. **(Probability)**

   The flight would have empty seats if $X < 120$. Thus, the probability that the flight departs with empty seats is

   $$P(X < 120) = 1 - P(X \geq 120) = 1 - \sum_{x=120}^{125} \binom{125}{x} 0.9^x 0.1^{125-x}$$
   $$= 1 - 0.011 = 0.988$$

3. **(Mean and Variance)**

   $$\mu = np = 125 \times 0.9 = 112.5 \text{ passengers}$$

   $$\sigma^2 = np(1-p) = 125 \times 0.9 \times (1-0.9) = 11.2 = 3.3^2$$

Exercise 3.6

1. **(Probability Mass Function; Geometric Distribution)**

   Since a successful optical alignment is a success, the probability of success is $p = 0.8$. Thus, the probability mass function of X is

   $$f(x) = (1-p)^{x-1} p = 0.2^{x-1} 0.8, \quad x = 1, 2, \ldots$$

2. **(Probability)**

   $$P(X \leq 4) = \sum_{x=1}^{4} 0.2^{x-1} 0.8 = 0.998$$

Exercise 3.6 (*cont.*)

3. **(Mean and Variance)**

$$\mu = \frac{1}{p} = \frac{1}{0.8} = 1.25 \text{ trials}$$

$$\sigma^2 = \frac{1-p}{p^2} = \frac{1-0.8}{0.8^2} = 0.31 = 0.56^2$$

Exercise 3.7

1. **(Probability Mass Function; Negative Binomial Distribution)**

$p = 0.01, \quad r = 3$

$$f(x) = \binom{x-1}{r-1}(1-p)^{x-r}p^r = \binom{x-1}{2}0.99^{x-3}0.01^3, \quad x = 3, 4, \ldots$$

2. **(Probability)**

$$P(X \geq 100) = 1 - P(X < 100) = 1 - \sum_{x=3}^{99}\binom{x-1}{2}0.99^{x-3}0.01^3$$
$$= 1 - 0.078 = 0.922$$

3. **(Mean and Variance)**

$$\mu = \frac{r}{p} = \frac{3}{0.01} = 300 \text{ fills}$$

$$\sigma^2 = \frac{r(1-p)}{p^2} = \frac{3(1-0.01)}{0.01^2} = 29{,}700 = 172.3^2$$

Exercise 3.8

1. **(Probability Mass Function; Hypergeometric Distribution)**

$N = 75, \quad K = 5, \quad n = 10$

$\max\{0, n-(N-K)\} = \max\{0, 10-(75-5)\} = \max\{0, -60\} = 0$

$\min\{K, n\} = \min\{5, 10\} = 5$

$$f(x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}} = \frac{\binom{5}{x}\binom{75-5}{10-x}}{\binom{75}{10}} = \frac{\binom{5}{x}\binom{70}{10-x}}{\binom{75}{10}}, \quad x = 0 \text{ to } 5$$

2. **(Probability)**

$$P(X \geq 1) = 1 - P(X < 1) = 1 - P(X = 0) = 1 - \binom{5}{0}\binom{70}{10-0}\Big/\binom{75}{10}$$
$$= 1 - 0.48 = 0.52$$

3. **(Mean and Variance)**

$$p = \frac{K}{N} = \frac{5}{75} = \frac{1}{15}$$

$$\mu = np = 10 \times \frac{1}{15} = 0.67 \text{ unacceptable washers in the sample of 10 washers}$$

$$\sigma^2 = np(1-p)\left(\frac{N-n}{N-1}\right) = 10 \times \frac{1}{15} \times \frac{14}{15} \times \frac{75-10}{75-1} = 0.55 = 0.74^2$$

Exercise 3.9

(Binomial Approximation; Hypergeometric Distribution)

The probability mass function of X is

$$f(x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}} = \frac{\binom{100}{x}\binom{500-100}{50-x}}{\binom{500}{50}} = \frac{\binom{100}{x}\binom{400}{50-x}}{\binom{500}{50}}, \quad x = 0 \text{ to } 50$$

(Note) $\max\{0, n-(N-K)\} = \max\{0, 50-(500-100)\}$
$= \max\{0, -350\} = 0$
$\min\{K, n\} = \min\{100, 50\} = 50$

Thus,

$$P(X=10) = f(10) = \frac{\binom{100}{x}\binom{400}{50-x}}{\binom{500}{50}} = \frac{\binom{100}{10}\binom{400}{50-10}}{\binom{500}{50}} = 0.147$$

The approximate binomial distribution of the hypergeometric random variable X is

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{50}{x} 0.2^x 0.8^{50-x}, \quad x = 0 \text{ to } 50$$

(Note) $p = \dfrac{K}{N} = \dfrac{100}{500} = 0.2$

Based on the binomial approximation,

$$P(X=10) = f(10) = \binom{50}{x} 0.2^x 0.8^{10-x} = \binom{50}{10} 0.2^{10} 0.8^{50-10} = 0.140$$

Since n is relatively small to N ($n/N = 50/500 = 0.1$), the binomial approximation of the hypergeometric random variable X is satisfactory.

Exercise 3.10

1. **(Probability Mass Function; Poisson Distribution)**

   $\lambda = E(X) = 2$ cracks/mile $\times$ 5 miles = 10 cracks every 5 miles

   $$f(x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-10}10^x}{x!}, \quad x = 0, 1, 2, \ldots$$

2. **(Probability)**

   $$P(X \geq 5) = 1 - P(X < 5) = 1 - \sum_{x=0}^{4} \frac{e^{-10}10^x}{x!} = 1 - 0.03 = 0.97$$

3. **(Mean and Variance)**

   $\mu = \lambda = 10$ cracks every 5 miles

   $\sigma^2 = \lambda = 10 = 3.2^2$

4 Continuous Random Variables and Probability Distributions

OUTLINE

4-1 Continuous Random Variables

See Section 2-8. Random Variables to review the notion of continuous random variable.

4-2 Probability Distributions and Probability Density Functions

Learning Goals

- ❑ Distinguish between probability density function and cumulative distribution function.
- ❑ Determine the probability of a continuous random variable by using the corresponding probability density function.

Probability Distribution Like the probability distribution of a discrete random variable, two types of functions are used to express the probability distribution of a continuous random variable X:

Probability Distribution (*cont.*)

1. **Probability Density Function (p.d.f.; $f(x)$)**: Describes the densities of probabilities over a range of X.

2. **Cumulative Distribution Function (c.d.f.; $F(x)$)**: Represents the probability of X that is less than or equal to a specified value, i.e., $P(X \le x)$.

Probability Density Distribution (p.d.f.)

The probability density function $f(x)$ of a continuous random variable X satisfies the following properties:

(1) $f(x) \ge 0$

(2) $\int_{-\infty}^{\infty} f(x)dx = 1$

(3) $P(a \le X \le b) = \int_a^b f(x)dx$

The integral of $f(x)$ from points a to b represents the probability that X belongs to the interval $[a, b]$ (i.e., the area under the density function between a and b as shown in Figure 4-1).

(4) $P(X = x) = 0$

The probability of every single point is zero because the width of any point is zero. Thus,

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

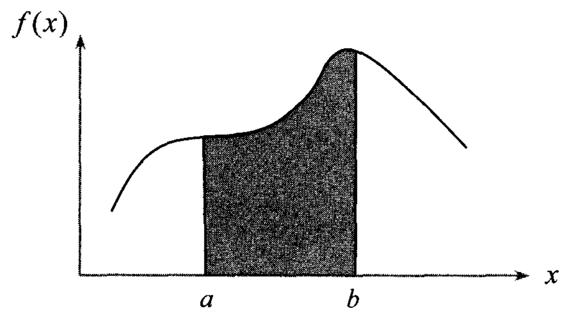


Figure 4-1 Probability density function. The probability $P(a \le X \le b)$ (shaded) is determined by integrating $f(x)$ from a to b.

Example 4.1

(**Probability Density Function**) Suppose that the probability density function of X is

$$f(x) = e^{-x}, \quad x > 0$$
$$= 0, \quad \text{elsewhere}$$

Determine $P(X < 2)$, $P(2 \le X < 4)$, and $P(X \ge 4)$.

(1) $P(X < 2) = \int_0^2 f(x)dx = \int_0^2 e^{-x}dx = -e^{-x}\Big|_0^2 = -e^{-2} + e^{-0} = 0.86$

(2) $P(2 \le X < 4) = \int_2^4 f(x)dx = \int_2^4 e^{-x}dx = -e^{-x}\Big|_2^4 = -e^{-4} + e^{-2} = 0.12$

(3) $P(X \ge 4) = \int_4^{\infty} f(x)dx = \int_4^{\infty} e^{-x}dx = -e^{-x}\Big|_4^{\infty} = -e^{-\infty} + e^{-4} = 0.02$

(Note) $P(X < 2) + P(2 \le X < 4) + P(X \ge 4) = 0.86 + 0.12 + 0.02 = 1$

Exercise 4.1

Suppose that the probability density function of X is

$$f(x) = 3x^2, \quad 0 < x < 1$$
$$= 0, \quad \text{elsewhere}$$

Determine $P(X < \frac{1}{3})$, $P(\frac{1}{3} \le X < \frac{2}{3})$, and $P(X \ge \frac{2}{3})$.

4-3 Cumulative Distribution Functions

Learning Goals

- ❑ Determine the cumulative distribution function of a continuous random variable.

Cumulative Distribution Function (c.d.f.)

The cumulative distribution function of a discrete random variable X is

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(u)du$$

and satisfies the following properties:

(1) $0 \le F(x) \le 1$

(2) $F(x) \le F(y)$ if $x \le y$

(Note) $f(x) = \dfrac{dF(x)}{dx}$

Example 4.2

(Cumulative Distribution Function) In Example 4-1, the probability density function of X is

$$f(x) = e^{-x}, \quad x > 0$$
$$= 0, \quad \text{elsewhere}$$

Determine the cumulative distribution function of X.

By using the probability density function of X,

$$F(x) = P(X \le x) = \int_0^x f(u)du = \int_0^x e^{-u}du = -e^{-u}\Big|_0^x = -e^{-x} + e^{-0} = 1 - e^{-x}$$

Therefore,

$$F(x) = \begin{cases} 0, & x < 0 \\ 1 - e^{-x}, & 0 \le x \end{cases}$$

Exercise 4.2

In Exercise 4-1, the probability density function of X is

$$f(x) = 3x^2, \quad 0 < x < 1$$
$$= 0, \quad \text{elsewhere}$$

Determine the cumulative distribution function of X.

4-4 Mean and Variance of a Continuous Random Variable

Learning Goals

- ☐ Calculate the mean, variance, and standard deviation of a continuous random variable.

Mean of X (μ)

The mean (expected value) of X is

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x)dx$$

Variance of X (σ^2)

The variance of X is

$$\sigma^2 = V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx = \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2$$

Standard Deviation of X (σ)

The standard deviation of X is

$$\sigma = \sqrt{V(X)}$$

Example 4.3

(Mean, Variance, and Standard Deviation) In Example 4-1, the probability density function of X is

$$f(x) = e^{-x}, \quad x > 0$$
$$= 0, \quad \text{elsewhere}$$

Determine the mean, variance, and standard deviation of X.

(1) The mean of X is

$$\mu = \int_0^{\infty} x f(x)dx = \int_0^{\infty} x e^{-x}dx$$

Apply the method of integration by parts as follows:

$$\int_a^b u\,dv = uv\Big|_a^b - \int_a^b v\,du$$

$u = x \qquad du = dx$

$v = -e^{-x} \qquad dv = e^{-x}dx$

Thus,

$$\mu = \int_0^{\infty} x e^{-x}dx = -xe^{-x}\Big|_0^{\infty} + \int_0^{\infty} e^{-x}dx = -e^{-x}\Big|_0^{\infty} = -e^{-\infty} + e^0 = 1$$

(2) The variance of X is

$$\sigma^2 = \int_0^{\infty} x^2 f(x)dx - \mu^2 = \int_0^{\infty} x^2 e^{-x}dx - \mu^2 = \int_0^{\infty} x^2 e^{-x}dx - 1$$

Apply the method of integration by parts as follows:

$$\int_a^b u\,dv = uv\Big|_a^b - \int_a^b v\,du$$

$u = x^2 \qquad du = 2xdx$

$v = -e^{-x} \qquad dv = e^{-x}dx$

Example 4.3 *(cont.)*

Then,

$$\int_0^\infty x^2 e^{-x} dx = -x^2 e^{-x}\Big|_0^\infty + \int_0^\infty 2xe^{-x} dx = 2\int_0^\infty xe^{-x} dx = 2\times 1 = 2$$

Therefore,

$$\sigma^2 = \int_0^\infty x^2 e^{-x} dx - 1 = 2 - 1 = 1$$

(3) The standard deviation of X is $\sigma = 1$.

Exercise 4.3

In Exercise 4-1, the probability density function of X is

$$f(x) = 3x^2, \quad 0 < x < 1$$
$$= 0, \quad \text{elsewhere}$$

Determine the mean, variance, and standard deviation of X.

4-5 Continuous Uniform Distribution

Learning Goals

- ❑ Describe the probability distribution of a continuous uniform random variable.
- ❑ Determine the probability density function, mean, and variance of a continuous uniform random variable.

Continuous Uniform Random Variable

A continuous uniform random variable X has a <u>constant</u> **probability density function** over the range of X (see Figure 4-2):

$$f(x) = \frac{1}{b-a}, \quad a \le x \le b$$

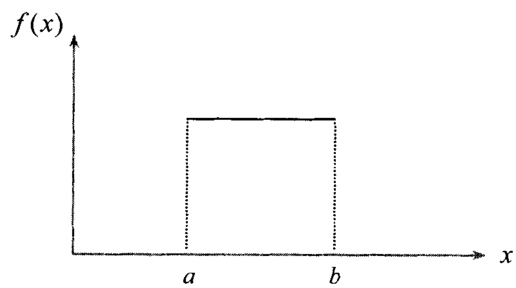


Figure 4-2 A continuous uniform distribution.

The **mean** and **variance** of X are

$$\mu = \frac{b+a}{2} \quad \text{and} \quad \sigma^2 = \frac{(b-a)^2}{12}, \text{ where } a \le x \le b$$

Example 4.4

Suppose that a random number generator produces real numbers that are uniformly distributed between 0 and 100.

1. (**Probability Density Function; Continuous Uniform Distribution**)
 Determine the probability density function of a random number (X) generated.

Example 4.4 (*cont.*)

$a = 0, \quad b = 100$

$$f(x) = \frac{1}{b-a} = \frac{1}{100-0} = \frac{1}{100}, \quad 0 \le x \le 100$$

2. (**Probability**) Find the probability that a random number (X) generated is between 10 and 90.

$$P(10 \le X \le 90) = \int_{0}^{90} f(x)dx = \int_{0}^{90} \frac{1}{100} dx = \frac{1}{100} x \Big|_{10}^{90} = \frac{1}{100}(90-10) = \frac{4}{5}$$

3. (**Mean and Variance**) Calculate the mean and variance of X.

$$\mu = \frac{b+a}{2} = \frac{100+0}{2} = 50$$

$$\sigma^2 = \frac{(b-a)^2}{12} = \frac{(100-0)^2}{12} = 833.3 = 28.9^2$$

Exercise 4.4 (MR 4-34)

Suppose that the thickness of a flange is uniformly distributed between 0.95 and 1.05 mm.

1. Determine the probability density function of flange thickness (X).
2. Find the probability that the thickness of a flange (X) exceeds 1.02 mm.
3. Calculate the mean and variance of X.

4-6 Normal Distribution

Learning Goals

- ❑ Describe the characteristics of a normal distribution.
- ❑ Read the z table.
- ❑ Standardize a normal random variable.

Normal Random Variable

A normal random variable with mean μ and variance σ^2 has the **probability density function**

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty \le x \le \infty$$

A normal distribution with **mean** μ and **variance** σ^2, denoted as $N(\mu, \sigma^2)$, is symmetric about μ and bell-shaped (see Figure 4-3). The symmetry of a normal curve implies

$$P(\mu < X) = P(X > \mu) = 0.5$$

The parameters μ and σ^2 determine the center and shape of the normal curve, respectively. As illustrated in Figure 4-3, the larger the value of μ, the more to the right the center of the normal curve is located; the smaller the value of σ^2, the sharper the normal curve.

Normal Random Variable (*cont.*)

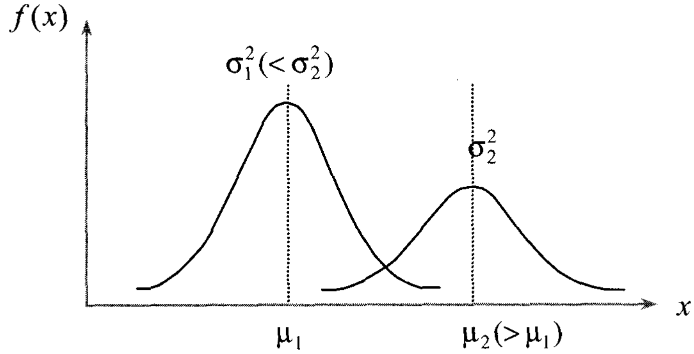


Figure 4-3 Normal curves with selected values of μ and σ^2.

Probabilities of Normal Distribution

Selected probabilities of a normal distribution are displayed in Figure 4-4. The area under the normal curve beyond $\pm 3\sigma$ is quite small (less than 0.01). Since 99.7% of possible values of X are within the interval (μ - 3σ, μ + 3σ), 6σ is referred to as the **width of a normal distribution**.

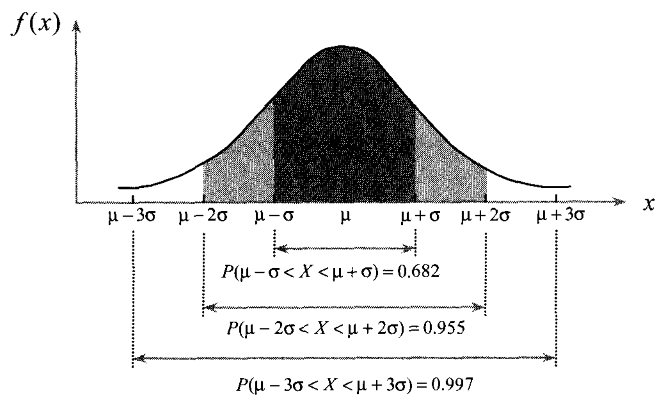


Figure 4-4 Probabilities of a normal distribution.

Standard Normal Random Variable

The standard normal random variable (denoted as Z) is a normal random variable with mean $\mu = 0$ and variance $\sigma^2 = 1$. The cumulative distribution function of Z is denoted as

$$\Phi(z) = P(Z \le z)$$

***z* Table**

The z table (see Appendix Table II in MR) provides values of $\Phi(z)$. In the table, the z column shows z values with tenth digits and the column headings indicate hundredth digits of z values.

(e.g.) Reading the z Table

(1) $\Phi(-1.57) = P(Z \le -1.57) = 0.058$

(2) $\Phi(1.57) = P(Z \le 1.57) = 0.942$

(3) $P(Z > 1.57) = 1 - P(Z \le 1.57) = 1 - 0.942 = 0.058$

(4) $P(-1.57 \le Z \le 1.57) = P(Z \le 1.57) - P(Z \le -1.57)$

$$= \Phi(1.57) - \Phi(-1.57) = 0.942 - 0.058 = 0.884$$

Standardization

An arbitrary normal random variable X with μ and σ^2 can be transformed into Z (with $\mu = 0$ and $\sigma^2 = 1$), as illustrated in Figure 4-5, by using the algebraic relationship between X and Z:

$$Z = \frac{X - \mu}{\sigma}$$

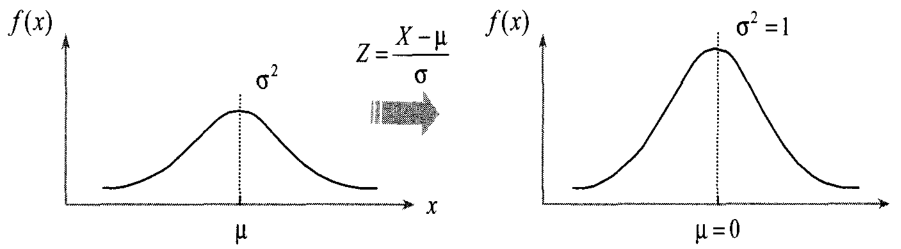


Figure 4-5 Standardizing a normal random variable.

After standardizing the normal random variable X, probabilities of X are determined by using the z table.

Example 4.5

(Standardization) Suppose that the life length of an INFINITY light bulb follows a normal distribution with $\mu = 750$ hours $\sigma^2 = 40^2$. Calculate $P(740 \le X \le 770)$.

$$P(740 \le X \le 770) = P\left(\frac{740-\mu}{\sigma} \le \frac{X-\mu}{\sigma} \le \frac{770-\mu}{\sigma}\right)$$
$$= P\left(\frac{740-750}{40} \le Z \le \frac{770-750}{40}\right) = P(-0.25 \le Z \le 0.5)$$
$$= P(Z \le 0.5) - P(Z \le -0.25) = 0.691 - 0.401 = 0.290$$

Exercise 4.5 (MR 4-57)

The sick-leave time of employees in a firm in a month is normally distributed with a mean of 100 hours and a standard deviation of 20 hours. Find the probability that the sick-leave time of an employee in a month exceeds 130 hours.

4-7 Normal Approximation to the Binomial and Poisson Distributions

Learning Goals

- ☐ Apply the normal approximation to the binomial, hypergeometric, and Poisson random variables.

Normal Approximation to Binomial Distribution

The binomial distribution $B(n, p)$ approximates to the normal distribution with $\mu = np$ and $\sigma^2 = np(1 - p)$ if $np > 5$ and $n(1 - p) > 5$ (see Figure 4-19 on page 119 of MR). However, when p is so close to 0 or 1 that $np < 5$ or $n(1 - p) < 5$, corresponding binomial distribution is significantly skewed (see Figure 4-20 on page 120 of MR); thus, the normal approximation is inappropriate.

Normal Approximation to Binomial Distribution (*cont.*)

When normal approximation is applicable, the probability of a binomial random variable X with $\mu = np$ and $\sigma^2 = np(1 - p)$ can be determined by using the standard normal random variable

$$Z = \frac{X - \mu}{\sigma} = \frac{X - np}{\sqrt{np(1-p)}}$$

Normal Approximation to Hypergeometric Distribution

Recall that the binomial approximation is applicable to a hypergeometric distribution if the sample size n is relatively small to the population size N, i.e., $n/N < 0.1$ (see Section 3-8. Hypergeometric Distribution). Consequently, the normal approximation can be applied to the hypergeometric distribution with $p = K/N$ (K: number of successes in N) if $n/N < 0.1$, $np > 5$, and $n(1 - p) > 5$.

Thus, when normal approximation is applicable, the probability of a hypergeometric random variable X with $\mu = np$ and $\sigma^2 = np(1 - p)(N - n)/(N - 1)$ can be determined by using the standard normal random variable

$$Z = \frac{X - \mu}{\sigma} = \frac{X - np}{\sqrt{np(1-p)(N-n)/(N-1)}}, \quad \text{where } p = \frac{K}{N}$$

Normal Approximation to Poisson Distribution

Recall that a Poisson distribution with $\mu = \lambda$ and $\sigma^2 = \lambda$ is the limiting form of a binomial distribution with $\mu = np$ and $\sigma^2 = np(1 - p)$ (see Section 3-9. Poisson Distribution). Therefore, the normal approximation is applicable to a Poisson distribution if $\lambda > 5$.

Accordingly, when normal approximation is applicable, the probability of a Poisson random variable X with $\mu = \lambda$ and $\sigma^2 = \lambda$ can be determined by using the standard normal random variable

$$Z = \frac{X - \mu}{\sigma} = \frac{X - \lambda}{\sqrt{\lambda}}$$

Example 4.6

1. (**Normal Approximation; Binomial Distribution**) Suppose that X is a binomial random variable with $n = 100$ and $p = 0.1$. Find the probability $P(X \leq 15)$ based on the corresponding binomial distribution and approximate normal distribution. Is the normal approximation reasonable?

The probability mass function, mean, and variance of the binomial random variable X are

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{100}{x} 0.1^x 0.9^{100-x}, \quad 0, 1, 2, \ldots, 100$$

$$\mu = np = 100 \times 0.1 = 10$$

$$\sigma^2 = np(1-p) = 100 \times 0.1 \times 0.9 = 9 = 3^2$$

Thus,

$$P(X \leq 15) = \sum_{x=0}^{15} \binom{100}{x} 0.1^x 0.9^{100-x} = 0.960$$

Example 4.6 (*cont.*)

By applying the normal approximation to the binomial random variable X,

$$P(X \le 15) = P\left(\frac{X-\mu}{\sigma} \le \frac{15-10}{3}\right) = P(Z \le 1.67) = 0.953$$

Since $np = 10 > 5$ and $n(1 - p) = 90 > 5$, the normal approximation to the binomial random variable X is satisfactory.

2. (**Normal Approximation; Hypergeometric Distribution**) Suppose that X has a hypergeometric distribution with $N = 1{,}000$, $K = 100$, and $n = 100$. Find the probability $P(X \le 15)$ based on the corresponding hypergeometric distribution and approximate normal distribution. Is the normal approximation reasonable?

The probability mass function, mean, and variance of the hypergeometric random variable X are

$$f(x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}} = \frac{\binom{100}{x}\binom{1000-100}{100-x}}{\binom{1000}{100}} = \frac{\binom{100}{x}\binom{900}{100-x}}{\binom{1000}{100}}, \; x = 0 \text{ to } 100$$

(Note) $\max\{0, n-(N-K)\} = \max\{0,\, 100-(1000-100)\}$
$= \max\{0, -800\} = 0$
$\min\{K, n\} = \min\{100, 100\} = 100$

$$\mu = np = 100 \times 0.1 = 10, \text{ where } p = \frac{K}{N} = \frac{100}{1000} = 0.1$$

$$\sigma^2 = np(1-p)\left(\frac{N-n}{N-1}\right) = 100 \times 0.1 \times 0.9 \times \frac{1000-100}{1000-1} = 8.11 = 2.85^2$$

Thus,

$$P(X \le 15) = \sum_{x=0}^{15} \frac{\binom{100}{x}\binom{900}{100-x}}{\binom{1000}{100}} = 0.968$$

By applying the normal approximation to the hypergeometric random variable X,

$$P(X \le 15) = P\left(\frac{X-\mu}{\sigma} \le \frac{15-10}{2.85}\right) = P(Z \le 1.75) = 0.960$$

Since $n/N = 0.1$, $np = 10 > 5$ and $n(1 - p) = 90 > 5$, the normal approximation to the hypergeometric random variable X is satisfactory.

3. (**Normal Approximation; Poisson Distribution**) Suppose that X has a Poisson distribution with $\lambda = 10$. Find the probability $P(X \le 15)$ based on the corresponding Poisson distribution and approximate normal distribution. Is the normal approximation reasonable?

Example 4.6 (*cont.*)

The probability mass function, mean, and variance of the Poisson random variable X are

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-10}10^x}{x!}, \quad x = 0, 1, 2, \ldots$$

$$\mu = \lambda = 10$$

$$\sigma^2 = \lambda = 10 = 3.16^2$$

Thus,

$$P(X \le 15) = \sum_{x=0}^{15} \frac{e^{-10}10^x}{x!} = 0.951$$

By applying the normal approximation to the Poisson random variable X,

$$P(X \le 15) = P\left(\frac{X-\mu}{\sigma} \le \frac{15-10}{3.16}\right) = P(Z \le 1.58) = 0.943$$

Since $\lambda = 10 > 5$, the normal approximation to the Poisson random variable X is satisfactory.

Exercise 4.6 (MR 4-61)

1. Suppose that X is a binomial random variable with $n = 200$ and $p = 0.4$. Find the probability $P(X \le 70)$ based on the corresponding binomial distribution and approximate normal distribution. Is the normal approximation reasonable?
2. Suppose that X has a hypergeometric distribution with $N = 2{,}000$, $K = 800$, and $n = 200$. Find the probability $P(X \le 70)$ based on the corresponding hypergeometric distribution and approximate normal distribution. Is the normal approximation reasonable?
3. Suppose that X has a Poisson distribution with $\lambda = 80$. Find the probability $P(X \le 70)$ based on the corresponding Poisson distribution and approximate normal distribution. Is the normal approximation reasonable?

4-9 Exponential Distribution

Learning Goals

- ❑ Describe the probability distribution of an exponential random variable.
- ❑ Explain the relationship between exponential and Poisson random variables.
- ❑ Explain the lack of memory property of an exponential random variable.
- ❑ Determine the probability density function, mean, and variance of an exponential random variable.

Exponential Random Variable

An exponential random variable represents the length of an interval (in time, space, etc.) from a certain point until the next success in a Poisson process. The exponential distribution parallels the geometric distribution by being concerned with the probability of the next success.

Exponential Random Variable (*cont.*)

The **probability density function** and **cumulative distribution function** of exponential random variable X with parameter λ (average number of successes per unit interval) are

$$f(x) = \lambda e^{-\lambda x} \quad \text{and} \quad F(x) = 1 - e^{-\lambda x}, \quad x \geq 0, \quad \lambda > 0$$

The **mean** and **variance** of X are

$$\mu = \frac{1}{\lambda} \quad \text{and} \quad \sigma^2 = \frac{1}{\lambda^2}$$

(Caution) Like the Poisson distribution, use consistent units to define an exponential random variable X and the corresponding parameter λ.

An exponential curve is skewed to the right, decreasing exponentially (see Figure 4-6; as λ increases, the skewness of the curve increases.

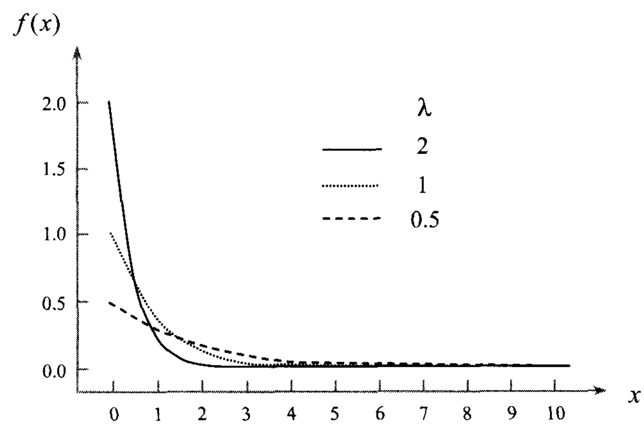


Figure 4-6 Exponential curves with selected values of λ.

Relationship between Exponential and Poisson Distributions

The exponential and Poisson distributions are interchangeable. Let an exponential random variable X with parameter λ (average number of successes during the unit length of time) denote the time until the next success. Let N denote the number of successes during a specific time interval x, i.e., N is a Poisson random variable with parameter λx.

Then, the following relationship exists between X and N:

$$P(X > x) = P(N = 0) = \frac{e^{-\lambda x}(\lambda x)^0}{0!} = e^{-\lambda x}$$

Therefore,

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

By differentiating $F(x)$, the probability density function of the exponential random variable X is

$$f(x) = \frac{dF(x)}{dx} = \lambda e^{-\lambda x}, \quad x \geq 0$$

Exponential vs. Poisson Distributions

In the exponential distribution the number of successes is fixed as one and the length of a time interval between successes is a random variable, whereas in the Poisson distribution the number of successes is a random variable and the length of a time interval is fixed (see Table 4-1).

Table 4-1 The Characteristics of Poisson and Exponential Distributions

Distribution	Population	Parameters		
		Rate of Success	Length of Time Interval	No. of Successes
Poisson	Infinite	Constant (λ)	Constant	Variable (X)
Exponential	Infinite	Constant (λ)	Variable (X)	Constant (1)

Lack of Memory Property

As introduced in Section 3-7.1 Geometric Distribution, the lack of memory property means that the system does not remember the history of previous outcomes so that the timer of the system can be reset at any time, i.e.,

$$P(X < t+\Delta t \mid X > t) = \frac{P(t < X < t+\Delta t)}{P(X > t)} = P(X < \Delta t)$$

The lack of memory property of an exponential distribution implies that the value of λ does not change with X. For example, in a reliability study of failure of a device, if the failure rate λ is constant (not increasing with time due to wear), the time to failure of the device has the lack of memory property. The exponential distribution is the only <u>continuous</u> distribution having the lack of memory. (Recall that the geometric distribution is the only <u>discrete</u> distribution having the lack of memory property.)

(Proof) The lack of memory property
The conditional probability $P(X < t+\Delta t \mid X > t)$ is

$$P(X < t+\Delta t \mid X > t) = \frac{P(t < X < t+\Delta t)}{P(X > t)} = \frac{F(t+\Delta t) - F(t)}{1 - F(t)}$$

$$= \frac{1 - e^{-\lambda(t+\Delta t)} - (1 - e^{-\lambda t})}{1 - (1 - e^{-\lambda t})} = \frac{e^{-\lambda t} - e^{-\lambda(t+\Delta t)}}{e^{-\lambda t}} = 1 - e^{-\lambda \Delta t}$$

The probability $P(X < \Delta t)$ is

$$P(X < \Delta t) = F(\Delta t) = 1 - e^{-\lambda \Delta t}$$

Therefore,

$$P(X < t+\Delta t \mid X > t) = P(X < \Delta t)$$

Example 4.7

The number of customers who come to the YUMMY donut store follows a Poisson process with a mean of 5 customers every 10 minutes.

1. (**Probability Density Function; Exponential Distribution**) Determine the probability density function of the time (X; unit: min.) until the next customer arrives.

Example 4.7 (*cont.*)

For a consistent unit for X and λ,

$\lambda = 5$ customers/10 min. = 0.5 customers/min.

Therefore, the probability density function of X is

$$f(x) = \lambda e^{-\lambda x} = 0.5e^{-0.5x}, \quad x \geq 0 \text{ (unit: min.)}$$

2. (**Probability**) Find the probability that there are no customers for <u>at least</u> 2 minutes by using the corresponding exponential and Poisson distributions.

By using the exponential distribution,

$$P(X > 2) = 1 - P(X \leq 2) = 1 - F(2) = 1 - (1 - e^{-1}) = e^{-1} = 0.368$$

The parameter of the corresponding Poisson distribution is

$\lambda x = 0.5$ customers/min. $\times$ 2 min. = 1

Therefore,

$$P(N = 0) = \frac{e^{-\lambda x}(\lambda x)^0}{0!} = e^{-1} = 0.368$$

As shown above, the probability from the exponential distribution is equal to the probability from the Poisson distribution.

3. (**Mean and Variance**) Find the mean and variance of X.

$\mu = \dfrac{1}{\lambda} = \dfrac{1}{0.5} = 2$ min. until the next customer arrival

$$\sigma^2 = \frac{1}{\lambda^2} = \frac{1}{(0.5)^2} = 4 = 2^2$$

Exercise 4.7 (MR 4-85)

The lifetime (X; unit: hour) of a mechanical assembly in a vibration test is exponentially distributed with a mean of 1 failure every 400 hours.

1. Determine the probability density function of the operating time (X) of the assembly before failure.
2. Find the probability that the assembly fails the vibration test in <u>less than</u> 100 hours by using the corresponding exponential and Poisson distributions.
3. Find the mean and variance of X.

4-10 Erlang and Gamma Distributions

4-10.1 Erlang Distribution

Learning Goals

- ❑ Describe the probability distribution of an Erlang random variable.
- ❑ Explain the relationship between Erlang and exponential random variables.
- ❑ Explain the relationship between Erlang and Poisson random variables.
- ❑ Determine the probability density function, mean, and variance of an Erlang random variable.

Erlang Random Variable

An Erlang random variable represents the length of an interval (in time, space, etc.) until the next r (positive integer) successes in a Poisson process. If X_1, $X_2, \ldots, X_r$ are mutually independent exponential random variables and each X_i represents the length of an interval for the i^{th} success after the $(i - 1)^{th}$ success, the Erlang random variable X is the sum of the r exponential random variables:

$$X = X_1 + X_2 + \cdots + X_r$$

The Erlang distribution parallels the negative binomial distribution by being concerned with the probability of the next r successes. (Recall that the exponential distribution parallels the geometric distribution by being concerned with the probability of the next success.)

The **probability density function** of an Erlang random variable X with parameters λ (average number of successes per unit interval) and r (number of successes) is

$$f(x) = \frac{\lambda^r x^{r-1} e^{-\lambda x}}{(r-1)!}, \quad x > 0$$

The **mean** and **variance** of X are

$$\mu = \frac{r}{\lambda} \quad \text{and} \quad \sigma^2 = \frac{r}{\lambda^2}$$

As illustrated in Figure 4-7 an Erlang curve is skewed to the right like an exponential curve; an Erlang curve with $r = 1$ becomes an exponential curve.

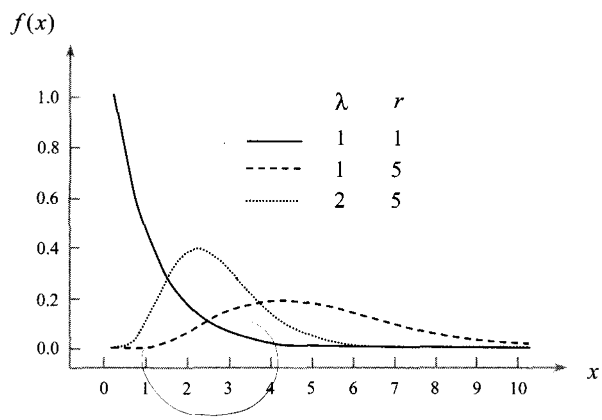


Figure 4-7 Erlang curves with selected values of λ and r.

Relationship between Erlang and Poisson Distributions

Like the exponential distribution, the Erlang and Poisson distributions are interchangeable. Let an Erlang random variable X with parameter λ (average number of successes during the unit length of time) denote the time until the next r successes. Let N denote the number of successes during a specific time interval x, i.e., N is a Poisson random variable with parameter λx.

Then, the following relationship exists between X and N:

$$P(X > x) = P(N < r) = \sum_{k=0}^{r-1} \frac{e^{-\lambda x} (\lambda x)^k}{k!}$$

Erlang vs. Poisson Distributions

In the Erlang distribution the number of successes is fixed as r and the length of a time interval until the next r successes is a random variable, whereas in the Poisson distribution the number of successes is a random variable and the length of a time interval is fixed (see Table 4-2).

Table 4-2 The Characteristics of Poisson and Erlang Distributions

Distribution	Population	Parameters		
		Rate of Success	Length of Interval	No. of Successes
Poisson	Infinite	Constant (λ)	Constant	Variable (X)
Erlang	Infinite	Constant (λ)	Variable (X)	Constant (r)

Example 4.8

The number of customers coming to the YUMMY donut store follows a Poisson process with a mean of 5 customers every 10 minutes.

1. (**Probability Density Function; Erlang Distribution**) Determine the probability density function of the time (X; unit: min.)) until $r = 3$ customers come to the donut store.

For a consistent unit for X and λ,

$\lambda = 5$ customers/10 min. = 0.5 customers/min.

Therefore, the probability density function of X with $r = 3$ customers is

$$f(x) = \frac{\lambda^r x^{r-1} e^{-\lambda x}}{(r-1)!} = \frac{0.5^3 x^{3-1} e^{-0.5x}}{(3-1)!} = 0.0625x^2 e^{-0.5x}, \quad x > 0 \text{ (unit: min.)}$$

2. (**Probability**) Find the probability that three customers come to the donut store in <u>less than</u> 10 min by using the corresponding Erlang and Poisson distributions.

By using the Erlang distribution,

$$P(X < 10) = \int_0^{10} 0.0625x^2 e^{-0.5x}\,dx = 0.875$$ (this integral is solved by integration by parts, which is not shown here)

The parameter of the corresponding Poisson distribution is

$\lambda x = 5$ customers/10 min. $\times$ 10 min. $= 5$

Therefore,

$$P(N \ge 3) = 1 - P(N < 3) = 1 - \sum_{k=0}^{2} \frac{e^{-\lambda x}(\lambda x)^k}{k!} = 1 - \sum_{k=0}^{2} \frac{e^{-5}(5)^k}{k!} = 0.875$$

The probability from the Erlang distribution is equal to the probability from the Poisson distribution.

3. (**Mean and Variance**) Find the mean and variance of X.

$$\mu = \frac{r}{\lambda} = \frac{3}{5/10} = 6 \text{ min. for } r = 3 \text{ customers}$$

$$\sigma^2 = \frac{r}{\lambda^2} = \frac{3}{(5/10)^2} = 12 = 3.5^2$$

Exercise 4.8 (MR 4-104)

The time between process problems in a manufacturing line is exponentially distributed with a mean of 30 days.

1. Determine the probability density function of the time (X; unit: day) until the fourth problem ($r = 4$).
2. Find the probability that the time until the fourth problem exceeds 120 days by using the corresponding Erlang and Poisson distributions.
3. Find the mean and variance of X.

4-10.2 Gamma Distribution

Learning Goals

- ❑ Explain the relationship between gamma and Erlang random variables.

Gamma Random Variable

A gamma random variable is the general case of an Erlang random variable by having a parameter r of a positive real number (recall that an Erlang random variable has a parameter r of a positive integer). In other words, a gamma distribution with a positive integer r becomes an Erlang distribution.

The **probability density function** of a gamma random variable X with parameters λ and r (> 0) is

$$f(x) = \frac{\lambda^r x^{r-1} e^{-\lambda x}}{\Gamma(r)}, \quad x > 0 \text{ and } r > 0$$

(Note) Gamma function $\Gamma(r)$

$$\Gamma(r) = \int_0^\infty x^{r-1} e^{-x} dx, \quad r > 0$$

A gamma function has the following properties:

(1) $\Gamma(r) = (r-1)\Gamma(r-1)$

(2) $\Gamma(r) = (r-1)!$ if r is a positive integer; $\Gamma(1) = 0! = 1$

(3) $\Gamma(0.5) = \sqrt{\pi} = 1.772$

(e.g.) Gamma function

(1) $\Gamma(4) = (r-1)! = (4-1)! = 3! = 6$

(2) $\Gamma(1.5) = (r-1)\Gamma(r-1) = (1.5-1)\Gamma(1.5-1) = 0.5\Gamma(0.5) = 0.5\sqrt{\pi} = 0.886$

The **mean** and **variance** of X are

$$\mu = \frac{r}{\lambda} \quad \text{and} \quad \sigma^2 = \frac{r}{\lambda^2}$$

Gamma vs. Erlang Distributions

While the Erlang distribution has a positive integer r, the gamma distribution has a positive real number integer r (see Table 4-3).

Gamma vs. Erlang Distributions (*cont.*)

Table 4-3 The Characteristics of Erlang and Gamma Distributions

Distribution	Population	Parameters: Rate of Success	Parameters: Length of Interval	Parameters: No. of Successes
Erlang	Infinite	Constant (λ)	Variable (X)	Positive integer r
Gamma	Infinite	Constant (λ)	Variable (X)	Positive real number r

4-11 Weibull Distribution

Learning Goals

- ❑ Describe the probability distribution of a Weibull random variable.
- ❑ Explain the relationship between Weibull and exponential random variables.
- ❑ Determine the probability density function, mean, and variance of a Weibull random variable.

Weibull Random Variable

A Weibull random variable represents the length of an interval (in time, space, etc.) until the next success for various types (increasing, decreasing, or constant) of success rates. Therefore, the Weibull random variable is more flexible than the exponential and gamma random variables by dealing with various types of success rates.

The **probability density function** and **cumulative distribution function** of a Weibull random variable X with shape parameter β (> 0) and scale parameter δ (> 0) are

$$f(x) = \frac{\beta}{\delta^{\beta}} x^{\beta-1} e^{-\left(\frac{x}{\delta}\right)^{\beta}}, x > 0 \quad \text{and} \quad F(x) = 1 - e^{-\left(\frac{x}{\delta}\right)^{\beta}}, x \geq 0$$

When the shape parameter $\beta = 1$, the Weibull distribution becomes the exponential distribution with $\lambda = 1/\delta$ (see Figure 4-8).

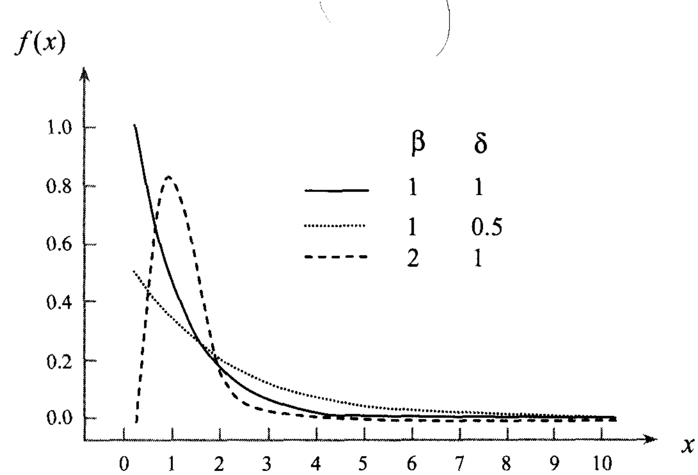


Figure 4-8 Weibull curves with selected values of β and δ.

Weibull Random Variable (*cont.*)

The **mean** and **variance** of X are

$$\mu = \delta\Gamma\left(1+\frac{1}{\beta}\right) \quad \text{and} \quad \sigma^2 = \delta^2\left\{\Gamma\left(1+\frac{2}{\beta}\right) - \left[\Gamma\left(1+\frac{1}{\beta}\right)\right]^2\right\}$$

Weibull vs. Exponential Distributions

In the Weibull distribution the rate of success can vary, whereas in the exponential distribution the rate of success is constant (see Table 4-4).

Table 4-4 The Characteristics of Exponential and Weibull Distributions

Distribution	Population	Parameters: Rate of Success	Parameters: Length of Interval	Parameters: No. of Successes
Exponential	Infinite	Constant (λ)	Variable (X)	Constant (1)
Weibull	Infinite	Changeable (β, δ)	Variable (X)	Constant (1)

Example 4.9

Assume that the time to failure of a hair braiding machine follows a Weibull distribution with $\alpha = 0.5$ and $\beta = 500$ hours.

1. (**Probability Density Function; Weibull Distribution**) Determine the probability density function of the time (X) until failure of the braiding machine.

$$f(x) = \frac{\beta}{\delta^\beta}x^{\beta-1}e^{-\left(\frac{x}{\delta}\right)^\beta} = \frac{0.5}{500^{0.5}}x^{0.5-1}e^{-\left(\frac{x}{500}\right)^{0.5}} = 0.02x^{-0.5}e^{-\left(\frac{x}{500}\right)^{0.5}}, x \geq 0$$

2. (**Probability**) Find the probability that the braiding machine lasts <u>longer than</u> 300 hours without failure.

$$P(X > 300) = 1 - P(X \leq 300) = 1 - [1 - e^{-\left(\frac{x}{\delta}\right)^\beta}] = e^{-\left(\frac{x}{\delta}\right)^\beta}$$
$$= e^{-(300/500)^{0.5}} = 0.461$$

3. (**Mean and Variance**) Find the mean and variance of X.

$$\mu = \delta\Gamma\left(1+\frac{1}{\beta}\right) = 500\Gamma\left(1+\frac{1}{0.5}\right) = 500\Gamma(3) = 500 \times 2! = 1{,}000 \text{ hrs until failure}$$

$$\sigma^2 = \delta^2\left\{\Gamma\left(1+\frac{2}{\beta}\right) - \left[\Gamma\left(1+\frac{1}{\beta}\right)\right]^2\right\} = 500^2\left\{\Gamma\left(1+\frac{2}{0.5}\right) - \left[\Gamma\left(1+\frac{1}{0.5}\right)\right]^2\right\}$$
$$= 500^2\{\Gamma(5) - [\Gamma(3)]^2\} = 500^2\{4! - (2!)^2\} = 5 \times 10^6 = 2{,}236^2$$

Exercise 4.9 (MR 4-114)

Suppose that the life of a recirculating pump follows a Weibull distribution with parameters $\alpha = 2$ and $\beta = 700$ hours.

1. Determine the probability density function of the time to failure (X) of the pump.
2. Find the probability that the pump lasts <u>less than</u> 700 hours.
3. Find the mean and variance of X.

4-12 Lognormal Distribution

Learning Goals

- ❑ Describe the probability distribution of a lognormal random variable.
- ❑ Determine the probability density function, mean, and variance of a lognormal random variable.

Lognormal Random Variable

A random variable X is called lognormally distributed if its natural logarithm, $W = \ln X$, follows a normal distribution with mean θ and variance ω^2. Like the Weibull distribution, the lognormal distribution is used to model the lifetime of a product whose failure rate is increasing over time.

The **probability density function** (see Figure 4-28 in MR) and **cumulative distribution function** of X with parameters θ and ω^2 are

$$f(x) = \frac{1}{x\omega\sqrt{2\pi}} e^{-\frac{(\ln x - \theta)^2}{2\omega^2}}, \; x > 0$$

$$F(x) = P(X \le x) = P(e^W \le x) = P(W \le \ln x) = P\left(Z \le \frac{\ln x - \theta}{\omega}\right) = \Phi\left(\frac{\ln x - \theta}{\omega}\right)$$

The **mean** and **variance** of X are

$$\mu = e^{\theta + \omega^2/2} \quad \text{and} \quad \sigma^2 = e^{2\theta + \omega^2}(e^{\omega^2} - 1)$$

Example 4.10

Assume that the time (unit: hours) to failure of a newly developed electronic device follows a lognormal distribution with parameters $\theta = -1$ and $\omega^2 = 2^2$.

1. (**Probability Density Function; Lognormal Distribution**) Determine the probability density function of the time (X) until failure of the device.

$$f(x) = \frac{1}{x\omega\sqrt{2\pi}} e^{-\frac{(\ln x - \theta)^2}{2\omega^2}} = \frac{1}{2\sqrt{2\pi}x} e^{-\frac{(\ln x + 1)^2}{8}}, \; x \ge 0$$

2. (**Probability**) Find the probability that the device lasts longer than 100 hours without failure.

$$P(X > 100) = 1 - P(X \le 100) = 1 - \Phi\left(\frac{\ln 100 - (-1)}{2}\right) = 1 - \Phi(2.80) = 0.003$$

3. (**Mean and Variance**) Find the mean and variance of X.

$$\mu = e^{\theta + \omega^2/2} = e^{-1 + 2^2/2} = 2.7 \text{ hour}$$

$$\sigma^2 = e^{2\theta + \omega^2}(e^{\omega^2} - 1) = e^{2\cdot(-1) + 2^2}(e^{2^2} - 1) = 19.9^2$$

Exercise 4.10 (MR 4-120)

Suppose that the length of time (in seconds) that a user views a page on a Web site before moving to another page is a lognormal random variable with parameters $\theta = 0.5$ and $\omega^2 = 1$.

1. Determine the probability density function of the time (X) until a user moves from the page.
2. Find the probability that a page is viewed for more than 10 seconds.
3. Find the mean and variance of X.

Summary of Continuous Probability Distributions

A. Probability Distributions

Distribution	Probability Density Function	Mean	Variance	Section
Uniform	$\frac{1}{b-a}, \quad a \le x \le b$	$\frac{b+a}{2}$	$\frac{(b-a)^2}{12}$	4-5
Normal	$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty \le x \le \infty$	μ	σ^2	4-6
Exponential	$\lambda e^{-\lambda x}, \quad x \ge 0, \lambda > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	4-9
Erlang	$\frac{\lambda^r x^{r-1} e^{-\lambda x}}{(r-1)!}, \quad x > 0, \lambda > 0, r = 1, 2, \ldots$	$\frac{r}{\lambda}$	$\frac{r}{\lambda^2}$	4-10.1
Gamma	$\frac{\lambda^r x^{r-1} e^{-\lambda x}}{\Gamma(r)}, \quad x > 0, \lambda > 0, r > 0$	$\frac{r}{\lambda}$	$\frac{r}{\lambda^2}$	4-10.2
Weibull	$\frac{\beta}{\delta^\beta} x^{\beta-1} e^{-\left(\frac{x}{\delta}\right)^\beta}, \quad x > 0, \beta > 0, \delta > 0$	$\delta\Gamma\left(1+\frac{1}{\beta}\right)$	$\delta^2\Gamma\left(1+\frac{2}{\beta}\right) - \delta^2\left[\Gamma\left(1+\frac{1}{\beta}\right)\right]^2$	4-11
Lognormal	$\frac{1}{x\omega\sqrt{2\pi}} e^{-\frac{(\ln x-\theta)^2}{2\omega^2}}, \quad x > 0$	$e^{\theta+\omega^2/2}$	$e^{2\theta+\omega^2}(e^{\omega^2}-1)$	4-12

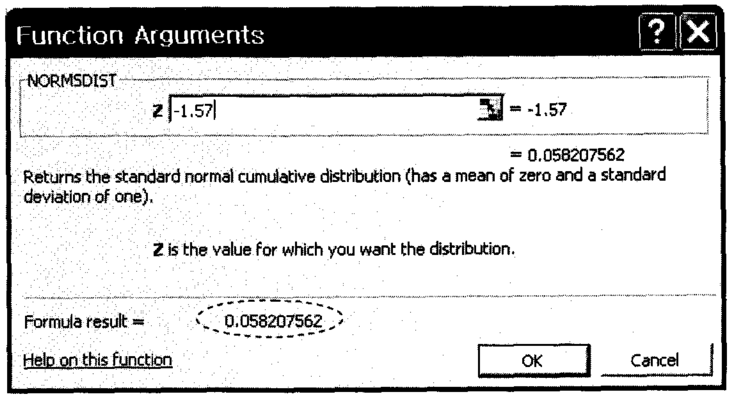
B. Comparison

Distribution	Population	Parameters: Rate of Success	Parameters: Length of Time Interval	Parameters: No. of Successes
Poisson	Infinite	Constant (λ)	Constant	Variable (X)
Exponential	Infinite	Constant (λ)	Variable (X)	Constant (1)
Erlang	Infinite	Constant (λ)	Variable (X)	Constant (positive integer r)
Gamma	Infinite	Constant (λ)	Variable (X)	Constant (positive real number r)
Weibull	Infinite	Changeable (β, δ)	Variable (X)	Constant (1)
Lognormal	Infinite	Changeable (θ, ω^2)	Variable (X)	Constant (1)

EXCEL Applications

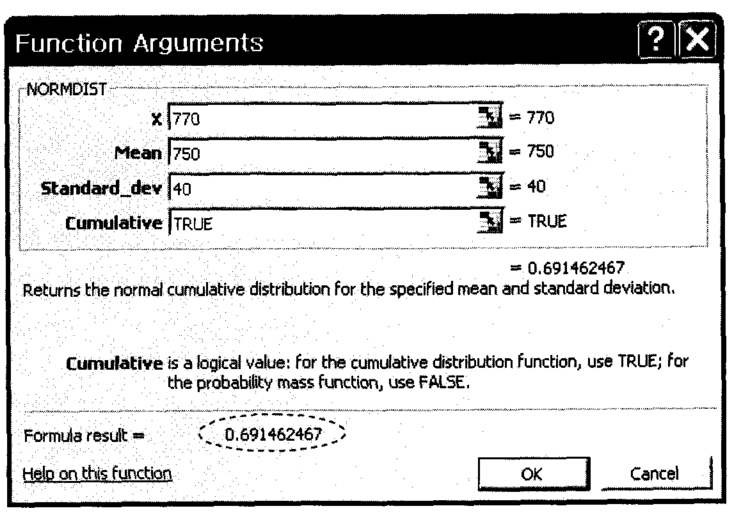
(Reading the z Table)

(1) Choose **Insert ➤ Function.**
(2) Select the category **Statistical** and the function **NORMSDIST**.
(3) Enter the value of Z (**Z**).
(4) Find the result circled with a dotted line below.



Example 4.5 **(Standardization)**

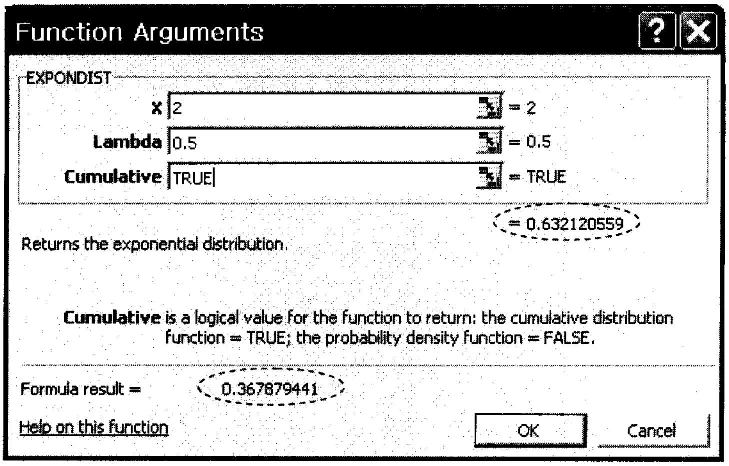
(1) Choose **Insert ➤ Function.**
(2) Select the category **Statistical** and the function **NORMDIST**.
(3) Enter the value of X (**X**), mean (**Mean**; μ), standard deviation (**Standard_dev**; σ), and logical value (**Cumulative**; true for the cumulative distribution function and false for the probability density function).
(4) Find the result circled with a dotted line below.



Example 4.7

2. **(Probability; Exponential Distribution)**

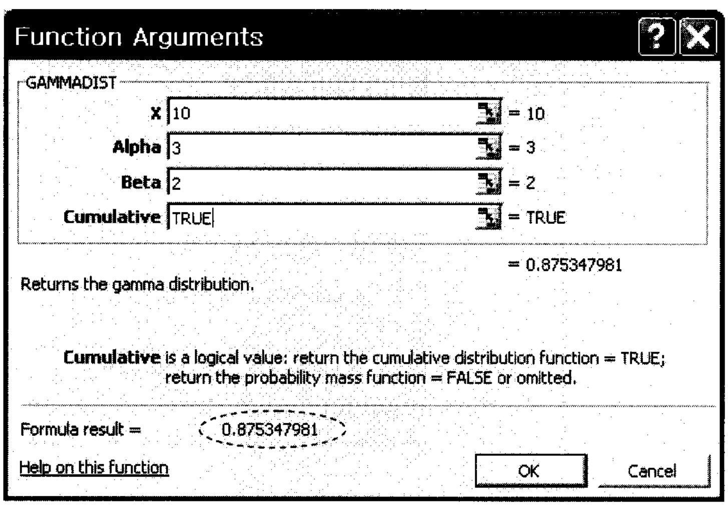
(1) Choose **Insert ➤ Function**.
(2) Select the category **Statistical** and the function **EXPONDIST**.
(3) Enter the value of X (**X**), average number of success per unit interval (**Lambda**; λ), and logical value (**Cumulative**; true for the cumulative distribution function and false for the probability density function).
(4) Find the results circled with a dotted line below.



Example 4.8

2. **(Probability; Erlang Distribution)**

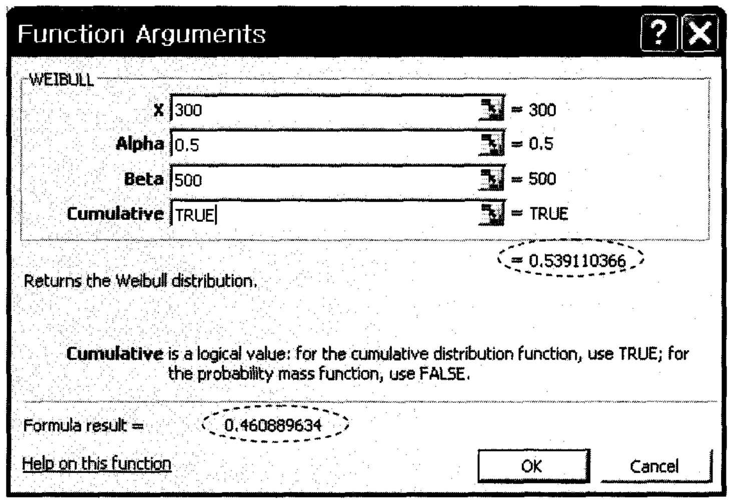
(1) Choose **Insert ➤ Function**.
(2) Select the category **Statistical** and the function **GAMMADIST**. Note that the gamma distribution is a general form of the Erlang distribution.
(3) Enter the value of X (**X**), number of successes (**Alpha**; r), average time until the next event (**Beta**; $1/\lambda$), and logical value (**Cumulative**; true for the cumulative distribution function and false for the probability density function). Note that a difference in notation exists between Excel and MR.
(4) Find the result circled with a dotted line below.



Example 4.9

2. (**Probability; Weibull Distribution**)

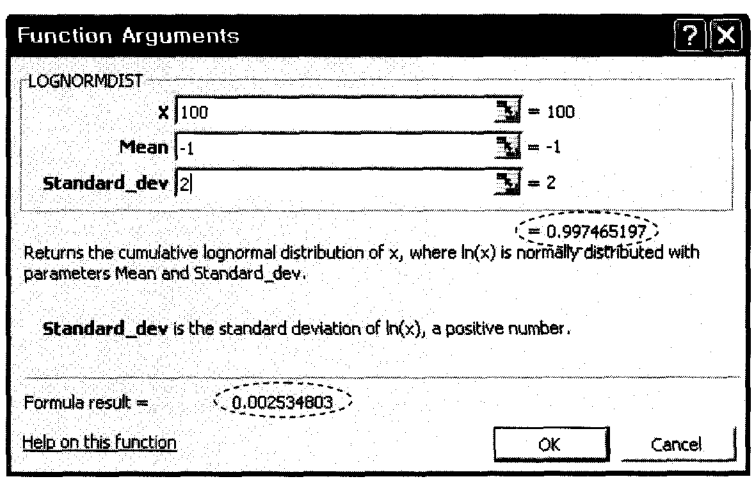
(1) Choose **Insert ➤ Function**.
(2) Select the category **Statistical** and the function **WEIBULL**.
(3) Enter the value of X (**X**), shape parameter (**Alpha**; β), scale parameter (**Beta**; δ), and logical value (**Cumulative**; true for the cumulative distribution function and false for the probability density function). Note that a difference in notation exists between Excel and MR.
(4) Find the results circled with a dotted line below.



Example 4.10

2. (**Probability; Lognormal Distribution**)

(1) Choose **Insert ➤ Function**.
(2) Select the category **Statistical** and the function **LOGNORMDIST**.
(3) Enter the values of X (**X**) and two parameters (**Mean**, θ; **Standard_dev**, ω) of the lognormal distribution.
(4) Find the results circled with a dotted line below.



Answers to Exercises

Exercise 4.1 **(Probability Density Function)**

$$(1)\ P\left(X<\frac{1}{3}\right)=\int_0^{1/3}f(x)dx=\int_0^{1/3}3x^2dx=x^3\Big|_0^{1/3}=\left(\frac{1}{3}\right)^3-0=\frac{1}{27}$$

$$(2)\ P\left(\frac{1}{3}\le X<\frac{2}{3}\right)=\int_{1/3}^{2/3}f(x)dx=\int_{1/3}^{2/3}3x^2dx=x^3\Big|_{1/3}^{2/3}=\left(\frac{2}{3}\right)^3-\left(\frac{1}{3}\right)^3=\frac{7}{27}$$

$$(3)\ P\left(X\ge\frac{2}{3}\right)=\int_{2/3}^{1}f(x)dx=\int_{2/3}^{1}3x^2dx=x^3\Big|_{2/3}^{1}=1-\left(\frac{2}{3}\right)^3=\frac{19}{27}$$

(Note) $P(X<\frac{1}{3})\ +\ P(\frac{1}{3}\le X<\frac{2}{3})\ +\ P(X\ge\frac{2}{3})\ =\ \frac{1}{27}+\frac{7}{27}+\frac{19}{27}=1$

Exercise 4.2 **(Cumulative Distribution Function)**

By using the probability density function of X,

$$F(x)=P(X\le x)=\int_0^x f(u)du=\int_0^x 3u^2du=u^3\Big|_0^x=x^3-0=x^3$$

Therefore,

$$F(x)=\begin{cases}0, & x<0\\ x^3, & 0\le x<1\\ 1, & 1\le x\end{cases}$$

Exercise 4.3 **(Mean, Variance, and Standard Deviation)**

(1) The mean of X is

$$\mu=\int_0^1 xf(x)dx=\int_0^1 x\cdot 3x^2dx=\int_0^1 3x^3dx=\frac{3}{4}x^4\Big|_0^1=\frac{3}{4}$$

(2) The variance of X is

$$\sigma^2=\int_0^1 x^2f(x)dx-\mu^2=\int_0^1 x^2\cdot 3x^2dx-\left(\frac{3}{4}\right)^2=\int_0^1 3x^4dx-\left(\frac{3}{4}\right)^2$$

$$=\frac{3}{5}x^5\Big|_0^1-\left(\frac{3}{4}\right)^2=\frac{3}{5}-\frac{9}{16}=\frac{3}{80}=0.038$$

(3) The standard deviation of X is

$$\sigma=\sqrt{V(X)}=\sqrt{0.038}=0.19$$

Exercise 4.4

1. **(Probability Density Function; Continuous Uniform Distribution)**

$a = 0.95,\quad b = 1.05$

$$f(x) = \frac{1}{b-a} = \frac{1}{1.05-0.95} = \frac{1}{0.1} = 10,\quad 0.95 \le x \le 1.05$$

2. **(Probability)**

$$P(X > 1.02) = \int_{1.02}^{1.05} f(x)dx = \int_{1.02}^{1.05} 10dx = 10x\Big|_{1.02}^{1.05} = 10(1.05-1.02) = 0.3$$

3. **(Mean and Variance)**

$$\mu = \frac{b+a}{2} = \frac{1.05+0.95}{2} = 1.0$$

$$\sigma^2 = \frac{(b-a)^2}{12} = \frac{(1.05-0.95)^2}{12} = 8.3\times 10^{-4} = 0.029^2$$

Exercise 4.5

(Standardization)

$$P(X > 130) = 1 - P(X \le 130) = 1 - P\left(\frac{X-\mu}{\sigma} \le \frac{130-\mu}{\sigma}\right)$$

$$= 1 - P\left(Z \le \frac{130-100}{20}\right) = 1 - P(Z \le 1.5) = 1 - 0.933 = 0.067$$

Exercise 4.6

1. **(Normal Approximation; Binomial Distribution)**

The probability mass function, mean, and variance of the binomial random variable X are

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{200}{x} 0.4^x 0.6^{200-x},\quad 0, 1, 2, \ldots, 100$$

$$\mu = np = 200 \times 0.4 = 80$$

$$\sigma^2 = np(1-p) = 200 \times 0.4 \times 0.6 = 48 = 6.93^2$$

Thus,

$$P(X \le 70) = \sum_{x=0}^{70} \binom{200}{x} 0.4^x 0.6^{200-x} = 0.084$$

By applying the normal approximation to the binomial random variable X,

$$P(X \le 70) = P\left(\frac{X-\mu}{\sigma} \le \frac{70-80}{6.93}\right) = P(Z \le -1.44) = 0.075$$

Since $np = 80 > 5$ and $n(1 - p) = 120 > 5$, the normal approximation to the binomial random variable X is satisfactory.

Exercise 4.6 (*cont.*)

2. **(Normal Approximation; Hypergeometric Distribution)**

The probability mass function, mean, and variance of X are

$$f(x)=\frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}=\frac{\binom{800}{x}\binom{2000-800}{200-x}}{\binom{2000}{200}}=\frac{\binom{800}{x}\binom{1200}{200-x}}{\binom{2000}{200}},\ x=0\text{ to }200$$

(Note) $\max\{0,n-(N-K)\}=\max\{0,200-(2000-800)\}$
$=\max\{0,-1000\}=0$
$\min\{K,n\}=\min\{800,200\}=200$

$$\mu=np=200\times0.4=80\text{, where } p=\frac{K}{N}=\frac{800}{2000}=0.4$$

$$\sigma^2=np(1-p)\left(\frac{N-n}{N-1}\right)=200\times0.4\times0.6\times\frac{2000-200}{2000-1}=43.2=6.57^2$$

Thus,

$$P(X\le70)=\sum_{x=0}^{70}\frac{\binom{800}{x}\binom{1200}{200-x}}{\binom{2000}{200}}=0.073$$

By applying the normal approximation to X

$$P(X\le70)=P\left(\frac{X-\mu}{\sigma}\le\frac{70-80}{6.57}\right)=P(Z\le-1.52)=0.064$$

Since $n/N=0.1$, $np=80>5$ and $n(1-p)=120>5$, the normal approximation to the hypergeometric random variable X is satisfactory.

3. **(Normal Approximation; Poisson Distribution)**

The probability mass function, mean, and variance of X are

$$f(x)=\frac{e^{-\lambda}\lambda^x}{x!}=\frac{e^{-80}80^x}{x!},\quad x=0,1,2,\ldots$$

$$\mu=\lambda=80$$

$$\sigma^2=\lambda=80=8.94^2$$

Thus,

$$P(X\le70)=\sum_{x=0}^{70}\frac{e^{-80}80^x}{x!}=0.143$$

By applying the normal approximation to X

$$P(X\le70)=P\left(\frac{X-\mu}{\sigma}\le\frac{70-80}{8.94}\right)=P(Z\le-1.12)=0.131$$

Since $\lambda=80>5$, the normal approximation to X is satisfactory.

Exercise 4.7

1. **(Probability Density Function; Exponential Distribution)**

For a consistent unit for X and λ,

λ = 1 failure/400 hrs = 0.0025 failure/hr

Therefore, the probability density function of X is

$$f(x) = \lambda e^{-\lambda x} = 0.0025e^{-0.0025x}, \quad x \geq 0 \text{ (unit: hr)}$$

2. **(Probability)**

By using the exponential distribution,

$$P(X < 100) = F(100) = 1 - e^{-0.25} = 0.221$$

The parameter of the corresponding Poisson distribution is

x = 100 hrs, λx = 0.0025 failure/hr × 100 hrs = 0.25

Therefore,

$$P(N \geq 1) = 1 - P(N = 0) = 1 - \frac{e^{-\lambda x}(\lambda x)^0}{0!} = 1 - e^{-\lambda x} = 1 - e^{-0.25} = 0.221$$

As shown above, the probability from the exponential distribution is equal to the probability from the Poisson distribution.

3. **(Mean and Variance)**

$$\mu = \frac{1}{\lambda} = \frac{1}{1/400} = 400 \text{ hrs until failure}$$

$$\sigma^2 = \frac{1}{\lambda^2} = \frac{1}{(1/400)^2} = 400^2$$

Exercise 4.8

1. **(Probability Density Function; Erlang Distribution)**

For a consistent unit for X and λ,

λ = 1 problem/30 days= 1/30 problem/day

Therefore, the probability density function of X with $r = 4$ problems is

$$f(x) = \frac{\lambda^r x^{r-1} e^{-\lambda x}}{(r-1)!} = \frac{0.033^4 x^{4-1} e^{-0.033x}}{(4-1)!} = 2.06 \times 10^{-7} x^3 e^{-0.033x},$$

where $x \geq 0$ (unit: day)

2. **(Probability)**

By using the Erlang distribution,

$$P(X > 120) = \int_{120}^{\infty} 2.06 \times 10^{-7} x^3 e^{-0.033x} dx = 0.433$$ (this integral is solved by integration by parts, which is not shown here)

The parameter of the corresponding Poisson distribution is

λx = 1/30 problem/day × 120 days = 4

Exercise 4.8 (*cont.*)

Therefore,

$$P(N<4)=\sum_{k=0}^{3}\frac{e^{-\lambda x}(\lambda x)^k}{k!}=\sum_{k=0}^{3}\frac{e^{-4}4^k}{k!}=0.433$$

The probability from the Erlang distribution is equal to the probability from the Poisson distribution.

3. **(Mean and Variance)**

$$\mu=\frac{r}{\lambda}=\frac{4}{1/30}=120 \text{ days}$$

$$\sigma^2=\frac{r}{\lambda^2}=\frac{4}{(1/30)^2}=60^2$$

Exercise 4.9

1. **(Probability Density Function; Weibull Distribution)**

$$f(x)=\frac{\beta}{\delta^\beta}x^{\beta-1}e^{-\left(\frac{x}{\delta}\right)^\beta}=\frac{2}{700^2}x^{2-1}e^{-\left(\frac{x}{700}\right)^2}=4.1\times10^{-6}xe^{-\left(\frac{x}{700}\right)^2},x\geq 0$$

2. **(Probability)**

$$P(X<700)=1-e^{-\left(\frac{x}{\delta}\right)^\beta}=1-e^{-\left(\frac{700}{700}\right)^2}=1-e^{-1}=0.632$$

3. **(Mean and Variance)**

$$\mu=\delta\Gamma\left(1+\frac{1}{\beta}\right)=700\Gamma\left(1+\frac{1}{2}\right)=700\Gamma(1.5)=700\times0.5\Gamma(0.5)$$

$$=700\times0.5\times\sqrt{\pi}=700\times0.886=620.4 \text{ hrs until failure}$$

$$\sigma^2=\delta^2\left\{\Gamma\left(1+\frac{2}{\beta}\right)-\left[\Gamma\left(1+\frac{1}{\beta}\right)\right]^2\right\}=700^2\left\{\Gamma\left(1+\frac{2}{2}\right)-\left[\Gamma\left(1+\frac{1}{2}\right)\right]^2\right\}$$

$$=700^2\left\{\Gamma(2)-[\Gamma(1.5)]^2\right\}=700^2\left\{1-0.886^2\right\}=105{,}155=324.3^2$$

Exercise 4.10

1. **(Probability Density Function; Lognormal Distribution)**

$$f(x)=\frac{1}{x\omega\sqrt{2\pi}}e^{-\frac{(\ln x-\theta)^2}{2\omega^2}}=\frac{1}{x\sqrt{2\pi}}e^{-\frac{(\ln x-0.5)^2}{2}},x\geq 0$$

2. **(Probability)**

$$P(X>10)=1-P(X\leq10)=1-\Phi\left(\frac{\ln 10-0.5}{1}\right)=1-\Phi(1.80)=1-0.96=0.04$$

3. **(Mean and Variance)**

$$\mu=e^{\theta+\omega^2/2}=e^{0.5+1/2}=2.7 \text{ sec.}$$

$$\sigma^2=e^{2\theta+\omega^2}(e^{\omega^2}-1)=e^{2\cdot0.5+1}(e^1-1)=3.6^2$$

5 Joint Probability Distributions

OUTLINE

5-1 Two Discrete Random Variables

Learning Goals

- ❑ Determine joint, marginal, and conditional probabilities for two discrete random variables X and Y by using corresponding probability distributions.
- ❑ Calculate the mean and variance of X (or Y) by using the corresponding marginal probability distribution.
- ❑ Calculate the conditional mean and conditional variance of X given $Y = y$ (or Y given $X = x$) by using the corresponding conditional probability distribution.
- ❑ Assess the independence of X and Y.

Probability Distributions of Two Random Variables

Three kinds of probability distributions are used to describe the stochastic characteristics of two random variables X and Y:

(1) Joint probability distribution
(2) Marginal probability distribution
(3) Conditional probability distribution

Joint Probability Mass Function

The joint probability mass function (p.m.f.) of two discrete random variables X and Y, denoted as $f_{XY}(x, y)$, satisfies the following conditions:

(1) $f_{XY}(x, y) = P(X = x, Y = y) \geq 0$

(2) $\sum_x \sum_y f_{XY}(x, y) = 1$

Marginal Probability Mass Function

The marginal p.m.f.'s of X and Y with the joint p.m.f. $f_{XY}(x, y)$ are

$$f_X(x) = P(X = x) = \sum_{R_x} f_{XY}(x, y)$$

$$f_Y(y) = P(Y = y) = \sum_{R_y} f_{XY}(x, y)$$

where R_x and R_y denote the set of all points in the range of (X, Y) for which $X = x$ and $Y = y$, respectively.

The marginal p.m.f. $f_X(x)$ satisfies the following:

(1) $f_X(x) = P(X = x) \geq 0$

(2) $\sum_x f_X(x) = 1$

The **mean** and **variance** of X are

$$E(X) = \mu_X = \sum_x x f_X(x)$$

$$V(X) = \sigma_X^2 = \sum_x (x - \mu_X)^2 f_X(x) = \sum_x x^2 f_X(x) - \mu_X^2$$

Conditional Probability Mass Function

Recall that $P(B \mid A) = P(A \cap B)/P(A)$ (see Section 2-4. Conditional Probability). In parallel, the conditional p.m.f. of X given $Y = y$, denoted as $f_{X|y}(x)$, with the joint p.m.f. $f_{XY}(x, y)$ is

$$f_{X|y}(x) = P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{f_{XY}(x, y)}{f_Y(y)}$$

The conditional p.m.f. $f_{X|y}(x)$ satisfies the following:

(1) $f_{X|y}(x) \geq 0$

(2) $\sum_x f_{X|y}(x) = 1$

The **conditional mean** and **conditional variance** of X given $Y = y$ are

$$E(X \mid y) = \mu_{X|y} = \sum_x x f_{X|y}(x)$$

$$V(X \mid y) = \sigma_{X|y}^2 = \sum_x (x - \mu_{X|y})^2 f_{X|y}(x) = \sum_x x^2 f_{X|y}(x) - \mu_{X|y}^2$$

Independence

Two random variables X and Y are independent if knowledge of the values of X does not affect any probabilities of the values of Y and vice versa. In other words, X and Y are not stochastically related.

Recall that $P(A \mid B) = P(A)$, $P(B \mid A) = P(B)$, and $P(A \cap B) = P(A)P(B)$ for two independent events A and B (see Section 2-4. Conditional Probability). Likewise, two independent discrete random variables X and Y satisfy any of the following:

(1) $f_{X|y}(x) = f_X(x)$

Independence (*cont.*)

(2) $f_{Y|x}(y) = f_Y(y)$

(3) $f_{XY}(x, y) = f_X(x) \cdot f_Y(y)$

Example 5.1

Grades in ergonomics (X) and grades in statistics (Y) are surveyed for 1,000 students as follows:

Ergonomics Grade (X)	Statistics Grade (Y) A ($y = 1$)	B ($y = 2$)	C ($y = 3$)	D or below ($y = 4$)	Totals
A ($x = 1$)	120	50	40	0	210
B ($x = 2$)	100	180	160	5	445
C ($x = 3$)	20	60	100	80	260
D or below ($x = 4$)	5	20	40	20	85
Totals	245	310	340	105	1,000

Let the values 1, 2, 3, and 4 of X and Y represent a letter grade A, B, C, and D or below, respectively.

1. **(Discrete Joint p.m.f.)** Determine the joint probability distribution of ergonomics grade (X) and statistics grade (Y). Find also the probability that a student earns an A for ergonomics ($X = 1$) and a B or above for statistics ($Y \le 2$).

The joint probabilities of X and Y are

$$f_{XY}(1,1) = 0.12,\ f_{XY}(1,2) = 0.05,\ f_{XY}(1,3) = 0.04,\ f_{XY}(1,4) = 0.0$$
$$f_{XY}(2,1) = 0.10,\ f_{XY}(2,2) = 0.18,\ f_{XY}(2,3) = 0.16,\ f_{XY}(2,4) = 0.005$$
$$f_{XY}(3,1) = 0.02,\ f_{XY}(3,2) = 0.06,\ f_{XY}(3,3) = 0.10,\ f_{XY}(3,4) = 0.08$$
$$f_{XY}(4,1) = 0.005,\ f_{XY}(4,2) = 0.02,\ f_{XY}(4,3) = 0.04,\ f_{XY}(4,4) = 0.02$$

Thus,

$$P(X = 1, Y \le 2) = \sum_{y=1}^{2} f_{XY}(1, y) = f_{XY}(1,1) + f_{XY}(1,2)$$
$$= 0.12 + 0.05 = 0.17$$

2. **(Discrete Marginal p.m.f.)** Determine the marginal probability distribution of ergonomics grade (X). Find also the mean and variance of X.

The marginal probabilities of X are

$$f_X(1) = \sum_{y=1}^{4} f_{XY}(1, y) = 0.21, \quad f_X(2) = \sum_{y=1}^{4} f_{XY}(2, y) = 0.445$$

$$f_X(3) = \sum_{y=1}^{4} f_{XY}(3, y) = 0.26, \quad f_X(4) = \sum_{y=1}^{4} f_{XY}(4, y) = 0.085$$

(Note) $\sum_{x} f_X(x) = f_X(1) + f_X(2) + f_X(3) + f_X(4) = 1$

Example 5.1 (*cont.*)

The mean and variance of X are

$$E(X) = \mu_X = \sum_{x=1}^{4} x f_X(x) = 1 \cdot f_X(1) + 2 \cdot f_X(2) + 3 \cdot f_X(3) + 4 \cdot f_X(4)$$
$$= 1 \times 0.21 + 2 \times 0.445 + 3 \times 0.26 + 4 \times 0.085 = 2.22$$

$$V(X) = \sigma_X^2 = \sum_x x^2 f_X(x) - \mu_X^2$$
$$= 1^2 \cdot f_X(1) + 2^2 \cdot f_X(2) + 3^2 \cdot f_X(3) + 4^2 \cdot f_X(4) - 2.22^2$$
$$= 1^2 \times 0.21 + 2^2 \times 0.445 + 3^2 \times 0.26 + 4^2 \times 0.085 - 2.22^2 = 0.87^2$$

3. (**Discrete Conditional p.m.f.**) Determine the conditional probability distribution of statistics grade (Y) given that ergonomics grade is an A ($X = 1$). Find also the conditional mean and conditional variance of Y given $X = 1$.

The conditional marginal probabilities of Y given $X = 1$ are

$$f_{Y|1}(1) = \frac{f_{XY}(1,1)}{f_X(1)} = \frac{0.12}{0.21} = 0.57, \quad f_{Y|1}(2) = \frac{f_{XY}(1,2)}{f_X(1)} = \frac{0.05}{0.21} = 0.24$$

$$f_{Y|1}(3) = \frac{f_{XY}(1,3)}{f_X(1)} = \frac{0.04}{0.21} = 0.19, \quad f_{Y|1}(4) = \frac{f_{XY}(1,4)}{f_X(1)} = \frac{0.0}{0.21} = 0.0$$

(Note) $\sum_y f_{Y|1}(y) = f_{Y|1}(1) + f_{Y|1}(2) + f_{Y|1}(3) + f_{Y|1}(4) = 1$

The conditional mean and conditional variance of Y given $X = 1$ are

$$E(Y\,|\,1) = \mu_{Y|1} = \sum_{y=1}^{4} y f_{Y|1}(y)$$
$$= 1 \cdot f_{Y|1}(1) + 2 \cdot f_{Y|1}(2) + 3 \cdot f_{Y|1}(3) + 4 \cdot f_{Y|1}(4)$$
$$= 1 \times 0.57 + 2 \times 0.24 + 3 \times 0.19 + 4 \times 0.0 = 1.62$$

$$V(Y\,|\,1) = \sigma_{Y|1}^2 = \sum_{y=1}^{4} y^2 f_{Y|1}(y) - \mu_{Y|1}^2$$
$$= 1^2 \cdot f_{Y|1}(1) + 2^2 \cdot f_{Y|1}(2) + 3^2 \cdot f_{Y|1}(3) + 4^2 \cdot f_{Y|1}(4) - 1.62^2$$
$$= 1^2 \times 0.57 + 2^2 \times 0.24 + 3^2 \times 0.19 + 4^2 \times 0.0 - 1.62^2 = 0.79^2$$

4. (**Independence**) Check if ergonomics grade (X) and statistics grade (Y) are independent.

Check if

$$f_{XY}(1,1) = f_X(1) \cdot f_Y(1)$$

By using the probabilities $f_{XY}(1,1) = 0.12$, $f_X(1) = 0.21$, and $f_Y(1) = 0.245$,

$$f_{XY}(1,1) = 0.12 \neq f_X(1) f_Y(1) = 0.21 \times 0.245 = 0.052$$

Thus, it is concluded that ergonomics grade (X) and statistics grade (Y) are not independent; in other words, ergonomics grade and statistics grade are related.

Exercise 5.1 (MR 5-5)

The number of defects on the front side (X) of a wooden panel and the number of defects on the rear side (Y) of the panel are under study.

1. Suppose that the joint p.m.f of X and Y is modeled as
   $$f_{XY}(x, y) = c(x + y), \quad x = 1, 2, 3 \quad \text{and} \quad y = 1, 2, 3$$
   Determine the value of c.
2. Determine the marginal p.m.f. of X. Find also the mean and variance of X.
3. Determine the conditional p.m.f. of Y given $X = 2$. Find also the conditional mean and conditional variance of Y given $X = 2$.
4. Check if the number of defects on the front side (X) of a wooden panel and the number of defects on the rear side (Y) of the panel are independent.

5-2 Multiple Discrete Random Variables

5-2.1 Joint Probability Distributions

Learning Goals

- ❑ Explain the joint, marginal, and conditional probability distributions of multiple discrete random variables.
- ❑ Explain the independence of multiple discrete random variables.

Joint Probability Mass Function

A joint p.m.f. of discrete random variables $X_1, X_2, \ldots, X_p$, which describes the probabilities of simultaneous outcomes of $X_1, X_2, \ldots, X_p$, is

$$f_R(x_1, x_2, \ldots, x_p) = P(X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p)$$

Next, the joint p.m.f. of $K = \{X_1, X_2, \ldots, X_k\}$, a subset of $R = \{X_1, X_2, \ldots, X_p\}$ with the joint p.m.f. $f_R(x_1, x_2, \ldots, x_p)$, is

$$f_K(x_1, x_2, \ldots, x_k) = \sum_{R_{x_1 x_2 \cdots x_k}} f_R(x_1, x_2, \ldots, x_p)$$

where $R_{x_1 x_2 \cdots x_k}$ denotes the set of all points in the range of R for which $X_1 = x_1$, $X_2 = x_2, \ldots, X_k = x_k$.

Marginal Probability Mass Function

A marginal p.m.f. of X_i, which belongs to $R = \{X_1, X_2, \ldots, X_p\}$ with the joint p.m.f. $f_R(x_1, x_2, \ldots, x_p)$, is

$$f_{X_i}(x_i) = P(X_i = x_i) = \sum_{R_{x_i}} f_R(x_1, x_2, \ldots, x_p)$$

where R_{x_i} denotes the set of all points in the range of R for which $X_i = x_i$.

Conditional Probability Mass Function

A conditional p.m.f. of $K = \{X_1, X_2, \ldots, X_k\}$ given $K' = \{X_{k+1}, X_{k+2}, \ldots, X_p\}$ describes the probabilities of outcomes of K at the given condition of K'. Note that K and K' are mutually exclusive.

As an extension of the bivariate conditional p.m.f., the conditional p.m.f. of K

Conditional Probability Mass Function (*cont.*)

given K' with the joint p.m.f. $f_R(x_1, x_2, \ldots, x_p)$ is

$$f_{K|K'}(x_1, x_2, \ldots, x_k) = \frac{P(X_1 = x_1, X_2 = x_2, \cdots, X_p = x_p)}{P(X_{k+1} = x_{k+1}, X_{k+2} = x_{k+2}, \cdots, X_p = x_p)}$$
$$= \frac{f_R(x_1, x_2, \ldots, x_p)}{f_{K'}(x_{k+1}, x_{k+2}, \ldots, x_p)}$$

Independence

As an extension of the bivariate independence conditions, multiple independent discrete random variables $R = \{X_1, X_2, \ldots, X_p\}$, consisting of two mutually exclusive subsets $K = \{X_1, X_2, \ldots, X_k\}$ and $K' = \{X_{k+1}, X_{k+2}, \ldots, X_p\}$, satisfy the following:

(1) $f_{K|K'}(x_1, x_2, \ldots, x_k \mid x_{k+1}, x_{k+2}, \ldots, x_p) = f_K(x_1, x_2, \ldots, x_k)$

(2) $f_{K'|K}(x_{k+1}, x_{k+2}, \ldots, x_p \mid x_1, x_2, \ldots, x_k) = f_{K'}(x_{k+1}, x_{k+2}, \ldots, x_p)$

(3) $f_R(x_1, x_2, \ldots, x_p) = f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_p}(x_p)$

5-2.2 Multinomial Probability Distributions

Learning Goals

- ❑ Describe the probability distribution of multinomial random variables.
- ❑ Determine the marginal probability distribution of multinomial random variables.

Multinomial Random Variables

Suppose that

(1) A random experiment consists of n trials.

(2) The outcome of each trial can be categorized into one of k classes.

(3) The multinomial random variables $X_1, X_2, \ldots, X_k$ represent the number of outcomes which belong to class 1, class 2, …, class k, respectively.

(4) $p_1, p_2, \ldots, p_k$ denote the probability of an outcome belonging to class 1, class 2, …, and class k, respectively.

(5) $p_1, p_2, \ldots, p_k$ are constant over n repeated trials.

Then, $X_1, X_2, \ldots, X_k$ have a multinomial distribution with the **joint p.m.f.**

$$f_K(x_1, x_2, \ldots, x_k) = \frac{n!}{x_1! x_2! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

where $x_1 + x_2 + \cdots + x_k = n$ and $p_1 + p_2 + \cdots + p_k = 1$

Note that the **marginal p.m.f.** of X_i, $i = 1$ to k, is a <u>binomial</u> distribution with mean and variance

$$E(X_i) = \mu_{X_i} = np_i \quad \text{and} \quad V(X_i) = \sigma_{X_i}^2 = np_i(1 - p_i)$$

Example 5.2

The maximum grip strength of an adult is categorized into one of three groups: 'low' for < 50 lbs., 'moderate' for 50 to 100 lbs., and 'high' for > 100 lbs. Suppose the probabilities that an adult belongs to the low, moderate, and high groups are 0.1, 0.7, and 0.2, respectively. A group of $n = 50$ people is measured for grip strength.

Example 5.2 (*cont.*)

1. (**Joint p.m.f.; Multinomial Distribution**) Find the probability that 6 people have low grip strength, 36 have moderate grip strength, and 8 people have high grip strength.

Let X_1, X_2, X_3 denote the number of people that have low, moderate, and high grip strength, respectively. Since $n = 50$, $p_1 = 0.1$, $p_2 = 0.7$, and $p_3 = 0.2$, the joint p.m.f. of X_1, X_2, X_3 is

$$f_{X_1X_2X_3}(x_1, x_2, x_3) = \frac{n!}{x_1!x_2!x_3!} p_1^{x_1} p_2^{x_2} p_3^{x_3} = \frac{50!}{x_1!x_2!x_3!} 0.1^{x_1} 0.7^{x_2} 0.2^{x_3},$$

where $x_1 + x_2 + \cdots + x_k = n$

Thus,

$$f_{X_1X_2X_3}(6, 36, 8) = \frac{50!}{6!36!8!} 0.1^6 \times 0.7^{36} \times 0.2^8 = 0.019$$

2. (**Marginal p.m.f.; Multinomial Distribution**) Find the probability that 36 people in the group have moderate grip strength.

The marginal p.m.f. of X_2 follows a binomial distribution with $n = 50$ and $p_2 = 0.7$, i.e.,

$$f_{X_2}(x_2) = \binom{n}{x_2} p_2^{x_2} (1 - p_2)^{n-x_2} = \binom{50}{x_2} 0.7^{x_2} 0.3^{50-x_2}, \quad x_2 = 0 \text{ to } 50$$

Thus,

$$f_{X_2}(36) = \binom{50}{36} 0.7^{36} 0.3^{50-36} = 0.119$$

Exercise 5.2 (MR 5-30)

In the transmission of digital information, the probabilities that a bit has high, moderate, and low distortion are 0.01, 0.04, and 0.95, respectively. Suppose that $n = 5$ bits are transmitted and the amount of distortion of each bit is independent.

1. Find the probability that one bit has high distortion, one has moderate distortion, and three bits have low distortion.
2. Find the probability that four bits have low distortion.

5-3 Two Continuous Random Variables

Learning Goals

- Determine joint, marginal, and conditional probabilities for two continuous random variables X and Y by using corresponding probability distributions.
- Calculate the mean and variance of X (or Y) a continuous random variable by using the corresponding marginal probability distribution.
- Calculate the conditional mean and conditional variance of X given $Y = y$ (or Y given $X = x$) by using the corresponding conditional probability distribution.
- Assess the independence of X and Y.

Joint Probability Density Function

The joint probability density function (p.d.f.) of two continuous random variables X and Y satisfies the following:

(1) $f_{XY}(x, y) \geq 0$

(2) $\int_y \int_x f_{XY}(x, y)\, dx\, dy = 1$

Marginal Probability Density Function

The marginal p.d.f.'s of X and Y with the joint p.d.f. $f_{XY}(x, y)$ are

$$f_X(x) = \int_y f_{XY}(x, y)\, dy$$

$$f_Y(y) = \int_x f_{XY}(x, y)\, dx$$

The marginal p.d.f. of X satisfies the following:

(1) $f_X(x) \geq 0$

(2) $\int_x f_X(x)\, dx = 1$

The **mean** and **variance** of X are

$$E(X) = \mu_X = \int_x x f_X(x)\, dx$$

$$V(X) = \sigma_X^2 = \int_x (x - \mu_X)^2 f_X(x)\, dx = \int_x x^2 f_X(x)\, dx - \mu_X^2$$

Conditional Probability Distribution

The conditional p.d.f. of X given $Y = y$, denoted as $f_{X|y}(x)$, with joint p.d.f. $f_{XY}(x, y)$ is

$$f_{X|y}(x) = \frac{f_{XY}(x, y)}{f_Y(y)}$$

The conditional p.d.f. $f_{X|y}(x)$ satisfies the following:

(1) $f_{X|y}(x) \geq 0$

(2) $\int_x f_{X|y}(x)\, dx = 1$

The **conditional mean** and **conditional variance** of X given $Y = y$ are

$$E(X \mid y) = \mu_{X|y} = \int_x x f_{X|y}(x)\, dx$$

$$V(X \mid y) = \sigma_{X|y}^2 = \int_x (x - \mu_{X|y})^2 f_{X|y}(x)\, dx = \int_x x^2 f_{X|y}(x)\, dx - \mu_{X|y}^2$$

Independence

Two continuous random variables X and Y are independent if any of the following is true:

(1) $f_{X|y}(x) = f_X(x)$

(2) $f_{Y|x}(y) = f_Y(y)$

(3) $f_{XY}(x, y) = f_X(x) f_Y(y)$

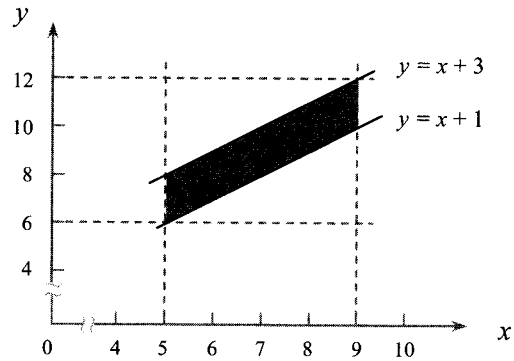
Example 5.3 The sizes of college books are under study in terms of width (X) and height (Y).

1. (**Continuous Joint p.d.f.**) Suppose that the joint p.d.f of X and Y is modeled as

$$f_{XY}(x, y) = cxy, \quad 5 < x < 9 \quad \text{and} \quad x + 1 < y < x + 3 \text{ (unit: inch)}$$

Determine the value of c.

The ranges of X and Y are displayed below.



The joint p.d.f. of X and Y should satisfy the following:

(1) $f_{XY}(x, y) = cxy \geq 0$

(2) $\int_y \int_x f_{XY}(x, y)\, dx\, dy = 1$

For the first condition, $c \geq 0$ because $x > 0$ and $y > 0$.

Next, for the second condition,

$$\int_y \int_x f_{XY}(x, y)\, dx\, dy = \int_5^9 \int_{x+1}^{x+3} cxy\, dy\, dx = c \int_5^9 \left(\tfrac{1}{2} y^2 \Big|_{x+1}^{x+3} \right) x\, dx$$

$$= \tfrac{1}{2} c \int_5^9 \left[(x+3)^2 - (x+1)^2 \right] x\, dx = 2c \int_5^9 (x^2 + 2x)\, dx$$

$$= 2c \left(\tfrac{1}{3} x^3 \Big|_5^9 + x^2 \Big|_5^9 \right) = 514.6c = 1 \quad \Rightarrow \quad c \cong 1/515$$

Therefore, the joint p.d.f. of X and Y is

$$f_{XY}(x, y) = \tfrac{1}{515} xy, \quad 5 < x < 9 \quad \text{and} \quad x + 1 < y < x + 3$$

2. (**Marginal Mean and Variance; Continuous Marginal p.d.f.**) Determine the marginal probability distribution of X. Find also the mean and variance of X.

The marginal p.d.f. of X is

$$f_X(x) = \int_y f_{XY}(x, y)\, dy = \tfrac{1}{515} \int_{x+1}^{x+3} xy\, dy = \tfrac{1}{515} x \left(\tfrac{1}{2} y^2 \Big|_{x+1}^{x+3} \right)$$

$$= \tfrac{1}{515} x \cdot \tfrac{1}{2} \left[(x+3)^2 - (x+1)^2 \right]$$

$$= \tfrac{2}{515} (x+2)x, \quad 5 < x < 9$$

The mean and variance of X are

$$E(X) = \mu_X = \int_x x f_X(x)\, dx = \tfrac{2}{515} \int_5^9 (x+2)x^2\, dx$$

Example 5.3
(*cont.*)

$$= \tfrac{2}{515}\left(\tfrac{1}{4}x^4\Big|_5^9 + \tfrac{2}{3}x^3\Big|_5^9\right) = \tfrac{2}{515}\times 1886.7 = 7.33 \text{ in.}$$

$$V(X) = \sigma_X^2 = \int_x x^2 f_X(x)\,dx - \mu_X^2 = \tfrac{2}{515}\int_5^9 (x+2)x^3\,dx - 7.33^2$$

$$= \tfrac{2}{515}\left(\tfrac{1}{5}x^5\Big|_5^9 + \tfrac{2}{4}x^4\Big|_5^9\right) - 7.33^2 = 1.23 = 1.13^2$$

3. (**Conditional Mean and Conditional Variance; Continuous Conditional p.d.f.**) Determine the conditional probability distribution of Y given $X = 8$. Find also the conditional mean and conditional variance of Y given $X = 8$.

The conditional p.d.f. of Y given $X = 8$ is

$$f_{Y|8}(y) = \frac{f_{XY}(8, y)}{f_X(8)} = \frac{\tfrac{1}{515}8y}{\tfrac{2}{515}\cdot(8+2)\cdot 8} = 0.05y, \quad 9 < y < 11$$

The conditional mean and conditional variance of Y given $X = 8$ are

$$E(Y\,|\,8) = \mu_{Y|8} = \int_9^{11} y f_{Y|8}(y)\,dy = 0.05\int_9^{11} y^2\,dy = 0.05\left(\tfrac{1}{3}y^3\Big|_9^{11}\right)$$

$$= 0.05\cdot\tfrac{1}{3}(11^3 - 9^3) = 10.03 \text{ in.}$$

$$V(Y\,|\,8) = \sigma_{Y|8}^2 = \int_9^{11} y^2 f_{Y|8}(y)\,dy - \mu_{Y|8}^2 = 0.05\int_9^{11} y^3 dy - 10.03^2$$

$$= 0.05\left(\tfrac{1}{4}y^4\Big|_9^{11}\right) - 10.03^2 = 0.39 = 0.63^2$$

4. (**Independence; Multiple Continuous Random Variables**) Check if the width (X) and height (Y) of a college book are independent.

Check if

$$f_{Y|x}(y) = \frac{f_{XY}(x, y)}{f_X(x)} = \frac{\tfrac{1}{515}xy}{\tfrac{2}{515}x(x+2)} = \frac{y}{2(x+2)} = f_Y(y)$$

The marginal p.d.f. of Y is

$$f_Y(y) = \int_x f_{XY}(x, y)\,dx = \int_x \tfrac{1}{515}xy\,dx = \tfrac{1}{515}y\int_x x\,dx$$

By the way, the range of X varies along y (see the graph on next page) as follows:

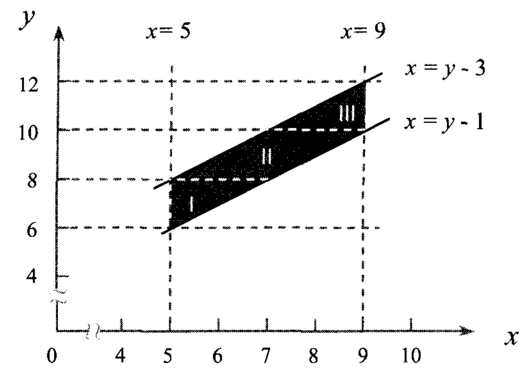
$5 < x < y - 3$	for $6 < y < 8$ (area I)
$y - 3 < x < y - 1$	for $8 < y < 10$ (area II)
$-3 < x < 9$	for $10 < y < 12$ (area III)

Thus, the marginal p.d.f. of Y is

$$f_Y(y) = \tfrac{1}{515}y\int_5^{y-1} x\,dx = \tfrac{1}{515}y\left(\tfrac{1}{2}x^2\Big|_5^{y-1}\right) = \tfrac{1}{1{,}030}y\left[(y-1)^2 - 5^2\right]$$

$$= \tfrac{1}{1{,}030}y(y^2 - 2y - 24) = \tfrac{1}{1{,}030}y(y-6)(y+4), \text{ for } 6 < y < 8$$

Example 5.3 (*cont.*)



$$f_Y(y) = \tfrac{1}{515}\, y \int_{y-3}^{y-1} x\,dx = \tfrac{1}{515}\, y\left(\tfrac{1}{2}x^2\Big|_{y-3}^{y-1}\right) = \tfrac{1}{1{,}030}\, y\left[(y-1)^2 - (y-3)^2\right]$$
$$= \tfrac{4}{1{,}030}\, y(y-2), \quad 8 < y < 10$$

$$f_Y(y) = \tfrac{1}{515}\, y \int_{y-3}^{9} x\,dx = \tfrac{1}{515}\, y\left(\tfrac{1}{2}x^2\Big|_{y-3}^{9}\right) = \tfrac{1}{1{,}030}\, y\left[9^2 - (y-3)^2\right]$$
$$= \tfrac{1}{1{,}030}\, y(-y^2 + 6y + 72) = \tfrac{1}{1{,}030}\, y(y+6)(12-y), \text{ for } 10 < y < 12$$

In summary,

$$f_Y(y) = \begin{cases} \tfrac{1}{1{,}030}\, y(y-6)(y+4), & 6 < y < 8 \\ \tfrac{4}{1{,}030}\, y(y-2), & 8 < y < 10 \\ \tfrac{1}{1{,}030}\, y(y+6)(12-y), & 10 < y < 12 \end{cases}$$

Since $f_{Y|x}(y) = \dfrac{y}{2(x+2)} \neq f_Y(y)$, the width ($X$) and height ($Y$) of a college book are not independent.

Exercise 5.3 (MR 5-49)

Two measurement methods are used to evaluate the surface smoothness of a paper product. Let X and Y denote the measurements of each of the two methods.

1. Suppose that the joint p.d.f of X and Y is modeled by
   $f_{XY}(x, y) = c, \quad 0 < x < 4, 0 < y \ \text{ and } \ x - 1 < y < x + 1$
   Determine the value of c.
2. Determine the marginal probability distribution of X. Find also the mean and variance of X.
3. Determine the conditional probability distribution of Y given $X = 2$. Find also the conditional mean and conditional variance of Y given $X = 2$.
4. Check if the measurements of the two methods X and Y are independent.

5-4 Multiple Continuous Random Variables

Learning Goals

- ❑ Explain the joint, marginal, and conditional probabilities of multiple continuous random variables.
- ❑ Explain the independence of multiple continuous random variables.

Joint Probability Density Function

As an extension of the bivariate joint p.d.f., the joint p.d.f. of multiple continuous random variables $R = \{X_1, X_2, \ldots, X_p\}$ satisfies the following:

(1) $f_R(x_1, x_2, \ldots, x_p) \geq 0$

(2) $\underset{R}{\int\int \cdots \int} f_R(x_1, x_2, \ldots, x_p)\, dx_1\, dx_2 \cdots dx_p = 1$

The joint p.d.f. of $K = \{X_1, X_2, \ldots, X_k\}$, a subset of $R = \{X_1, X_2, \ldots, X_p\}$ with the joint p.d.f. $f_R(x_1, x_2, \ldots, x_p)$, is

$$f_K(x_1, x_2, \ldots, x_k) = \underset{R_{x_1 x_2 \cdots x_k}}{\int \int \cdots \int} f_R(x_1, x_2, \ldots, x_p)\, dx_{k+1} dx_{k+2} \cdots dx_p$$

where $R_{x_1 x_2 \cdots x_k}$ denotes the set of all points in the range of R for which $X_1 = x_1$, $X_2 = x_2, \ldots, X_k = x_k$.

Marginal Probability Distribution

A marginal p.d.f. of X_i, which belongs to $R = \{X_1, X_2, \ldots, X_p\}$ with the joint p.d.f. $f_R(x_1, x_2, \ldots, x_p)$, is

$$f_{X_i}(x_i) = \underset{R_{x_i}}{\int \int \cdots \int} f_R(x_1, x_2, \ldots, x_p)\, dx_1\, dx_2 \cdots dx_{i-1} dx_{i+1} \cdots dx_p$$

where R_{x_i} denotes the set of all points in the range of R for which $X_i = x_i$.

Conditional Probability Distribution

As an extension of the bivariate conditional p.d.f., the conditional p.d.f. of $K = \{X_1, X_2, \ldots, X_k\}$ for a given condition of $K' = \{X_{k+1}, X_{k+2}, \ldots, X_p\}$ with the joint p.d.f. $f_R(x_1, x_2, \ldots, x_p)$ is

$$f_{K|K'}(x_1, x_2, \ldots, x_k) = \frac{f_R(x_1, x_2, \ldots, x_p)}{f_{K'}(x_{k+1}, x_{k+2}, \ldots, x_p)}$$

Independence

As an extension of the bivariate independence conditions, multiple continuous random variables $R = \{X_1, X_2, \ldots, X_p\}$, consisting of two mutually exclusive subsets $K = \{X_1, X_2, \ldots, X_k\}$ and $K' = \{X_{k+1}, X_{k+2}, \ldots, X_p\}$, are independent if

(1) $f_{K|K'}(x_1, x_2, \ldots, x_k) = f_K(x_1, x_2, \ldots, x_k)$

(2) $f_{K'|K}(x_{k+1}, x_{k+2}, \ldots, x_p) = f_{K'}(x_{k+1}, x_{k+2}, \ldots, x_p)$

(3) $f_R(x_1, x_2, \ldots, x_p) = f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_p}(x_p)$

5-5 Covariance and Correlation

Learning Goals

- ❑ Explain the terms *covariance* and *correlation* between two random variables X and Y.
- ❑ Calculate the covariance and correlation of X and Y.

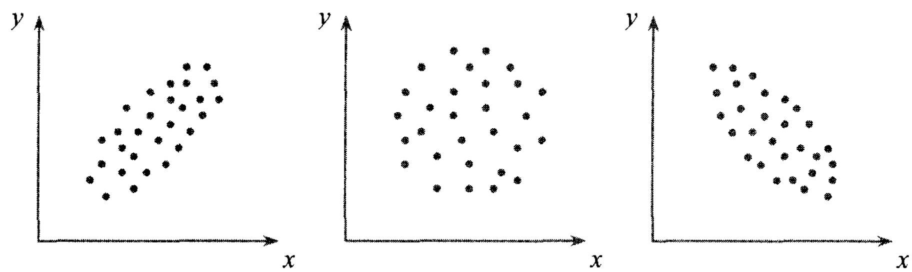
Covariance (σ_{XY})

The covariance between two random variables X and Y (denoted as cov(X, Y) or σ_{XY}) indicates the linear relationship between X and Y:

$$\sigma_{XY} = E[(X-\mu_X)(Y-\mu_Y)] = E(XY) - \mu_X\mu_Y, \quad -\infty < \sigma_{XY} < \infty$$

Note that the covariance may not be an appropriate measure to represent a relationship between random variables if they are related nonlinearly.

The positive, negative, and zero values of σ_{XY} represent positive, negative, and zero linear relationships between X and Y, respectively (see Figure 5-1).



(a) (b) (c)

Figure 5-1 Covariance and correlation between X and Y: (a) positive linearity ($\sigma_{XY} > 0$; $0 < \rho_{XY} < 1$); (b) zero linearity ($\sigma_{XY} = 0$; $\rho_{XY} = 0$); (c) negative linearity ($\sigma_{XY} < 0$; $-1 < \rho_{XY} < 0$).

(Derivation) $E[(X-\mu_X)(Y-\mu_Y)] = E(XY) - \mu_X\mu_Y$

$$\begin{aligned} E[(X-\mu_X)(Y-\mu_Y)] &= E[XY - \mu_X Y - \mu_Y X + \mu_X\mu_Y] \\ &= E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X\mu_Y \\ &= E(XY) - \mu_X\mu_Y - \mu_Y\mu_X + \mu_X\mu_Y \\ &= E(XY) - \mu_X\mu_Y \end{aligned}$$

Correlation (ρ_{XY})

The correlation between X and Y (denoted as ρ_{XY}) represents the normalized linear relationship between X and Y (σ_{XY} normalized by σ_X and σ_Y):

$$\rho_{XY} = \frac{\text{cov}(X,Y)}{\sqrt{V(X)V(Y)}} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}, \quad -1 < \rho_{XY} < 1$$

While σ_{XY} depends on the units of X and Y, ρ_{XY} is dimensionless. Like σ_{XY}, the positive, negative, and zero values of ρ_{XY} represent positive, negative, and zero linear relationships, respectively (see Figure 5-1).

Independence

If X and Y are independent (not related), then

$$\sigma_{XY} = \rho_{XY} = 0$$

However, $\sigma_{XY} = \rho_{XY} = 0$ is a <u>necessary (not sufficient) condition</u> for the independence of X and Y. In other words, even if $\sigma_{XY} = \rho_{XY} = 0$, we cannot immediately conclude that X and Y are independent.

Example 5.4

(Covariance and Correlation) In Example 5-1, determine the covariance and correlation between ergonomics grade (X) and statistics grade (Y).

The mean and variance of X are

$$E(X) = \mu_X = 2.22 \quad \text{and} \quad V(X) = \sigma_X^2 = 0.87^2$$

The mean and variance of Y are

$$E(Y) = \mu_Y = \sum_{y=1}^{4} y f_Y(y) = 1 \cdot f_Y(1) + 2 \cdot f_Y(2) + 3 \cdot f_Y(3) + 4 \cdot f_Y(4)$$
$$= 1 \times 0.245 + 2 \times 0.31 + 3 \times 0.34 + 4 \times 0.105 = 2.31$$
$$V(Y) = \sigma_Y^2 = \sum_{y} y^2 f_Y(y) - \mu_Y^2$$
$$= 1^2 \cdot f_Y(1) + 2^2 \cdot f_Y(2) + 3^2 \cdot f_Y(3) + 4^2 \cdot f_Y(4) - 2.31^2$$
$$= 1^2 \times 0.245 + 2^2 \times 0.31 + 3^2 \times 0.34 + 4^2 \times 0.105 - 2.31^2 = 0.95^2$$

The mean of XY is

$$E(XY) = \sum_{y=1}^{4}\sum_{x=1}^{4} xy f_{XY}(x,y) = 1 \cdot 1 \cdot f_{XY}(1,1) + 2 \cdot 1 \cdot f_{XY}(2,1) + \cdots + 4 \cdot 1 \cdot f_{XY}(4,1)$$
$$+ \cdots + 1 \cdot 4 \cdot f_{XY}(4,1) + 2 \cdot 4 \cdot f_{XY}(2,4) + \cdots + 4 \cdot 4 \cdot f_{XY}(4,4)$$
$$= 1 \cdot 1 \cdot 0.12 + 2 \cdot 1 \cdot 0.1 + \cdots + 4 \cdot 1 \cdot 0.05$$
$$+ \cdots + 1 \cdot 4 \cdot 0 + 2 \cdot 4 \cdot 0.05 + \cdots + 4 \cdot 4 \cdot 0.02 = 5.52$$

Therefore,

$$\sigma_{XY} = E(XY) - \mu_X \mu_Y = 5.52 - 2.22 \times 2.31 = 0.39$$
$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{0.39}{0.87 \times 0.95} = 0.47$$

Since σ_{XY} and $\rho_{XY} > 0$, it can be concluded that ergonomics grade (X) and statistics grade (Y) have a positive linear relationship.

Exercise 5.4 (MR 5-69)

In Exercise 5-1, the joint p.m.f. of X (the number of defects on the front side of a wooden panel) and Y (the number of defects on the rear side of the panel) is

$$f_{XY}(x, y) = \tfrac{1}{36}(x + y), \quad x = 1, 2, 3 \quad \text{and} \quad y = 1, 2, 3$$

The means and variances of X and Y are

$$E(X) = E(Y) = 2.17 \quad \text{and} \quad V(X) = V(Y) = 0.80^2$$

Determine the covariance and correlation of X and Y.

5-6 Bivariate Normal Distribution

Learning Goals

- ☐ Explain the joint probability distribution of bivariate normal random variables.
- ☐ Determine the joint, marginal, and conditional probabilities of bivariate normal random variables.

Bivariate Normal Random Variables

The **joint probability density function** of two normal random variables X and Y with means μ_X and μ_Y, variances σ_X^2 and σ_Y^2, and correlation ρ_{XY} $(-1 < \rho_{XY} < 1)$ is

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{XY}}}\exp\left\{\frac{-1}{2(1-\rho_{XY}^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho_{XY}(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right\}, \quad -\infty < x, y < \infty$$

Bivariate normal distributions with different values of ρ_{XY} are shown in Figure 5-2. The contour plots of the binomial curves are elliptical. The center of each ellipse is located at the point (μ_X, μ_Y) in the xy plane. If $\rho_{XY} > 0$ ($\rho_{XY} < 0$), the major axis of each ellipse is positive (negative); if $\rho_{XY} = 0$, the ellipses become circles.

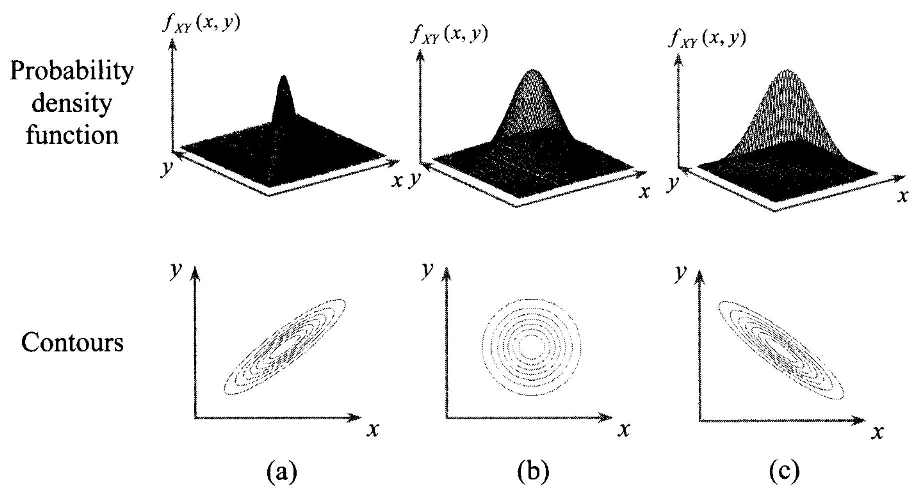


Figure 5-2 Bivariate normal distributions with different values of ρ_{XY}: (a) $0 < \rho_{XY} < 1$; (b) $\rho_{XY} = 0$; (c) $-1 < \rho_{XY} < 0$.

The **marginal probability distributions** of X and Y are normal with means μ_X and μ_Y and variances σ_X^2 and σ_Y^2, respectively.

The **conditional probability distribution** of Y given $X = x$ is normal with mean and variance

$$E(Y \mid x) = \mu_Y + \rho_{XY}\frac{\sigma_Y}{\sigma_X}(x-\mu_X) \quad \text{and} \quad V(Y \mid x) = \sigma_Y^2(1-\rho_{XY})$$

Example 5.5 (**Bivariate Normal Distribution**) The heights (X; unit: in.) and weights (Y; unit: lbs.) of adult males in the US are normally distributed with means $\mu_X = 68.3$ and $\mu_Y = 162.8$, variances $\sigma_X^2 = 2.3^2$ and $\sigma_Y^2 = 23.0^2$, and correlation $\rho_{XY} = 0.49$. Determine the probability that the weight (Y) of an adult male is between 160 and 180 lbs given that his height (X) is 70 in.

The conditional probability density function of Y given $X = 70$ is normal with mean and variance

$$E(Y \mid 70) = \mu_Y + \rho_{XY} \frac{\sigma_Y}{\sigma_X}(x - \mu_X) = 162.8 + 0.49 \times \frac{23.0}{2.3} \times (70 - 68.3)$$
$$= 162.8 + 0.49 \times \frac{23.0}{2.3} \times (70 - 68.3) = 171.1$$
$$V(Y \mid 70) = \sigma_Y^2 (1 - \rho_{XY}) = 23.0^2 \times (1 - 0.49) = 16.4^2$$

Therefore,

$$P(160 \le Y \le 180 \mid X = 70) = P\left(\frac{160 - 171.1}{16.4} \le \frac{Y - \mu_{Y|70}}{\sigma_{Y|70}} \le \frac{180 - 171.1}{16.4}\right)$$
$$= P\left(\frac{160 - 171.1}{16.4} \le Z \le \frac{180 - 171.1}{16.4}\right)$$
$$= P(-0.68 \le Z \le 0.54)$$
$$= P(Z \le 0.54) - P(Z \le -0.68)$$
$$= 0.706 - 0.249 = 0.457$$

Exercise 5.5 (MR 5-79) Let X and Y represent two dimensions of an injection molded part. Suppose that X and Y have a binomial distribution with means $\mu_X = 3.00$ and $\mu_Y = 7.70$, and variances $\sigma_X^2 = 0.04^2$ and $\sigma_Y^2 = 0.08^2$. Assume that X and Y are independent, i.e., $\rho_{XY} = 0$. Determine $P(2.95 < X < 3.05, 7.60 < Y < 7.80)$.

5-7 Linear Combinations of Random Variables

Learning Goals

- Explain the term *linear combination* of random variables.
- Determine the mean and variance of a linear combination of random variables.

Linear Combination A random variable Y is sometimes defined by a linear combination of several random variables $X_1, X_2, \ldots, X_p$:

$$Y = c_1 X_1 + c_2 X_2 + \cdots + c_p X_p, \quad \text{where } c_i\text{'s are constants}$$

Rules for Linear Combination

The following rules are useful to determine the mean and variance of a linear combination of X and Y:

1. **Rules for Means**
   (1) If b is a constant, then $E(b) = 0$.
   (2) If a is a constant, then $E(aX) = aE(X)$.
   (3) $E(aX \pm bY) = aE(X) \pm bE(Y)$

2. **Rules for Variances**
   (1) $V(b) = 0$.
   (2) $V(aX) = a^2V(X)$.
   (3) $V(aX \pm bY) = a^2V(X) + b^2V(Y) \pm 2ab\,\text{cov}(X,Y)$
   $= a^2V(X) + b^2V(Y)$, if X and Y are independent.
   (Note) $\text{cov}(X,Y) = \sigma_{XY} = \rho_{XY}\sigma_X\sigma_Y$

The rules 1.3 and 2.3 can be extended for $Y = c_1X_1 + c_2X_2 + \cdots + c_pX_p$ as follows:

$$E(Y) = c_1E(X_1) + c_2E(X_2) + \cdots + c_pE(X_p)$$

$$V(Y) = c_1^2V(X_1) + c_2^2V(X_2) + \cdots + c_p^2V(X_p) + 2\sum_{i=1}^{p-1}\sum_{j=i+1}^{p} c_ic_j\,\text{cov}(X_i, X_j)$$

$$= c_1^2V(X_1) + c_2^2V(X_2) + \cdots + c_p^2V(X_p)\text{, if } X_i\text{'s are independent}$$

Example 5.6

(**Linear Combination**) For adult males in the US, the length of the shoulder-elbow (X; unit: in.) and the length of the elbow-fingertip (Y; unit: in.) have a bivariate normal distribution with means $\mu_X = 14.4$ and $\mu_Y = 18.9$, variances $\sigma_X^2 = \sigma_Y^2 = 0.8^2$, and correlation $\rho_{XY} = 0.737$. Determine the mean and variance of the length of the shoulder-fingertip ($X + Y$).

The mean and variance of the length of the shoulder-fingertip $Z = X + Y$ is

$$E(X+Y) = E(X) + E(Y) = 14.4 + 18.9 = 33.3$$

$$V(X+Y) = V(X) + V(Y) + 2\,\text{cov}(X,Y) = 0.8^2 + 0.8^2 + 2\,\text{cov}(X,Y)$$

Note that $\text{cov}(X,Y) = \rho_{XY}\sigma_X\sigma_Y = 0.737 \times 0.8 \times 0.8 = 0.472$. Thus,

$$V(X+Y) = 0.8^2 + 0.8^2 + 2 \times 0.472 = 1.49^2$$

Exercise 5.6 (MR 5-91)

The width of a casing (X; unit: in.) and the width of a door (Y; unit: in.) are normally distributed with means $\mu_X = 24$ and $\mu_Y = 23\frac{7}{8}$, and standard deviations $\sigma_X = \frac{1}{8}$ and $\sigma_Y = \frac{1}{16}$, respectively. Assume that the width of the casing (X) and the width of the door (X) are independent. Determine the mean and standard deviation of the difference between the width of the casing (X) and the width of the door (Y).

5-9 Moment Generating Functions

Learning Goals

- ❑ Explain the term *moment generating function.*
- ❑ Determine the moment generating function of a random variable X and the mth moments of X.
- ❑ Find the mean and variance of X by using the first and second moments of X.

Moment Generating Function

The moment generating function of a random variable X (denoted as $M(t)$) is the expected value of e^{tX}, i.e.,

$$M(t) = E(e^{tX}) = \begin{cases} \sum_x e^{tx} f(x), & \text{if } X \text{ is a discrete random variable} \\ \int_{-\infty}^{\infty} e^{tx} f(x)dx, & \text{if } X \text{ is a continuous random variable} \end{cases}$$

The moment generating function of X is unique if it exists and completely determines the probability distribution of X. Thus, if two random variables have the same moment generating function, they have the same probability distribution.

Moment

If $M^{(m)}(t)$ denotes the mth derivative of $M(t)$, the mth moment of X about the origin (zero) (denoted by $E(X^m)$) is

$$E(X^m) = M^{(m)}(0) = \begin{cases} \sum_x x^m f(x), & \text{if } X \text{ is a discrete random variable} \\ \int_{-\infty}^{\infty} x^m f(x)dx, & \text{if } X \text{ is a continuous random variable} \end{cases}$$

(Derivation) $E(X^m) = M^{(m)}(0)$

The mth derivative of $M(t)$ is

$$M^{(m)}(t) = \frac{d^m M(t)}{dt^m} = \begin{cases} \sum_x x^m e^{tx} f(x), & \text{if } X \text{ is a discrete random variable} \\ \int_{-\infty}^{\infty} x^m e^{tx} f(x)dx, & \text{if } X \text{ is a continuous random variable} \end{cases}$$

By setting t = 0,

$$M^{(m)}(0) = \begin{cases} \sum_x x^m f(x), & \text{if } X \text{ is a discrete random variable} \\ \int_{-\infty}^{\infty} x^m f(x)dx, & \text{if } X \text{ is a continuous random variable} \end{cases}$$

Application of Moments

The mean and variance of X can be determined by using the first and second moments of X:

$$\mu = E(X) = M'(0)$$

$$\sigma^2 = E(X^2) - [E(X)]^2 = M''(0) - [M'(0)]^2$$

Example 5.7

Suppose that X has an exponential distribution

$$f(x) = \lambda e^{-\lambda x}, \quad x > 0$$

1. (**Moment Generating Function**) Find the moment generating function of X.

From the definition of moment generating function,

$$M(t) = \int_{-\infty}^{\infty} e^{tx} f(x)dx = \int_{0}^{\infty} \lambda e^{(t-\lambda)x} dx = \frac{\lambda}{t-\lambda} e^{(t-\lambda)x}\Big|_{0}^{\infty}$$

$$= \frac{\lambda}{\lambda - t} \quad \text{for } t - \lambda < 0$$

2. (**Application of Moments**) Determine the mean and variance of X by using the first and second moments of X about zero.

The first moment of X about the origin is

$$M'(0) = \frac{dM(t)}{dt}\Big|_{t=0} = \frac{d[\lambda(\lambda - t)^{-1}]}{dt}\Big|_{t=0} = \lambda(\lambda - t)^{-2}\Big|_{t=0} = \frac{1}{\lambda}$$

The second moment of X about the origin is

$$M''(0) = \frac{d^2 M(t)}{dt^2}\Big|_{t=0} = \frac{dM'(t)}{dt}\Big|_{t=0} = \frac{d[\lambda(\lambda - t)^{-2}]}{dt}\Big|_{t=0} = 2\lambda(\lambda - t)^{-3}\Big|_{t=0}$$

$$= \frac{2}{\lambda^2}$$

Therefore, the mean and variance of X are

$$\mu = E(X) = M'(0) = \frac{1}{\lambda}$$

and

$$\sigma^2 = E(X^2) - [E(X)]^2 = M''(0) - [M'(0)]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

Exercise 5.7 (MR S5-15)

The geometric random variable X has probability distribution

$$f(x) = (1-p)^{x-1} p, \quad x = 1, 2, \ldots, n$$

1. Find the moment generating function of X.
2. Determine the mean and variance of X by using the first and second moments of X about zero.

5-10 Chebyshev's Inequality

Learning Goals

- ❑ Explain the use of Chebyshev's inequality rule.
- ❑ Bound the probability of a random variable by using Chebyshev's inequality rule and compare the bound probability with the corresponding actual probability.

Chebyshev's Inequality

A relationship between the <u>mean</u> and <u>variance</u> of a random variable X having a certain <u>probability</u> distribution is formulated by Chebyshev as follows:

$$P(|X-\mu| \geq c\sigma) \leq 1/c^2, \quad c > 0$$

By using Chebyshev's inequality rule, a bound probability of any random variable can be determined. For example, Table 6-1 presents bound probabilities of a normal random variable X with μ and σ^2 (by using Chebyshev's inequality rule) and corresponding actual probabilities; the table indicates that a bound probability always includes the corresponding actual probability.

Table 6-1 Bound Probabilities and Actual Probabilities of a Normal Random Variable X

Probability Condition	c	Bound Probability ($1/c^2$)	Actual Probability
$P(\|X-\mu\| \geq 1.5\sigma)$	1.5	0.444	0.134
$P(\|X-\mu\| \geq 2\sigma)$	2	0.250	0.046
$P(\|X-\mu\| \geq 3\sigma)$	3	0.111	0.003
$P(\|X-\mu\| \geq 4\sigma)$	4	0.063	< 0.001

Example 5.8

(**Chebyshev's Inequality**) The power grip forces (X) of adult females in the US are normally distributed with a mean of 62.7 and a standard deviation of 17.1 lbs. Bound the probability that the grip strength of an adult female differs from the mean more than 2.5 times standard deviation. Compare also the bound probability with its actual probability.

By applying Chebyshev's inequality rule,

$$P(|X-\mu| \geq 2.5\sigma) \leq 1/2.5^2 = 0.16$$

The actual probability is

$$P(|X-\mu| \geq 2.5\sigma) = P\left(\frac{|X-\mu|}{\sigma} \geq 2.5\right) = P(|Z| \geq 2.5)$$
$$= P(Z < -2.5) + P(Z > 2.5) = 2 \times 0.006 = 0.012$$

The actual probability is less than the bound probability.

Exercise 5.8 (MR S5-25)

Suppose that the photoresist thickness (X) in semiconductor manufacturing has a continuous uniform distribution with a mean of 10 μm and a standard deviation of 2.31 μm over the range $6 < x < 14$ μm. Bound the probability that the photoresist thickness is less than 7 or greater than 13 μm. Compare also the bound probability with its actual probability.

Answers to Exercises

Exercise 5.1

1. **(Discrete Joint p.m.f.)**

The joint p.m.f. $f_{XY}(x, y)$, $x = 1, 2, 3$ and $y = 1, 2, 3$, should satisfy the following:

(1) $f_{XY}(x, y) = c(x + y) \geq 0$

(2) $\sum_x \sum_y f_{XY}(x, y) = 1$

For the first condition, $c \geq 0$ because $x > 0$ and $y > 0$.

Next, for the second condition,

$$\sum_{y=1}^{3}\sum_{x=1}^{3} f(x, y) = \sum_{y=1}^{3}\sum_{x=1}^{3} c(x + y) = 36c = 1$$

$\Rightarrow \quad c = 1/36$

Therefore, the joint p.m.f. of X and Y is

$$f_{XY}(x, y) = \tfrac{1}{36}(x + y), \quad x = 1, 2, 3 \quad \text{and} \quad y = 1, 2, 3$$

2. **(Discrete Marginal p.m.f.)**

The marginal probabilities of X are

$$f_X(1) = \sum_{y=1}^{3} f_{XY}(1, y) = \sum_{y=1}^{3} \frac{1}{36}(1 + y) = \frac{1}{36} \cdot 9 = \frac{1}{4}$$

$$f_X(2) = \sum_{y=1}^{3} f_{XY}(2, y) = \sum_{y=1}^{3} \frac{1}{36}(2 + y) = \frac{1}{36} \cdot 12 = \frac{1}{3}$$

$$f_X(3) = \sum_{y=1}^{3} f_{XY}(3, y) = \sum_{y=1}^{3} \frac{1}{36}(3 + y) = \frac{1}{36} \cdot 15 = \frac{5}{12}$$

(Note) $\sum_x f_X(x) = f_X(1) + f_X(2) + f_X(3) = 1$

The mean and variance of X are

$$E(X) = \mu_X = \sum_{x=1}^{3} x f_X(x) = 1 \cdot f_X(1) + 2 \cdot f_X(2) + 3 \cdot f_X(3)$$

$$= 1 \times \tfrac{1}{4} + 2 \times \tfrac{1}{3} + 3 \times \tfrac{5}{12} = 2.17$$

$$V(X) = \sigma_X^2 = \sum_x x^2 f_X(x) - \mu_X^2$$

$$= 1^2 \cdot f_X(1) + 2^2 \cdot f_X(2) + 3^2 \cdot f_X(3) - 2.17^2$$

$$= 1^2 \times \tfrac{1}{4} + 2^2 \times \tfrac{1}{3} + 3^2 \times \tfrac{5}{12} - 2.17^2 = 0.80^2$$

Exercise 5.1 (cont.)

3. **(Discrete Conditional p.m.f.)**

The conditional marginal probabilities of Y given $X = 2$ are

$$f_{Y|2}(1) = \frac{f_{XY}(2,1)}{f_X(2)} = \frac{\frac{1}{36}(2+1)}{\frac{1}{3}} = \frac{1}{4}$$

$$f_{Y|2}(2) = \frac{f_{XY}(2,2)}{f_X(2)} = \frac{\frac{1}{36}(2+2)}{\frac{1}{3}} = \frac{1}{3}$$

$$f_{Y|2}(3) = \frac{f_{XY}(2,3)}{f_X(2)} = \frac{\frac{1}{36}(2+3)}{\frac{1}{3}} = \frac{5}{12}$$

(Note) $\sum_y f_{Y|2}(y) = f_{Y|2}(1) + f_{Y|2}(2) + f_{Y|2}(3) = 1$

The conditional mean and conditional variance of Y given $X = 2$ are

$$E(Y\,|\,2) = \mu_{Y|2} = \sum_{y=1}^{3} y f_{Y|2}(y)$$

$$= 1 \cdot f_{Y|2}(1) + 2 \cdot f_{Y|2}(2) + 3 \cdot f_{Y|2}(3)$$

$$= 1 \times \tfrac{1}{4} + 2 \times \tfrac{1}{3} + 3 \times \tfrac{5}{12} = 2.17$$

$$V(Y\,|\,2) = \sigma^2_{Y|2} = \sum_{y=1}^{3} y^2 f_{Y|2}(y) - \mu^2_{Y|2}$$

$$= 1^2 \cdot f_{Y|2}(1) + 2^2 \cdot f_{Y|2}(2) + 3^2 \cdot f_{Y|2}(3) - 2.17^2$$

$$= 1^2 \times \tfrac{1}{4} + 2^2 \times \tfrac{1}{3} + 3^2 \times \tfrac{5}{12} - 2.17^2 = 0.80^2$$

4. **(Independence)**

Check if

$$f_{XY}(1,1) = f_X(1) \cdot f_Y(1)$$

By using $f_{XY}(1,1) = \frac{1}{36}(1+1) = \frac{1}{18}$, $f_X(1) = \frac{1}{4}$, and $f_Y(1) = \frac{1}{4}$,

$$f_{XY}(1,1) = \tfrac{1}{18} \neq f_X(1) f_Y(1) = \tfrac{1}{4} \times \tfrac{1}{4} = \tfrac{1}{16}$$

Thus, the number of defects on the front side (X) of a wooden panel and the number of defects on the rear side (Y) of the panel are not independent.

Exercise 5.2

1. **(Joint p.m.f.; Multinomial Distribution)**

Let X_1, X_2, X_3 denote the number of bits that have high, moderate, and low distortion, respectively. Since $n = 5$, $p_1 = 0.01$, $p_2 = 0.04$, and $p_3 = 0.95$, the joint p.m.f. of X_1, X_2, X_3 is

$$f_{X_1X_2X_3}(x_1, x_2, x_3) = \frac{n!}{x_1!x_2!x_3!} p_1^{x_1} p_2^{x_2} p_3^{x_3} = \frac{5!}{x_1!x_2!x_3!} 0.01^{x_1} 0.04^{x_2} 0.95^{x_3}$$

Thus,

$$f_{X_1X_2X_3}(1,1,3) = \frac{5!}{1!1!3!} 0.01 \times 0.04 \times 0.95^3 = 0.007$$

Exercise 5.2 (*cont.*)

2. **(Marginal p.m.f.)**

The marginal p.m.f. of X_3 follows a binomial distribution with $n = 5$ and $p_3 = 0.95$, i.e.,

$$f_{X_3}(x_3) = \binom{n}{x_3} p_3^{x_3}(1-p_3)^{n-x_3} = \binom{5}{x_3} 0.95^{x_3} 0.05^{5-x_3}, \quad x_3 = 0 \text{ to } 5$$

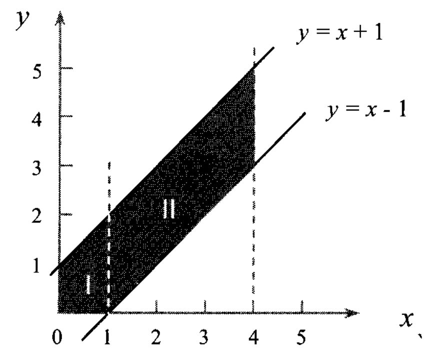
Thus,

$$f_{X_3}(4) = \binom{5}{4} 0.95^4 0.05^{5-4} = 0.204$$

Exercise 5.3

1. **(Continuous Joint p.d.f.)**

The ranges of X and Y are displayed below. Note that integration for $0 < x \le 1$ (area I) and that for $1 < x < 4$ (area II) should be conducted separately because they have different low limits.



The joint p.d.f. of X and Y should satisfy the following:

(1) $f_{XY}(x, y) = c \ge 0$

(2) $\int_y \int_x f_{XY}(x, y)\, dx\, dy = 1$

For the first condition, $c \ge 0$ because $x > 0$ and $y > 0$.

Next, for the second condition,

$$\int_y \int_x f_{XY}(x, y)\, dx\, dy = \int_0^1 \int_0^{x+1} c\, dy\, dx + \int_1^4 \int_{x-1}^{x+1} c\, dy\, dx = c\int_0^1 (x+1)\, dx + c\int_1^4 2\, dx$$

$$= c\left(\tfrac{1}{2}x^2\Big|_0^1 + x\Big|_0^1\right) + c\left(2x\Big|_1^4\right) = \tfrac{3}{2}c + 6c = 7.5c = 1$$

$\Rightarrow \quad c = 1/7.5 = 2/15$

Therefore, the joint p.d.f. of X and Y is

$$f_{XY}(x, y) = \tfrac{2}{15}, \quad 0 < x < 4, 0 < y \quad \text{and} \quad x - 1 < y < x + 1$$

Exercise 5.3 (*cont.*)

2. **(Continuous Marginal p.d.f.)**

The marginal p.d.f. of X is

$$f_X(x) = \int_y f_{XY}(x,y)\,dy = \begin{cases} \int_0^{x+1} \frac{2}{15}\,dy = \frac{2}{15}(x+1), & 0 < x \le 1 \\ \int_{x-1}^{x+1} \frac{2}{15}\,dy = \frac{2}{15}\times 2 = \frac{4}{15}, & 1 < x < 4 \end{cases}$$

The mean and variance of X are

$$E(X) = \mu_X = \int_x x f_X(x)\,dx = \int_0^1 \tfrac{2}{15}x(x+1)\,dx + \int_1^4 \tfrac{4}{15}x\,dx$$

$$= \tfrac{2}{15}\left(\tfrac{1}{3}x^3\Big|_0^1 + \tfrac{1}{2}x^2\Big|_0^1\right) + \tfrac{2}{15}\left(x^2\Big|_1^4\right) = \tfrac{2}{15}\cdot\tfrac{5}{6} + \tfrac{2}{15}\cdot 15 = 2.11$$

$$V(X) = \sigma_X^2 = \int_x x^2 f_X(x)\,dx - \mu_X^2 = \int_0^1 \tfrac{2}{15}x^2(x+1)\,dx + \int_1^4 \tfrac{4}{15}x^2\,dx - 2.11^2$$

$$= \tfrac{2}{15}\left(\tfrac{1}{4}x^4\Big|_0^1 + \tfrac{1}{3}x^3\Big|_0^1\right) + \tfrac{4}{15}\left(\tfrac{1}{3}x^3\Big|_1^4\right) - 2.11^2$$

$$= \tfrac{2}{15}\cdot\tfrac{7}{12} + \tfrac{4}{15}\cdot\tfrac{63}{3} - 2.11^2 = 1.23 = 1.11^2$$

3. **(Continuous Conditional p.d.f.)**

The conditional p.d.f. of Y given $X = 2$ is

$$f_{Y|2}(y) = \frac{f_{XY}(2,y)}{f_X(2)} = \frac{2/15}{4/15} = \frac{1}{2}, \quad 1 < y < 3$$

The conditional mean and conditional variance of Y given $X = 2$ are

$$E(Y\,|\,2) = \mu_{Y|2} = \int_0^4 y f_{Y|2}(y)\,dy = \tfrac{1}{2}\int_1^3 y\,dy = \tfrac{1}{2}\left(\tfrac{1}{2}y^2\Big|_1^3\right) = \tfrac{1}{4}\cdot(3^2 - 1) = 2$$

$$V(Y\,|\,2) = \sigma_{Y|2}^2 = \int_1^3 y^2 f_{Y|2}(y)\,dy - \mu_{Y|2}^2 = \tfrac{1}{2}\int_1^3 y^2 dy - 2^2$$

$$= \tfrac{1}{2}\left(\tfrac{1}{3}y^3\Big|_1^3\right) - 2^2 = \tfrac{1}{6}\cdot(3^3 - 1) - 2^2 = 0.33 = 0.58^2$$

4. **(Independence)**

Check if

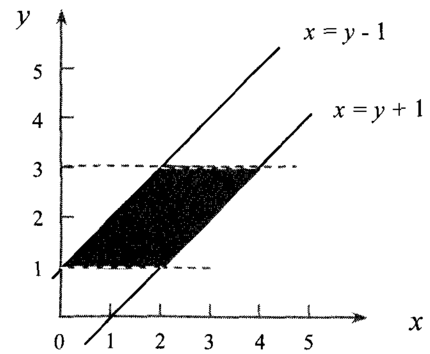
$$f_{Y|2}(y) = \frac{f_{XY}(2,y)}{f_X(2)} = \frac{1}{2} = f_Y(y), \quad 1 < y < 3$$

Note that the range of X for $1 < y < 3$ is $y - 1 < x < y + 1$ (see the graph on next page). Thus, tthe marginal p.d.f. of Y for $1 < y < 3$ is

$$f_Y(y) = \int_x f_{XY}(x,y)\,dx = \int_{y-1}^{4} \tfrac{2}{15}\,dx = \tfrac{2}{15}\left(x\Big|_{y-1}^{y+1}\right) = \tfrac{4}{15}, \quad 1 < y < 3$$

Since $f_{Y|2}(y) = \frac{1}{2} \ne f_Y(y) = \frac{4}{15}$, the measurements of the two methods X and Y are not independent.

Exercise 5.3 (*cont.*)



Exercise 5.4

(Covariance and Correlation)

The mean of XY is

$$E(XY)=\sum_{y=1}^{3}\sum_{x=1}^{3}xyf_{XY}(x,y)=\tfrac{1}{36}\sum_{y=1}^{3}\sum_{x=1}^{3}xy(x+y)$$

$$=\tfrac{1}{36}[1\cdot1\cdot(1+1)+2\cdot1\cdot(2+1)+3\cdot1\cdot(3+1)+\cdots$$
$$+1\cdot3\cdot(1+3)+2\cdot3\cdot(2+3)+3\cdot3\cdot(3+3)]=\tfrac{1}{36}\times168=4.67$$

Therefore,

$$\sigma_{XY}=E(XY)-\mu_X\mu_Y=4.67-2.17\times2.17=-0.04$$

$$\rho_{XY}=\frac{\sigma_{XY}}{\sigma_X\sigma_Y}=\frac{-0.04}{0.80\times0.80}=-0.06$$

There is a weak, negative correlation between the number of defects on the front side (X) of a wooden panel and the number of defects on the rear side (Y) of the panel.

Exercise 5.5

(Bivariate Normal Distribution)

Since X and Y are independent,

$$P(2.95<X<3.05, 7.60<Y<7.80)=P(2.95<X<3.05)P(7.60<Y<7.80)$$

By standardizing X and Y each,

$$P(2.95<X<3.05)=P\left(\frac{2.95-3.00}{0.04}<\frac{X-\mu_X}{\sigma_X}<\frac{3.05-3.00}{0.04}\right)$$
$$=P(-1.25<Z<1.25)=P(Z<1.25)-P(Z<-1.25)$$
$$=0.894-0.106=0.789$$

$$P(7.60<Y<7.80)=P\left(\frac{7.60-7.70}{0.08}<\frac{Y-\mu_Y}{\sigma_Y}<\frac{7.80-7.70}{0.08}\right)$$
$$=P(-1.25<Z<1.25)=0.789$$

Thus,

$$P(2.95<X<3.05, 7.60<Y<7.80)=0.789\times0.789=0.623$$

Exercise 5.6

(Linear Combination)

$X \sim N(24, \frac{1}{8}^2)$, $Y \sim N(23\frac{7}{8}, \frac{1}{16}^2)$, and $\text{cov}(X,Y) = 0$ because X and Y are independent.

Therefore,

$$E(X - Y) = E(X) - E(Y) = 24 - 23\tfrac{7}{8} = \tfrac{1}{8}$$

$$V(X - Y) = V(X) + V(Y) - 2\,\text{cov}(X,Y) = \tfrac{1}{8}^2 + \tfrac{1}{16}^2 - 0 = 0.14^2$$

Exercise 5.7

1. (**Moment Generating Function**)

From the definition of moment generating function,

$$M(t) = \sum_x e^{tx} f(x) = \sum_{x=1}^{\infty} e^{tx}(1-p)^{x-1} p = \sum_{x=1}^{\infty} pe^t [(1-p)e^t]^{x-1}$$

Note that the sum of the infinite geometric sequence ($a, ar, ar^2, \dots$) is

$$S = \sum_{n=1}^{\infty} ar^{n-1} = \frac{a}{1-r}, \quad \text{where } |r| < 1$$

Thus,

$$M(t) = \frac{pe^t}{1-(1-p)e^t}$$

2. (**Application of Moments**)

The first moment of X about the origin is

$$M'(0) = \left.\frac{dM(t)}{dt}\right|_{t=0} = \left.\frac{d\left\{pe^t[1-(1-p)e^t]^{-1}\right\}}{dt}\right|_{t=0}$$

$$= \left.\frac{pe^t}{1-(1-p)e^t} + \frac{p(1-p)e^{2t}}{[1-(1-p)e^t]^2}\right|_{t=0} = \frac{p}{1-(1-p)} + \frac{p(1-p)}{[1-(1-p)]^2}$$

$$= \frac{p}{p} + \frac{p(1-p)}{p^2} = \frac{1}{p}$$

The second moment of X about the origin is

$$M''(0) = \left.\frac{d^2M(t)}{dt^2}\right|_{t=0} = \left.\frac{dM'(t)}{dt}\right|_{t=0}$$

$$= \left.\frac{d\left\{pe^t[1-(1-p)e^t]^{-1}\right\}}{dt}\right|_{t=0} + \left.\frac{d\left\{p(1-p)e^{2t}[1-(1-p)e^t]^{-2}\right\}}{dt}\right|_{t=0}$$

$$= \left.\frac{1}{p} + \frac{2p(1-p)e^{2t}}{[1-(1-p)e^t]^2} + \frac{2p(1-p)^2e^{2t}}{[1-(1-p)e^t]^3}\right|_{t=0}$$

$$= \frac{1}{p} + \frac{2p(1-p)}{[1-(1-p)]^2} + \frac{2p(1-p)^2}{[1-(1-p)]^3}$$

Exercise 5.7 (*cont.*)

$$= \frac{1}{p} + \frac{2p(1-p)}{p^2} + \frac{2p(1-p)^2}{p^3}$$
$$= \frac{1}{p} + \frac{2(1-p)}{p} + \frac{2(1-p)^2}{p^2}$$
$$= \frac{2-p}{p^2}$$

Therefore, the mean and variance of X are

$$\mu = E(X) = M'(0) = np$$

and

$$\sigma^2 = E(X^2) - [E(X)]^2 = M''(0) - [M'(0)]^2$$
$$= \frac{2-p}{p^2} - \left(\frac{1}{p}\right)^2 = \frac{1-p}{p^2}$$

Exercise 5.8 **(Chebyshev's Inequality)**

The probability density function of the uniform random variable X are

$$f(x) = \frac{1}{b-a} = \frac{1}{14-6} = \frac{1}{8}, \quad 6 < x < 14$$

By applying Chebyshev's inequality rule,

$$P(X < 7) + P(X > 13) = P(X - 10 < 7 - 10) + P(X - 10 > 13 - 10)$$
$$= P(|X - 10| > 3) = P(|X - 10| > c\sigma) < 1/c^2$$

Therefore, the bound of $P(X < 7) + P(X > 13)$ is

$$P(|X - 10| > 3) = P(|X - 10| > 1.30 \times 2.31) < 1/1.3^2 = 0.59$$

The actual value of $P(X < 7) + P(X > 13)$ is

$$P(X < 7) + P(X > 13) = 1 - P(7 < x < 13) = 1 - \int_7^{13} \frac{1}{8}\,dx$$
$$= 1 - \tfrac{1}{8}x\Big|_7^{13} = 1 - \tfrac{6}{8} = 0.25$$

The actual probability is less than the bound probability, which supports Chebyshev's inequality rule.

6 Random Sampling and Data Description

OUTLINE

6-1 Data Summary and Display

Learning Goals

- ❑ Determine the mean, variance, standard deviation, minimum, maximum, and range of a set of data.

Mean The arithmetic mean of a set of data indicates the central tendency of the data.

1. **Sample Mean** (denoted as $\bar{x}$): The average of all the observations in a sample, which is selected from larger population of observations.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

2. **Population Mean** (denoted as μ): The average of all observations in the population.

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}, \quad \text{for a finite population with size } N$$

Variance and Standard Deviation The variance of a set of data indicates the dispersion of the data along the mean: the larger the variance, the wider the spread of the data about the mean.

Variance and Standard Deviation (*cont.*)

1. **Sample Variance**

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

2. **Population Variance**

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} = \frac{\sum_{i=1}^{N} x_i^2}{N} - \mu^2, \quad \text{for a finite population with size } N$$

Note that the divisor for the sample variance is n-1, while that for the population variance is N.

The **sample standard deviation** and **population standard deviation** are

$$s = \sqrt{s^2} \quad \text{and} \quad \sigma = \sqrt{\sigma^2}$$

Minimum and Maximum

The minimum and maximum of a set of data are the smallest and largest values of the data set, respectively.

Sample Range

The difference between the minimum and maximum of the sample.

$$R = \max(x_i) - \min(x_i), \quad i = 1, 2, \ldots, n$$

Example 6.1

(**Mean, Variance, and Range**) The scores (X; integers) of n = 30 students in a statistics test are as follows (arranged in ascending order):

65	68	70	72	72	73	74	75	75	79
81	83	84	84	85	86	87	88	88	88
88	88	90	92	92	94	96	96	97	98

Determine the (1) sample mean, (2) sample variance, (3) minimum, (4) maximum, and (5) sample range of the test scores. The following summary quantities have been calculated:

$$n = 30, \ \sum_{i=1}^{30} x_i = 2{,}508, \text{ and } \sum_{i=1}^{30} x_i^2 = 212{,}194$$

(1) $\bar{x} = \sum_{i=1}^{30} x_i / n = 2{,}508/30 = 83.6$

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

$$= \frac{212{,}194 - 2{,}508^2/30}{30-1} = 87.1 = 9.3^2$$

(3) minimum = 65
(4) maximum = 98
(5) $R = 98 - 65 = 33$

Exercise 6.1 (MR 6-15)

The numbers of cycles (X; integers) to failure of $n = 70$ aluminum test coupons under repeated alternating stress at 21,000 psi and 18 cycles/sec. are as follows (arranged in ascending order):

375	706	758	785	798	845	865	885	910	990
1000	1015	1016	1018	1020	1055	1085	1102	1102	1109
1115	1120	1203	1223	1238	1258	1260	1269	1270	1310
1315	1315	1330	1416	1421	1452	1468	1481	1501	1502
1512	1522	1535	1540	1560	1567	1578	1594	1605	1608
1642	1674	1730	1750	1750	1764	1781	1782	1792	1820
1883	1888	1890	1910	1940	2023	2100	2130	2215	2265

Determine the (1) sample mean, (2) sample variance, (3) minimum, (4) maximum, and (5) sample range of the test results. The following summary quantities have been calculated:

$$n = 70, \sum_{i=1}^{70} x_i = 98{,}256, \text{ and } \sum_{i=1}^{70} x_i^2 = 149{,}089{,}794$$

6-2 Random Sampling

Learning Goals

- ❑ Distinguish between population and sample.
- ❑ Describe the terms *random sample* and *statistic*.
- ❑ Explain why selecting a representative random sample is important in research.
- ❑ Describe the characteristics of a random sample.

Population vs. Sample

1. **Population**: The collection of all the elements of a universe of interest. The size of a population may be either (1) small, (2) large but finite, or (3) infinite.
   (e.g.) Population size
   (1) Students in a statistics class: Small
   (2) College students in the United States: Large but finite
   (3) College students in the world: Infinite

2. **Sample**: A set of elements taken from the population under study, i.e., a <u>subset of the population</u> (see Figure 6-1). Since it is impossible and/or impractical to examine the entire population in most cases, a sample is often used for statistical inference about the population.

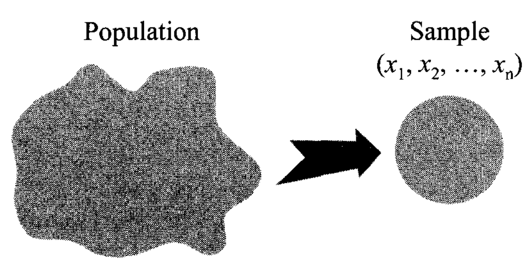


Figure 6-1 Population versus sample.

Random Sample

A random sample refers to a sample selected from the population under study by certain chance mechanism to avoid bias (over- or underestimation). Thus, selecting a random sample that properly represents the population is important for valid statistical inference about the population.

A random sample of size n is denoted by random variables $X_1, X_2, \ldots, X_n$ where X_i represents the i^{th} observation of the sample. The random sample $X_1, X_2, \ldots, X_n$ has the following characteristics:

(1) The X_i's are independent of each other, and

(2) Every X_i has the same probability distribution $f(x)$.

In other words, $X_1, X_2, \ldots, X_n$ are **independently and identically distributed (*i.i.d.*)**. Thus, the joint probability density function of $X_1, X_2, \ldots, X_n$ is

$$f_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n)=f(x_1)f(x_2)\cdots f(x_n)$$

Example 6.2

(**Random Sample**) A public opinion poll on contraception is conducted by calling people's homes, which are randomly selected, between 8 AM and 5 PM on weekdays. Discuss this study design.

Even if the telephone numbers are randomly chosen, the survey sample is biased because most working class people are not at home during the survey time period. Therefore, the survey result is most likely an over- or underestimate.

Exercise 6.2

Defects on sheet metal panels in three categories of sizes (small, medium, and large) are under study. The metal panels are made of the same material. The analyst randomly selects 20 large-size panels (not including small- and medium-size panels) and draws a conclusion regarding the average number of defects/ft^2. Discuss this study design.

Statistic

A statistic is a function of random variables $X_1, X_2, \ldots, X_n$; therefore, a statistic is also a random variable.

(e.g.) Statistic

(1) Sample mean $\overline{X}=\sum_{i=1}^{n} X_i / n$

(2) Sample variance $S^2=\sum_{i=1}^{n}(X_i-\overline{X})^2/(n-1)$

6-3 Stem-and-Leaf Diagrams

Learning Goals

- ❑ Explain the use of a stem-and-leaf diagram.
- ❑ Construct a stem-and-leaf diagram to visualize a set of data.
- ❑ Determine the $100k^{\text{th}}$ percentiles, quartiles, interquartile range, median, and mode of a set of data.
- ❑ Compare the variability measures variance, range, and interquartile range with each other.

Stem-and-Leaf Diagram

A stem-and-leaf diagram is a graphical display that presents both the individual values and corresponding frequency distribution simultaneously by using 'stems' and 'leaves.'

The following procedures are applied to construct a stem-and-leaf diagram (see Figure 6-2):

Step 1: Determine **stems** and **leaves**.

- The digits of each data is divided into two parts: (1) stem for the 'significant' digits (one or two digits in most cases) and (2) leaf for the 'less significant' digit (last digit usually). The analyst should exercise his/her own discretion to determine which digits are most significant with consideration of the range of data.
- The number of stems between 5 and 20 is commonly used. A stem can be further subdivided as necessary. As an example, the stem 5 can be divided into two stems 5L and 5U for the leaves 0 to 4 and for the leaves 5 to 9, respectively.

Step 2: **Arrange** the stems and leaves.

The stems are arranged in ascending (or descending) order. Then, beside each stem, corresponding leaves are listed next to each other in a row.

Step 3: Summarize the **frequency** of leaves for each stem.

Stem	Leaf	Frequency
3	5	1
4	0 3 9	3
5	2 5 9	3
6	1 3 4 8	4
7	2 3 7 8 8 8 9	7
8	0 2 2 7 9	5
9	1 3 8	3

Figure 6-2 A stem-and-leaf diagram.

Example 6.3

(Stem-and-Leaf Diagram) In Example 6-1, construct a stem-and-leaf diagram of the test scores.

Step 1: Determine **stems** and **leaves**.

The test scores range from 65 to 98. Each test score consists of two digits and the first digit has higher significance. Thus, the first and last digits of a test score are defined as the stem and leaf of the data, respectively: 6 to 9 for stems and 0 to 9 for leaves. To better represent the distribution of the test scores, the stems are further split into

6U, 7L, 7U, 8L, 8U, 9L, and 9U

where *L* is for the lower half leaves (0 to 4) and *U* is for the upper half leaves (5 to 9).

Example 6.3 (*cont.*)

Step 2: **Arrange** the stems and leaves.

Stem	Leaf	Frequency
6U	5 8	
7L	0 2 2 3 4	
7U	5 5 9	
8L	1 3 4 4	
8U	5 6 7 8 8 8 8 8	
9L	0 2 2 4	
9U	6 6 7 8	

Step 3: Summarize the **frequency** of leaves for each stem.

Stem	Leaf	Frequency
6U	5 8	2
7L	0 2 2 3 4	5
7U	5 5 9	3
8L	1 3 4 4	4
8U	5 6 7 8 8 8 8 8	8
9L	0 2 2 4	4
9U	6 6 7 8	4

Exercise 6.3 In Exercise 6-1, construct a stem-and-leaf diagram of the test results.

Percentile (p_k)

The $100k^{\text{th}}$ percentile ($0 < k \le 1$; denoted as p_k) of a set of data means the $100k^{\text{th}}$ largest value of the data set.

The procedures to find p_k from n observations, $x_1, x_2, \ldots, x_n$, are as follows:

Step 1: Arrange data in ascending order, say $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$.

Step 2: Calculate the position number r by using n and k

$$r = \begin{cases} nk, & \text{if } n \text{ is odd} \\ (n+1)k, & \text{if } n \text{ is even} \end{cases}$$

Step 3: Determine p_k based on r

$$p_k = \begin{cases} x_{(r)}, & \text{if } r \text{ is an integer} \\ x_{(\lfloor r \rfloor)} + (x_{(\lceil r \rceil)} - x_{(\lfloor r \rfloor)})(r - \lfloor r \rfloor), & \text{if } r \text{ is not an integer} \end{cases}$$

(Note) The symbols $\lceil r \rceil$ (called ceiling) and $\lfloor r \rfloor$ (called floor) denote 'round-up' and 'round-down' r to the nearest integer, respectively.

Quartile (q_k)

The following three quartiles divide a set of data into four equal parts in ascending order:

(1) First (lower) quartile: $q_1 = p_{0.25}$

(2) Second quartile (**median**): $q_2 = p_{0.50}$

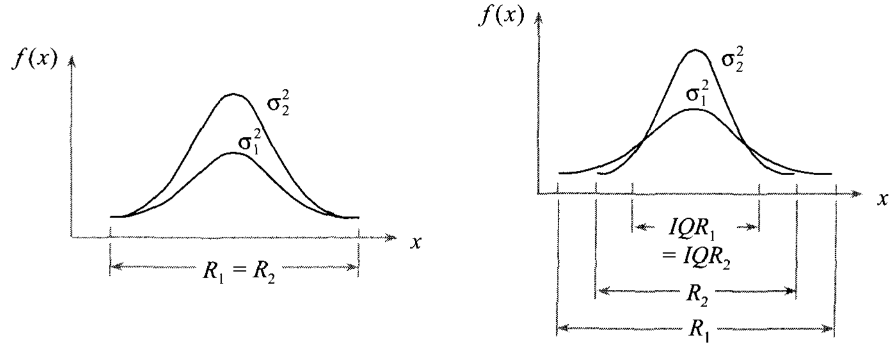
(3) Third (upper) quartile: $q_3 = p_{0.75}$

Interquartile Range (*IQR*)

The interquartile range (denoted as *IQR*) of a set of data is the difference between the upper and lower quartiles of the dataset, i.e.,

$$IQR = q_3 - q_1 = p_{0.75} - p_{0.25}$$

Comparison of Variability Measures The variance, range, and interquartile range of a data set have different sensitivity (stability): variance is most sensitive (least stable) and interquartile is least sensitive (most stable). For example, Figure 6-3(a) shows that two different data sets with the same range can have different variances; and Figure 6-3(b) displays that two different data sets with the same interquartile range can have different ranges and variances. These illustrations imply that variance, range, and interquartile range have high (low), moderate, and low (high) sensitivities (stabilities), respectively.



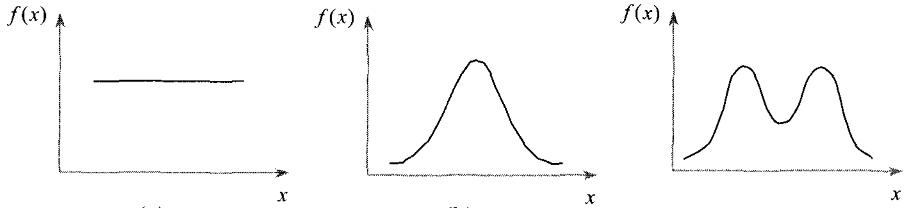
(a) (b)

Figure 6-3 Sensitivities of variance, range, and interquartile range: (a) $R_1 = R_2$, while $\sigma_1^2 \neq \sigma_2^2$; (b) $IQR_1 = IQR_2$, while $R_1 \neq R_2$ and $\sigma_1^2 \neq \sigma_2^2$.

Of the three variability measures, the appropriate variability measure is selected depending on the analyst's judgment on which measure best represents the variability of the data.

Median The median of a set of data is the value that divides its ranked data set into two equal halves, i.e., $p_{0.50} = q_2$.

Mode The mode of a set of data is the value that is observed most frequently. A data set can have no mode, one mode (unimodal), or multiple modes (bimodal, trimodal, etc.), as illustrated in Figure 6-4.



(a) (b) (c)

Figure 6-4 Modes of data distributions: (a) no mode; (b) unimodal; (c) bimodal.

Example 6.4 **(Percentile, Quartile, Interquartile Range, Median, and Mode)** In Example 6-1, determine the (1) 80^{th} percentile, (2) first quartile, (3) interquartile range, (4) median, and (5) mode of the test scores. Note that $n = 30$ and $q_3 = 90.5$.

Example 6.4 (*cont.*)

(1) Since $n = 30$ is even, $r = (n+1)k = 31 \times 0.8 = 24.8$, which is not an integer. Therefore,

$$\begin{aligned} p_{0.8} &= x_{(\lfloor r \rfloor)} + (x_{(\lceil r \rceil)} - x_{(\lfloor r \rfloor)})(r - \lfloor r \rfloor) \\ &= x_{(\lfloor 24.8 \rfloor)} + (x_{(\lceil 24.8 \rceil)} - x_{(\lfloor 24.8 \rfloor)})(24.8 - \lfloor 24.8 \rfloor) \\ &= x_{(24)} + (x_{(25)} - x_{(24)})(24.8 - 24) = 92 + (92 - 92) \times 0.8 = 92 \end{aligned}$$

(2) Since $r = (n+1)k = 31 \times 0.25 = 7.75$ is not an integer,

$$\begin{aligned} q_1 = p_{0.25} &= x_{(\lfloor r \rfloor)} + (x_{(\lceil r \rceil)} - x_{(\lfloor r \rfloor)})(r - \lfloor r \rfloor) \\ &= x_{(\lfloor 7.75 \rfloor)} + (x_{(\lceil 7.75 \rceil)} - x_{(\lfloor 7.75 \rfloor)})(7.75 - \lfloor 7.75 \rfloor) \\ &= x_{(7)} + (x_{(8)} - x_{(7)})(7.75 - 7) = 74 + (75 - 74) \times 0.75 = 74.75 \end{aligned}$$

(3) $IQR = q_3 - q_1 = p_{0.75} - p_{0.25} = 90.5 - 74.75 = 15.75$

(4) Since $r = (n+1)k = 31 \times 0.5 = 15.5$ is not an integer,

$$\begin{aligned} p_{0.5} &= x_{(\lceil r \rceil)} + (x_{(\lfloor r \rfloor)} - x_{(\lceil r \rceil)})(r - \lceil r \rceil) \\ &= x_{(\lceil 15.5 \rceil)} + (x_{(\lfloor 15.5 \rfloor)} - x_{(\lceil 15.5 \rceil)})(15.5 - \lceil 15.5 \rceil) \\ &= x_{(15)} + (x_{(16)} - x_{(15)})(15.5 - 15) = 85 + (86 - 85) \times 0.5 = 85.5 \end{aligned}$$

(5) Mode = 88 (unimodal)

Exercise 6.4

In Exercise 6-1, determine the (1) 90^{th} percentile, (2) third quartile, (3) interquartile range, (4) median, and (5) mode of the test results. Note that $n = 70$ and $q_1 = 1{,}097.75$.

6-4 Frequency Distributions and Histograms

Learning Goals

- ❑ Explain the terms *frequency*, *cumulative frequency*, *relative frequency*, and *relative cumulative frequency*.
- ❑ Construct a frequency table of a set of data.
- ❑ Construct a bar graph and histogram to visualize a set of data.
- ❑ Explain the term *skewness* and describe the relationships between the central tendency measures mean, median, and mode depending on the skewness of data.

Frequency

A frequency indicates the number of observations that belong to a bin (class interval, cell, or category). The frequency of the i^{th} bin is denoted by f_i.

The sum of the frequencies of bins is called **cumulative frequency** (denoted by F_i):

$$F_i = \sum_{j=1}^{i} f_j, \quad i = 1, 2, \ldots, k$$

Relative Frequency

A relative frequency is the ratio of a frequency to the total number of observations (n). The relative frequency of the i^{th} bin (denoted by p_i) is

$$p_i = \frac{f_i}{n}, \quad n = \text{total number of observations}$$

The sum of the relative frequencies of bins is called **cumulative relative frequency** (denoted by P_i):

$$P_i = \sum_{j=1}^{i} p_j, \quad i = 1, 2, \ldots, k$$

Example 6.5

(Frequency Measures) In Example 6-1, prepare a frequency table (including frequency, relative frequency, and cumulative relative frequency) of the test scores with five class intervals: below 60, 60 to 69, 70 to 79, 80 to 89, and 90 or above.

Class Interval (X)	Frequency	Relative Frequency	Cumulative Relative Frequency
$x < 60$	0	0.00	0.00
$60 \le x < 70$	2	0.07	0.07
$70 \le x < 80$	8	0.27	0.33
$80 \le x < 90$	12	0.40	0.73
$90 \le x$	8	0.27	1.00

Exercise 6.5

In Exercise 6-1, prepare a frequency table (including frequency, relative frequency, and cumulative relative frequency) of the failure test results with five class intervals: below 500, 500 to 999, 1,000 to 1,499, 1,500 to 1,999, and 2,000 or above.

Bar Graph

A bar graph is a graphical display with bars, the length of each bar being proportional to the frequency and/or relative frequency of the corresponding category. A bar graph is used to provide a visual comparison of the frequencies and/or relative frequencies of multiple categories.

Histogram

A histogram is a bar graph in which the length and width of each bar are proportional to the frequency (or relative frequency) and size of the corresponding bin, respectively. The shape of a histogram of a small set of data may vary significantly as the number of bins and corresponding bin width change. As the size of a data set becomes large (say, 75 or above), the shape of the histogram becomes stable.

Different histograms can be constructed for the same data set depending on the analyst's judgment. The following procedures may be of use to construct a histogram with bins of an equal width for a data set with size n:

Step 1: Determine the **number of bins** (k).

The following formula can be used for k:

$$k \cong \sqrt{n}$$

(Note) The number of bins between 5 and 20 is used in most cases.

Histogram (*cont.*)

Step 2: Determine the **width of a bin** (w).

The bin width is selected based on the range of the data set and the number of bins:

$$w > \frac{R}{k} = \frac{\max - \min}{k}$$

Step 3: Determine the **limits of the histogram**.

The lower limit (l) of the first bin and the upper limit (u) of the last bin are slightly lower than the minimum and larger than the maximum, respectively, i.e.,

$$l \cong \min - \frac{kw - R}{2}, \ u = l + kw; \text{ or } u \cong \max + \frac{kw - R}{2}, \ l = u - kw$$

Step 4: Prepare a **frequency table** based on the bins determined.

Count the number of observations that belong to each bin and calculate relative frequencies accordingly.

Step 5: Develop a **histogram**.

Construct a bar graph by using the frequencies (**frequency histogram**) and/or relative frequencies (**relative frequency histogram**) of the bins.

Example 6.6

(**Bar Graph/Histogram**) In Example 6-1, construct a histogram of the test scores by using the following summary quantities:

$n = 30$, minimum $= 65$, and maximum $= 98$

Step 1: Determine the **number of bins** (k).

Since $\sqrt{n} = \sqrt{30} = 5.5$, $k = 6$.

Step 2: Determine the **width of a bin** (w).

$$\text{Since } \frac{R}{k} = \frac{\max - \min}{k} = \frac{98 - 65}{6} = 5.5, \ w = 6.$$

Step 3: Determine the **limits of the histogram**.

$$l \cong \min - \frac{kw - R}{2} = 65 - \frac{6 \times 6 - 33}{2} = 63.5 \cong 64$$

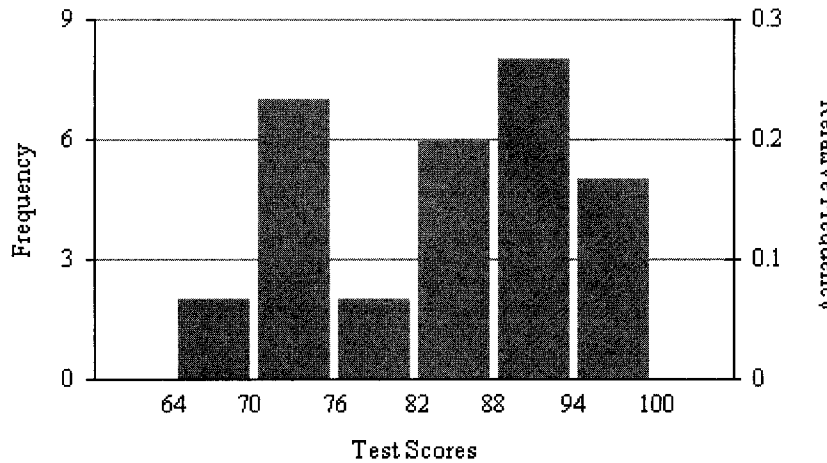
$$u = l + kw = 64 + 6 \times 6 = 100$$

Step 4: Prepare a **frequency table** based on the bins determined.

Class Interval (X)	Frequency	Relative Frequency	Cumulative Relative Frequency
$64 \le x < 70$	2	0.07	0.07
$70 \le x < 76$	7	0.23	0.30
$76 \le x < 82$	2	0.07	0.37
$82 \le x < 88$	6	0.20	0.57
$88 \le x < 94$	8	0.27	0.83
$94 \le x < 100$	5	0.17	1.00

Example 6.6 (*cont.*)

Step 5: Develop a **histogram**.



Exercise 6.6

In Exercise 6-1, construct a histogram of the test results by using the following summary quantities:

$n = 70$, minimum = 375, and maximum = 2,265

Skewness

The skewness (denoted by γ) of a set of data indicates the degree of asymmetry of the data set around the mean.

As illustrated in Figure 6-5, the distribution with a long tail to the left and distribution with a long tail to the right have negative and positive values of skewness, respectively; the symmetric curve has a zero skewness.

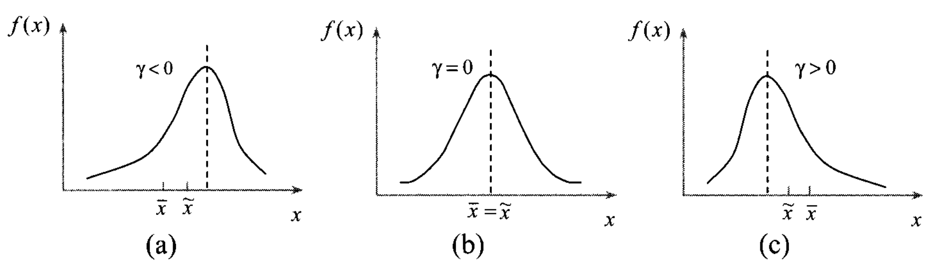


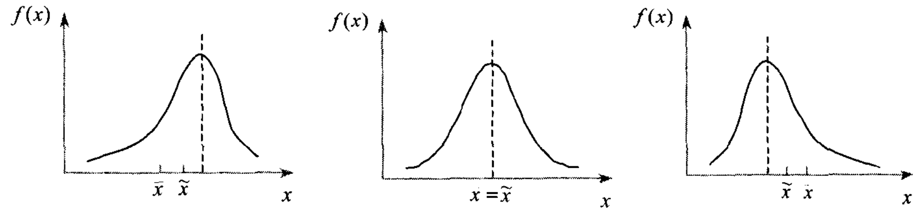
Figure 6-5 Distributions with different values of skewness (γ): (a) $\gamma < 0$ (left-skewed); (b) $\gamma = 0$ (symmetric); (c) $\gamma > 0$ (right-skewed).

Comparison of Central Tendency Measures

The mean, median, and mode of a data set can be different depending on the skewness of the data distribution, as illustrated in Figure 6-6. Both the mean and mode shift toward the direction of skewness with high sensitivity, whereas the median moves with low sensitivity.

Of the three central tendency measures, the appropriate measure is selected depending on the analyst's judgment on which measure best suits the purpose of analysis: mean for arithmetic average, median for relative standing, and mode for frequency.

Comparison of Central Tendency Measures (*cont.*)



mean < median < mode (a) mean = mode = median (b) mode < median < mean (c)

Figure 6-6 Comparisons of means, medians, and modes of various distributions: (a) left-skewed; (b) symmetric; (c) right-skewed.

6-5 Box Plots

Learning Goals

- ❑ Explain the use of a box plot.
- ❑ Construct a box plot to visualize a set of data.

Box Plot A box plot is a graphical display that presents the center, dispersion, symmetry, and outliers of a set of data simultaneously. Note that outliers are observations that are unusually low or high compared to most of the data.

A box plot consists of three components (see Figure 6-7):

(1) **Box**: The left edge, middle line, and right edge of the box indicate the first (q_1), second (q_2), and third (q_3) quartiles, respectively; therefore, the length of the box indicates the interquartile range (*IQR*) of the data set.

(2) **Whiskers**: The left whisker extends from the left edge of the box to the smallest data point within 1.5 *IQR* from q_1; and the right whisker extends from the right edge of the box to the largest data point within 1.5 *IQR* from q_3.

(3) **Outliers**: Points beyond the whiskers are outliers.

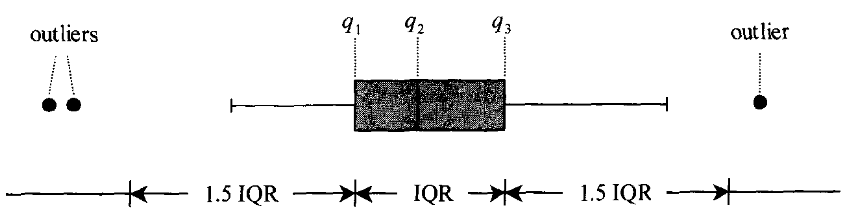
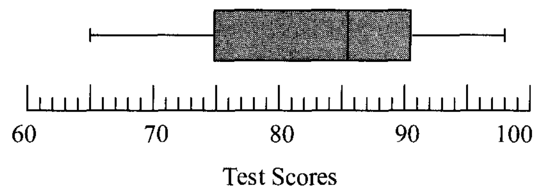


Figure 6-7 A box plot.

Example 6.7 (**Box Plot**) In Example 6-1, construct a box plot of the test scores by using the following summary quantities:

minimum = 65, maximum = 98, $q_1 = 74.75$, $q_2 = 85.5$, $q_3 = 90.5$, and $IQR = 15.75$

Since the minimum = 65 is greater than $q_1 - 1.5 \times IQR = 51.1$ and the maximum = 98 is less than $q_3 + 1.5 \times IQR = 114.1$, the left and right whiskers extend to 65 and 98, respectively. Thus, there are no outliers in the test scores which exceed 1.5 IQR from the edges of the box.



Exercise 6.7 In Exercise 6-1, construct a box plot of the test results by using the following summary quantities:

minimum = 375, maximum = 2,265, $q_1 = 1{,}097.75$, $q_2 = 1{,}436.5$, $q_3 = 1{,}735$, and $IQR = 637.25$

6-6 Time Sequence Plots

Learning Goals

- ❑ Explain the use of a time sequence plot.

Time Sequence Plot A time sequence plot is a graphical display that presents a time sequence (time series; data recorded in the order of time) to show the trend of the data over time.

As an example, Figure 6-8 shows two time sequence plots of a company's sales of three years by year and by quarter. Each horizontal axis denotes time (years or quarters; t) and each vertical axis denotes sales (in year or quarter; x). Figure 6-8(a) displays the annual sales are increasing over the years and Figure 6-8(b) indicates that the annual sales have a cyclic variation by quarter—the first- and second-quarter sales are higher than the sales of the other quarters.

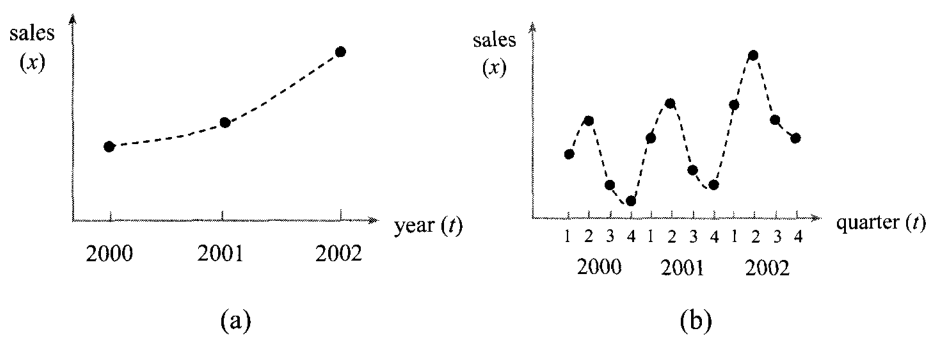


Figure 6-8 Time sequence plots of a company' sales: (a) annual sales; (b) quarterly sales.

6-7 Probability Plots

Learning Goals

- ☐ Explain the use of a probability plot.

Probability Plot

A probability plot is a graphical display that presents the cumulative relative frequencies of ordered data $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ on the probability paper of a hypothetical distribution (such as normal, Weibull, and gamma distributions) to examine if the hypothesized distribution fits the data set.

For example, Figure 6-9 shows a normal probability plot of 20 ordered observations $x_{(j)}$ against corresponding cumulative relative frequencies $(j - 0.5)/n$. Statistical software is usually used to produce a probability plot. As the plotted points closely lie along a straight line (chosen subjectively), the adequacy of the hypothesized distribution is supported.

The probability plot method is useful to examine if the data conforms to a hypothetical distribution, but lacks objectivity by relying on subjective judgment. A formal method to test the goodness-of-fit of a probability distribution is presented in Section 9-7.

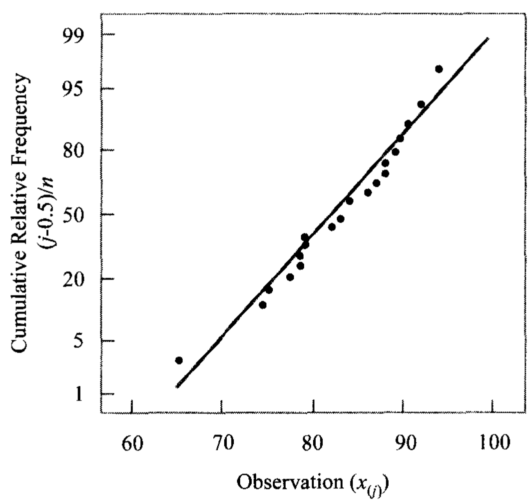


Figure 6-9 A normal probability plot.

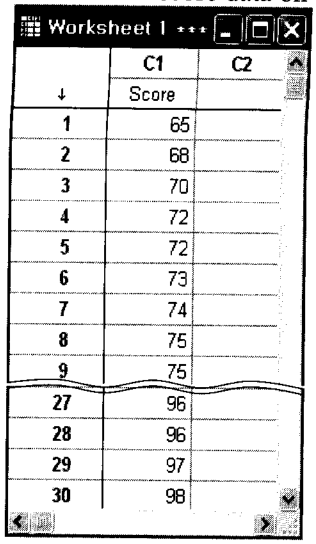
MINITAB Applications

Examples 6.1 and 6.4

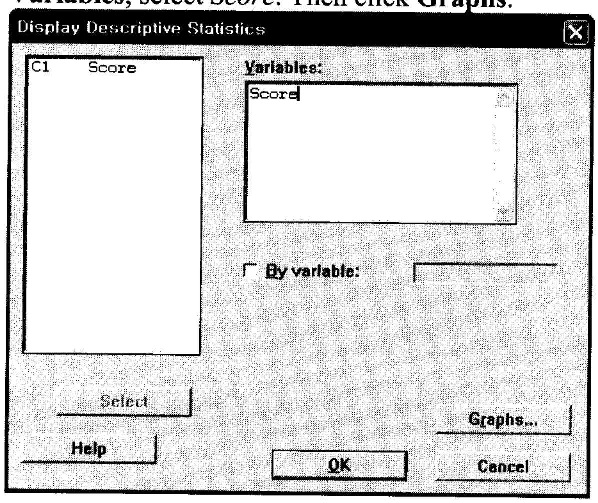
(Descriptive Statistics)

(1) Choose **File ➤ New**, click **Minitab Project**, and click **OK**.

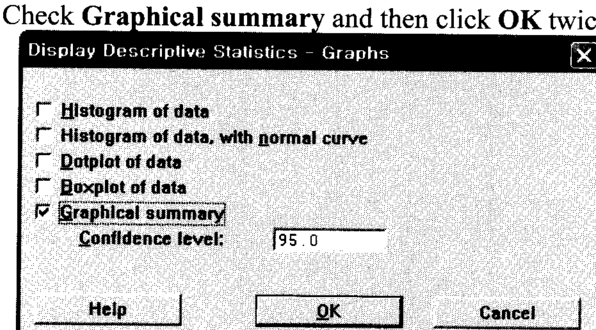
(2) Enter the test score data on the worksheet.



(3) Choose **Stat ➤ Basic Statistics ➤ Display Descriptive Statistics.** In **Variables**, select *Score*. Then click **Graphs**.

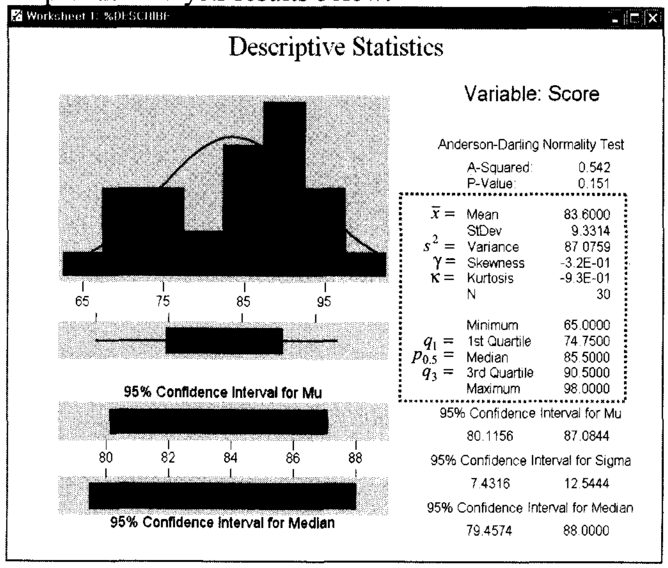


(4) Check **Graphical summary** and then click **OK** twice.



Examples 6.2 and 6.4 (*cont.*)

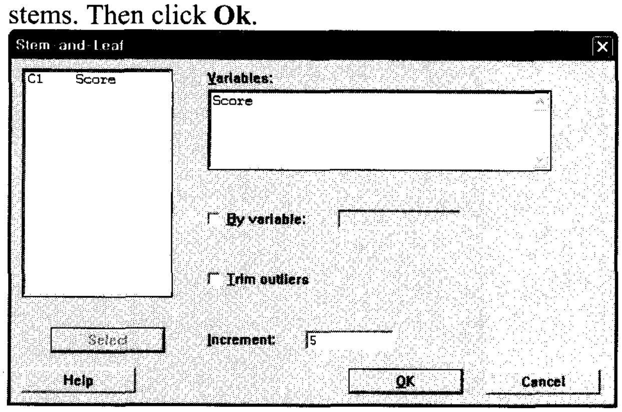
(5) Interpret the analysis results below.



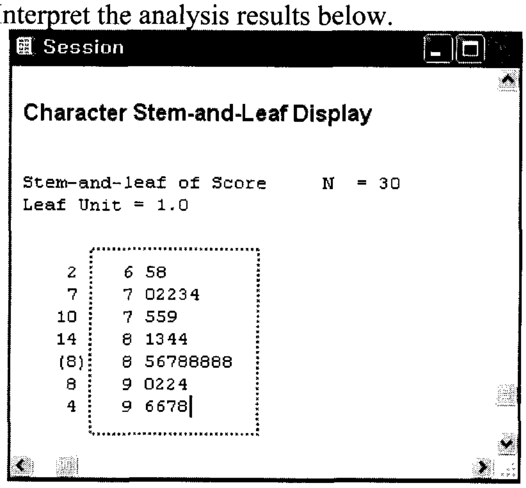
Example 6.3

(Stem-and-Leaf Diagram)

(1) Choose **Graph ➤ Stem-and-Leaf**. In **Variables**, select *Score*. In **Increment**, enter the difference between the smallest values of adjacent stems. Then click **Ok**.

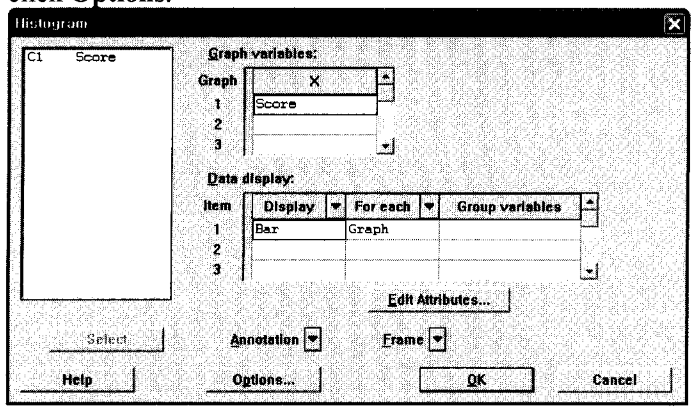


(2) Interpret the analysis results below.

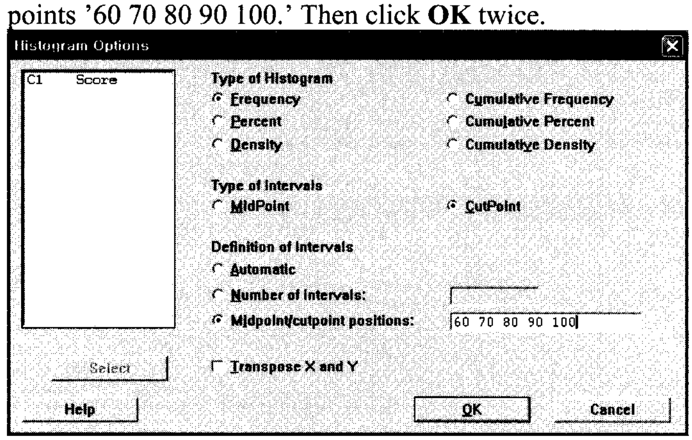


Example 6.5 **(Frequency Measures)**

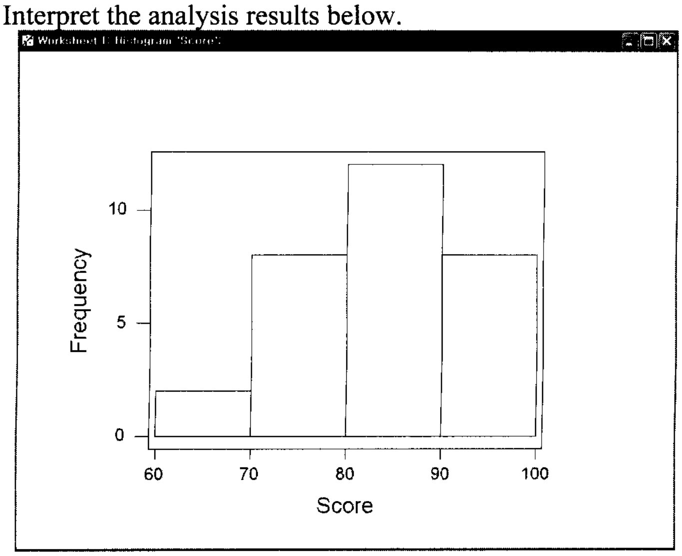
(1) Choose **Graph ➤ Histogram**. In **Graph variables**, select *Score*. Then click **Options**.



(2) Under **Type of Histogram**, check **Frequency** or **Cumulative Frequency**. Under **Type of Intervals**, check **CutPoint**. Under **Definition of Intervals**, check **Midpoint/cutpoint positions** and enter the set of cut points '60 70 80 90 100.' Then click **OK** twice.

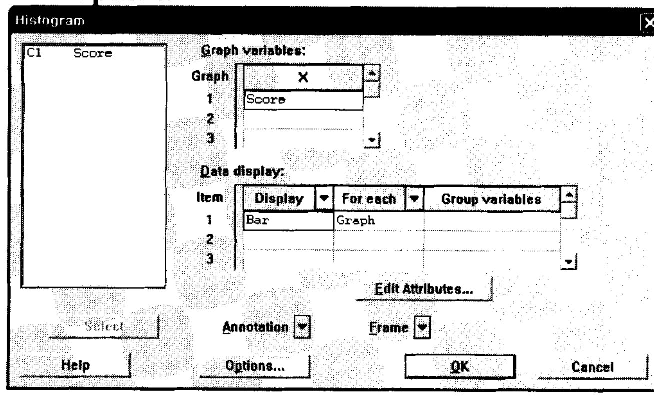


(3) Interpret the analysis results below.

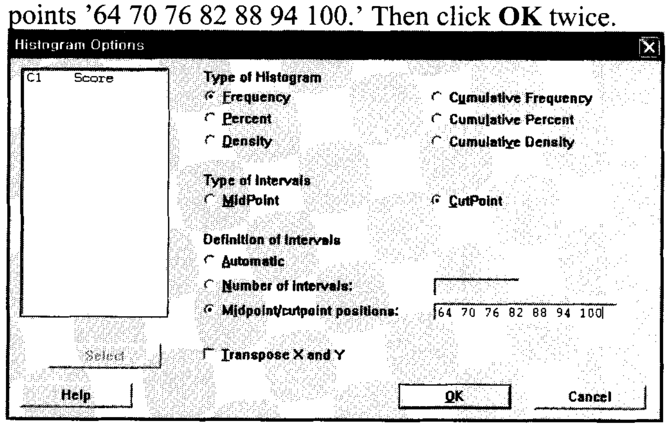


Example 6.6 **(Bar Graph/Histogram)**

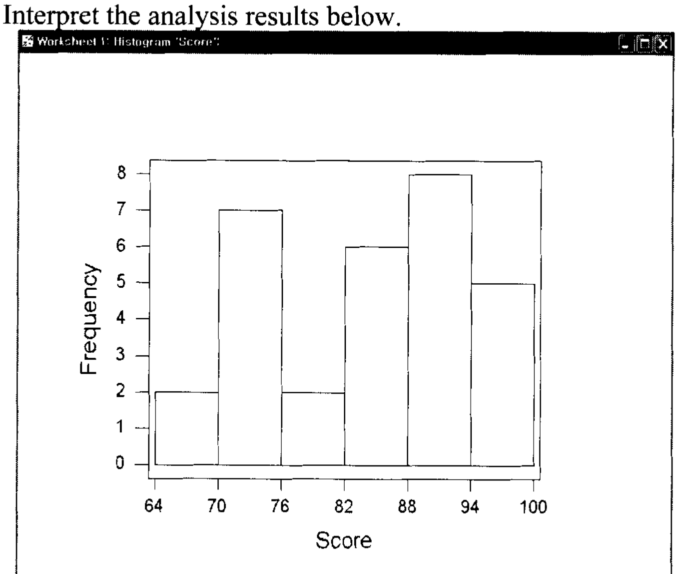
(1) Choose **Graph ➤ Histogram**. In **Graph variables**, select *Score*. Then click **Options**.



(2) Under **Type of Histogram**, check **Frequency** or **Cumulative Frequency**. Under **Type of Intervals**, check **CutPoint**. Under **Definition of Intervals**, check **Midpoint/cutpoint positions** and enter the set of cut points '64 70 76 82 88 94 100.' Then click **OK** twice.

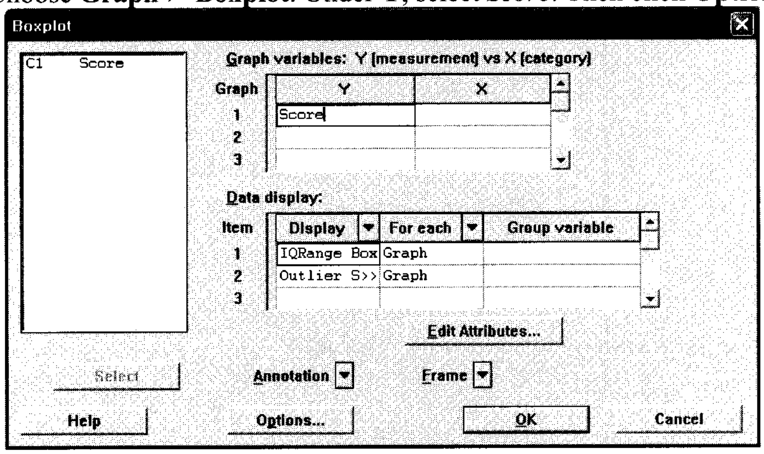


(3) Interpret the analysis results below.

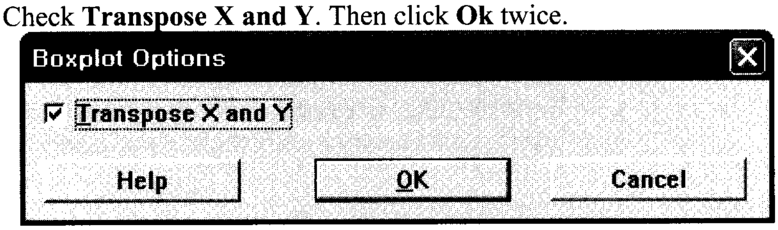


Example 6.7 **(Box Plot)**

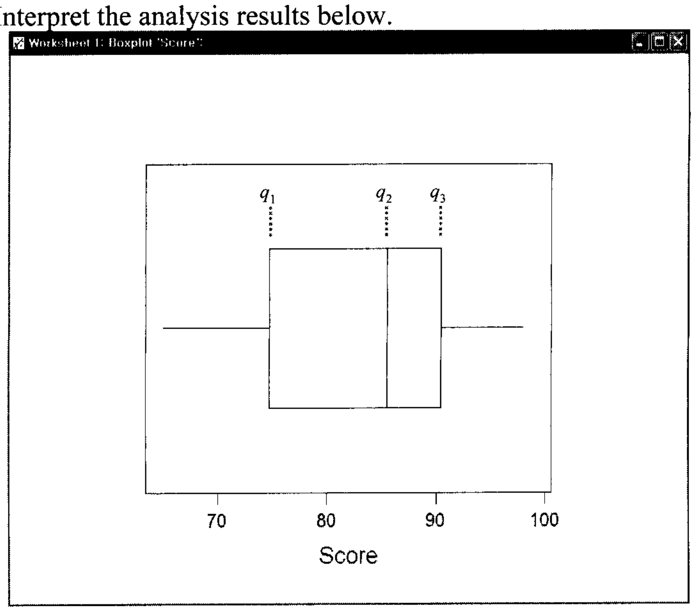
(1) Choose **Graph ➤ Boxplot**. Under **Y**, select *Score*. Then click **Options**.



(2) Check **Transpose X and Y**. Then click **Ok** twice.



(3) Interpret the analysis results below.



Answers to Exercises

Exercise 6.1

(Mean, Variance, and Range)

(1) $\bar{x} = \sum_{i=1}^{70} x_i / n = 98{,}256 / 70 = 1{,}403.7$

(2) $s^2 = \dfrac{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 / n}{n-1} = \dfrac{149{,}089{,}794 - 98{,}256^2 / 70}{70-1}$

$= 161{,}913.9 = 402.4^2$

(3) Minimum = 375

(4) Maximum = 2,265

(5) $R = 2{,}265 - 375 = 1{,}890$

Exercise 6.2

(Random Sample)

Since the same material is used for the sheet metal panels in various sizes, the number of defects per unit panel area will not be significantly different depending on the size of the panel. Thus, the random sample selected only from the large-size panels will properly represent the entire group of the sheet metal panels to study the average number of defects/ft^2.

Exercise 6.3

(Stem-and-Leaf Diagram)

Step 1: Determine **stems** and **leaves**.

The test results range from 375 to 2,265. For simplicity, round each data at the 10s place and use only the 1,000s and 100s digits for plot. Since the 1,000s digit is most meaningful, the 1,000s and 100s digits of a test result are defined as the stem and leaf of the data, respectively: 0 to 2 for stems and 0 to 9 for leaves. However, the number of the stems is only three, therefore, the stems are further divided into

0f, 0s, 0e, 1z, 1t, 1f, 1s, 1e, 2z, and 2t

where 'z' is for the leaves 0 and 1, 't' is for 2 and 3, 'f' is for 4 and 5, 's' is for 6 and 7, and 'e' is for 8 and 9.

Step 2: **Arrange** the stems and leaves.

Stem	Leaf	Frequency
0f	4	1
0s	7	1
0e	8 8 8 8 9 9 9	7
1z	0 0 0 0 0 0 1 1 1 1 1 1 1	13
1t	2 2 2 3 3 3 3 3 3 3 3	11
1f	4 4 5 5 5 5 5 5 5 5 5	11
1s	6 6 6 6 6 6 6 7 7	9
1e	8 8 8 8 8 8 8 9 9 9 9 9	12
2z	0 1 1	3
2t	2 3	2

Step 3: Summarize the **frequency** of leaves for each stem as shown above.

Exercise 6.4 **(Percentile, Quartile, Interquartile Range, Median, and Mode)**

(1) Since $n = 70$ is even, $r = (n+1)k = 71\times 0.9 = 63.9$, which is not an integer. Therefore,

$$\begin{aligned} p_{0.9} &= x_{(\lfloor r \rfloor)} + (x_{(\lceil r \rceil)} - x_{(\lfloor r \rfloor)})(r - \lfloor r \rfloor) \\ &= x_{(\lfloor 63.9 \rfloor)} + (x_{(\lceil 63.9 \rceil)} - x_{(\lfloor 63.9 \rfloor)})(63.9 - \lfloor 63.9 \rfloor) \\ &= x_{(63)} + (x_{(64)} - x_{(63)})(63.9 - 63) = 1{,}890 + (1{,}910 - 1{,}890)\times 0.9 \\ &= 1{,}908 \end{aligned}$$

(2) Since $r = (n+1)k = 71\times 0.75 = 53.25$ is not an integer,

$$\begin{aligned} q_3 = p_{0.75} &= x_{(\lfloor r \rfloor)} + (x_{(\lceil r \rceil)} - x_{(\lfloor r \rfloor)})(r - \lfloor r \rfloor) \\ &= x_{(\lfloor 53.25 \rfloor)} + (x_{(\lceil 53.25 \rceil)} - x_{(\lfloor 53.25 \rfloor)})(53.25 - \lfloor 53.25 \rfloor) \\ &= x_{(53)} + (x_{(54)} - x_{(53)})(53.25 - 53) = 1{,}730 + (1{,}750 - 1{,}730)\times 0.25 \\ &= 1{,}735 \end{aligned}$$

(3) $IQR = q_3 - q_1 = p_{0.75} - p_{0.25} = 1{,}735 - 1{,}097.75 = 637.25$

(4) Since $r = (n+1)k = 71\times 0.5 = 35.5$ is not an integer,

$$\begin{aligned} p_{0.5} &= x_{(\lfloor r \rfloor)} + (x_{(\lceil r \rceil)} - x_{(\lfloor r \rfloor)})(r - \lfloor r \rfloor), \\ &= x_{(\lfloor 35.5 \rfloor)} + (x_{(\lceil 35.5 \rceil)} - x_{(\lfloor 35.5 \rfloor)})(35.5 - \lfloor 35.5 \rfloor) \\ &= x_{(35)} + (x_{(36)} - x_{(35)})(35.5 - 35) = 1{,}421 + (1{,}452 - 1{,}421)\times 0.5 \\ &= 1{,}436.5 \end{aligned}$$

(5) Modes = 1,102, 1,315, and 1,750 (trimodal)

Exercise 6.5 **(Frequency Measures)**

Class Interval (X)	Frequency	Relative Frequency	Cumulative Relative Frequency
$x < 500$	1	0.01	0.01
$500 \le x < 1{,}000$	9	0.13	0.14
$1{,}000 \le x < 1{,}500$	28	0.40	0.54
$1{,}500 \le x < 2{,}000$	27	0.39	0.93
$2{,}000 \le x$	5	0.07	1.00

Exercise 6.6 **(Bar Graph/Histogram)**

Step 1: Determine the **number of bins** (k).

Since $\sqrt{n} = \sqrt{70} = 8.4$, $k = 9$.

Step 2: Determine the **width of a bin** (w).

Since $\frac{R}{k} = \frac{\max - \min}{k} = \frac{2{,}265 - 375}{9} = 210$, use $w = 220$.

Step 3: Determine the **limits of the histogram**.

$$l \cong \min - \frac{kw - R}{2} = 375 - \frac{9\times 220 - 1890}{2} = 330$$

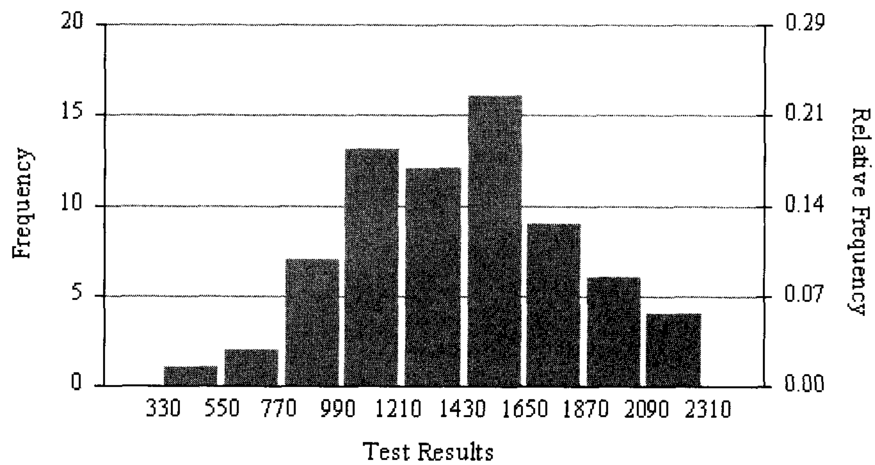
$$u = l + kw = 330 + 9\times 220 = 2{,}310$$

Exercise 6.6 (*cont.*)

Step 4: Prepare a **frequency table** based on the bins determined.

Class Interval (X)	Frequency	Relative Frequency	Cumulative Relative Frequency
$330 < x \leq 550$	1	0.01	0.01
$550 < x \leq 770$	2	0.03	0.04
$770 < x \leq 990$	3	0.10	0.14
$990 < x \leq 1{,}210$	13	0.19	0.33
$1{,}210 < x \leq 1{,}430$	12	0.17	0.50
$1{,}430 < x \leq 1{,}650$	16	0.23	0.73
$1{,}650 < x \leq 1{,}870$	9	0.13	0.86
$1{,}870 < x \leq 2{,}090$	6	0.09	0.94
$2{,}090 < x \leq 2{,}310$	4	0.06	1.00

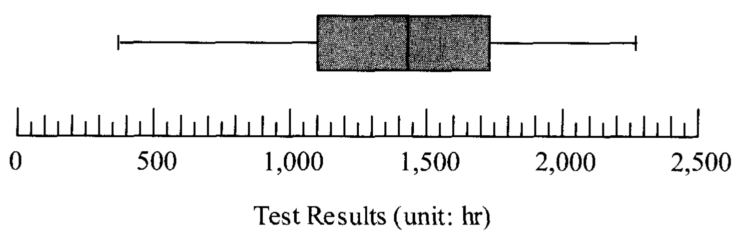
Step 5: Develop a **histogram**.



Exercise 6.7

(Box Plot)

Since the minimum = 375 is greater than $q_1 - 1.5 \times IQR = 141.9$ and the maximum = 2,265 is less than $q_3 + 1.5 \times IQR = 2{,}690.9$, the left and right whiskers extend to the minimum and maximum, respectively. Thus, there are no outliers in the test results which exceed 1.5 IQR from the edges of the box.



7 Point Estimation of Parameters

OUTLINE

7-1 Introduction

Learning Goals

- ❑ Describe the terms *parameter*, *point estimator*, and *point estimate*.
- ❑ Identify two major areas of statistical inference.
- ❑ Distinguish between point estimation and interval estimation.
- ❑ Determine the point estimate of a parameter.

Parameter (θ)

A parameter (denoted by θ) represents a characteristic of the population under study. It is constant but unknown in most cases.

(e.g.) Parameters

Mean (μ), variance (σ^2), proportion (p), correlation coefficient (ρ), and regression coefficient (β)

Statistical Inference

Statistical inference refers to making decisions or drawing conclusions about a population by analyzing a sample from the population. Two major areas of statistical inference can be defined:

1. **Parameter estimation**: Estimates the value of θ. (e.g.) $\mu = 150$
2. **Hypothesis testing**: Tests an assertion of θ. (e.g.) H_0: $\mu = 150$

Parameter Estimation

Parameter estimation is further divided into two areas:

1. **Point estimation**: Estimates the exact location of θ. (e.g.) $\mu = 150$
2. **Interval estimation**: Establishes an interval that includes the true value of θ with a designated probability. (e.g.) $145 < \mu < 155$

Point Estimator ($\hat{\Theta}$)

A point estimator (denoted by $\hat{\Theta}$) is a statistic used to estimate θ. It is a random variable because a statistic is a random variable.

(e.g.) Point estimator of μ: sample mean $\overline{X} = \sum_{i=1}^{n} X_i / n$

A <u>single numerical value</u> of $\hat{\Theta}$ determined by a particular random sample is called a **point estimate** of θ (denoted by $\hat{\theta}$).

Example 7.1

(**Point Estimate**) Suppose that the life length of an INFINITY light bulb (X) has a normal distribution with mean μ and variance σ^2. A random sample of size $n = 25$ light bulbs was examined and the sum of the life lengths was 14,900 hours. Estimate the mean life length (μ) of an INFINITY light bulb.

$$\hat{\mu} = \overline{x} = \frac{\sum_i x_i}{n} = \frac{14{,}900}{25} = 745 \text{ hrs.}$$

Exercise 7.1

The line width (X) of a tool used for semiconductor manufacturing is assumed to be normally distributed with mean μ and variance σ^2. A random sample of size $n = 20$ tools was measured and the sum of the line width measurements was 13.2 micrometers (μm). Estimate the mean line width (μ) of the tool.

7-2 General Concepts of Point Estimation

Learning Goals

- ❑ Explain the terms *unbiased estimator*, *minimum variance unbiased estimator* (MVUE), *standard error*, and *mean square error* (MSE) *of an estimator*.
- ❑ Select the appropriate point estimator of θ in terms of unbiasedness, minimum variance, and minimum mean square error each.

Unbiased Estimator

An unbiased estimator is a point estimator ($\hat{\Theta}$) whose expected value is equal to the true value of θ, i.e.,

$$E(\hat{\Theta}) = \theta .$$

Note that several unbiased estimators can be defined for a single parameter θ.

As shown in Figure 7-1, the **bias** of $\hat{\Theta}$ is the difference of the expected value of $\hat{\Theta}$ from the true value of θ:

$$\text{bias} = E(\hat{\Theta}) - \theta$$

Unbiased Estimator (*cont.*)

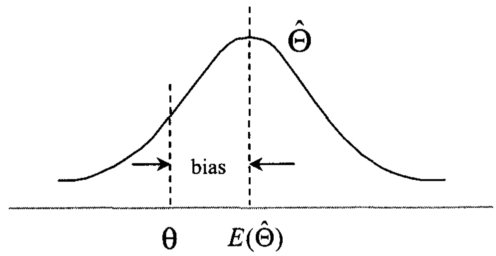


Figure 7-1 Bias of a point estimator $\hat{\Theta}$.

Example 7.2

(**Unbiased Estimator**) Let $X_1, X_2, \ldots, X_n$ denote a random sample of size n from a probability distribution with $E(X) = \mu$ and $V(X) = \sigma^2$. Show if the sample mean

$$\overline{X} = \sum_{i=1}^{n} X_i / n$$

is an unbiased estimator of the population mean μ.

$$E(\overline{X}) = E\left(\sum_{i=1}^{n} X_i / n\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\mu = \mu$$

Since $E(\overline{X}) = \mu$, $\overline{X}$ is an unbiased estimator of μ.

Exercise 7.2 (MR 7-2)

Let $X_1, X_2, \ldots, X_7$ denote a random sample from a population having mean μ and variance σ^2. Show if the following estimator

$$\hat{\Theta} = \frac{2X_1 - X_6 + X_4}{2}$$

is an unbiased estimator of μ.

Minimum Variance Unbiased Estimator (MVUE)

A minimum variance unbiased estimator (MVUE) of θ is the unbiased $\hat{\Theta}$ with the smallest variance (see Figure 7-2). By using the MVUE of θ, the unknown parameter θ can be estimated accurately and precisely.

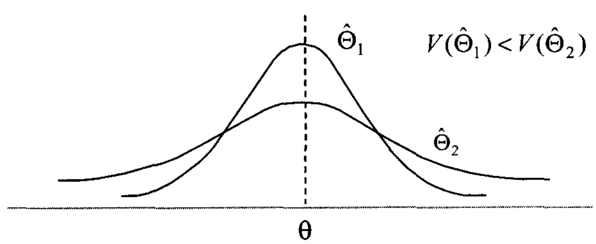


Figure 7-2 Variances of unbiased estimators of θ.

Example 7.3

(**Minimum Variance Unbiased Estimator**) Let $X_1, X_2, \ldots, X_n$ denote a random sample of size n (>1) from a population with $E(X) = \mu$ and $V(X) = \sigma^2$. Both X_1 and $\overline{X}$ are unbiased estimators of μ because their expected values are equal to μ. Of the two estimators, which is preferred to estimate μ and why?

Example 7.3 (*cont.*)

The variances of X_1 and $\overline{X}$ are

$$V(X_1) = \sigma^2$$

$$V(\overline{X}) = V\left(\sum_i X_i / n\right) = \frac{1}{n^2}\sum_i V(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Since $V(X_1) = \sigma^2 > V(\overline{X}) = \sigma^2 / n$ for $n > 1$, $\overline{X}$ is preferred to estimate μ with higher accuracy and precision.

Exercise 7.3

Let $X_1, X_2, \dots, X_n$ denote a random sample of size n (>3) from a population with $E(X) = \mu$ and $V(X) = \sigma^2$. Both $(X_1 + X_2 + X_3)/3$ and $\overline{X}$ are unbiased estimators of μ because their expected values are equal to μ. Of the two estimators, which is preferred to estimate μ and why?

Standard Error ($\sigma_{\hat{\Theta}}$)

The standard error of a point estimator (denoted by $\sigma_{\hat{\Theta}}$) is the standard deviation of a point estimator $\hat{\Theta}$. It can be used as a measure to indicate the **precision** of parameter estimation.

If $\sigma_{\hat{\Theta}}$ includes unknown parameters that can be estimated, use of the estimates of the parameters in calculating $\sigma_{\hat{\Theta}}$ produces an **estimated standard error** (denoted by $s_{\hat{\Theta}}$ or $se(\hat{\Theta})$).

Example 7.4

In Example 7-1, X has a normal distribution with mean μ and variance σ^2 and a random sample of size $n = 25$ was examined.

1. (**Standard Error**) Assuming $\sigma^2 = 40^2$, determine the standard error of the sample mean ($\overline{X}$). Note that $\overline{X} = \sum_{i=1}^{n} X_i / n \sim N(\mu, \sigma^2 / n)$.

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{40}{\sqrt{25}} = 8.0 \text{ hrs.}$$

2. (**Estimated Standard Error**) Suppose that σ^2 is unknown and the sample variance $s_X^2 = 35^2$. Calculate the estimated standard error of the sample mean ($\overline{X}$).

$$s_{\overline{X}} = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{s_X}{\sqrt{n}} = \frac{35}{\sqrt{25}} = 7.0 \text{ hrs.}$$

Exercise 7.4

In Exercise 7-1, X is normally distributed with mean μ and variance σ^2 and a random sample of size $n = 20$ was measured.

1. Assuming $\sigma^2 = 0.05^2$, determine the standard error of the sample mean ($\overline{X}$).
2. Suppose that σ^2 is unknown and the sample variance $s_X^2 = 0.04^2$. Calculate the estimated standard error of the sample mean ($\overline{X}$).

Mean Square Error (MSE) of Estimator

The mean square error (MSE) of a point estimator $\hat{\Theta}$ is the expected squared difference between $\hat{\Theta}$ and θ:

$$MSE(\hat{\Theta}) = E(\hat{\Theta} - \theta)^2 = E\left[\hat{\Theta} - E(\hat{\Theta})\right]^2 + \left[\theta - E(\hat{\Theta})\right]^2$$
$$= V(\hat{\Theta}) + (bias)^2$$
$$= V(\hat{\Theta}) \quad \text{for an unbiased } \hat{\Theta} \text{ because bias} = 0.$$

Note that a biased estimator of θ with the smallest MSE is sometimes used for precise estimation of θ while sacrificing the accuracy of estimation on θ (see Figure 7-3).

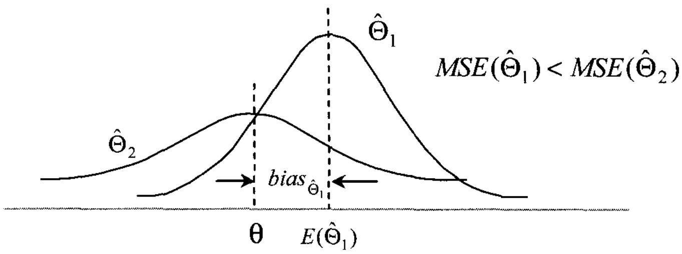


Figure 7-3 A biased $\hat{\Theta}_1$ with a smaller mean square error than that of the unbiased $\hat{\Theta}_2$.

Example 7.5

(Mean Square Error of Estimator) Suppose that the means and variances of $\hat{\Theta}_1$ and $\hat{\Theta}_2$ are $E(\hat{\Theta}_1) = \theta$, $E(\hat{\Theta}_2) = 0.9\theta$, $V(\hat{\Theta}_1) = 5$, and $V(\hat{\Theta}_2) = 4$, respectively. Which estimator is preferred to estimate θ and why?

For $\hat{\Theta}_1$,

$$bias_{\hat{\Theta}_1} = E(\hat{\Theta}_1) - \theta = \theta - \theta = 0$$
$$MSE(\hat{\Theta}_1) = V(\hat{\Theta}_1) + (bias_{\hat{\Theta}_1})^2 = 5 - 0 = 5$$

For $\hat{\Theta}_2$,

$$bias_{\hat{\Theta}_2} = E(\hat{\Theta}_2) - \theta = 0.9\theta - \theta = -0.1\theta$$
$$MSE(\hat{\Theta}_2) = V(\hat{\Theta}_2) + (bias_{\hat{\Theta}_2})^2 = 4 + 0.01\theta^2$$

By subtracting $MSE(\hat{\Theta}_2)$ from $MSE(\hat{\Theta}_1)$,

$$MSE(\hat{\Theta}_1) - MSE(\hat{\Theta}_2) = 5 - (4 + 0.01\theta^2) = 1 - 0.01\theta^2$$

The preferred estimator of θ for precise estimation depends on the range of θ as follows:

$$\begin{cases} \hat{\Theta}_1, & \text{if } \theta \geq 10 \text{ because } MSE(\hat{\Theta}_1) \leq MSE(\hat{\Theta}_2) \text{ and } \hat{\Theta}_1 \text{ is unbiased} \\ \hat{\Theta}_2, & \text{if } \theta < 10 \text{ because } MSE(\hat{\Theta}_1) > MSE(\hat{\Theta}_2) \end{cases}$$

Exercise 7.5

Let $X_1, X_2, \ldots, X_5$ denote a random sample of size $n = 5$ from a population with $E(X) = 70$ and $V(X) = 5^2$. Two estimators of μ are proposed:

$$\hat{\Theta}_1 = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{5} \quad \text{and} \quad \hat{\Theta}_2 = \frac{2X_1 + X_3 + X_5}{4}$$

Of the two estimators, which is preferred to estimate μ and why?

7-3 Methods of Point Estimation

Learning Goals

- ❑ Explain the utility of the maximum likelihood method.
- ❑ Find a point estimator of θ by using the maximum likelihood method.

Maximum Likelihood Method

The method of maximum likelihood is used to derive a point estimator of θ. This method finds a **maximum likelihood estimator** of θ which maximizes the likelihood function of a random sample $X_1, X_2, \ldots, X_n$

$$L(\theta) = f(x_1;\theta)f(x_2;\theta)\cdots f(x_n;\theta), \quad \text{where } X_1, X_2, \ldots, X_n \sim i.i.d.\ f(x;\theta)$$

Example 7.6

(**Maximum Likelihood Estimator**) Let $X_1, X_2, \ldots, X_n$ denote a random sample of size n from an exponential distribution with the parameter λ. Find the maximum likelihood estimator of λ.

The probability density function of an exponential distribution is

$$f(x;\lambda) = \lambda e^{-\lambda x}$$

Thus, the likelihood function of $X_1, X_2, \ldots, X_n$ is

$$L(\lambda) = f(x_1;\lambda)f(x_2;\lambda)\cdots f(x_n;\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}$$

Then, the log likelihood function is

$$\ln L(\lambda) = n \ln \lambda - \lambda \sum_{i=1}^{n} x_i$$

Now, the derivative of $\ln L(\lambda)$ is

$$\frac{d \ln L(\lambda)}{d\lambda} = \frac{d}{d\lambda}\left(n \ln \lambda - \lambda \sum_{i=1}^{n} x_i \right) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i$$

By equating this derivative of $\ln L(\lambda)$ to zero, the point estimator of λ which maximizes $L(\lambda)$ is

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\overline{X}}$$

Exercise 7.6 (MR 7-21) Let $X_1, X_2, \dots, X_n$ represent a random sample of size n from a geometric distribution with parameter p. Find the maximum likelihood estimators of p.

7-4 Sampling Distribution

Learning Goals

- ☐ Explain the term *sampling distribution.*

Sampling Distribution A sampling distribution is the probability distribution of a statistic (a function of random variables such as sample mean and sample variance). The sampling distribution of a statistic depends on the following:

(1) The distribution of the population
(2) The size of the sample
(3) The method of sample selection

(e.g.) Sampling distribution of a sample mean

$$\overline{X} = \sum_{i=1}^{n} X_i / n \sim N\left(\mu, \frac{\sigma^2}{n}\right), \text{ where } X_1, X_2, \dots, X_n \sim i.i.d.\ N(\mu, \sigma^2)$$

7-5 Sampling Distributions of Means

Learning Goals

- ☐ Explain the central limit theorem (CLT).
- ☐ Determine the distribution of a sample mean by applying the central limit theorem.

Sampling Distribution of $\overline{X}$ Suppose that a random sample of size n is taken from a normal distribution with mean μ and variance σ^2. Then, the sampling distribution of the sample mean is

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \sim N(\mu, \frac{\sigma^2}{n})$$

(Derivation) $\overline{X} = \sum_{i=1}^{n} X_i / n \sim N(\mu, \frac{\sigma^2}{n})$, where $X \sim N(\mu, \sigma^2)$

Since $X_1, X_2, \dots, X_n$ are independent and normally distributed with the same $E(X) = \mu$ and $V(X) = \sigma^2$, the distribution of $\overline{X}$ is normal with mean and variance

$$E(\overline{X}) = E\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{1}{n}[E(X_1) + E(X_2) + \cdots + E(X_n)]$$

$$= \frac{1}{n}[\mu + \mu + \cdots + \mu] = \frac{1}{n} \times n\mu = \mu$$

Sampling Distribution of $\overline{X}$ (*cont.*)

$$V(\overline{X}) = V\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{1}{n^2}[V(X_1) + V(X_2) + \cdots + V(X_n)]$$
$$= \frac{1}{n^2}\left[\sigma^2 + \sigma^2 + \cdots + \sigma^2\right] = \frac{1}{n^2} \times n\sigma^2 = \frac{\sigma^2}{n}$$

Central Limit Theorem (CLT)

Let $X_1, X_2, \ldots, X_n$ denote a random sample of size n taken from a population (X) with mean μ and variance σ^2. Then, the limiting form of the distribution of the sample mean $\overline{X}$ is

$$\overline{X} = \sum_{i=1}^{n} X_i / n \sim N(\mu, \frac{\sigma^2}{n}) \quad \Rightarrow \quad Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

as n approaches ∞. This normal approximation of $\overline{X}$ is called the central limit theorem (CLT).

As displayed in Figure 7-4, the distributions of the sample means from uniform, binomial, and exponential distributions become normal distributions as their sample sizes n become large. In most cases, if $\boldsymbol{n \geq 30}$, normal approximation of $\overline{X}$ will be satisfactory regardless of the distribution of X. In case that $\boldsymbol{n < 30}$ but the distribution of X is **close to the normal**, normal approximation of $\overline{X}$ will be still held.

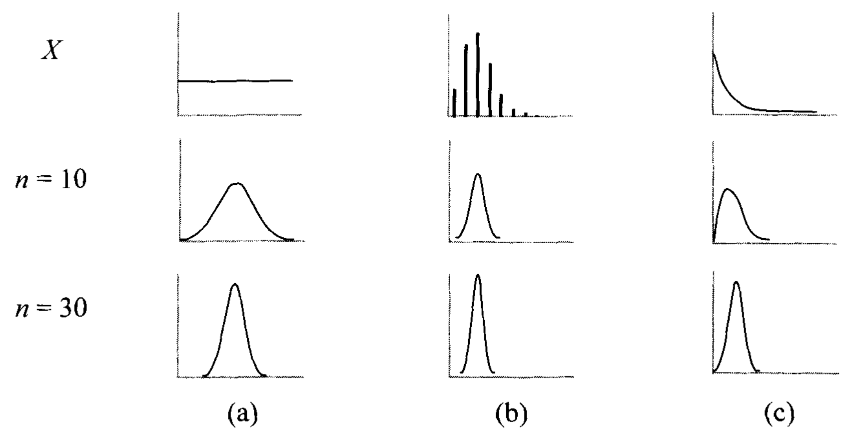


Figure 7-4 Sampling distribution of $\overline{X}$: (a) $X \sim$ uniform; (b) $X \sim$ binomial with $p = 0.2$; (c) $X \sim$ exponential with $\lambda = 1$.

Based on the central limit theorem, the sampling distribution of $\overline{X}$ where X is normal or non-normal is as follows:

1. **Case 1: Normal population,** $X \sim N(\mu, \sigma^2)$

$$\overline{X} = \sum_{i=1}^{n} X_i / n \sim N(\mu, \frac{\sigma^2}{n}), \quad \text{where } X_1, X_2, \ldots, X_n \sim i.i.d.\ N(\mu, \sigma^2)$$

Central Limit Theorem (CLT) (*cont.*)

2. **Case 2: Non-normal population** with μ and σ^2
   (1) **Case 2.1: Normal approximation applicable**

   $$\overline{X} = \sum_{i=1}^{n} X_i / n \sim N(\mu, \frac{\sigma^2}{n}), \quad \text{if } n \geq 30 \text{ or the distribution of } X \text{ is close to the normal}$$

   (2) **Case 2.2: Normal approximation inapplicable**
   It would be difficult to find the distribution of $\overline{X}$ if $n < 30$ and the distribution of X is significantly deviated from the normal. In this case, use <u>non-parametric statistics</u> for statistical inference (see Chapter 15).

Example 7.7

(**Central Limit Theorem; Sampling Distribution of Sample Mean**) Suppose that the waiting time (X; unit: min.) of a customer to pick up his/her prescription at a drug store follows an exponential distribution with $E(X) = 20$ min. and $V(X) = 5^2$. A random sample of size $n = 40$ customers is observed. What is the distribution of the sample mean?

Since $n \geq 30$, normal approximation is applicable to $\overline{X}$ even if X is exponentially distributed. Thus, the sampling distribution of $\overline{X}$ is

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n}) = N(20, \frac{5^2}{40}) = N(20, 0.8^2)$$

Exercise 7.7

Suppose that the number of messages (X) arriving to a computer server follows a Poisson distribution with $E(X) = 10$ messages/hour and $V(X) = 10$. The number of messages per hour was recorded for $n = 48$ hours. What is the distribution of the sample mean?

Answers to Exercises

Exercise 7.1

1. **(Point Estimate)**

$$\hat{\mu} = \bar{x} = \frac{\sum_i x_i}{n} = \frac{13.2}{20} = 0.66 \ \mu\text{m}$$

Exercise 7.2

(Unbiased Estimator)

$$E(\hat{\Theta}) = E\left(\frac{2X_1 - X_6 + X_4}{2}\right) = E(X_1) - \frac{1}{2}E(X_6) + \frac{1}{2}E(X_4) = \mu$$

Since $E(\hat{\Theta}) = \mu$, $\hat{\Theta}$ is an unbiased estimator of μ.

Exercise 7.3

(Minimum Variance Unbiased Estimator)

The variances of $(X_1 + X_2 + X_3)/3$ and $\bar{X}$ are

$$V\left(\frac{X_1 + X_2 + X_3}{3}\right) = \frac{1}{9}[V(X_1) + V(X_2) + V(X_3)] = \frac{1}{9} \times 3\sigma^2 = \frac{\sigma^2}{3}$$

$$V(\bar{X}) = V\left(\sum_i X_i / n\right) = \frac{1}{n^2}\sum_i V(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Since $V\left(\dfrac{X_1 + X_2 + X_3}{3}\right) = \dfrac{\sigma^2}{3} > V(\bar{X}) = \dfrac{\sigma^2}{n}$ for $n > 3$, $\bar{X}$ is preferred as the point estimator of μ for more accurate and precise estimation on μ.

Exercise 7.4

1. **(Standard Error)**

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{0.05}{\sqrt{20}} = 0.011 \ \mu\text{m}$$

2. **(Estimated Standard Error)**

$$s_{\bar{X}} = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{s_X}{\sqrt{n}} = \frac{0.04}{\sqrt{20}} = 0.009 \ \mu\text{m}$$

Exercise 7.5

(Mean Square Error of Estimator)

For $\hat{\Theta}_1$,

$$E(\hat{\Theta}_1) = E\left(\frac{X_1 + X_2 + X_3 + X_4 + X_5}{5}\right) = \frac{1}{5} \times 5E(X) = 70$$

$$bias_{\hat{\Theta}_1} = E(\hat{\Theta}_1) - \mu = 70 - 70 = 0$$

$$V(\hat{\Theta}_1) = V\left(\frac{X_1 + X_2 + X_3 + X_4 + X_5}{5}\right) = \frac{1}{5^2} \times 5V(X) = 5$$

$$MSE(\hat{\Theta}_1) = V(\hat{\Theta}_1) + (bias_{\hat{\Theta}_1})^2 = 5 + 0 = 5$$

Exercise 7.5 (*cont.*)

For $\hat{\Theta}_2$,

$$E(\hat{\Theta}_2) = E\left(\frac{2X_1 + X_3 + X_5}{4}\right) = \frac{1}{4} \times 4E(X) = 70$$

$$bias_{\hat{\Theta}_2} = E(\hat{\Theta}_2) - \mu = 70 - 70 = 0$$

$$V(\hat{\Theta}_2) = V\left(\frac{2X_1 + X_3 + X_5}{4}\right) = \frac{1}{4^2} \times 6V(X) = 9.4$$

$$MSE(\hat{\Theta}_2) = V(\hat{\Theta}_2) + (bias_{\hat{\Theta}_2})^2 = 9.4 + 0 = 9.4$$

Since $MSE(\hat{\Theta}_1) = 5.0 < MSE(\hat{\Theta}_2) = 9.4$, $\hat{\Theta}_1$ is selected for more accurate and precise estimation on μ.

Exercise 7.6

(Maximum Likelihood Estimator)

The probability mass function of a geometric distribution is

$$f(x; p) = (1-p)^{x-1} p$$

Thus, the likelihood function of $X_1, X_2, \ldots, X_n$ is

$$L(p) = \prod_{i=1}^{n} (1-p)^{x_i - 1} p = (1-p)^{\sum_{i=1}^{n} x_i - n} p^n$$

Then, the log likelihood function is

$$\ln L(p) = \left(\sum_{i=1}^{n} x_i - n\right) \ln(1-p) + n \ln p$$

Now, the derivative of $\ln L(p)$ is

$$\frac{d \ln L(p)}{dp} = -\frac{\sum_{i=1}^{n} x_i - n}{1-p} + \frac{n}{p}$$

By equating the derivative of $\ln L(p)$ to zero, the point estimator of p is

$$\hat{p} = \frac{1}{\sum_{i=1}^{n} x_i / n} = \frac{1}{\overline{X}}$$

The maximum likelihood estimator of p is the reciprocal of $\overline{X}$.

Exercise 7.7

(Central Limit Theorem; Sampling Distribution of Sample Mean)

Since $n = 48 \geq 30$, normal approximation is applicable to $\overline{X}$ even if X follows a Poison distribution. Therefore, the approximate distribution of $\overline{X}$ is

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n}) = N(10, \frac{10}{48}) = N(10, 0.7^2)$$

8 Statistical Intervals for a Single Sample

OUTLINE

8-1 Introduction

Learning Goals

❑ Distinguish between confidence, prediction, and tolerance intervals.

Statistical Intervals

While point estimation estimates the exact location of a parameter (θ), interval estimation establishes bounds of plausible values for θ. Three types of statistical intervals are defined:

1. **Confidence Interval (CI)**: Bounds a parameter of the population distribution. (e.g.) A 90% CI on μ where $X \sim N(\mu, \sigma^2)$ indicates that the CI contains the value of μ with 90% confidence (see Section 8.2).
2. **Prediction Interval (PI)**: Bounds a future observation. (e.g.) A 90% PI on a new observation where $X \sim N(\mu, \sigma^2)$ indicates that the PI contains a new observation with 90% confidence (see Section 8.6).
3. **Tolerance Interval (TI)**: Bounds a selected proportion of the population distribution. (e.g.) A 95% TI on X with 90% confidence indicates that the TI contains 95% of X values with 90% confidence (see Section 8.7).

8-2 Confidence Interval on the Mean of a Normal Distribution, Variance Known

Learning Goals

- ❑ Interpret a 100(1 - α)% confidence interval (CI).
- ❑ Explain the relationship between the length of a CI and the precision of estimation.
- ❑ Determine a sample size to satisfy a predetermined level of error (E) in estimating μ.
- ❑ Identify the point estimator of μ when σ^2 is known and the sampling distribution of the point estimator.
- ❑ Establish a 100(1 - α)% CI on μ when σ^2 is known.

Confidence Interval

A 100(1 - α)% confidence interval on a parameter (θ) has lower and/or upper bounds ($l \le \theta \le u$), which are computed by using a sample from the population. Since different samples will produce different values of l and u, the lower- and upper-confidence limits are considered values of random variables L and U which satisfy the following:

$$P(L \le \theta \le U) = 1 - \alpha, \quad 0 \le \alpha \le 1$$

Since L and U are random variables, a CI is a random interval. A 100(1 - α)% CI indicates that, if CIs are established from an infinite number of random samples, 100(1 - α)% of the CIs will contain the true value of θ.

The types of confidence intervals include:

1. **Two-Sided CI**: Specifies <u>both the lower- and upper-confidence limits</u> of θ, such as $l \le \theta \le u$.
   (e.g.) $730 \le \mu_X \le 750$ hrs, where X = life length of an INFINITY light bulb
2. **One-Sided CI**: Defines <u>either the lower- or upper-confidence limit</u> of θ.
   (e.g.) $730 \le \mu_X$ or $\mu_X \le 750$ hrs, where X = life length of an INFINITY light bulb

Length of CI and Precision of Estimation

The **length of a CI** refers to the distance between the upper and lower limits ($u - l$). The wider the CI, the more confident we are that the interval actually contains the true value of θ (see Figure 8-1), but less informed we are about the true value of θ.

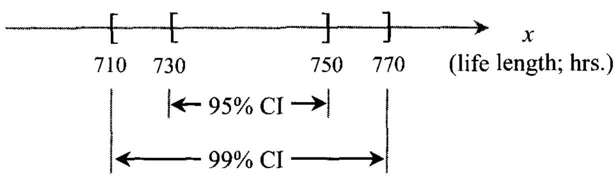


Figure 8-1 Confidence intervals on the mean life length ($\theta = \mu_X$) of an INFINITY light bulb with selected levels of confidence.

The length of a CI on θ is inversely related to the **precision** of estimation on θ: the wider the CI, the less precise the estimation of θ.

Sample Size Selection (*cont.*)

The distance of a point estimate of θ from the true value of θ represents an **error** (denoted by E) in parameter estimation:

$$E = |\hat{\theta} - \theta|$$
$$= |\bar{x} - \mu| \quad \text{in estimating } \mu$$

To establish a 100(1 - α)% CI on μ which does not exceed a predefined level of E, the sample size is determined by the formula

$$n = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2 \quad \text{(in case } n \text{ is not an integer, round up the value)}$$

Inference Context

Parameter of interest: μ

Point estimator of μ: $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$, σ^2 known

$$\Rightarrow Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

Confidence Interval Formula

A 100(1 - α)% CI on μ when σ^2 is known is

$$\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \text{for two-sided CI}$$

$$\bar{X} - z_{\alpha}\frac{\sigma}{\sqrt{n}} \le \mu \quad \text{for lower-confidence bound}$$

$$\mu \le \bar{X} + z_{\alpha}\frac{\sigma}{\sqrt{n}} \quad \text{for upper-confidence bound}$$

(Derivation) Two-sided confidence interval on μ, σ^2 known

By using $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$,

$$P(L \le \mu \le U) = 1 - \alpha$$
$$\Rightarrow P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) = 1 - \alpha$$
$$\Rightarrow P\left(-z_{\alpha/2} \le \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \le z_{\alpha/2}\right) = 1 - \alpha$$
$$\Rightarrow P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Therefore,

$$L = \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \text{and} \quad U = \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Note that, for a one-sided CI, use z_{α} instead of $z_{\alpha/2}$ to derive the corresponding limit.

Example 8.1

Suppose that the life length of an INFINITY light bulb (X; unit: hour) follows the normal distribution $N(\mu, 40^2)$. A random sample of n = 30 bulbs is tested as shown below, and the sample mean is found to be $\bar{x}$ = 780 hours.

No.	Life Length	No.	Life Length	No.	Life Length
1	727	11	831	21	725
2	755	12	742	22	735
3	714	13	784	23	770
4	840	14	807	24	792
5	772	15	820	25	765
6	750	16	812	26	749
7	814	17	804	27	829
8	820	18	754	28	821
9	753	19	715	29	816
10	796	20	845	30	743

1. (**Confidence Interval on μ, σ^2 Known; Two-Sided CI**) Construct a 95% two-sided confidence interval on the mean life length (μ) of an INFINITY light bulb.

$1 - \alpha = 0.95 \quad \Rightarrow \quad \alpha = 0.05$

95% two-sided CI on μ:

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$\Rightarrow \quad 780 - z_{0.05/2}\frac{40}{\sqrt{30}} \le \mu \le 780 + z_{0.05/2}\frac{40}{\sqrt{30}}$$

$$\Rightarrow \quad 780 - 1.96 \times \frac{40}{\sqrt{30}} \le \mu \le 780 + 1.96 \times \frac{40}{\sqrt{30}}$$

$$\Rightarrow \quad 765.7 \le \mu \le 794.3$$

2. (**Sample Size Selection**) Find a sample size n to construct a two-sided confidence interval on μ with an error = 20 hours from the true mean life length. Use $\alpha = 0.05$.

$$n = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2 = \left(\frac{z_{0.05/2} \times 40}{20}\right)^2 = \left(\frac{1.96 \times 40}{20}\right)^2 = 15.4 \cong 16$$

Exercise 8.1 (MR 8-10)

The diameter of a hole (X; unit: in.) for a cable harness is normal with $\sigma^2 = 0.01^2$. A random sample of n = 10 yields an average diameter ($\bar{x}$) of 1.5045 in.

1. Construct a 99% upper-confidence bound on the mean diameter (μ) of the hole.
2. Find a sample size n to construct a two-sided confidence interval on μ with an error = 0.003 inch from the true mean hole diameter. Use $\alpha = 0.05$.

8-3 Confidence Interval on the Mean of a Normal Distribution, Variance Unknown

Learning Goals

- ❑ Explain the characteristics of a t distribution and read the t table.
- ❑ Identify the point estimator of μ when σ^2 is unknown and the sampling distribution of the point estimator.
- ❑ Establish a 100(1 - α)% CI on μ when σ^2 is unknown.

t Distribution

The probability density function of a t distribution with ν degrees of freedom is

$$f(x) = \frac{\Gamma[(\nu+1)/2]}{\sqrt{\pi\nu}\,\Gamma(\nu/2)} \frac{1}{[1+x^2/\nu]^{(\nu+1)/2}}, \quad -\infty < x < \infty$$

The mean and variance of the t distribution are

$$E(X) = 0 \quad \text{and} \quad V(X) = \nu/(\nu-2) \ \text{ for } \nu > 2$$

A t distribution is symmetric about zero, unimodal, and bell-shaped like the standard normal, but has thicker tails than the standard normal (see Figure 8-4 in MR). As ν approaches ∞, the t distribution is close to the standard normal.

t Table

The t table (see Appendix Table IV in MR) provides 100α **upper-tail percentage points** ($t_{\alpha,\nu}$) of t distributions with various degrees of freedom (ν):

$$P(T_\nu > t_{\alpha,\nu}) = \alpha$$

Since a t distribution is symmetric about zero,

$$t_{1-\alpha,\nu} = -t_{\alpha,\nu}$$

(e.g.) Reading the t table

(1) $P(T_{12} > t) = 0.05$, $t_{0.05,12} = 1.782$

(2) $P(T_{12} < t) = 0.05 \Rightarrow P(T_{12} > t) = 0.95$

Since $t_{0.95,12} = t_{1-0.05,12} = -t_{0.05,12}$, $t_{0.95,12} = -1.782$

Inference Context

Parameter of interest: μ

Point estimator of μ: $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$, σ^2 unknown

$$\Rightarrow T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(\nu), \ \nu = n - 1$$

Confidence Interval Formula

A 100(1 - α)% CI on μ when σ^2 is known is

$$\overline{X} - t_{\alpha/2,\nu}\frac{S}{\sqrt{n}} \le \mu \le \overline{X} + t_{\alpha/2,\nu}\frac{S}{\sqrt{n}} \qquad \text{for two-sided CI}$$

$$\overline{X} - t_{\alpha,\nu}\frac{S}{\sqrt{n}} \le \mu \qquad \text{for lower-confidence bound}$$

$$\mu \le \overline{X} + t_{\alpha,\nu}\frac{S}{\sqrt{n}} \qquad \text{for upper-confidence bound}$$

CI Formula (*cont.*)

(Derivation) Two-sided confidence interval on μ, σ^2 unknown

By using $T = \dfrac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(\nu)$,

$$P(L \le \mu \le U) = 1 - \alpha$$

$$\Rightarrow P(-t_{\alpha/2,\nu} \le T \le t_{\alpha/2,\nu}) = 1 - \alpha$$

$$\Rightarrow P\left(-t_{\alpha/2,\nu} \le \frac{\overline{X} - \mu}{S/\sqrt{n}} \le t_{\alpha/2,\nu}\right) = 1 - \alpha$$

$$\Rightarrow P\left(\overline{X} - t_{\alpha/2,\nu}\frac{S}{\sqrt{n}} \le \mu \le \overline{X} + t_{\alpha/2,\nu}\frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

Therefore,

$$L = \overline{X} - t_{\alpha/2,\nu}\frac{S}{\sqrt{n}} \quad \text{and} \quad U = \overline{X} + t_{\alpha/2,\nu}\frac{S}{\sqrt{n}}$$

Note that, for a one-sided CI, use t_α instead of $t_{\alpha/2}$ to derive the corresponding limit.

Large Sample CI

If the sample size is large ($n \ge 30$), the <u>t-based</u> CI formula on μ is similar to the <u>z-based</u> CI formula on μ below, regardless of whether the underlying population is normal or non-normal according to the central limit theorem (presented in Section 7-5).

$$\overline{X} - z_{\alpha/2}\frac{S}{\sqrt{n}} \le \mu \le \overline{X} + z_{\alpha/2}\frac{S}{\sqrt{n}} \qquad \text{for two-sided CI}$$

Example 8.2

(Confidence Interval on μ, σ^2 Unknown; Two-Sided CI) For the light bulb life length data in Example 8.1, the following quantities have been obtained:

$n = 30$, $\bar{x} = 780$, and $s^2 = 40^2$

Assume that the variance σ^2 is unknown. Construct a 95% two-sided confidence interval on the mean life length (μ) of an INFINITY light bulb.

$1 - \alpha = 0.95 \ \Rightarrow\ \alpha = 0.05;\quad \nu = n - 1 = 30 - 1 = 29$

95% two-sided CI on μ:

$$\bar{x} - t_{\alpha/2,\nu}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{\alpha/2,\nu}\frac{s}{\sqrt{n}}$$

$$\Rightarrow 780 - t_{0.05/2,29}\frac{40}{\sqrt{30}} \le \mu \le 780 + t_{0.05/2,29}\frac{40}{\sqrt{30}}$$

$$\Rightarrow 780 - 2.045 \times \frac{40}{\sqrt{30}} \le \mu \le 780 + 2.045 \times \frac{40}{\sqrt{30}}$$

$$\Rightarrow 765.1 \le \mu \le 794.9$$

Note that this t-based CI is <u>wider</u> than the corresponding z-based CI $765.7 \le \mu \le 794.3$ in Example 8.1.

Exercise 8.2 (MR 8-23)

An Izod impact test was performed on $n = 20$ specimens of a PVC pipe product. The sample average and sample standard deviation were $\bar{x} = 1.25$ and $s = 0.25$, respectively. Assume that Izod impact strength (X) is normally distributed. Construct a 95% upper-confidence bound on the mean Izod impact strength (μ).

8-4 Confidence Interval on the Variance and Standard Deviation of a Normal Population

Learning Goals

- ❑ Explain the characteristics of a χ^2 distribution.
- ❑ Read the χ^2 table.
- ❑ Identify the point estimator of σ^2 and the sampling distribution of the point estimator.
- ❑ Establish a $100(1 - \alpha)\%$ CI on σ^2.

χ^2 Distribution

The probability density function of a χ^2 distribution with ν degrees of freedom is

$$f(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1}e^{-x/2}, \quad 0 < x$$

The mean and variance of the χ^2 distribution are

$$E(X) = \nu \text{ and } V(X) = 2\nu$$

A χ^2 distribution is unimodal and skewed to the right (see Figure 8-8 in MR). As ν approaches ∞, the χ^2 distribution is close to a normal distribution.

χ^2 Table

The χ^2 table (see Appendix Table III in MR) provides 100α **upper-tail percentage points** ($\chi^2_{\alpha,\nu}$) of χ^2 distributions with various degrees of freedom (ν), i.e.,

$$P(X^2_\nu > \chi^2_{\alpha,\nu}) = \alpha$$

(e.g.) Reading the χ^2 Table

(1) $P(X^2_{14} > \chi^2) = 0.1$, $\chi^2 = \chi^2_{0.1,14} = 21.06$

(2) $P(X^2_{14} < \chi^2) = 0.1 \Rightarrow 1 - P(X^2_{14} > \chi^2) = 0.9$, $\chi^2 = \chi^2_{0.9,14} = 7.79$

Inference Context

Parameter of interest: σ^2

Point estimator of σ^2: $S^2 = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$, $X_1, X_2, \ldots, X_n \sim i.i.d.\ N(\mu, \sigma^2)$

$$\Rightarrow \quad X^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(\nu),\ \nu = n-1$$

Confidence Interval Formula

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2,v}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,v}} \quad \text{for two-sided CI}$$

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2,v}} \le \sigma^2 \quad \text{for lower-confidence bound}$$

$$\sigma^2 \le \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,v}} \quad \text{for upper-confidence bound}$$

(Derivation) Two-sided confidence interval on σ^2

By using the test statistic $X^2 = \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(v)$,

$$P\left(L \le \sigma^2 \le U\right) = 1-\alpha$$

$$\Rightarrow P\left(\chi^2_{1-\alpha/2,v} \le X^2 \le \chi^2_{\alpha/2,v}\right) = 1-\alpha$$

$$\Rightarrow P\left(\chi^2_{1-\alpha/2,v} \le \frac{(n-1)S^2}{\sigma^2} \le \chi^2_{\alpha/2,v}\right) = 1-\alpha$$

$$\Rightarrow P\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2,v}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,v}}\right) = 1-\alpha$$

Therefore,

$$L = \frac{(n-1)S^2}{\chi^2_{\alpha/2,v}} \quad \text{and} \quad U = \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,v}}$$

Note that, for a one-sided CI, use χ^2_{α} instead of $\chi^2_{\alpha/2}$ to derive the corresponding limit.

Example 8.3 **(Confidence Interval on σ^2; Two-Sided CI)** Suppose that the life length of an INFINITY light bulb (X) follows a normal distribution with mean μ and variance σ^2. A random sample of $n = 16$ bulbs is tested, and the sample variance is found to be $s^2 = 44^2$. Construct a 95% two-sided confidence interval on the variance of the life length of an INFINITY light bulb (σ^2).

$1 - \alpha = 0.95 \Rightarrow \alpha = 0.05$; $v = n - 1 = 16 - 1 = 15$

95% two-sided CI on σ^2:

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2,v}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,v}}$$

$$\Rightarrow \frac{(16-1)\times 44^2}{\chi^2_{0.05/2,15}} \le \sigma^2 \le \frac{(16-1)\times 44^2}{\chi^2_{1-0.05/2,15}}$$

$$\Rightarrow \frac{29{,}040}{27.49} \le \sigma^2 \le \frac{29{,}040}{6.26}$$

$$\Rightarrow 32.5^2 \le \sigma^2 \le 68.1^2$$

Exercise 8.3 (MR 8-35)

A rivet is to be inserted into a hole. A random sample of $n = 15$ holes is selected, resulting in the sample standard deviation of the hold diameter = 0.008 mm. Assume that the hole diameter is normally distributed. Construct a 99% upper-confidence bound on the variance of the hole diameter (σ^2).

8-5 A Large-Sample Confidence Interval for a Population Proportion

Learning Goals

- ❑ Identify the point estimator of p and the sampling distribution of the point estimator.
- ❑ Establish a $100(1 - \alpha)\%$ CI on p.
- ❑ Determine a sample size to satisfy a predetermined level of error (E) in estimating p.

Inference Context

Parameter of interest: p

Point estimator of p: $\hat{P} = \dfrac{X}{n}$, where $X \sim B(n, p)$

$$\Rightarrow \; Z = \frac{\hat{P} - p}{\sqrt{p(1-p)/n}} \sim N(0,1) \text{ if } np \text{ and } n(1-p) > 5$$

Sampling Distribution of $\hat{P}$

The mean and variance of a binomial random variable $X \sim B(n, p)$ are

$$E(X) = np \text{ and } V(X) = np(1-p)$$

Thus,

$$E(\hat{P}) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n}np = p$$

$$V(\hat{P}) = V\left(\frac{X}{n}\right) = \frac{1}{n^2}V(X) = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$$

Recall that in Section 4-7 a binomial distribution $B(n, p)$ approximates to a normal distribution if p is not close to either zero or one and n is relatively large so that both np and $n(1-p)$ are greater than five.

Therefore, if np and $n(1 - p) > 5$, the approximate distribution of $\hat{P}$ is

$$\hat{P} \sim N\left(p, \frac{p(1-p)}{n}\right) \quad \Rightarrow \quad Z = \frac{\hat{P} - p}{\sqrt{p(1-p)/n}} \sim N(0,1)$$

Confidence Interval Formula

$$\hat{P} - z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \le p \le \hat{P} + z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \quad \text{for two-sided CI}$$

$$\hat{P} - z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \le p \quad \text{for lower-confidence bound}$$

$$p \le \hat{P} + z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \quad \text{for upper-confidence bound}$$

(Derivation) Two-sided confidence interval* on p

By using the test statistic $Z = \dfrac{\hat{P} - p}{\sqrt{p(1-p)/n}} \sim N(0,1)$,

$$P(L \le p \le U) = 1 - \alpha$$

$$\Rightarrow P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) = 1 - \alpha$$

$$\Rightarrow P\left(-z_{\alpha/2} \le \frac{\hat{P} - p}{\sqrt{p(1-p)/n}} \le z_{\alpha/2}\right) = 1 - \alpha$$

$$\Rightarrow P\left(\hat{P} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \le p \le \hat{P} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right) = 1 - \alpha$$

By using $\hat{P}$ in place of p because the upper and lower limits of the CI include the unknown parameter p,

$$\Rightarrow P\left(\hat{P} - z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \le p \le \hat{P} + z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}\right) = 1 - \alpha$$

Therefore,

$$L = \hat{P} - z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \quad \text{and} \quad U = \hat{P} + z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$$

* For a one-sided CI, use z_α instead of $z_{\alpha/2}$ to derive the corresponding limit.

Sample Size Selection

In estimating p, the following formulas are used to select a sample size n for a predefined level of error:

$$n = \left(\frac{z_{\alpha/2}}{E}\right)^2 p(1-p) \quad \text{if information of } p \text{ is available}$$

$$= \left(\frac{z_{\alpha/2}}{E}\right)^2 \times 0.25 \quad \text{if information of } p \text{ is unavailable}$$

Example 8.4

A sample of $n = 40$ bridges in HAPPY County is tested for metal corrosion, and $x = 28$ bridges are found corroded.

1. **(Confidence Interval on p; Two-Sided CI)** Construct a 95% two-sided confidence interval on the proportion of corroded bridges (p) in the county.

Example 8.4 (*cont.*)

$1 - \alpha = 0.95 \Rightarrow \alpha = 0.05;\quad \hat{p} = \frac{x}{n} = \frac{28}{40} = 0.7$

Since both $n\hat{p} = 40 \times 0.7 = 28$ and $n(1-\hat{p}) = 40 \times 0.3 = 12$ are greater than five, the sampling distribution of $\hat{P}$ is approximately normal.

95% two-sided CI on p:

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\Rightarrow\ 0.7 - z_{0.05/2}\sqrt{\frac{0.7(1-0.7)}{40}} \le p \le 0.7 + z_{0.05/2}\sqrt{\frac{0.7(1-0.7)}{40}}$$

$$\Rightarrow\ 0.7 - 1.96 \times 0.07 \le p \le 0.7 + 1.96 \times 0.07$$

$$\Rightarrow\ 0.56 \le p \le 0.84$$

2. (**Sample Size Selection**) Determine a sample size n to establish a 95% confidence interval on p with an error = 0.05 from the true proportion.

$$n = \left(\frac{z_{\alpha/2}}{E}\right)^2 \times 0.25 = \left(\frac{z_{0.05/2}}{0.05}\right)^2 \times 0.25 = \left(\frac{1.96}{0.05}\right)^2 \times 0.25 = 384.2 \cong 385$$

Exercise 8.4 (MR 8-48)

The proportion of defective integrated circuits produced in a photolithography process is being studied. A random sample of $n = 300$ circuits is tested, revealing $x = 13$ defectives.

1. Construct a 95% lower-confidence bound on the proportion of defective integrated circuits (p).
2. Determine a sample size n to establish a 95% confidence interval on p with an error = 0.02 from the true proportion. Use $\hat{p} = 0.05$ as an initial estimate of p for sample size calculation.

8-6 A Prediction Interval for a Future Observation

Learning Goals

- Identify the distribution of the prediction error $X_{n+1} - \overline{X}$.
- Establish a 100(1 - α)% prediction interval (PI) for a new observation.

Sampling Distribution of $E = X_{n+1} - \overline{X}$

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from a normal population with mean μ and variance σ^2 and we wish to predict a new observation X_{n+1}. If $\overline{X}$ is used as a point estimator of X_{n+1}, then the distribution of the corresponding prediction error E is

$$E = X_{n+1} - \overline{X} \sim N\left(0, \sigma^2\left[1 + \frac{1}{n}\right]\right),\ \text{since } X_{n+1} \sim N(0,\sigma^2),\ \overline{X} \sim N\left(0, \frac{\sigma^2}{n}\right),$$

X_{n+1} and $\overline{X}$ are independent

Sampling Distribution of $E = X_{n+1} - \overline{X}$ (*cont.*)

Thus,

$$Z = \frac{X_{n+1} - \overline{X}}{\sigma\sqrt{1 + \frac{1}{n}}} \sim N(0,1^2), \quad \text{if } \sigma^2 \text{ is known}$$

$$T = \frac{X_{n+1} - \overline{X}}{S\sqrt{1 + \frac{1}{n}}} \sim t(n-1), \quad \text{if } \sigma^2 \text{ is unknown}$$

Prediction Interval Formula

$$\overline{x} - z_{\alpha/2}\sigma\sqrt{1 + \frac{1}{n}} \le X_{n+1} \le \overline{x} + z_{\alpha/2}\sigma\sqrt{1 + \frac{1}{n}} \qquad \text{for two-sided CI, } \sigma^2 \text{ known}$$

$$\overline{x} - t_{\alpha/2,n-1}s\sqrt{1 + \frac{1}{n}} \le X_{n+1} \le \overline{x} + t_{\alpha/2,n-1}s\sqrt{1 + \frac{1}{n}} \qquad \text{for two-sided CI, } \sigma^2 \text{ unknown}$$

Example 8.5

(Prediction Interval on X_{n+1}, σ^2 Unknown; Two-Sided CI) For the light bulb life length data in Example 8.2, the following quantities have been obtained:

$n = 30,\ \overline{x} = 780$, and $s^2 = 40^2$

Construct a 95% two-sided prediction interval on the life length of the next light bulb tested (X_{31}).

$1 - \alpha = 0.95 \quad \Rightarrow \quad \alpha = 0.05$

95% two-sided PI on X_{31}:

$$\overline{x} - t_{\alpha/2,n-1}s\sqrt{1 + \frac{1}{n}} \le X_{n+1} \le \overline{x} + t_{\alpha/2,n-1}s\sqrt{1 + \frac{1}{n}}$$

$$\Rightarrow \quad 780 - t_{0.05/2,29} \times 40 \times \sqrt{1 + \frac{1}{30}} \le X_{31} \le 780 + t_{0.05/2,29} \times 40 \times \sqrt{1 + \frac{1}{30}}$$

$$\Rightarrow \quad 780 - 2.045 \times 40.7 \le X_{31} \le 780 + 2.045 \times 40.7$$

$$\Rightarrow \quad 696.8 \le X_{31} \le 863.2$$

Note that this t-based PI is wider than the corresponding t-based CI on μ $765.1 \le \mu \le 794.9$ in Example 8.2.

Exercise 8.5 (MR 8-50)

For the Izod impact test in Exercise 8.2, the following quantities have been obtained:

$n = 20,\ \overline{x} = 1.25$, and $s^2 = 0.25^2$

Construct a 95% two-sided prediction interval on the impact strength of the next specimen of PVC pipe tested (X_{21}).

8-7 Tolerance Intervals for a Normal Distribution

Learning Goals

- ❑ Establish a γ tolerance interval with 100(1 - α)% confidence for a normal population.

Tolerance Interval

Suppose that the life length (X) of an INFINITY light bulb follows a normal distribution with mean μ and variance σ^2. Then, the interval ($\mu - 1.96\sigma$, $\mu + 1.96\sigma$), called <u>tolerance interval</u>, includes 95% (γ, coverage percent) of the light bulb population in terms of life length.

When μ and σ^2 are unknown, the 95% tolerance interval may be established as ($\bar{x} - 1.96s$, $\bar{x} + 1.96s$) by using $\bar{x}$ and s^2. However, due to sampling variability in $\bar{x}$ and s^2, the estimated tolerance interval includes less than 95% of the population.

Thus, when μ and σ^2 are unknown, a γ tolerance interval with confidence level 100(1 - α)% is established by ($\bar{x} - ks$, $\bar{x} + ks$), where k is found in Appendix Table XI of MR. Note that as the sample size n approaches ∞, the value of k is close to the z-value associated with γ.

Example 8.6

(Tolerance Interval on *X*; Two-Sided CI) For the light bulb life length data in Example 8.2, the following quantities have been obtained:

$n = 30,\ \bar{x} = 780$, and $s^2 = 40^2$

Construct a 95% two-sided tolerance interval with 95% confidence for the life length measurements of light bulbs.

$k = 2.529$ for $n = 30$, $\gamma = 0.95$, and confidence level = 0.95

95% two-sided TI is

$(\bar{x} - ks,\ \bar{x} + ks)$

$\Rightarrow (780 - 2.529 \times 40,\ 780 + 2.529 \times 40)$

$\Rightarrow (678.8, 881.2)$

Note that this TI is <u>wider</u> than the corresponding t-based CI on μ $765.1 \le \mu \le 794.9$ in Example 8.2.

Exercise 8.6 (MR 8-61)

For the Izod impact test in Exercise 8.2, the following quantities have been obtained:

$n = 20,\ \bar{x} = 1.25$, and $s^2 = 0.25^2$

Construct a 95% two-sided tolerance interval on the impact strength of PVC pipe with 90% confidence.

Summary of Statistical Intervals for a Single Sample

A. Confidence Intervals

Case	Point Estimator	100(1 - α)% Two-Sided CI	Section
μ, σ^2 known	$\overline{X}$	$\overline{X} - z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$	8-2
μ, σ^2 unknown	$\overline{X}$	$\overline{X} - t_{\alpha/2,n-1}\dfrac{S}{\sqrt{n}} \le \mu \le \overline{X} + t_{\alpha/2,n-1}\dfrac{S}{\sqrt{n}}$	8-3
σ^2 of a normal distribution	S^2	$\dfrac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \le \sigma^2 \le \dfrac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}$	8-4
p of a binomial distribution	$\hat{P} = \dfrac{X}{n}$	$\hat{P} - z_{\alpha/2}\sqrt{\dfrac{\hat{P}(1-\hat{P})}{n}} \le p \le \hat{P} + z_{\alpha/2}\sqrt{\dfrac{\hat{P}(1-\hat{P})}{n}}$	8-5

B. Prediction Intervals

Case	Point Estimator	100(1 - α)% Two-Sided PI	Section
X_{n+1}, σ^2 known	$\overline{X}$	$\overline{x} - z_{\alpha/2}\sigma\sqrt{1+\dfrac{1}{n}} \le X_{n+1} \le \overline{x} + z_{\alpha/2}\sigma\sqrt{1+\dfrac{1}{n}}$	8-6
X_{n+1}, σ^2 unknown	$\overline{X}$	$\overline{x} - t_{\alpha/2,n-1}s\sqrt{1+\dfrac{1}{n}} \le X_{n+1} \le \overline{x} + t_{\alpha/2,n-1}s\sqrt{1+\dfrac{1}{n}}$	

C. Tolerance Intervals

Case	γ* Two-Sided TI with 100(1 - α)% confidence	Section
$X \sim N(\mu, \sigma^2)$, σ^2 known	$\overline{x} - z_{(1-\gamma)/2}\sigma \le X \le \overline{x} + z_{(1-\gamma)/2}\sigma$	8-7
$X \sim N(\mu, \sigma^2)$, σ^2 unknown	$\overline{x} - ks \le X \le \overline{x} + ks$ **	

* γ: coverage percent

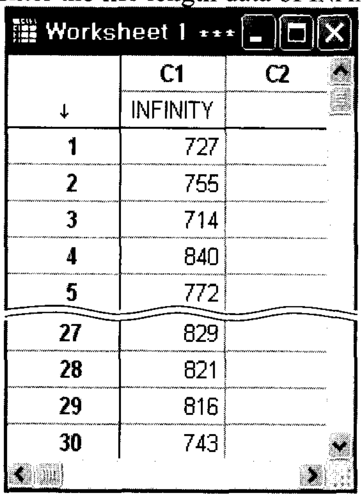
** k: tolerance level factor found in Appendix Table XI of MR

MINITAB Applications

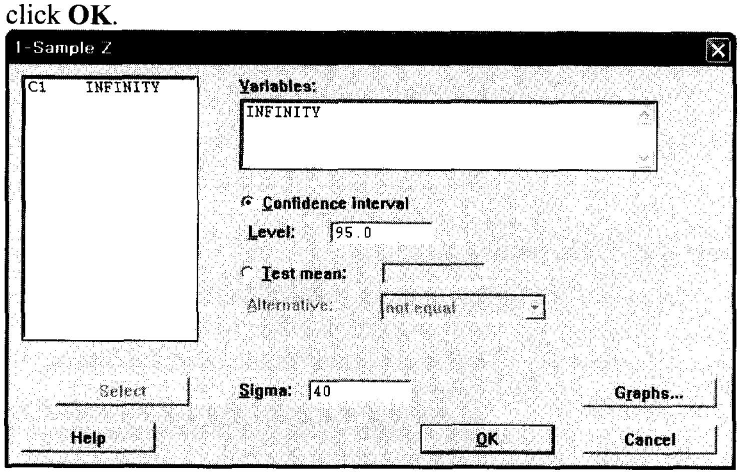
Example 8.1 **(Inference on μ, σ^2 Known)**

(1) Choose **File ➤ New**, click **Minitab Project**, and click **OK**.

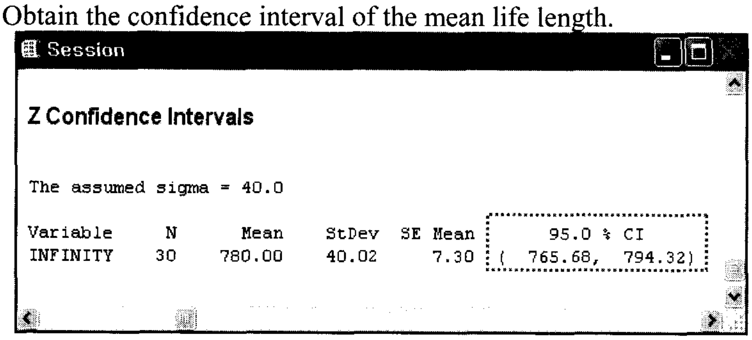
(2) Enter the life length data of INFINITY light bulbs on the worksheet.



(3) Choose **Stat ➤ Basic Statistics ➤ 1-Sample Z**. In **Variables**, select INFINITY. Click **Confidence Interval** and enter the level of confidence in **Level**. In **Sigma**, enter the population standard deviation assumed. Then click **OK**.



(4) Obtain the confidence interval of the mean life length.

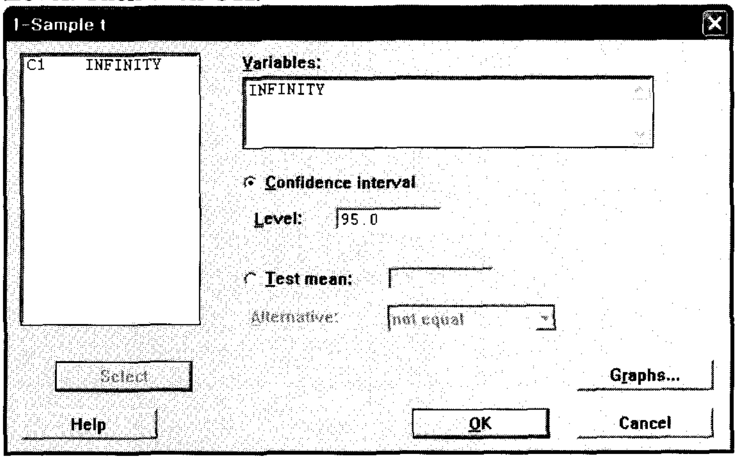


Example 8.2 **(Inference on μ, σ^2 Unknown)**

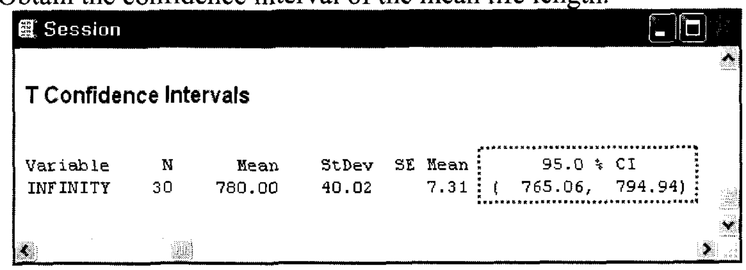
(1) Choose **File ➤ New**, click **Minitab Project**, and click **OK**.

(2) Enter the life length data of INFINITY light bulbs on the worksheet.

(3) Choose **Stat ➤ Basic Statistics ➤ 1-Sample t**. In **Variables**, select INFINITY. Click **Confidence Interval** and enter the level of confidence in **Level**. Then click **OK**.

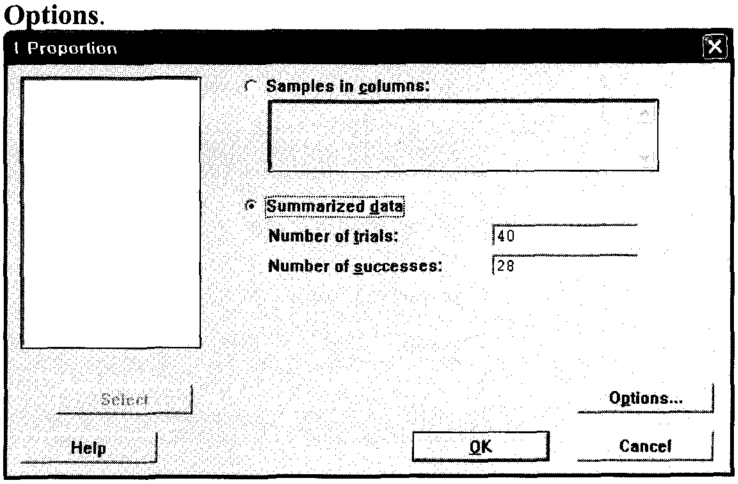


(4) Obtain the confidence interval of the mean life length.



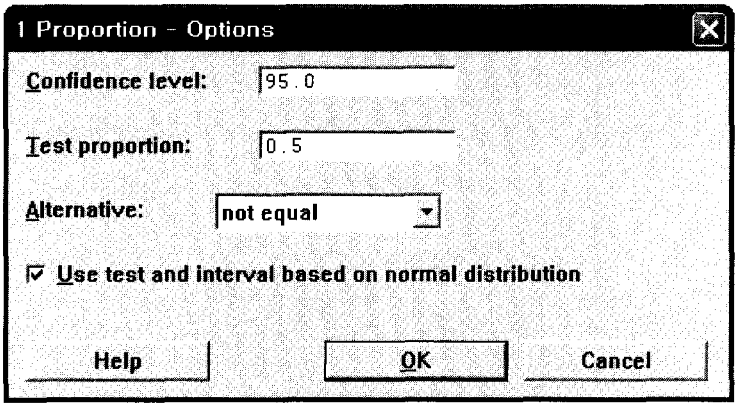
Example 8.4 **(Inference on *p*)**

(1) Choose **Stat ➤ Basic Statistics ➤ 1 Proportion**. Click **Summarized data** and enter the number of bridges surveyed in **Number of trials** and the number of bridges corroded in **Number of successes**. Then click **Options**.

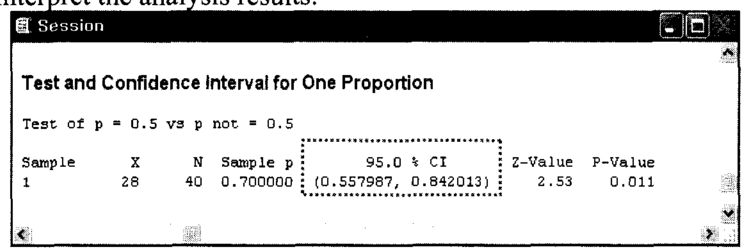


Example 8.4 (*cont.*)

(2) Enter the level of confidence in **Confidence level** and the hypothesized proportion in **Test proportion**. Check **Use test and interval based on normal distribution**. Then click **OK** twice.



(3) Interpret the analysis results.



Answers to Exercises

Exercise 8.1

1. **(Confidence Interval on μ, σ^2 Known; Upper-Confidence Bound)**

$1 - \alpha = 0.99 \quad \Rightarrow \quad \alpha = 0.01$

99% upper-confidence bound on μ:

$$\mu \leq \bar{x} + z_{\alpha} \frac{\sigma}{\sqrt{n}}$$

$$\Rightarrow \quad \mu \leq 1.5045 + z_{0.01} \frac{0.01}{\sqrt{10}}$$

$$\Rightarrow \quad \mu \leq 1.5045 + 2.57 \times \frac{0.01}{\sqrt{10}}$$

$$\Rightarrow \quad \mu \leq 1.51$$

2. **(Sample Size Selection)**

$$n = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2 = \left(\frac{z_{0.05/2} \times 0.01}{0.003}\right)^2 = \left(\frac{1.96 \times 0.01}{0.003}\right)^2 = 42.7 \cong 43$$

Exercise 8.2

(Confidence Interval on μ, σ^2 Unknown; Upper-Confidence Bound)

$1 - \alpha = 0.95 \quad \Rightarrow \quad \alpha = 0.05; \quad \nu = n - 1 = 20 - 1 = 19$

95% upper-confidence bound on μ:

$$\mu \leq \bar{x} + t_{\alpha,\nu} \frac{s}{\sqrt{n}}$$

$$\Rightarrow \quad \mu \leq 1.25 + t_{0.05,19} \frac{0.25}{\sqrt{20}}$$

$$\Rightarrow \quad \mu \leq 1.25 + 1.73 \times \frac{0.25}{\sqrt{20}}$$

$$\Rightarrow \quad \mu \leq 1.35$$

Exercise 8.3

(Confidence Interval on σ^2; Upper-Confidence Bound)

$1 - \alpha = 0.99 \quad \Rightarrow \quad \alpha = 0.01; \quad \nu = n - 1 = 15 - 1 = 14$

99% two-sided CI on σ^2:

$$\sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha,\nu}}$$

$$\Rightarrow \quad \sigma^2 \leq \frac{(15-1) \times 0.008^2}{\chi^2_{1-0.01,14}}$$

$$\Rightarrow \quad \sigma^2 \leq \frac{0.0009}{4.66}$$

$$\Rightarrow \quad \sigma^2 \leq 0.014^2$$

Exercise 8.4

1. **(Confidence Interval on p; Lower-Confidence Bound)**

$$1 - \alpha = 0.95 \quad \Rightarrow \quad \alpha = 0.05; \quad \hat{p} = \frac{x}{n} = \frac{13}{300} = 0.043$$

Since both $n\hat{p} = 300 \times 0.043 = 13$ and $n(1-\hat{p}) = 300 \times 0.957 = 287$ are greater than five, the sampling distribution of $\hat{P}$ is approximately normal.

95% lower-confidence bound on p:

$$\hat{p} - z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p$$

$$\Rightarrow \quad 0.043 - z_{0.05}\sqrt{\frac{0.043(1-0.043)}{300}} \le p$$

$$\Rightarrow \quad 0.043 - 1.64 \times 0.012 \le p$$

$$\Rightarrow \quad 0.024 \le p$$

2. **(Sample Size Selection)**

$$n = \left(\frac{z_{\alpha/2}}{E}\right)^2 p(1-p) = \left(\frac{z_{0.05/2}}{0.02}\right)^2 \times 0.05 \times (1-0.05)$$

$$= \left(\frac{1.96}{0.02}\right)^2 \times 0.05 \times 0.95 = 456.2 \cong 457$$

Exercise 8.5

(Prediction Interval on X_{n+1}, σ^2 Unknown; Two-Sided CI)

$1 - \alpha = 0.95 \quad \Rightarrow \quad \alpha = 0.05$

95% two-sided PI on X_{21}:

$$\bar{x} - t_{\alpha/2,n-1}s\sqrt{1+\frac{1}{n}} \le X_{n+1} \le \bar{x} + t_{\alpha/2,n-1}s\sqrt{1+\frac{1}{n}}$$

$$\Rightarrow \quad 1.25 - t_{0.05/2,19} \times 0.25 \times \sqrt{1+\frac{1}{20}} \le X_{21} \le 1.25 + t_{0.05/2,19} \times 0.25 \times \sqrt{1+\frac{1}{20}}$$

$$\Rightarrow \quad 1.25 - 2.093 \times 0.25 \le X_{21} \le 1.25 + 2.093 \times 0.25$$

$$\Rightarrow \quad 0.72 \le X_{21} \le 1.78$$

Exercise 8.6

(Tolerance Interval on X; Two-Sided CI)

$k = 2.564$ for $n = 20$, $\gamma = 0.95$, and confidence level $= 0.90$

95% two-sided TI is

$$(\bar{x} - ks,\ \bar{x} + ks)$$

$$\Rightarrow \quad (1.25 - 2.564 \times 0.25,\ 1.25 + 2.564 \times 0.25)$$

$$\Rightarrow \quad (0.61,\ 1.89)$$

9 Tests of Hypotheses for a Single Sample

OUTLINE

9-1 Hypothesis Testing

Learning Goals

- ☐ Explain the terms *null hypothesis*, *alternative hypothesis*, *test statistic*, *acceptance region*, *rejection region*, *critical value*, *type I error probability* (α), *type II error probability* (β), and *power of a test* (1-β).
- ☐ Establish the acceptance and rejection regions of a hypothesis test at α.
- ☐ Determine the type II error probability and power of a test.
- ☐ Explain the relationships between α and β.
- ☐ Explain the difference between strong conclusion and weak conclusion.
- ☐ Identify the procedure of hypothesis testing.

Hypothesis A hypothesis is an assertion about the parameters (θ) of one or more populations under study. There are largely two kinds of hypotheses:

1. **Null hypothesis (H_0)**: States the presumed condition of θ (based on experience, theory, design specification, regulation, or contractual obligation) that will be held unless there is strong evidence against it. Note that H_0 should always specify an exact value of θ.

   (e.g.) H_0: $\mu_X = 750$ hrs, where X is the life length of an INFINITY light bulb

Hypothesis (*cont.*)

2. **Alternative hypothesis (H_1)**: States the condition of θ that would be concluded if H_0 is rejected. In parallel to the types of confidence intervals (see Section 8-2), the following types of H_1 are defined depending on the directionality of θ assumed in H_1.
   (1) **Two-sided H_1**: Indicates <u>no</u> directionality of θ.
   (e.g.) H_1: $\mu_X \neq 750$ hrs
   (2) **One-sided H_1**: Indicates the directionality of θ.
   (a) **Lower-sided H_1**: Includes the <u>lower-side</u> inequality sign.
   (e.g.) H_1: $\mu_X < 750$ hrs
   (b) **Upper-sided H_1**: Includes the <u>upper-side</u> inequality sign.
   (e.g.) H_1: $\mu_X > 750$ hrs

Test Statistic

A test statistic refers to a <u>statistic</u> used for <u>statistical inference</u> about θ.

(e.g.) Test statistic for inference on μ_X where $X \sim N(\mu, \sigma^2)$ and σ^2 is known:

$$\overline{X} = \sum_{i=1}^{n} X_i / n \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{or} \quad Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1^2)$$

Test Regions

Two regions of a test statistic (see Figure 9-1) are established for testing H_0 against H_1:

1. **Acceptance region**: The region of a test statistic that will lead to <u>failure to reject</u> H_0.
2. **Rejection (critical) region**: The region of a test statistic that will lead to <u>rejection</u> of H_0.

Note that the boundaries between the acceptance and rejection regions are called **critical values**.

For example, in Figure 9-1, in testing H_0: $\mu_X = 750$ vs. H_1: $\mu_X \neq 750$ where X denotes the life length of an INFINITY light bulb, the test statistic of μ is the sample mean and the critical values of the sample mean are 730 and 770 hours. Therefore, we will conclude the following:

(1) If $730 \leq \bar{x} \leq 770$ (acceptance region), fail to reject H_0.
(2) if $\bar{x} < 730$ or $\bar{x} > 770$ (rejection region), reject H_0.

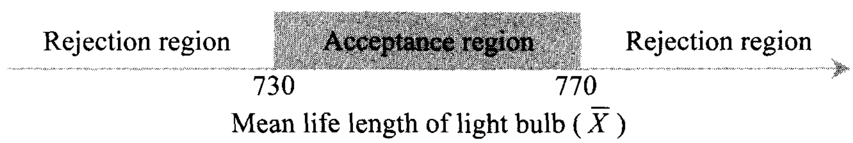


Figure 9-1 Acceptance and rejection rejections for H_0: $\mu_X = 750$ vs. H_1: $\mu_X \neq 750$, where X (unit: hours) is the life length of an INFINITY light bulb.

Test Errors

The truth or falsity of a hypothesis can <u>never be known with certainty</u> unless the entire population is examined accurately and thoroughly. Therefore, a hypothesis test based on a random sample may lead to one of the two types of wrong conclusions (see Table 9-1):

(1) **Type I error**: Reject H_0 when H_0 is true.
(2) **Type II error**: Fail to reject H_0 when H_0 is false.

Test Errors (cont.)

Table 9-1 Decision Matrix of Hypothesis Testing

		Decision: Fail to reject H_0	Decision: Reject H_0
Truth	H_0 is true (H_1 is false)	Correct decision (Correct acceptance)	**Type I error** (False rejection)
	H_0 is false (H_1 is true)	**Type II error** (False acceptance)	Correct decision (Correct rejection)

***The probability of type I error** (denoted as α) and **the probability of type II** error (denoted as β) are conditional probabilities as follows.*

$\alpha = P(\text{type I error}) = P(\text{reject } H_0 \mid H_0 \text{ is true})$
$\beta = P(\text{type II error}) = P(\text{fail to reject } H_0 \mid H_0 \text{ is false})$

Note that the type I error probability (α) is also called the **level of significance** of a test. Small values of α such as <u>0.05 and 0.01</u> are typically used to make it difficult to reject H_0 if it is true.

Power of Test

The **power** of a statistical test indicates the probability of rejecting H_0 when H_0 is false, indicating the ability (sensitivity) of the test to detect evidence against H_0:

$$\begin{aligned} power &= P(\text{reject } H_0 \mid H_0 \text{ is false}) \\ &= 1 - P(\text{fail to reject } H_0 \mid H_0 \text{ is false}) = 1 - \beta \end{aligned}$$

Test Regions, Hypotheses, α, β, and Power

The test regions, hypotheses, α, β, and power of a test are related to each other. The <u>acceptance and rejection regions</u> of a test statistic $\hat{\Theta}$ are determined <u>based on α</u> and the <u>hypothesized value of θ</u> (denoted as θ_0) in H_0 (note that α and θ_0 are specified by the analyst). If L and U denote the lower and upper limits of an acceptance region, respectively, and we are testing H_0: $\theta = \theta_0$, the acceptance region of $\hat{\Theta}$ is determined as follows:

$$\begin{aligned} 1 - \alpha &= P(\text{fail to reject } H_0 \mid H_0 \text{ is true}) = P(\text{fail to reject } H_0 \mid \theta = \theta_0) \\ &= \begin{cases} P(L \le \hat{\Theta} \le U \mid \theta = \theta_0), & \text{for two-sided } H_1 \\ P(L \le \hat{\Theta} \mid \theta = \theta_0), & \text{for lower-sided } H_1 \\ P(\hat{\Theta} \le U \mid \theta = \theta_0), & \text{for upper-sided } H_1 \end{cases} \end{aligned}$$

On the other hand, the <u>β and power</u> of the test are determined <u>based on the acceptance region</u> of the test and the <u>true value of θ</u> as follows:

$$\begin{aligned} \beta &= P(\text{fail to reject } H_0 \mid H_0 \text{ is false}) = P(\text{fail to reject } H_0 \mid \theta \ne \theta_0) \\ &= \begin{cases} P(L \le \hat{\Theta} \le U \mid \theta \ne \theta_0), & \text{for two-sided } H_1 \\ P(L \le \hat{\Theta} \mid \theta \ne \theta_0), & \text{for lower-sided } H_1 \\ P(\hat{\Theta} \le U \mid \theta \ne \theta_0), & \text{for upper-sided } H_1 \end{cases} \end{aligned}$$

and

$$power = 1 - \beta$$

Test Errors (*cont.*)

Table 9-1 Decision Matrix of Hypothesis Testing

		Decision	
		Fail to reject H_0	Reject H_0
Truth	H_0 is true (H_1 is false)	Correct decision (Correct acceptance)	**Type I error** (False rejection)
	H_0 is false (H_1 is true)	**Type II error** (False acceptance)	Correct decision (Correct rejection)

The probability of type I error (denoted as α) and **the probability of type II error** (denoted as β) are conditional probabilities as follows:

$\alpha = P(\text{type I error}) = P(\text{reject } H_0 \mid H_0 \text{ is true})$

$\beta = P(\text{type II error}) = P(\text{fail to reject } H_0 \mid H_0 \text{ is false})$

Note that the type I error probability (α) is also called the **level of significance** of a test. Small values of α such as 0.05 and 0.01 are typically used to make it difficult to reject H_0 if it is true.

Power of Test

The **power** of a statistical test indicates the probability of rejecting H_0 when H_0 is false, indicating the ability (sensitivity) of the test to detect evidence against H_0:

$$\begin{aligned} power &= P(\text{reject } H_0 \mid H_0 \text{ is false}) \\ &= 1 - P(\text{fail to reject } H_0 \mid H_0 \text{ is false}) = 1 - \beta \end{aligned}$$

Test Regions, Hypotheses, α, β, and Power

The test regions, hypotheses, α, β, and power of a test are related to each other. The acceptance and rejection regions of a test statistic $\hat{\Theta}$ are determined based on α and the hypothesized value of θ (denoted as θ_0) in H_0 (note that α and θ_0 are specified by the analyst). If L and U denote the lower and upper limits of an acceptance region, respectively, and we are testing H_0: $\theta = \theta_0$, the acceptance region of $\hat{\Theta}$ is determined as follows:

$$1 - \alpha = P(\text{fail to reject } H_0 \mid H_0 \text{ is true}) = P(\text{fail to reject } H_0 \mid \theta = \theta_0)$$

$$= \begin{cases} P(L \le \hat{\Theta} \le U \mid \theta = \theta_0), & \text{for two-sided } H_1 \\ P(L \le \hat{\Theta} \mid \theta = \theta_0), & \text{for lower-sided } H_1 \\ P(\hat{\Theta} \le U \mid \theta = \theta_0), & \text{for upper-sided } H_1 \end{cases}$$

On the other hand, the β and power of the test are determined based on the acceptance region of the test and the true value of θ as follows:

$$\beta = P(\text{fail to reject } H_0 \mid H_0 \text{ is false}) = P(\text{fail to reject } H_0 \mid \theta \ne \theta_0)$$

$$= \begin{cases} P(L \le \hat{\Theta} \le U \mid \theta \ne \theta_0), & \text{for two-sided } H_1 \\ P(L \le \hat{\Theta} \mid \theta \ne \theta_0), & \text{for lower-sided } H_1 \\ P(\hat{\Theta} \le U \mid \theta \ne \theta_0), & \text{for upper-sided } H_1 \end{cases}$$

and

$$power = 1 - \beta$$

Example 9.1

Suppose that the life length of an INFINITY light bulb (X; unit: hour) is normally distributed with $\sigma^2 = 40^2$. We wish to test H_0: $\mu_X = 750$ versus H_1: $\mu_X \neq 750$ with a random sample of size $n = 30$ light bulbs.

1. (**Acceptance/Rejection Regions**) Construct the acceptance and rejection regions of the test on μ at $\alpha = 0.05$.

The test statistic of μ is the sample mean with the following sampling distribution:

$$\bar{X} \sim N\left(\mu, \frac{40^2}{30}\right)$$

The acceptance region $l \leq \bar{X} \leq u$ for H_0: $\mu_X = 750$ vs. H_1: $\mu_X \neq 750$ satisfies the following:

$$\begin{aligned}
1 - \alpha = 1 - 0.05 = 0.95 &= P(\text{fail to reject } H_0 \mid \mu = 750) \\
&= P(l \leq \bar{X} \leq u \mid \mu = 750) \\
&= P\left(\frac{l - \mu}{40/\sqrt{30}} \leq \frac{\bar{X} - \mu}{40/\sqrt{30}} \leq \frac{u - \mu}{40/\sqrt{30}} \mid \mu = 750\right) \\
&= P\left(\frac{l - 750}{40/\sqrt{30}} \leq Z \leq \frac{u - 750}{40/\sqrt{30}}\right) \\
&= P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = P(-z_{0.025} \leq Z \leq z_{0.025}) \\
&= P(-1.96 \leq Z \leq 1.96)
\end{aligned}$$

Accordingly, the critical values are

$$\frac{l - 750}{40/\sqrt{30}} = -1.96 \Rightarrow l = 750 - 1.96 \times \frac{40}{\sqrt{30}} = 735.7$$

$$\frac{u - 750}{40/\sqrt{30}} = 1.96 \Rightarrow u = 750 + 1.96 \times \frac{40}{\sqrt{30}} = 764.3$$

Therefore,

acceptance region: $735.7 \leq \bar{x} \leq 764.3$

rejection region: $\bar{x} < 735.7$ and $\bar{x} > 764.3$

2. (**β and Power of Test**) Assume that the true value of $\mu_X = 730$ hrs. Find the β and power of the test if the acceptance region is $735.7 \leq \bar{x} \leq 764.3$.

$$\begin{aligned}
\beta &= P(\text{fail to reject } H_0 \mid H_0 \text{ is false}) \\
&= P(735.7 \leq \bar{X} \leq 764.3 \mid \mu = 730) \\
&= P\left(\frac{735.7 - \mu}{40/\sqrt{30}} \leq \frac{\bar{X} - \mu}{40/\sqrt{30}} \leq \frac{764.3 - \mu}{40/\sqrt{30}} \mid \mu = 730\right) \\
&= P\left(\frac{735.7 - 730}{40/\sqrt{30}} \leq Z \leq \frac{764.3 - 730}{40/\sqrt{30}}\right) \\
&= P(0.78 \leq Z \leq 4.70) = P(Z \leq 4.77) - P(Z \leq 0.78) \\
&= 1 - 0.782 = 0.218
\end{aligned}$$

power $= 1 - \beta = 0.782$

Exercise 9.1 (MR 9-6)

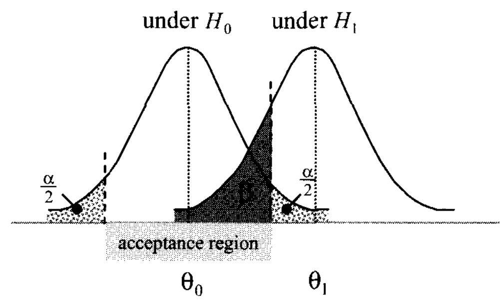
The heat evolved from a cement mixture (X; unit: cal/g) is approximately normally distributed with $\sigma^2 = 2^2$. We wish to test H_0: $\mu = 100$ versus H_1: $\mu \neq 100$ with a sample of $n = 9$ specimens.

1. Find the acceptance region of the test at $\alpha = 0.01$.
2. Suppose that the acceptance region is defined as $98.3 \leq \bar{x} \leq 101.7$. Find the β and power of the test if the true mean heat evolved is 103.

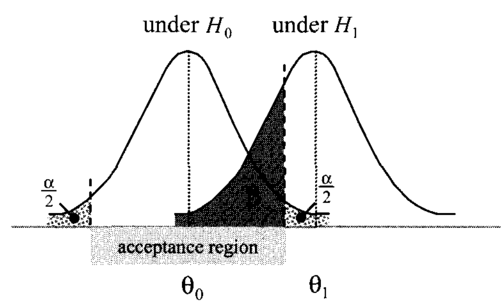
Relationship between α and β

The relationships between α and β are explained with illustrations in Figure 9-2.

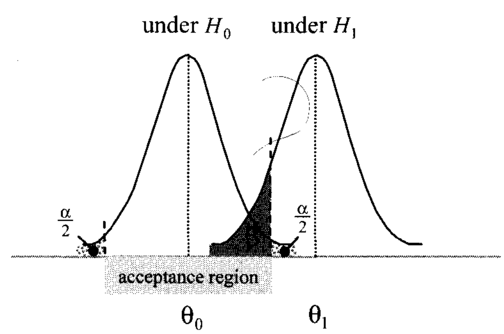
Hypothetical distributions of $\bar{X} \sim N(\mu, \sigma^2 / n)$ for H_0: $\theta = \theta_0$ and H_1: $\theta \neq \theta_0$



1. As the acceptance region widens, α decreases but β increases, provided that the sample size n does not change.



2. As n increases, both α and β decrease, provided that the critical values are held constant.



3. When H_0: $\theta = \theta_0$ is false, β increases as the true value of θ approaches to θ_0 and vice versa.

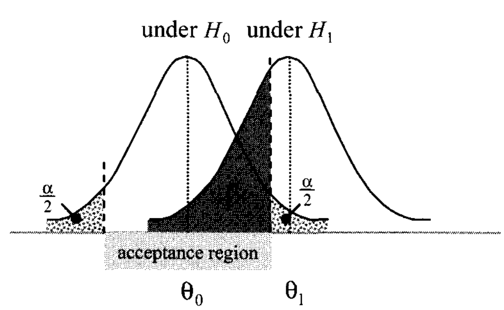


Figure 9-2 Relationships between α and β.

Strong vs. Weak Conclusion

The analyst can directly control both α (probability of rejecting H_0 when H_0 is true) and θ_0 (hypothetical value of θ), but not β (probability of accepting H_0 when H_0 is false; β depends on the true value of θ). Thus, it is considered that "rejection of H_0" is a **strong conclusion** and "acceptance of H_0" is a **weak conclusion**. Consequently, rather than saying we "accept H_0," we prefer saying "fail to reject H_0," which indicates that we have not found sufficient evidence to reject H_0.

Hypothesis Testing Procedure

The formal procedure to make a statistical decision about hypotheses consists of the following four steps throughout the rest of the chapters:

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.
Step 2: Determine a **test statistic and its value.**
Step 3: Determine a **critical value(s) for** $\boldsymbol{\alpha}$.
Step 4: Make a **conclusion**, such as *we reject* or *fail to reject* H_0 *at* α.

(Note) The four-step procedure of hypothesis testing is basically same with the eight-step procedure of hypothesis testing in MR.

9-2 Tests on the Mean of a Normal Distribution, Variance Known

Learning Goals

- ❑ Test a hypothesis on μ when σ^2 is known (z-test).
- ❑ Calculate the P-value of a z-test.
- ❑ Compare the α-value approach with the P-value approach in reporting hypothesis test results.
- ❑ Explain the relationship between confidence interval estimation and hypothesis testing.
- ❑ Determine the sample size of a z-test for statistical inference on μ by applying an appropriate sample size formula and operating characteristic (OC) curve.
- ❑ Explain the effect of the sample size n on the statistical significance and power of the test.
- ❑ Distinguish between statistical significance and practical significance.

Inference Context

Parameter of interest: μ

Point estimator of μ: $\overline{X} \sim N(\mu, \dfrac{\sigma^2}{n})$, σ^2 known

Test statistic of μ: $Z = \dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

Test Procedure (z-test)

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $\mu = \mu_0$
H_1: $\mu \neq \mu_0$ for two-sided test
$\mu > \mu_0$ or $\mu < \mu_0$ for one-sided test

Step 2: Determine a **test statistic and its value.**

$$Z_0 = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$$

***z*-test (*cont.*)**

Step 3: Determine a **critical value(s) for α.**

$z_{\alpha/2}$ for two-sided test; z_α for one-sided test

Step 4: Make a **conclusion**. Reject H_0 if

$|z_0| > z_{\alpha/2}$ for two-sided test

$z_0 > z_\alpha$ for upper-sided test

$z_0 < -z_\alpha$ for lower-sided test

Significance Probability (*P*-value)

The significance probability (denoted as P) of a test statistic refers to the smallest level of significance (α) that would lead to rejection of H_0. Thus, the smaller the P-value, the higher the statistical significance (an inverse relationship between P-value and statistical significance).

For example, the P-value of a test statistic z_0 can be computed by using the following formulas:

$$P = \begin{cases} 2[1-\Phi(|z_0|)] & \text{for two-sided test} \\ 1-\Phi(z_0) & \text{for upper-sided test} \\ \Phi(z_0) & \text{for lower-sided test} \end{cases}$$

(Note) $\Phi(z) = P(Z \le z)$: cumulative distribution function of Z

α-value vs. *P*-value Approach

Two approaches are available to use in reporting the result of a hypothesis test:

1. **α-value approach**: States the test result at the value of α preselected. This reporting method lacks providing the weight of evidence against H_0.
2. ***P*-value approach**: Specifies how far the test statistic is from the critical value(s). Once the P-value is known, the decision maker can draw a conclusion at any specified level of significance (α) as follows:

$$\begin{cases} \text{Reject } H_0 \text{ at } \alpha, & \text{if } P \le \alpha \\ \text{Fail to reject } H_0 \text{ at } \alpha, & \text{otherwise} \end{cases}$$

Note that the P-value approach is more <u>flexible and informative</u> than the α-value approach.

Interval Estimation and Hypothesis Test

A close relationship exists between a <u>two-sided 100(1 - α)% confidence interval (CI) on θ</u> ($l \le \theta \le u$) and the <u>test of a null hypothesis about θ</u> (H_0: $\theta = \theta_0$) at α as follows:

$$\begin{cases} \text{Reject } H_0 \text{ at } \alpha, & \text{if the two- sided CI does not include } \theta_0 \\ \text{Fail to reject } H_0 \text{ at } \alpha, & \text{otherwise} \end{cases}$$

For example, suppose that a 95% CI on μ is $751 \le \mu \le 779$ and we are testing H_0: $\mu = 750$ vs. H_1: $\mu \ne 750$ at $\alpha = 0.05$. Since the CI of μ does not include the hypothesized value $\mu = 750$, we will fail to reject H_0 at $\alpha = 0.05$.

Sample Size Formula

Formulas are used to determine the sample size of a particular test for particular levels of β (or power = 1 - β), α, and δ (= |θ - θ₀|, difference between the true value of θ and its hypothesized value). For a z- test on a single sample, the following formulas are applied:

Sample Size Formula (*cont.*)

$$n = \frac{(z_{\alpha/2} + z_\beta)^2 \sigma^2}{\delta^2}, \text{ where } \delta = \mu - \mu_0 \quad \text{for two-sided test}$$

$$= \frac{(z_\alpha + z_\beta)^2 \sigma^2}{\delta^2} \quad \text{for one-sided test}$$

Note that the sample size requirement increases as α, β, and δ decrease and σ increases.

Operating Characteristic (OC) Curve

Operating characteristic (OC) curves in Appendix A of MR plot β against d (for z- and t-tests) or λ (for χ^2- and F-tests) for various sample sizes n at $\alpha = 0.01$ and 0.05, i.e.,

$\beta = f(n, d \text{ or } \lambda, \alpha)$

By using an appropriate OC chart, an adequate sample size can be determined for a particular test to satisfy a predefined test condition.

Table 9-2 displays a list of OC charts in MR and a formula of the OC parameter d for a z-test on μ. By using the table, the appropriate OC chart for a particular z-test is chosen (e.g., for a one-sided z-test at $\alpha = 0.05$, chart VIc is selected).

Table 9-2 Operating Characteristic Charts for z-test – Single Sample

Test		α	Chart VI (Appendix A in MR)	OC parameter
z-test	Two-sided	0.05	(a)	$d = \frac{\lvert\mu - \mu_0\rvert}{\sigma} = \frac{\lvert\delta\rvert}{\sigma}$
		0.01	(b)	
	One-sided	0.05	(c)	
		0.01	(d)	

Effect of Sample Size

As the sample size n increases, both the statistical significance (inverse to P-value) and power (= 1 - β) of the test increase. For example, Table 9-3 presents P-values and powers of testing on μ for the following conditions:

- $\overline{X} \sim N\left(\mu, \frac{2^2}{n}\right)$ and $\overline{x} = 50.5$
- H_0: $\mu_X = 50$ vs. H_1: $\mu_X \neq 50$
- $\alpha = 0.05$ and the true value of $\mu = 50.5$

The P-value column indicates that, for the same value of $\overline{x} = 50.5$, H_0 is rejected at $\alpha = 0.05$ when $n = 100$ because $P \leq \alpha$, while H_0 is not rejected at $\alpha = 0.05$ when $n \leq 50$ because $P \not\leq \alpha$.

Table 9-3 The P-values and Powers of Testing on μ for Selected Sample Sizes

Sample Size (n)		P-value		Power
	10		0.43	0.124
	25		0.21	0.240
	50		0.08	0.424
	100		0.01	0.705
increasing sample size	400	increasing statistical significance	5.73×10^{-7}	increasing power 0.998
	1000		2.57×10^{-15}	>0.999

Statistical vs. Practical Significance

The statistical significance of a test does not necessarily indicate its practical significance. Since the power of a test increases as the sample size increases, any small departure of θ from the hypothesized value θ_0 will be detected (in other words, H_0: $\theta = \theta_0$ will be rejected) for a large sample, even when the departure is of little practical significance. Therefore, the analyst should check if the statistical test result has also practical significance.

Example 9.2

For the light bulb life length data in Example 8.1, the following results have been obtained:

$n = 30,\ \bar{x} = 780,\ \sigma^2 = 40^2$

95% two-sided CI on μ: $765.7 \le \mu \le 794.3$

1. (**Hypothesis Test on μ, σ^2 Known; Two-Sided Test**) Test H_0: $\mu = 765$ hrs vs. H_1: $\mu \neq 765$ hrs at $\alpha = 0.05$.

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $\mu = 765$
H_1: $\mu \neq 765$

Step 2: Determine a **test statistic and its value**.

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{780 - 765}{40 / \sqrt{30}} = 2.05$$

Step 3: Determine a **critical value(s) for** $\boldsymbol{\alpha}$.

$$z_{\alpha/2} = z_{0.05/2} = z_{0.025} = 1.96$$

Step 4: Make a **conclusion**.

Since $|z_0| = 2.05 > z_{0.025} = 1.96$, reject H_0 at $\alpha = 0.05$.

2. (***P*-value Approach**) Find the P-value for this two-sided z-test.

$P = 2[1 - \Phi(|z_0|)] = 2[1 - \Phi(|2.05|)] = 2[1 - 0.980] = 0.04$

Since $P = 0.04 \le \alpha = 0.05$, reject H_0 at $\alpha = 0.05$.

3. (**Relationship Between CI and Hypothesis Test**) Test H_0: $\mu = 765$ hrs vs. H_1: $\mu \neq 765$ hrs at $\alpha = 0.05$ based on the 95% two-sided CI on μ.

Since the 95% two-sided CI on μ, $765.7 \le \mu \le 794.3$, does not include the hypothesized value 765 hrs, reject H_0 at $\alpha = 0.05$.

4. (**Sample Size Determination**) Determine the sample size n required for this two-sided z-test to detect the true mean as high as 785 hours with power of 0.9. Apply an appropriate sample size formula and OC curve.

(1) Sample Size Formula

power = P(reject H_0 | H_0 is false) = 1 - β = 0.9 $\Rightarrow \beta = 0.1$

$\delta = \mu - \mu_0 = 785 - 765 = 20$

$$n = \frac{(z_{\alpha/2} + z_\beta)^2 \sigma^2}{\delta^2} = \frac{(z_{0.05/2} + z_{0.1})^2 40^2}{20^2} = \frac{(1.96 + 1.28)^2 \times 40^2}{20^2} \cong 42$$

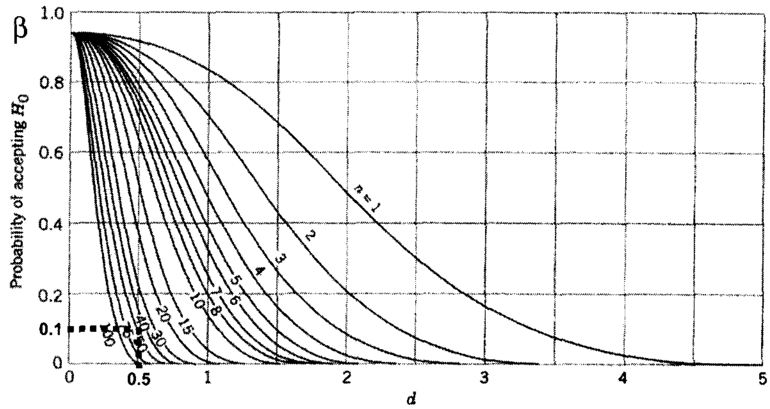
Example 9.2 (*cont.*)

(2) OC Curve

To design a two-sided z-test at $\alpha = 0.05$ for a single sample, OC Chart VI*a* is applicable with the parameter

$$d = \frac{|\mu - \mu_0|}{\sigma} = \frac{|\delta|}{\sigma} = \frac{|20|}{40} = 0.5$$

By using $d = 0.5$ and $\beta = 0.1$, the required sample size is determined as $n = 44$ as displayed below. Note that this sample size is similar with the sample size $n = 42$ determined by using a sample size formula.



(a) O.C. curves for different values of n for the two-sided normal test for a level of significance $\alpha = 0.05$.

Exercise 9.2 (MR 9-28)

For the hole diameter data in Exercise 8.1, the following results have been obtained:

$$n = 10,\ \bar{x} = 1.5045,\ \text{and}\ \sigma^2 = 0.01^2$$

1. Test the hypothesis that the true mean hole diameter is greater than 1.50 in. at $\alpha = 0.01$.
2. Find the P-value for this one-sided z-test.
3. Determine the sample size n required for this one-sided z-test to detect the true mean as high as 1.505 inches with power of 0.9. Apply an appropriate sample size formula and OC curve.

9-3 Tests on the Mean of a Normal Distribution, Variance Unknown

Learning Goals

- ❑ Test a hypothesis on μ when σ^2 is unknown (t-test).
- ❑ Determine the sample size of a t-test for statistical inference on μ by using an appropriate operating characteristic (OC) curve.

Inference Context

Parameter of interest: μ

Point estimator of μ: $\overline{X} \sim N(\mu, \dfrac{\sigma^2}{n})$, σ^2 unknown

Test statistic of μ: $T = \dfrac{\overline{X} - \mu}{S / \sqrt{n}} \sim t(\nu)$, $\nu = n - 1$

Test Procedure (t-test)

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $\mu = \mu_0$

H_1: $\mu \neq \mu_0$ for two-sided test,

$\mu > \mu_0$ or $\mu < \mu_0$ for one-sided test.

Step 2: Determine a **test statistic and its value.**

$$T_0 = \frac{\overline{X} - \mu_0}{S / \sqrt{n}} \sim t(n-1)$$

Step 3: Determine a **critical value(s) for α.**

$t_{\alpha/2,n-1}$ for two-sided test; $t_{\alpha,n-1}$ for one-sided test

Step 4: Make a **conclusion**. Reject H_0 if

$|t_0| > t_{\alpha/2,n-1}$ for two-sided test

$t_0 > t_{\alpha,n-1}$ for upper-sided test

$t_0 < -t_{\alpha,n-1}$ for lower-sided test

Operating Characteristic (OC) Curve

Table 9-4 displays a list of OC charts in MR and a formula of the OC parameter d for a t-test on μ. By using the table, the appropriate OC chart for a particular t-test is selected (e.g., for a two-sided t-test at $\alpha = 0.01$, chart VIf is selected).

Table 9-4 Operating Characteristic Charts for t-test – Single Sample

Test		α	Chart VI (Appendix A in MR)	OC parameter
t-test	Two-sided	0.05	(e)	$d = \dfrac{\lvert\mu - \mu_0\rvert}{\hat{\sigma}} = \dfrac{\lvert\delta\rvert}{\hat{\sigma}}$
		0.01	(f)	
	One-sided	0.05	(g)	
		0.01	(h)	

(Note) For $\hat{\sigma}$, use a sample standard deviation or subjective estimate.

Example 9.3

For the life length data in Example 8.2, the following results have been obtained:

$n = 30$, $\bar{x} = 780$, and $s^2 = 40^2$

1. (**Hypothesis Test on μ, σ^2 Unknown; Two-Sided Test**) Test H_0: $\mu = 765$ hrs vs. H_1: $\mu \neq 765$ hrs at $\alpha = 0.05$.

Example 9.3 (*cont.*)

Step 1: State H_0 and H_1.

H_0: $\mu = 765$
H_1: $\mu \neq 765$

Step 2: Determine a **test statistic and its value.**

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{780 - 765}{40/\sqrt{30}} = 2.05$$

Step 3: Determine a **critical value(s) for α.**

$$t_{\alpha/2,n-1} = t_{0.05/2,30-1} = t_{0.025,29} = 2.045$$

Step 4: Make a **conclusion.**

Since $|t_0| = 2.05 > t_{0.025,29} = 2.045$, reject H_0 at $\alpha = 0.05$.

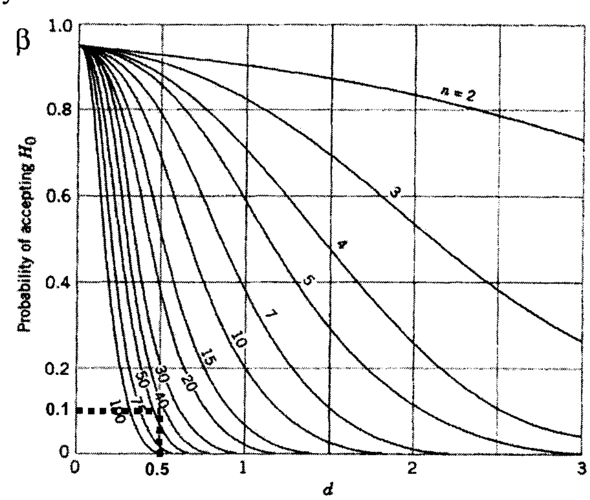
2. (**Sample Size Determination**) Determine the sample size n required for this two-sided t-test to detect the true mean as high as 785 hours with power of 0.9. Apply an appropriate OC curve.

power = P(reject H_0 | H_0 is false) = $1 - \beta = 0.9 \Rightarrow \beta = 0.1$
$\delta = \mu - \mu_0 = 785 - 765 = 20$

To design a two-sided t-test at $\alpha = 0.05$ for a single sample, OC Chart VI*e* is applicable with the parameter

$$d = \frac{|\mu - \mu_0|}{\hat{\sigma}} = \frac{|\delta|}{s} = \frac{|20|}{40} = 0.5$$

By using $d = 0.5$ and $\beta = 0.1$, the required sample size is determined as n = 45 as displayed below.



(*e*) *O.C.* curves for different values of n for the two-sided t-test for a level of significance $\alpha = 0.05$.

Exercise 9.3 (MR 9-35)

For the Izod impact test data of a PVC pipe product in Exercise 8.2, the following results have been obtained:

$n = 20,\ \bar{x} = 1.25$, and $s^2 = 0.25^2$

The ASTM standard requires that Izod impact strength must be greater than 1.0 ft-lb/in.

1. Test if the Izod impact test results satisfy the ASTM standard at $\alpha = 0.05$.
2. Determine the sample size n required for this one-sided t-test to detect the true mean as high as 1.10 with power of 0.8. Apply an appropriate OC curve.

9-4 Hypothesis Tests on the Variance and Standard Deviation of a Normal Population

Learning Goals

- Test a hypothesis on σ^2 (χ^2-test).
- Determine the sample size of a χ^2-test for statistical inference on σ^2 by using an appropriate operating characteristic (OC) curve.

Inference Context

Parameter of interest: σ^2

Point estimator of σ^2: $S^2 = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$, $X_1, X_2, \ldots, X_n \sim i.i.d.\ N(\mu, \sigma^2)$

Point estimator of σ^2: $X^2 = \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(\nu)$, $\nu = n-1$

Test Procedure (χ^2-test)

Step 1: State H_0 and H_1.

H_0: $\sigma^2 = \sigma_0^2$

H_1: $\sigma^2 \neq \sigma_0^2$ for two-sided test,

$\sigma^2 > \sigma_0^2$ for upper-sided test

$\sigma^2 < \sigma_0^2$ for lower-sided test.

Step 2: Determine a **test statistic and its value.**

$$X_0^2 = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1)$$

Step 3: Determine a **critical value(s) for α.**

$\chi^2_{\alpha/2,n-1}$ and $\chi^2_{1-\alpha/2,n-1}$ for two-sided test

$\chi^2_{\alpha,n-1}$ for upper-sided test

$\chi^2_{1-\alpha,n-1}$ for lower-sided test

χ^2-test (*cont.*)

Step 4: Make a **conclusion.** Reject H_0 if

$$\chi_0^2 > \chi^2_{\alpha/2,n-1} \quad \text{or} \quad \chi_0^2 < \chi^2_{1-\alpha/2,n-1} \qquad \text{for two-sided test}$$

$$\chi_0^2 > \chi^2_{\alpha,n-1} \qquad \text{for upper-sided test}$$

$$\chi_0^2 < \chi^2_{1-\alpha,n-1} \qquad \text{for lower-sided test}$$

Operating Characteristic (OC) Curve

Table 9-5 displays a list of OC charts in MR and a formula of the OC parameter λ for a χ^2-test on σ^2. By using the table, the appropriate OC chart for a particular χ^2-test is selected (e.g., for an upper-sided χ^2-test at $\alpha = 0.05$, chart VI*k* is selected).

Table 9-5 Operating Characteristic Charts for χ^2-test

	Test	α	Chart VI (Appendix A in MR)	OC parameter
χ^2-test	Two-sided	0.05	(i)	$\lambda = \dfrac{\sigma}{\sigma_0}$
		0.01	(j)	
	Upper-sided	0.05	(k)	
		0.01	(l)	
	Lower-sided	0.05	(m)	
		0.01	(n)	

Example 9.4

For the light bulb life length data in Example 8.3, the following results have been obtained:

$$n = 16,\ s^2 = 44^2$$

$$95\%\ \text{two-sided CI on } \sigma^2\text{:}\ 32.5^2 \le \sigma^2 \le 68.1^2$$

1. (**Hypothesis Test on σ^2; Two-Sided Test**) Test H_0: $\sigma^2 = 40^2$ vs. H_1: $\sigma^2 \neq 40^2$ at $\alpha = 0.05$.

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

$$H_0\text{:}\quad \sigma^2 = 40^2$$

$$H_1\text{:}\quad \sigma^2 \neq 40^2$$

Step 2: Determine a **test statistic and its value.**

$$\chi_0^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(16-1)\times 44^2}{40^2} = 18.15$$

Step 3: Determine a **critical value(s) for α.**

$$\chi^2_{0.05/2,16-1} = \chi^2_{0.025,15} = 28.65 \text{ and } \chi^2_{1-0.05/2,16-1} = \chi^2_{0.975,15} = 6.91$$

Step 4: Make a **conclusion.**

Since $\chi_0^2 = 18.15 \not> \chi^2_{0.025,15} = 28.65$ and

$\chi_0^2 = 18.15 \not< \chi^2_{0.975,15} = 6.91$, fail to reject H_0 at $\alpha = 0.05$.

2. (**Relationship Between CI and Hypothesis Test**) Test H_0: $\sigma^2 = 40^2$ vs. H_1: $\sigma^2 \neq 40^2$ at $\alpha = 0.05$ based on the 95% two-sided CI on σ^2.

Example 9.4 (*cont.*)

Since the 95% two-sided CI on σ^2, $32.5^2 \le \sigma^2 \le 68.1^2$, includes the hypothesized value 40^2, fail to reject H_0 at $\alpha = 0.05$.

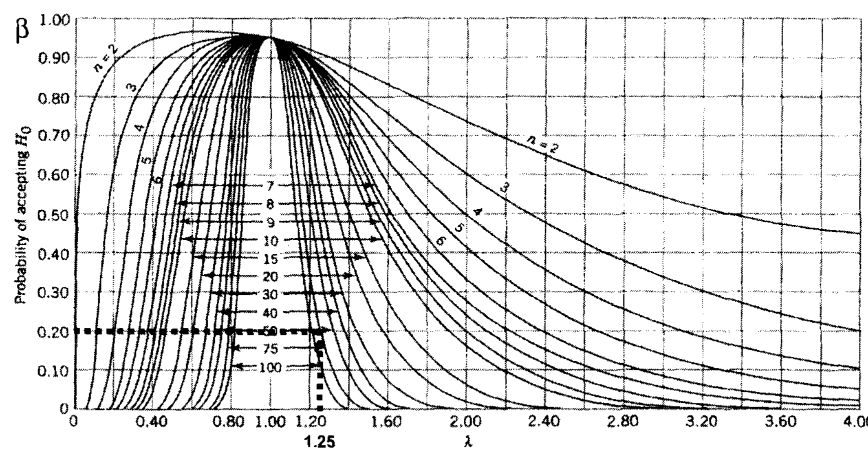
3. (**Sample Size Determination**) Determine the sample size n required for this two-sided χ^2-test to detect the true standard deviation as high as 50 with power of 0.8. Apply an appropriate OC curve.

power = $P(\text{reject } H_0 \mid H_0 \text{ is false}) = 1 - \beta = 0.8 \Rightarrow \beta = 0.2$

To design a two-sided χ^2-test at $\alpha = 0.05$, OC Chart VI*i* is applicable with the parameter

$$\lambda = \frac{\sigma}{\sigma_0} = \frac{50}{40} = 1.25$$

By using $\lambda = 1.25$ and $\beta = 0.2$, the required sample size is determined as n = 75 as displayed below.



(*i*) *O.C.* curves for different values of *n* for the two-sided chi-square test for a level of significance $\alpha = 0.05$.

Exercise 9.4 (MR 9-48)

For the hole diameter data in Exercise 8.3, the following results have been obtained:

$$n = 15 \quad \text{and} \quad s^2 = 0.008^2$$

1. Test if there is strong evidence to indicate that the standard deviation of the hole diameter <u>exceeds</u> 0.01 mm. Use $\alpha = 0.01$.
2. Determine the sample size n required for this upper-sided χ^2-test to detect the true standard deviation mean which exceeds the hypothesized value by 50% with power of 0.9. Apply an appropriate OC curve.

9-5 Tests on a Population Proportion

Learning Goals

- ☐ Test a hypothesis on p (z-test) for a large sample.
- ☐ Determine the sample size for statistical inference on p by using an appropriate sample size formula.

Inference Context

Parameter of interest: p

Point estimator of p: $\hat{P} = \dfrac{X}{n}$, where $X \sim B(n, p)$

Test statistic of p: $Z = \dfrac{\hat{P} - p}{\sqrt{p(1-p)/n}} \sim N(0,1)$ if np and $n(1-p) > 5$

Test Procedure (z-test)

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $p = p_0$

H_1: $p \neq p_0$ for two-sided test,

$p > p_0$ or $p < p_0$ for one-sided test.

Step 2: Determine a **test statistic and its value.**

$$Z_0 = \frac{\hat{P} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$$

Step 3: Determine a **critical value(s) for α.**

$z_{\alpha/2}$ for two-sided test; z_α for one-sided test

Step 4: Make a **conclusion**. Reject H_0 if

$|z_0| > z_{\alpha/2}$ for two-sided test

$z_0 > z_\alpha$ for upper-sided test

$z_0 < -z_\alpha$ for lower-sided test

Sample Size Formula

For a hypothesis test on p, the following formulas are applied to determine the sample size:

$$n = \left(\frac{z_{\alpha/2}\sqrt{p_0(1-p_0)} + z_\beta\sqrt{p(1-p)}}{p - p_0} \right)^2 \quad \text{for two-sided test}$$

$$= \left(\frac{z_{\alpha}\sqrt{p_0(1-p_0)} + z_\beta\sqrt{p(1-p)}}{p - p_0} \right)^2 \quad \text{for one-sided test}$$

Example 9.5

For the corroded bridge data in Example 8.4, the following results have been obtained:

$$n = 40 \quad \text{and} \quad \hat{p} = \frac{x}{n} = \frac{28}{40} = 0.7$$

1. (**Hypothesis Test on p; Two-Sided Test**) Test H_0: $p = 0.5$ vs. H_1: $p \neq 0.5$ at $\alpha = 0.05$.

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $p = 0.5$

H_1: $p \neq 0.5$

Example 9.5 (*cont.*)

Step 2: Determine a **test statistic and its value.**

$$z_0 = \frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} = \frac{0.7 - 0.5}{\sqrt{\dfrac{0.5 \times (1-0.5)}{40}}} = 2.53$$

Step 3: Determine a **critical value(s) for α.**

$$z_{\alpha/2} = z_{0.05/2} = z_{0.025} = 1.96$$

Step 4: Make a **conclusion.**

Since $|z_0| = 2.53 > z_{0.025} = 1.96$, reject H_0 at $\alpha = 0.05$.

2. (**Sample Size Determination**) Determine the sample size n required for this two-sided z-test to detect the true proportion as high as 70% with power of 0.9. Apply an appropriate sample size formula.

power = P(reject H_0 | H_0 is false) = 1 - β = 0.9 ⇒ β = 0.1

$$n = \left(\frac{z_{\alpha/2}\sqrt{p_0(1-p_0)} + z_\beta \sqrt{p(1-p)}}{p - p_0} \right)^2$$

$$= \left(\frac{z_{0.05/2}\sqrt{0.5(1-0.5)} + z_{0.1}\sqrt{0.7(1-0.7)}}{0.7 - 0.5} \right)^2$$

$$= \left(\frac{1.96 \times 0.5 + 1.28 \times 0.46}{0.2} \right)^2 \cong 62$$

Exercise 9.5 (MR 9-53)

For the circuit test data in Exercise 8.4, the following results have been obtained:

$$n = 300 \quad \text{and} \quad \hat{p} = \frac{x}{n} = \frac{13}{300} = 0.043$$

1. Test if the fraction of defective units produced is <u>less</u> than 0.05. Use $\alpha = 0.05$.
2. Determine the sample size n required for this one-sided z-test to detect the true proportion as low as 2% with power of 0.8. Apply an appropriate sample size formula.

9-7 Testing for Goodness of Fit

Learning Goals

- ❑ Explain the term *categorical variable*.
- ❑ Distinguish between nominal and ordinal variables.
- ❑ Explain why the expected frequency of each class interval should be at least three in the goodness-of-fit test.
- ❑ Conduct a goodness-of-fit test on a hypothesized distribution.

Categorical Variable

A categorical variable is used to represent a set of categories. Two types of categorical variables are defined depending on the significance of the order of the category listing:

1. **Nominal variable**: The order of listing of categories is not meaningful.
   (e.g.) gender (male and female)
   hand dominance (left-handed, right-handed, and ambidextrous)
2. **Ordinal variable**: The order of listing of categories is meaningful.
   (e.g.) education (< 9 years, 9 – 12 years, and > 12 years)
   symptom severity (none, mild, moderate, and severe)

Inference Context

The underlying probability distribution of the population is unknown. Thus, we wish to test if a particular distribution fits the population.

(e.g.) H_0: $X \sim$ Poisson distribution with λ (discrete distribution)

H_0: $X \sim N(\mu, \sigma^2)$ (continuous distribution)

Test Statistic

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2(k - p - 1)$$

where: k = Number of class intervals (bins)
O_i = Observed frequency of class interval i
E_i = Expected frequency of class interval i
p = Number of parameters of the hypothesized distribution that are estimated by sample statistics

As the observed frequencies are close to the corresponding expected frequencies, the statistic X^2 becomes small. Thus, X^2 is used to test the null hypothesis H_0: X follows a particular distribution. Table 9-6 can be used to calculate the test statistic X^2.

Table 9-6 Goodness-of-Fit Test Table

Class Intervals i	Observed Frequency O_i	Probability p_i	Expected Frequency $E_i(=np_i)$	$O_i - E_i$	$(O_i - E_i)^2 / E_i$
1					
2					
⋮					
k					

(Caution) Minimum expected frequency

If an expected frequency is too small, X^2 can be improperly large by a small departure of the corresponding observed frequency from the expected frequency. Although there is no consensus on the minimum value of an expected frequency, the value 3, 4, or 5 is widely used as the minimum.

To avoid this undesirable case, when an expected frequency is very small (say, < 3), the corresponding class interval should be combined with an adjacent class interval and the number of class intervals k is reduced by one.

Test Procedure (χ^2-test)

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $X \sim$ A particular distribution.
H_1: $X \nsim$ The hypothesized distribution.

Step 2: Determine a **test statistic and its value.**

1. Estimate the parameter(s) of the hypothesized distribution if their values are not provided.

2. Define class intervals and summarize observed frequencies (O_i's) accordingly.

3. Estimate the probabilities (p_i's) of the class intervals.

4. Calculate the expected frequencies ($E_i = np_i$) of the class intervals. If an expected frequency of a class interval is too small (< 3), combine it to an adjacent class interval. Then, repeat steps 2.2 to 2.4.

5. Calculate the test statistic $X_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$

Step 3: Determine a **critical value for $\boldsymbol{\alpha}$.**

$\chi^2_{\alpha,k-p-1}$

Step 4: Make a **conclusion**. Reject H_0 if

$\chi_0^2 > \chi^2_{\alpha,k-p-1}$

(Caution) Upper-sided critical region
Since the test statistic X_0^2 becomes smaller as the hypothesized distribution fits better, no lower limit is set as a critical value in the goodness-of-fit test.

Example 9.6

(Goodness-of-Fit Test; Discrete Distribution)
The number of e-mails per hour (X) coming to Ms. Young's e-mail account is assumed to follow a Poisson distribution. The following hourly e-mail arrival data are obtained during n = 100 hours:

No. e-mails/hour (X)	Frequency
0	60
1	28
2	7
3	5

Conduct a goodness-of-fit test at α = 0.05.

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $X \sim$ Poisson distribution with λ (Note) $E(X) = V(X) = \lambda$
H_1: $X \nsim$ Poisson distribution

Example 9.6 (*cont.*)

Step 2: Determine a **test statistic and its value.**

1. Estimate the parameter of the hypothesized distribution.

$$\hat{\lambda} = E(X) = \frac{0\times 60 + 1\times 28 + 2\times 7 + 3\times 5}{100} = 0.57$$

The number of parameters estimated is $p = 1$.

2. Define class intervals and summarize observed frequencies (O_i's) accordingly.

# e-mails per hour X	Observed Frequency O_i	Probability $\hat{p}_i$	Expected Frequency $E_i (= n\hat{p}_i)$	$O_i - E_i$	$(O_i - E_i)^2 / E_i$
0	60	0.57	57		
1	28	0.32	32		
2	7	0.09	9		
3 or more	5	0.02	2		

3. Estimate the probabilities (p_i's) of the class intervals.

$$\hat{p}_1 = P(X=0) = \frac{e^{-0.57}(0.57)^0}{0!} = 0.57$$

$$\hat{p}_2 = P(X=1) = \frac{e^{-0.57}(0.57)^1}{1!} = 0.32$$

$$\hat{p}_3 = P(X=2) = \frac{e^{-0.57}(0.57)^2}{2!} = 0.09$$

$$\hat{p}_4 = P(X \ge 3) = 1 - (\hat{p}_1 + \hat{p}_2 + \hat{p}_3) = 0.02$$

4. Calculate the expected frequencies ($E_i = np_i$) of the class intervals. If an expected frequency is too small (< 3), adjust the class intervals.

Since the expected frequency of the last class interval in the above table is less than three, combine the last two cells as follows:

# e-mails per hour	Observed Frequency O_i	Probability $\hat{p}_i$	Expected Frequency $E_i (= n\hat{p}_i)$	$O_i - E_i$	$(O_i - E_i)^2 / E_i$
0	60	0.57	57	-3	0.16
1	25	0.32	32	-7	1.53
2 or more	15	0.11	11	4	1.45

5. Calculate the test statistic: $\chi_0^2 = \sum_{i=1}^{3} \frac{(O_i - E_i)^2}{E_i} = 3.14$

Step 3: Determine a **critical value for α.**

$$\chi^2_{\alpha,k-p-1} = \chi^2_{0.05,3-1-1} = \chi^2_{0.05,1} = 3.84$$

Step 4: Make a **conclusion.**

Since $\chi_0^2 = 3.14 \not> \chi^2_{0.05,1} = 3.84$, fail to reject H_0 at $\alpha = 0.05$. In other words, the number of e-mail arrivals per hour follows a Poisson distribution at $\alpha = 0.05$.

Exercise 9.6 (MR 9-59)

Consider the following frequency table (n = 100) of a random variable X:

X	0	1	2	3	4
Frequency	24	30	31	11	4

Test if a Poisson distribution with $\lambda = 1.2$ is an appropriate model for X at α = 0.05.

Example 9.7

(Goodness-of-Fit Test; Continuous Distribution)
The final scores (X) of n = 40 students in a statistics class are summarized as follows:

Bins in X	Frequency
$x < 60$	3
$60 \le x < 70$	2
$70 \le x < 80$	9
$80 \le x < 90$	12
$90 \le x$	14

The mean and variance of the scores are 82.2 and 13.3^2, respectively. Test if a normal distribution fits the test scores at α = 0.05.

Step 1: State H_0 and H_1.

$$H_0:\ X \sim N(82.2, 13.3^2)$$

$$H_1:\ X \nsim N(82.2, 13.3^2)$$

Step 2: Determine a **test statistic and its value**.

1. Estimate the parameter of the hypothesized distribution.
   Since μ and σ^2 are known, skip this step and the number of parameters estimated is p = 0.

2. Define class intervals and summarize observed frequencies (O_i's) accordingly.

Bins in X	Bins in Z $\frac{x_i - \mu}{\sigma}$	Observed Frequency O_i	Probability p_i	Expected Frequency $E_i (= np_i)$
$x < 60$	$z < -1.67$	3	0.05	2
$60 \le x < 70$	$-1.67 \le z < -0.92$	2	0.13	5
$70 \le x < 80$	$-0.92 \le z < -0.17$	9	0.25	10
$80 \le x < 90$	$-0.17 \le z < 0.59$	12	0.29	12
$90 \le x$	$0.59 \le z$	14	0.28	11

3. Estimate the probabilities (p_i's) of the class intervals.

$$p_1 = P(X < 60) = P(Z < -1.67) = 0.05$$

$$p_2 = P(60 \le X < 70) = P(-1.67 \le Z < -0.92) = 0.13$$

$$p_3 = P(70 \le X < 80) = P(-0.92 \le Z < -0.17) = 0.25$$

$$p_4 = P(80 \le X < 90) = P(-0.17 \le Z < 0.59) = 0.29$$

$$p_5 = P(90 \le X) = P(0.59 \le Z) = 1 - P(Z < 0.59) = 0.28$$

Example 9.7 (*cont.*)

4. Calculate the expected frequencies ($E_i = np_i$) of the class intervals. If an expected frequency is too small (< 3), adjust the class intervals.
   Since the expected frequency of the first class interval in the previous table is less than three, combine the first two class intervals as follows:

Bins in X	Bins in Z $\frac{x_i - \mu}{\sigma}$	Observed Frequency O_i	Probability p_i	Expected Frequency $E_i(=np_i)$
$x < 70$	$z < -0.92$	5	0.18	7
$70 \le x < 80$	$-0.92 \le z < -0.17$	9	0.25	10
$80 \le x < 90$	$-0.17 \le z < 0.59$	12	0.29	12
$90 \le x$	$0.59 \le z$	14	0.28	11

5. Calculate the test statistic: $\chi_0^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = 1.49$

Bins in X	Observed Frequency O_i	Expected Frequency $E_i(=np_i)$	$O_i - E_i$	$(O_i - E_i)^2 / E_i$
$x < 70$	5	7	-2	0.57
$70 \le x < 80$	9	10	-1	0.10
$80 \le x < 90$	12	12	0	0.00
$90 \le x$	14	11	3	0.82

Step 3: Determine a **critical value for** α.

$$\chi^2_{\alpha,k-p-1} = \chi^2_{0.05,4-0-1} = \chi^2_{0.05,3} = 7.81$$

Step 4: Make a **conclusion**.

Since $\chi_0^2 = 1.49 \not> \chi^2_{0.05,3} = 7.81$, fail to reject H_0 at $\alpha = 0.05$.

Exercise 9.7

A random number generator that uniformly produces numbers (X) between zero and one is tested. The distribution of $n = 100$ numbers from the generator is summarized as follows:

Bins in X	Frequency
$x < 0.2$	18
$0.2 \le x < 0.4$	22
$0.4 \le x < 0.6$	23
$0.6 \le x < 0.8$	16
$0.8 \le x$	21

Does this generator work properly? Use $\alpha = 0.05$.

9-8 Contingency Table Tests

Learning Goals

- ❑ Describe a contingency table.
- ❑ Conduct a contingency table test for independence/homogeneity of categorical variables.

$r \times c$ Contingency Table

An $r \times c$ contingency table (see Table 9-7) refers to the table having r rows (for the categories of X) and c columns (for the categories of Y) where each cell ij (category i of X and category j of Y) contains the corresponding observed frequency O_{ij}.

Table 9-7 An $r \times c$ Contingency Table

		Y					
		1	2	…	j	…	c
	1	O_{11}	O_{12}	…	O_{1j}	…	O_{1c}
	2	O_{21}	O_{22}	…	O_{2j}	…	O_{2c}
	⋮	⋮	⋮	…	⋮	…	⋮
X	i	O_{i1}	O_{i2}	…	O_{ij}	…	O_{ic}
	⋮	⋮	⋮	…	⋮	…	⋮
	r	O_{r1}	O_{r2}	…	O_{rj}	…	O_{rc}

Inference Context

We wish to test the association between two categorical variables (X and Y) by using an $r \times c$ contingency table for independence or homogeneity as follows:

1. **Independence**: To examine if X and Y are independent, a representative sample is selected from a single population and then each element in the sample is classified into one of r categories in X and one of c categories in Y.
   (e.g.) Classifying a sample of residents in HAPPY County in terms of gender (X) and occupation (Y).
2. **Homogeneity**: To examine if r populations (X) are homogeneous in terms of Y, representative samples are selected from the r populations and then the elements of each sample are classified into c categories in Y.
   (e.g.) Classifying five samples of residents from different counties (X) in terms of occupation (Y).

Recall that $P(A \mid B) = P(A)$, $P(B \mid A) = P(B)$, and $P(A \cap B) = P(A)P(B)$ for two independent events A and B (see Section 2-6). Likewise, the relationship of two categorical variables is considered independent/homogeneous if

(1) $P(X = x_i \mid Y) = P(X = x_i) = p_{i.}$,

(2) $P(Y = y_j \mid X) = P(Y = y_j) = p_{.j}$, or

(3) $P(X = x_i, Y = y_j) = p_{ij} = P(X = x_i)P(Y = y_j) = p_{i.}p_{.j}$

where: $P(X = x_i \mid Y)$ and $P(Y = y_j \mid X)$ are conditional probabilities,
$P(X = x_i)$ and $P(Y = y_j)$ are marginal probabilities, and
$P(X = x_i, Y = y_j)$ is the joint probability of X and Y.

Test Statistic

$$X^2 = \sum_{j=1}^{c}\sum_{i=1}^{r}\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2(\nu),\ \nu = (r-1)(c-1)$$

where: O_{ij} = observed frequency of cell ij

E_{ij} = expected frequency of cell ij = $np_{ij} = np_{i.}p_{.j} = n\dfrac{n_{i.}}{n}\dfrac{n_{.j}}{n} = \dfrac{n_{i.}n_{.j}}{n}$

Test Statistic (*cont.*)

(Derivation) Determination of degrees of freedom

$\nu = k - p - 1 = rc - [(r - 1) + (c - 1)] - 1 = rc - r - c + 1 = (r - 1)(c - 1)$

since: k = # cells = rc

p = # marginal probabilities that should be estimated by sample statistics to determine E_{ij}'s = $(r - 1) + (c - 1)$

As the observed frequencies are close to the corresponding expected frequencies, the value of X^2 becomes small. Thus, the statistic X^2 is used to test the null hypothesis H_0: X and Y are independent for independence or H_0: X's are homogeneous in terms of Y for homogeneity. Table 9-8 can be used to calculate the test statistic X^2.

Table 9-8 Independence/Homogeneity Test Table

		Y						Totals
		1	2	...	j	...	c	
X	1	O_{11} E_{11}	O_{12} E_{12}	...	O_{1j} E_{1j}	...	O_{1c} E_{1c}	$n_{1\cdot}$
	2	O_{21} E_{21}	O_{22} E_{22}	...	O_{2j} E_{2j}	...	O_{2c} E_{2c}	$n_{2\cdot}$
	$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$	$\vdots$
	i	O_{i1} E_{i1}	O_{i2} E_{i2}	...	O_{ij} E_{ij}	...	O_{ic} E_{ic}	$n_{i\cdot}$
	$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$	$\vdots$
	r	O_{r1} E_{r1}	O_{r2} E_{r2}	...	O_{rj} E_{rj}	...	O_{rc} E_{rc}	$n_{r\cdot}$
Totals		$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot j}$	...	$n_{\cdot c}$	n

(Caution) Minimum expected frequency

Like the minimum expected frequency for the goodness-of-fit test (see Section 9-2), if an expected frequency is too small (say, < 3), X^2 can be improperly large for a small departure of the observed frequency from the expected one. Thus, any category whose expected frequency is small (< 3) should be combined with an adjacent category.

Test Procedure (χ^2-test)

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

(1) **Case 1:** Testing for **Independence**.

H_0: X and Y are independent.
H_1: X and Y are not independent.

(2) **Case 2:** Testing for **Homogeneity**.

H_0: X_i's (r populations) are homogenous in terms of Y.
H_1: X_i's (r populations) are not homogenous in terms of Y.

Step 2: Determine a **test statistic and its value.**

$$X_0^2 = \sum_{j=1}^{c}\sum_{i=1}^{r}\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2((r-1)(c-1)), \quad \text{where } E_{ij} = \frac{n_{i.}n_{.j}}{n}$$

χ^2-test (*cont.*)

Step 3: Determine a **critical value for α.**

$$\chi^2_{\alpha,(r-1)(c-1)}$$

Step 4: Make a **conclusion**. Reject H_0 if

$$\chi^2_0 > \chi^2_{\alpha,(r-1)(c-1)}$$

(Caution) Upper-sided critical region
Since the test statistic X_0^2 becomes small as the null hypothesis of independence/homogeneity is true, no lower limit is set as a critical value in the independence/homogeneity test.

Example 9.8

(Contingency Table Test; Independence)
Grades in ergonomics (*X*) and grades in statistics (*Y*) of n = 100 students are summarized as follows:

		Statistics Grade (*Y*)		
		A	*B*	Others
Ergonomics Grade (*X*)	*A*	12	5	4
	B	10	19	17
	Others	4	8	21

Test if grades in ergonomics (*X*) and grades in statistics (*Y*) are independent at α = 0.05.

Step 1: State H_0 and H_1.

H_0: Grades in ergonomics (*X*) and grades in statistics (*Y*) are independent.

H_1: Grades in ergonomics (*X*) and grades in statistics (*Y*) are <u>not</u> independent.

Step 2: Determine a **test statistic and its value.**

		Statistics Grade (*Y*)			Totals
		A	*B*	Others	
Ergonomics Grade (*X*)	*A*	12 *5.5*	5 *6.7*	4 *8.8*	21
	B	10 *12.0*	19 *14.7*	17 *19.3*	46
	Others	4 *8.6*	8 *10.6*	21 *13.9*	33
Totals		26	32	42	100

$$\chi^2_0 = \sum_{j=1}^{3}\sum_{i=1}^{3}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = 19.5$$

Step 3: Determine a **critical value for α.**

$$\chi^2_{\alpha,(r-1)(c-1)} = \chi^2_{0.05,(3-1)(3-1)} = \chi^2_{0.05,4} = 9.49$$

Step 4: Make a **conclusion.**

Since $\chi^2_0 = 19.5 > \chi^2_{0.05,4} = 9.49$, reject H_0 at α = 0.05. It is concluded that grades in ergonomics (*X*) and grades in statistics (*Y*) are not independent at α = 0.05.

Exercise 9.8 (MR 9-69)

The failures of an electronic component are under study. There are four types of failures (Y) and two mounting positions (X) at the device. A summary of $n = 134$ failures are as follows:

		Failure Type (Y)			
		A	*B*	*C*	*D*
Mounting	Front	22	46	18	9
Position (X)	Back	4	17	6	12

Test if the type of failure (Y) is independent of the mounting position (X) at $\alpha = 0.05$.

Example 9.9

(Contingency Table Test; Homogeneity)

A random sample of $n = 300$ adults with different hand sizes (X) evaluates two mouse designs (Y). The evaluation results are summarized as follows:

		Mouse Designs (Y)	
		Conventional	New
Hand Size Groups (X)	Small	35	65
	Medium	20	80
	Large	30	70

Test if users in different hand-size groups (X) have homogeneous opinions on the mouse designs at $\alpha = 0.05$.

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: Users in different hand-size groups are homogeneous in terms of opinions on the mouse designs.

H_1: Users in different hand-size groups are not homogeneous in terms of opinions on the mouse designs.

Step 2: Determine a **test statistic and its value.**

		Mouse Designs (Y)		Totals
		Conventional	New	
Hand Size Groups (X)	Small	35 / *28.3*	65 / *71.7*	100
	Medium	20 / *28.3*	80 / *71.7*	100
	Large	30 / *28.3*	70 / *71.7*	100
Totals		85	215	300

$$\chi_0^2 = \sum_{j=1}^{2}\sum_{i=1}^{3}\frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 5.7$$

Step 3: Determine a **critical value for α.**

$$\chi^2_{\alpha,(r-1)(c-1)} = \chi^2_{0.05,(3-1)(2-1)} = \chi^2_{0.05,2} = 5.99$$

Step 4: Make a **conclusion.**

Since $\chi_0^2 = 5.7 \not> \chi^2_{0.05,2} = 5.99$, fail to reject H_0 at $\alpha = 0.05$. It is concluded that users in different hand-size groups do not have significantly different opinions on the mouse designs at $\alpha = 0.05$.

Exercise 9.9 (MR 9-70)

A random sample of $n = 630$ students in different class standings (X) is asked their opinions (Y) on a proposed change in core curriculum. The survey results are as follows:

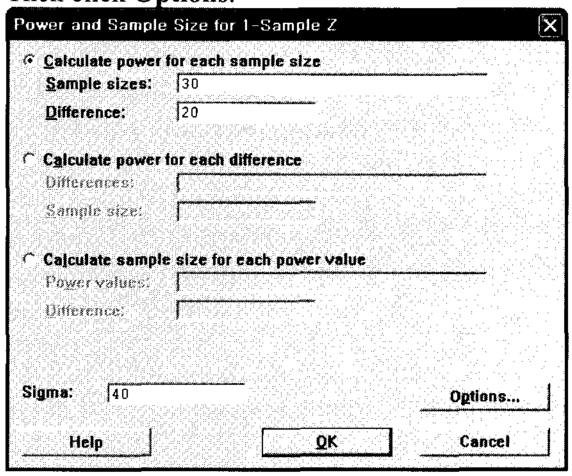
		Opinion (Y)	
		Favoring	Opposing
	Freshmen	120	80
Class	Sophomore	70	130
Standing (X)	Junior	60	70
	Senior	40	60

Test if students in different class standings (X) have homogeneous opinions on the curriculum change at $\alpha = 0.05$.

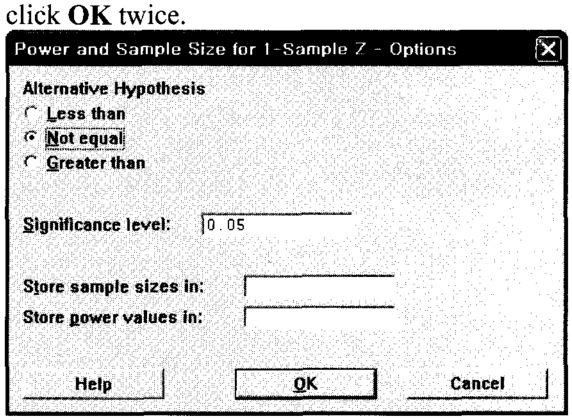
MINITAB Applications

Example 9.1 **(Power of Test)**

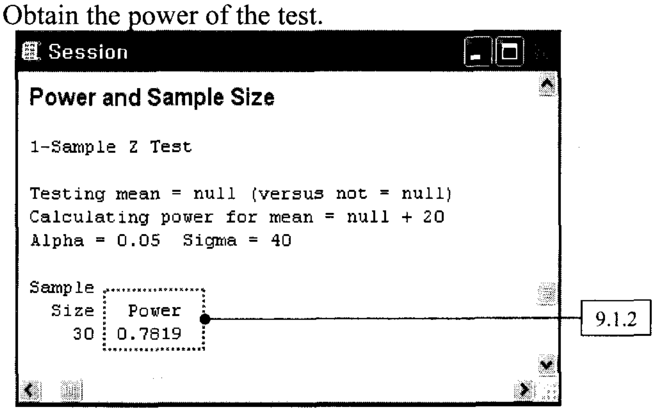
(1) Choose **Stat ➤ Power and Sample Size ➤ 1-Sample Z**. Click **Calculate power for each sample size**. Enter the sample size selected in **Sample sizes,** the difference between the true mean and hypothesized mean to be detected in **Difference**, and the population standard deviation in **Sigma**. Then click **Options**.



(2) Under **Alternative Hypothesis** select the type of alternative hypothesis, and in **Significance level** enter the probability of type I error (α).Then click **OK** twice.



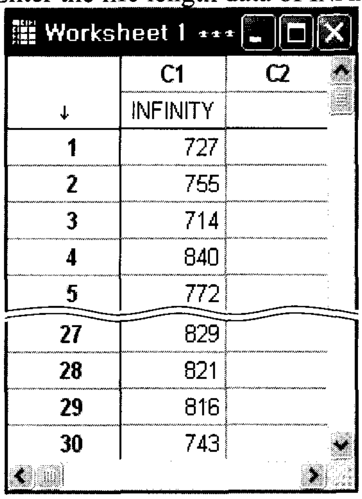
(3) Obtain the power of the test.



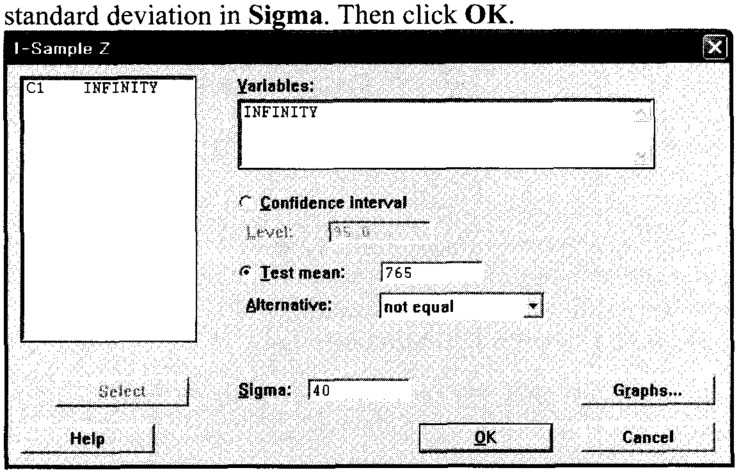
Example 9.2 **(Inference on μ, σ^2 Known)**

(1) Choose **File ➤ New**, click **Minitab Project**, and click **OK**.

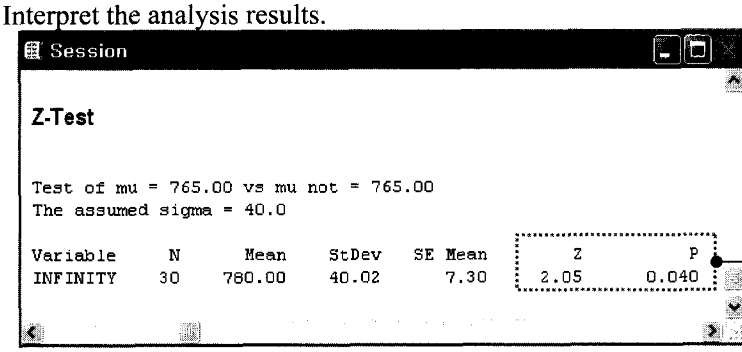
(2) Enter the life length data of INFINITY light bulbs on the worksheet.



(3) Choose **Stat ➤ Basic Statistics ➤ 1-Sample Z**. Click **Test mean**, enter the hypothesized mean life length, and select **not equal** (type of the alternative hypothesis) in **Alternative**. Enter the assumed population standard deviation in **Sigma**. Then click **OK**.



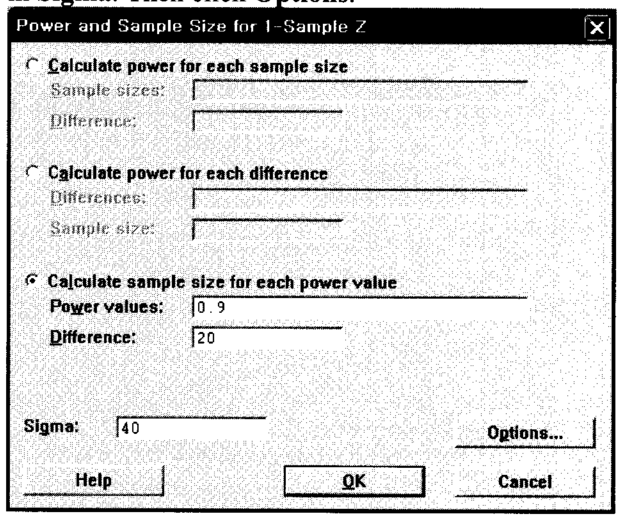
(4) Interpret the analysis results.



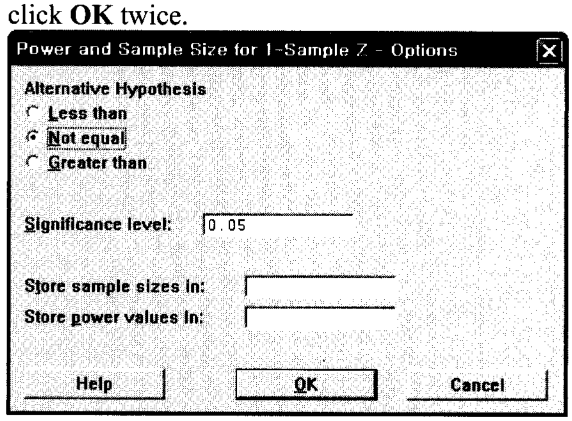
9.2.1,2

Example 9.2 (*cont.*)

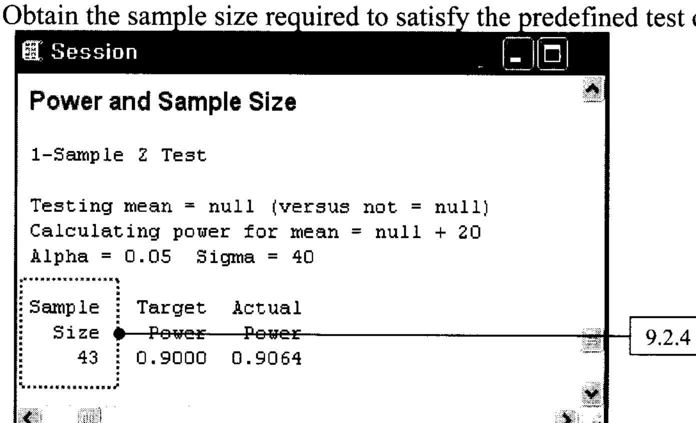
(5) Choose **Stat ➤ Power and Sample Size ➤ 1-Sample Z**. Click **Calculate sample size for each power value**. Enter the power of the test predefined in **Power values,** the difference between the true mean and hypothesized mean to be detected in **Difference**, and the population standard deviation in **Sigma**. Then click **Options**.



(6) Under **Alternative Hypothesis** select the type of alternative hypothesis, and in **Significance level** enter the probability of type I error (α).Then click **OK** twice.



(7) Obtain the sample size required to satisfy the predefined test condition.

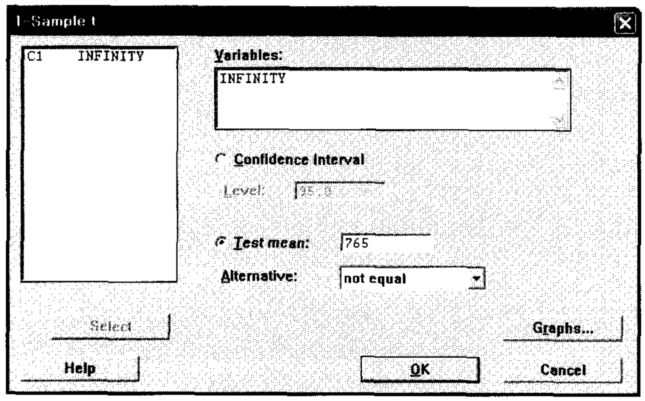


Example 9.3 **(Inference on μ, σ^2 Unknown)**

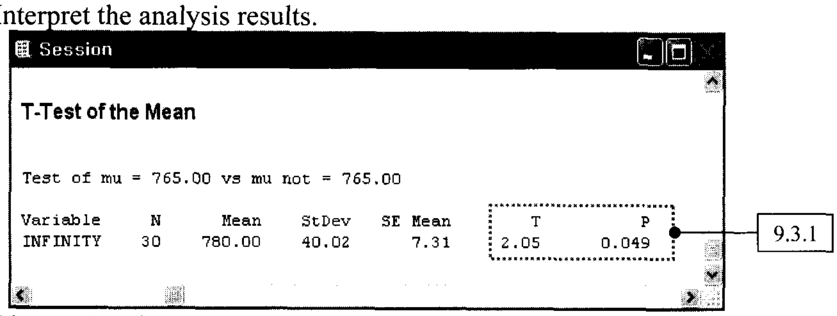
(1) Choose **File ➤ New**, click **Minitab Project**, and click **OK**.

(2) Enter the life length data of INFINITY light bulbs on the worksheet.

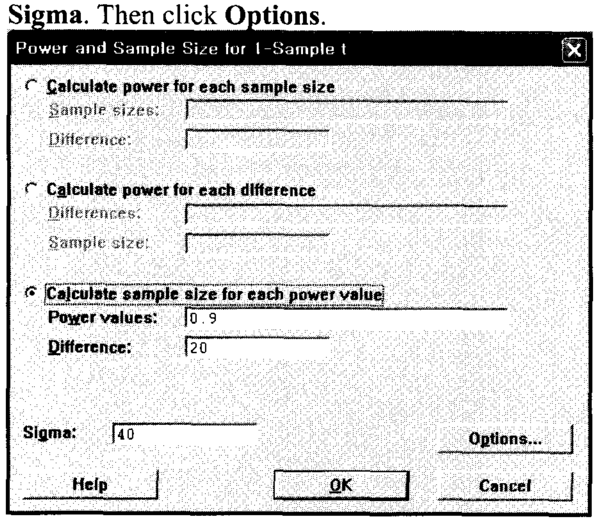
(3) Choose **Stat ➤ Basic Statistics ➤ 1-Sample t**. Click **Test mean**, enter the hypothesized mean life length, and select **not equal** in **Alternative**. Then click **OK**.



(4) Interpret the analysis results.

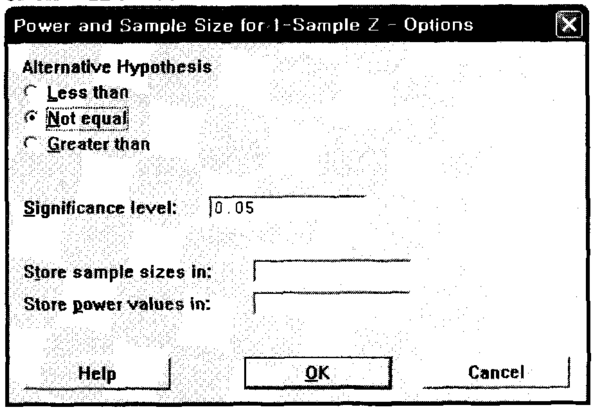


(5) Choose **Stat ➤ Power and Sample Size ➤ 1-Sample t**. Click **Calculate sample size for each power value**. Enter the power of the test predefined in **Power values,** the difference between the true mean and hypothesized mean to be detected in **Difference**, and the sample standard deviation in **Sigma**. Then click **Options**.

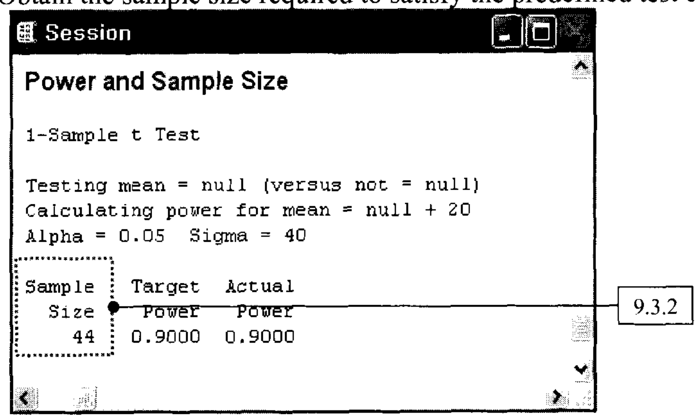


Example 9.3 (*cont.*)

(6) Under **Alternative Hypothesis** select the type of alternative hypothesis, and in **Significance level** enter the probability of type I error (α).Then click **OK** twice.



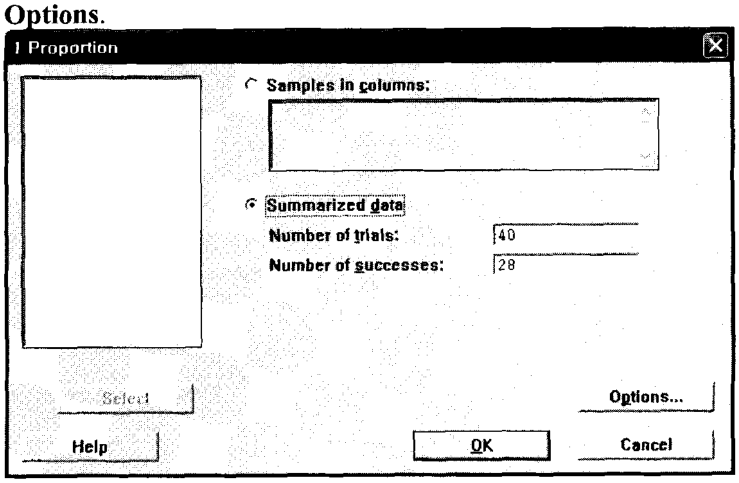
(7) Obtain the sample size required to satisfy the predefined test condition.



Example 9.5

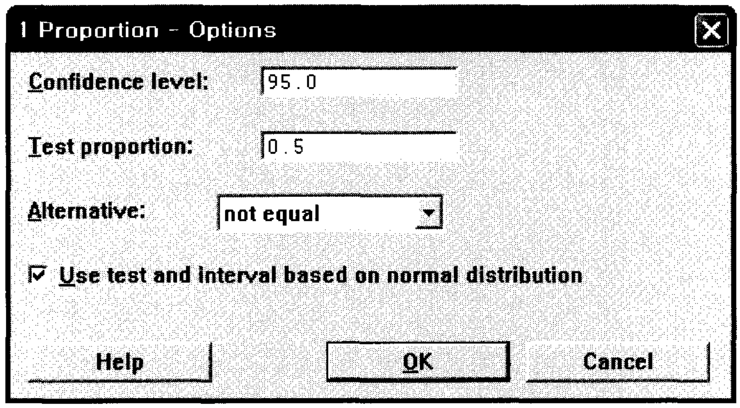
(Inference on p)

(1) Choose **Stat ➤ Basic Statistics ➤ 1 Proportion**. Click **Summarized data** and enter the number of bridges surveyed in **Number of trials** and the number of bridges corroded in **Number of successes**. Then click **Options**.

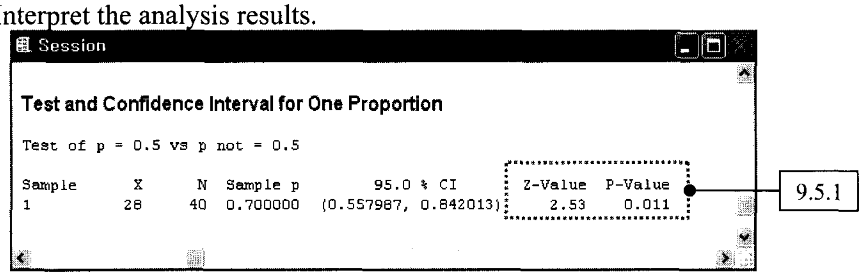


Example 9.5
(cont.)

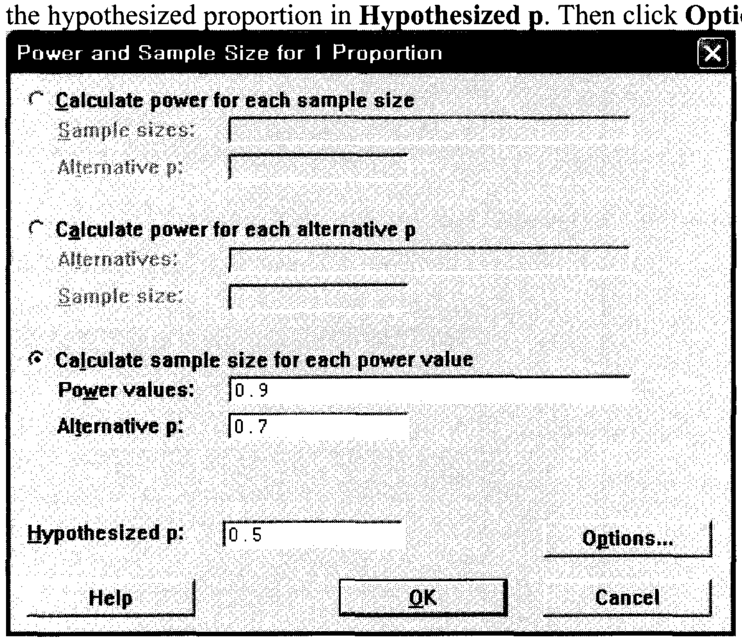
(2) Enter the level of confidence in **Confidence level**, the hypothesized proportion in **Test proportion**, and **not equal** in **Alternative**. Check **Use test and interval based on normal distribution**. Then click **OK** twice.



(3) Interpret the analysis results.

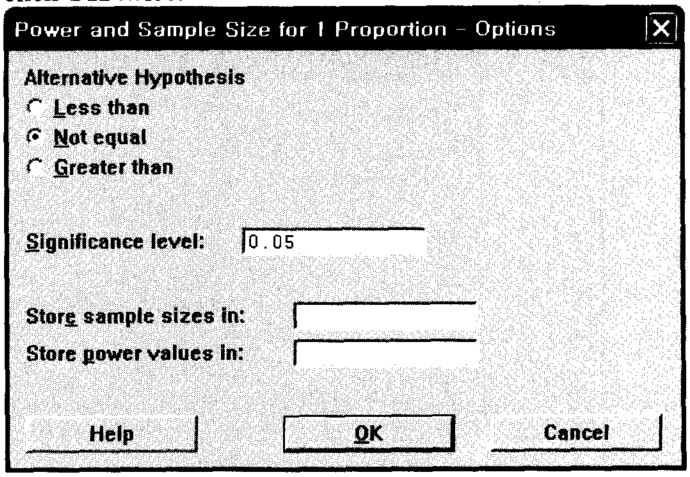


(4) Choose **Stat ➤ Power and Sample Size ➤ 1 Proportion**. Click **Calculate sample size for each power value**. Enter the power of the test predefined in **Power values,** the true proportion to be detected in **Alternative p**, and the hypothesized proportion in **Hypothesized p**. Then click **Options**.

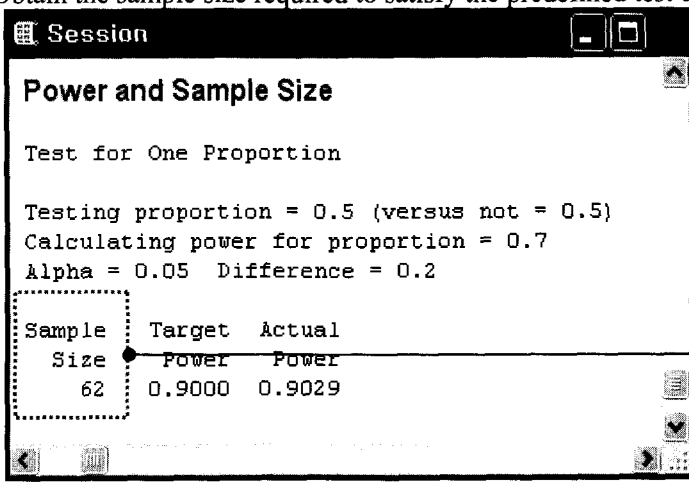


Example 9.5 (*cont.*)

(5) Under **Alternative Hypothesis** select the type of alternative hypothesis, and in **Significance level** enter the probability of type I error (α).Then click **OK** twice.



(6) Obtain the sample size required to satisfy the predefined test condition.



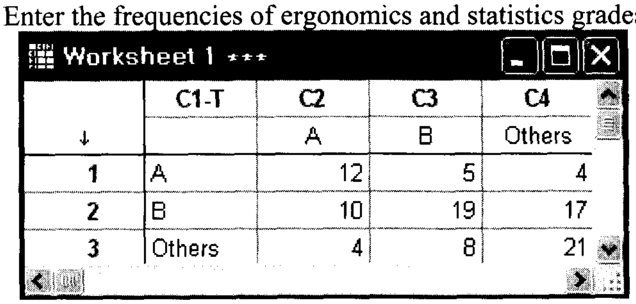
9.5.2

Example 9.8

(Goodness-of-Fit Test; Independence)

(1) Choose **File ➤ New**, click **Minitab Project**, and click **OK**.

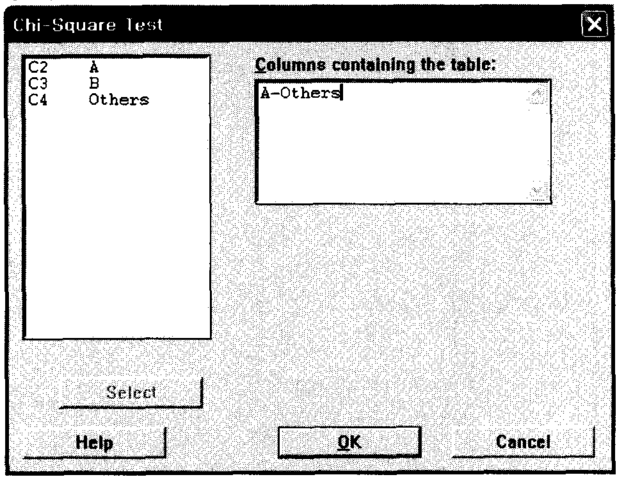
(2) Enter the frequencies of ergonomics and statistics grades on the worksheet.


Worksheet 1 ***

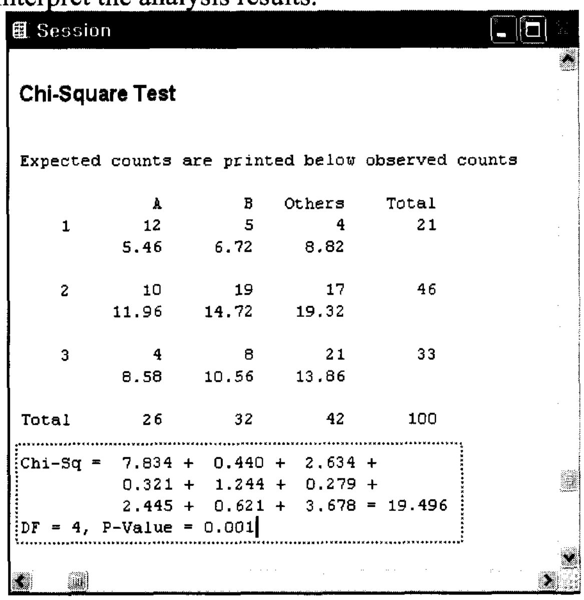
	C1-T	C2	C3	C4
↓		A	B	Others
1	A	12	5	4
2	B	10	19	17
3	Others	4	8	21

Example 9.8 (*cont.*)

(3) Choose **Stat ➤ Tables ➤ Chi-Square Test**. In **Columns containing the table**, select *A*, *B*, and *Others* which contain the frequencies. Then click **OK**.



(4) Interpret the analysis results.



Session

Chi-Square Test

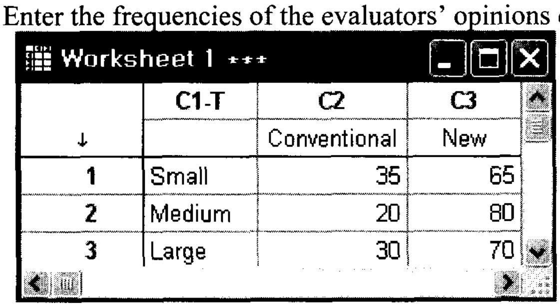
```
Expected counts are printed below observed counts

            A        B   Others    Total
    1      12        5        4       21
         5.46     6.72     8.82

    2      10       19       17       46
        11.96    14.72    19.32

    3       4        8       21       33
         8.58    10.56    13.86

Total      26       32       42      100

Chi-Sq =  7.834 +  0.440 +  2.634 +
          0.321 +  1.244 +  0.279 +
          2.445 +  0.621 +  3.678 = 19.496
DF = 4, P-Value = 0.001
```

Example 9.9

(Goodness-of-Fit Test; Homogeneity)

(1) Choose **File ➤ New**, click **Minitab Project**, and click **OK**.

(2) Enter the frequencies of the evaluators' opinions on the mouse designs.

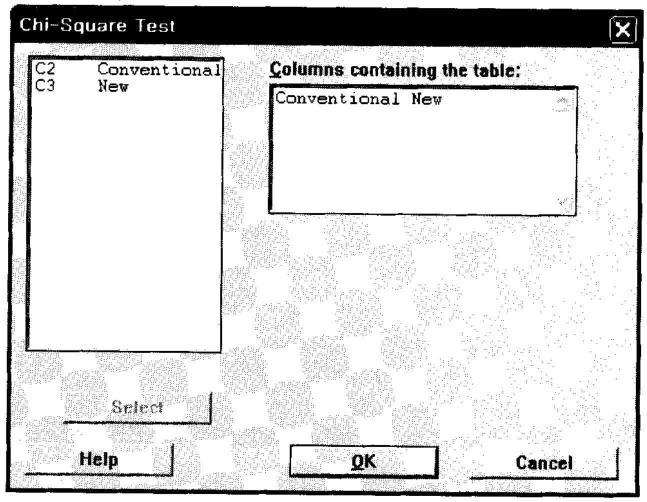


Worksheet 1 ***

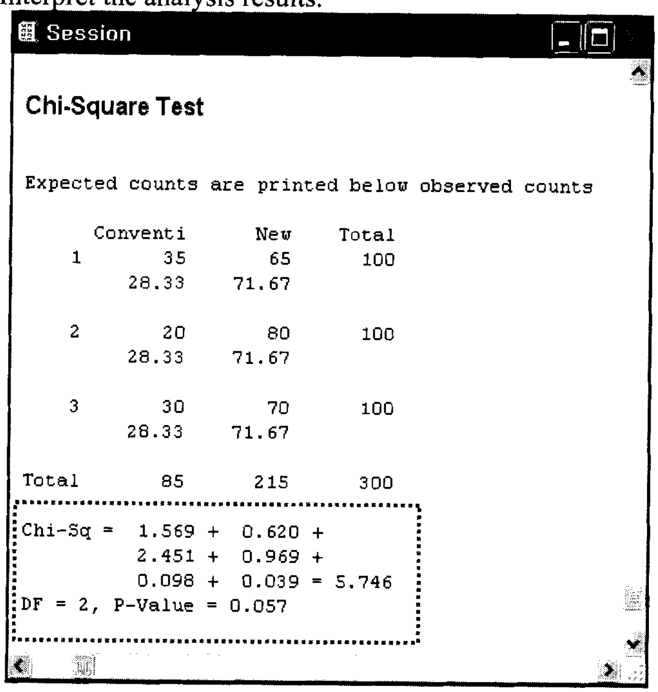
	C1-T	C2	C3
↓		Conventional	New
1	Small	35	65
2	Medium	20	80
3	Large	30	70

Example 9.9
(cont.)

(3) Choose **Stat ➤ Tables ➤ Chi-Square Test**. In **Columns containing the table**, select *Conventional* and *New* which contain the frequencies. Then click **OK**.



(4) Interpret the analysis results.



Answers to Exercises

Exercise 9.1

1. **(Acceptance/Critical Regions)**

The test statistic of μ is the sample mean with the following sampling distribution:

$$\bar{X} \sim N\left(\mu, \frac{2^2}{9}\right)$$

The acceptance region $l \le \bar{X} \le u$ for H_0: $\mu_X = 100$ vs. H_1: $\mu_X \ne 100$ satisfies the following:

$$\begin{aligned}
1 - \alpha = 1 - 0.01 = 0.99 &= P(\text{fail to reject } H_0 \mid \mu = 100) \\
&= P(l \le \bar{X} \le u \mid \mu = 100) \\
&= P\left(\frac{l-\mu}{2/3} \le \frac{\bar{X}-\mu}{2/3} \le \frac{u-\mu}{2/3} \mid \mu = 100\right) \\
&= P\left(\frac{l-100}{2/3} \le Z \le \frac{u-100}{2/3}\right) \\
&= P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) = P(-z_{0.005} \le Z \le z_{0.005}) \\
&= P(-2.58 \le Z \le 2.58)
\end{aligned}$$

Accordingly, the critical values are

$$\frac{l-100}{2/3} = -2.58 \Rightarrow l = 100 - 2.58 \times \frac{2}{3} = 98.3$$

$$\frac{u-100}{2/3} = 2.58 \Rightarrow u = 100 + 2.58 \times \frac{2}{3} = 101.7$$

Thus,

acceptance region: $98.3 \le \bar{x} \le 101.7$

rejection region: $\bar{x} < 98.3$ and $\bar{x} > 101.7$

2. **(β and Power of Test)**

$$\begin{aligned}
\beta &= P(\text{fail to reject } H_0 \mid H_0 \text{ is false}) \\
&= P(98.3 \le \bar{X} \le 101.7 \mid \mu = 103) \\
&= P\left(\frac{98.3-\mu}{2/3} \le \frac{\bar{X}-\mu}{2/3} \le \frac{101.7-\mu}{2/3} \mid \mu = 103\right) \\
&= P\left(\frac{98.3-103}{2/3} \le Z \le \frac{101.7-103}{2/3}\right) \\
&= P(-7.08 \le Z \le -1.92) = P(Z \le -1.92) - P(Z \le -7.08) \\
&= 0.027 - 0 = 0.027
\end{aligned}$$

power = 1 - β = 0.973

Exercise 9.2

1. **(Hypothesis Test on μ, σ^2 Known; Upper-Sided Test)**

   Step 1: State H_0 and H_1.

   H_0: $\mu = 1.50$

   H_1: $\mu > 1.50$

   Step 2: Determine a **test statistic and its value.**

   $$z_0 = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{1.5045 - 1.50}{0.01 / \sqrt{10}} = 1.42$$

   Step 3: Determine a **critical value(s) for α.**

   $$z_\alpha = z_{0.01} = 2.33$$

   Step 4: Make a **conclusion.**

   Since $z_0 = 1.42 \not> z_{0.01} = 2.33$, fail to reject H_0 at $\alpha = 0.01$.

2. **(*P*-value Approach)**

   $$P = 1 - \Phi(z_0) = 1 - \Phi(1.42) = 1 - 0.922 = 0.078$$

   Since $P = 0.078 \not\leq \alpha = 0.01$, fail to reject H_0 at $\alpha = 0.01$.

3. **(Sample Size Determination)**

   (1) Sample Size Formula

   power = P(reject H_0 | H_0 is false) = 1 - β = 0.9 $\Rightarrow$ β = 0.1

   $\delta = \mu - \mu_0 = 1.50 - 1.505 = 0.005$

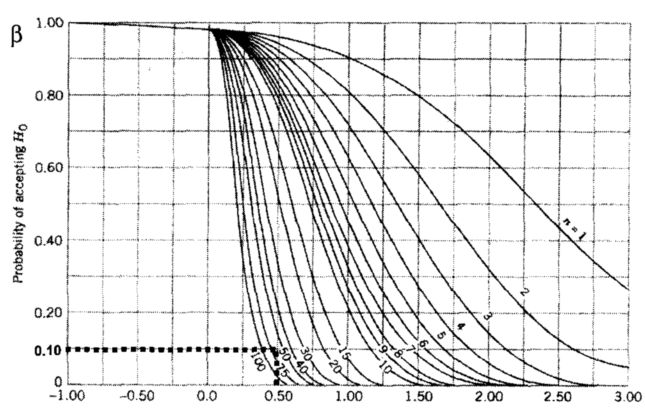
   $$n = \frac{(z_\alpha + z_\beta)^2 \sigma^2}{\delta^2} = \frac{(z_{0.01} + z_{0.1})^2 0.01^2}{0.005^2} = \frac{(2.33 + 1.28)^2 \times 0.01^2}{0.005^2} \cong 53$$

   (2) OC Curve

   To design a one-sided z-test at $\alpha = 0.01$ for a single sample, OC Chart VI*d* is applicable with the parameter

   $$d = \frac{|\mu - \mu_0|}{\sigma} = \frac{|\delta|}{\sigma} = \frac{|0.005|}{0.01} = 0.5$$

   By using $d = 0.5$ and $\beta = 0.1$, the required sample size is determined as n = 55 as displayed below. Note that this sample size is similar with the sample size $n = 53$ determined by using a sample size formula.



(*d*) *O.C.* curves for different values of *n* for the one-sided normal test for a level of significance α = 0.01.

Exercise 9.3

1. **(Hypothesis Test on μ, σ^2 Unknown; Upper-Sided Test)**

   Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

   H_0: $\mu = 1.0$
   H_1: $\mu > 1.0$

   Step 2: Determine a **test statistic and its value.**

   $$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{1.25 - 1.0}{0.25/\sqrt{20}} = 4.47$$

   Step 3: Determine a **critical value(s) for α.**

   $$t_{\alpha,n-1} = t_{0.05,20-1} = t_{0.05,19} = 1.73$$

   Step 4: Make a **conclusion**.

   Since $t_0 = 4.47 > t_{0.05,19} = 1.73$, reject H_0 at $\alpha = 0.05$. We conclude that the PVC pipe product meets the ASTM standard at $\alpha = 0.05$.

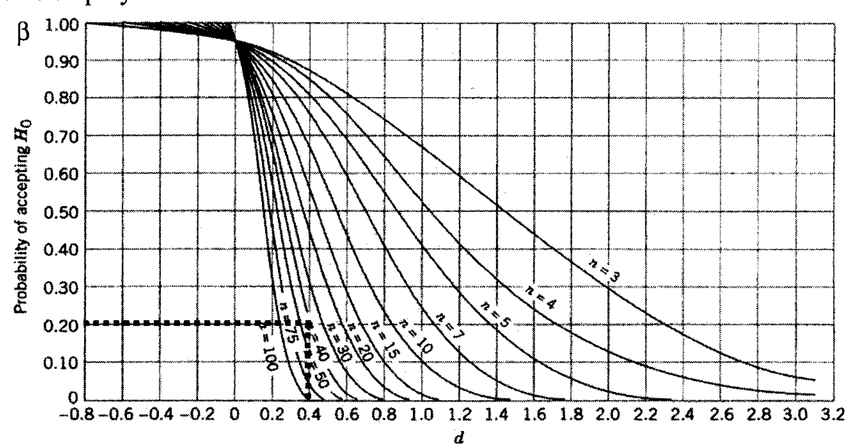
2. **(Sample Size Determination)**

   power = P(reject H_0 | H_0 is false) = 1 - $\beta = 0.8 \Rightarrow \beta = 0.2$
   $\delta = \mu - \mu_0 = 1.10 - 1.0 = 0.10$

   To design a one-sided t-test at $\alpha = 0.05$ for a single sample, OC Chart VIg is applicable with the parameter

   $$d = \frac{|\mu - \mu_0|}{\hat{\sigma}} = \frac{|\delta|}{s} = \frac{|0.10|}{0.25} = 0.4$$

   By using $d = 0.4$ and $\beta = 0.2$, the required sample size is determined as n = 40 as displayed below.



(g) *O.C.* curves for different values of n for the one-sided t-test for a level of significance $\alpha = 0.05$.

Exercise 9.4

1. (**Hypothesis Test on σ^2; Upper-Sided Test**)

Step 1: State H_0 and H_1.

H_0: $\sigma^2 = 0.01^2$

H_1: $\sigma^2 > 0.01^2$

Step 2: Determine a **test statistic and its value.**

$$\chi_0^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(15-1)\times 0.008^2}{0.01^2} = 8.96$$

Step 3: Determine a **critical value(s) for α.**

$$\chi^2_{\alpha,n-1} = \chi^2_{0.01,15-1} = \chi^2_{0.01,14} = 29.14$$

Step 4: Make a **conclusion.**

Since $\chi_0^2 = 8.96 \not> \chi^2_{0.01,14} = 29.14$, fail to reject H_0 at $\alpha = 0.01$.
There is no significant evidence which indicates that the standard deviation of the hole diameter is greater than 0.01 mm at $\alpha = 0.01$.

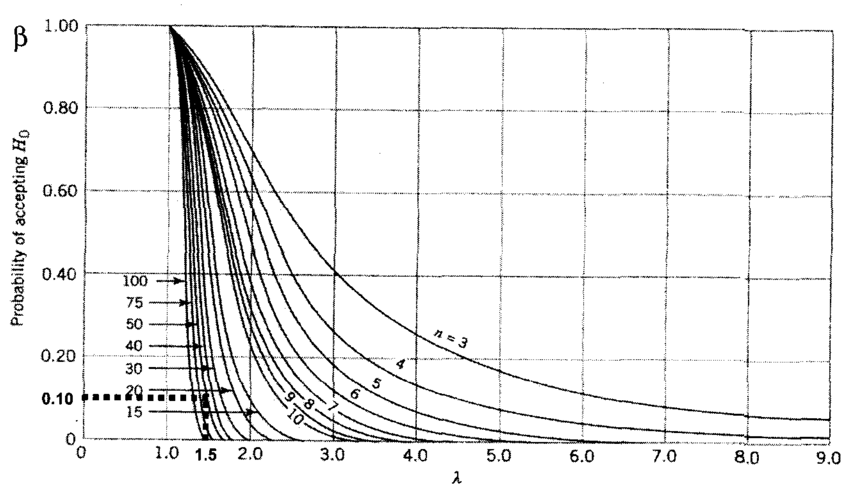
2. (**Sample Size Determination**)

power = P(reject H_0 | H_0 is false) = 1 - β = 0.9 $\Rightarrow \beta = 0.1$

To design an upper-sided χ^2-test at $\alpha = 0.01$, OC Chart VI*l* is applicable with the parameter

$$\lambda = \frac{\sigma}{\sigma_0} = 1.5$$

By using $\lambda = 1.5$ and $\beta = 0.1$, the required sample size is determined as n = 50 as displayed below.



(*l*) *O.C.* curves for different values of *n* for the one-sided (upper tail) chi-square test for a level of significance $\alpha = 0.01$.

Exercise 9.5

1. **(Hypothesis Test on p; Lower-Sided Test)**

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $p = 0.05$
H_1: $p < 0.05$

Step 2: Determine a **test statistic and its value.**

$$z_0 = \frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} = \frac{0.043 - 0.05}{\sqrt{\dfrac{0.05 \times (1-0.05)}{300}}} = -0.53$$

Step 3: Determine a **critical value(s) for $\boldsymbol{\alpha}$.**

$$z_\alpha = z_{0.05} = 1.64$$

Step 4: Make a **conclusion.**

Since $z_0 = -0.53 \nless -z_{0.05} = -1.64$, fail to reject H_0 at $\alpha = 0.05$.

2. **(Sample Size Determination)**

power = P(reject H_0 | H_0 is false) = 1 - β = 0.8 $\Rightarrow$ β = 0.2

$$n = \left(\frac{z_\alpha\sqrt{p_0(1-p_0)} + z_\beta\sqrt{p(1-p)}}{p - p_0}\right)^2$$

$$= \left(\frac{z_{0.05}\sqrt{0.05(1-0.05)} + z_{0.2}\sqrt{0.02(1-0.02)}}{0.02 - 0.05}\right)^2$$

$$= \left(\frac{1.64 \times 0.22 + 0.84 \times 0.14}{-0.03}\right)^2 \cong 253$$

Exercise 9.6

(Goodness-of-Fit Test; Discrete Distribution)

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $X \sim$ Poisson distribution with $\lambda = 1.2$
H_1: $X \nsim$ Poisson distribution

Step 2: Determine a **test statistic and its value.**

1. Estimate the parameter of the hypothesized distribution.
   Since the value of λ is given, skip this step. The number of parameters estimated is $p = 0$.

2. Define class intervals and summarize observed frequencies (O_i's) accordingly.

Exercise 9.6 (*cont.*)

X	Observed Frequency O_i	Probability p_i	Expected Frequency $E_i(=np_i)$	$O_i - E_i$	$(O_i - E_i)^2/E_i$
0	24	0.30	30	-6	1.2
1	30	0.36	36	-6	1.0
2	31	0.22	22	9	3.7
3	11	0.09	9	2	0.4
4 or more	4	0.03	3	1	0.3

3. Estimate the probabilities (p_i's) of the class intervals.

$$p_1 = P(X=0) = \frac{e^{-1.2}(1.2)^0}{0!} = 0.30$$

$$p_2 = P(X=1) = \frac{e^{-1.2}(1.2)^1}{1!} = 0.36$$

$$p_3 = P(X=2) = \frac{e^{-1.2}(1.2)^2}{2!} = 0.22$$

$$p_4 = P(X=3) = \frac{e^{-1.2}(1.2)^3}{3!} = 0.09$$

$$p_5 = P(X \ge 4) = 1-(p_1+p_2+p_3+p_4) = 0.03$$

4. Calculate the expected frequencies ($E_i = np_i$) of the class intervals. If an expected frequency is too small (< 3), adjust the class intervals.
   All the expected frequencies are three or above.

5. Calculate the test statistic: $\chi_0^2 = \sum_{i=1}^{5} \frac{(O_i - E_i)^2}{E_i} = 6.6$

Step 3: Determine a **critical value for α.**

$$\chi^2_{\alpha,k-p-1} = \chi^2_{0.05,5-0-1} = \chi^2_{0.05,4} = 9.49$$

Step 4: Make a **conclusion.**

Since $\chi_0^2 = 6.6 \not> \chi^2_{0.05,4} = 9.49$, fail to reject H_0 at $\alpha = 0.05$.

Exercise 9.7

(Goodness-of-Fit Test; Continuous Distribution)

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $X \sim$ Uniform distribution $f(x) = 1$, $0 \le x \le 1$

H_1: $X \nsim$ Uniform distribution

Step 2: Determine a **test statistic and its value.**

1. Estimate the parameter of the hypothesized distribution.
   Since the uniform distribution does not have any parameter, skip this step. The number of parameters estimated is $p = 0$.

2. Define class intervals and summarize observed frequencies (O_i's) accordingly.

Exercise 9.7 (*cont.*)

Bins in X	Observed Frequency O_i	Probability p_i	Expected Frequency $E_i(=np_i)$	$O_i - E_i$	$(O_i - E_i)^2 / E_i$
$X < 0.2$	18	0.2	20	-2	0.20
$0.2 \le x < 0.4$	22	0.2	20	2	0.20
$0.4 \le x < 0.6$	23	0.2	20	3	0.45
$0.6 \le x < 0.8$	16	0.2	20	-4	0.80
$0.8 \le x$	21	0.2	20	1	0.05

3. Estimate the probabilities (p_i's) of the class intervals.

$$p_1 = P(X < 0.2) = \int_0^{0.2} 1dx = x\Big|_0^{0.2} = 0.2$$

Since the sizes of the bins are the same, the corresponding probabilities are equal to each other as 0.2.

4. Calculate the expected frequencies ($E_i = np_i$) of the class intervals. If an expected frequency is too small (< 3), adjust the class intervals.
   None of the expected frequencies are less than three.

5. Calculate the test statistic: $\chi_0^2 = \sum_{i=1}^{5} \frac{(O_i - E_i)^2}{E_i} = 1.70$

Step 3: Determine a **critical value for α.**

$$\chi^2_{\alpha,k-p-1} = \chi^2_{0.05,5-0-1} = \chi^2_{0.05,4} = 9.49$$

Step 4: Make a **conclusion**.

Since $\chi_0^2 = 1.49 \not> \chi^2_{0.05,3} = 7.81$, fail to reject H_0 at $\alpha = 0.05$. The data does not indicate any abnormal functioning of the random number generator at $\alpha = 0.05$.

Exercise 9.8

(Contingency Table Test; Independence)

Step 1: State H_0 and H_1.

H_0: The type of failure is independent of the mounting position.
H_1: The type of failure is not independent of the mounting position.

Step 2: Determine a **test statistic and its value**:

		Failure Type (Y) A	B	C	D	Totals
Mounting Position (X)	Front	22 *18.4*	46 *44.7*	18 *17.0*	9 *14.9*	95
	Back	4 *7.6*	17 *18.3*	6 *7.0*	12 *6.1*	39
Totals		26	63	24	21	134

$$\chi_0^2 = \sum_{j=1}^{4} \sum_{i=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 10.8$$

Exercise 9.8 (*cont.*)

Step 3: Determine a **critical value for α.**

$$\chi^2_{\alpha,(r-1)(c-1)} = \chi^2_{0.05,(2-1)(4-1)} = \chi^2_{0.05,3} = 7.81$$

Step 4: Make a **conclusion.**

Since $\chi_0^2 = 10.8 > \chi^2_{0.05,3} = 7.81$, reject H_0 at $\alpha = 0.05$.

Exercise 9.9

(Contingency Table Test; Homogeneity)

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: Students in different class standings are homogeneous in terms of opinions on the curriculum change.

H_1: Students in different class standings are not homogeneous in terms of opinions on the curriculum change.

Step 2: Determine a **test statistic and its value.**

		Opinion (Y) Favoring	Opinion (Y) Opposing	Totals
Class Standing (X)	Freshmen	120 *92.1*	80 *107.9*	200
	Sophomore	70 *92.1*	130 *107.9*	200
	Junior	60 *59.8*	70 *70.2*	130
	Senior	40 *46.0*	60 *54.0*	100
Totals		290	340	630

$$\chi_0^2 = \sum_{j=1}^{2}\sum_{i=1}^{4}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = 27.0$$

Step 3: Determine a **critical value for α.**

$$\chi^2_{\alpha,(r-1)(c-1)} = \chi^2_{0.05,(4-1)(2-1)} = \chi^2_{0.05,4} = 9.49$$

Step 4: Make a **conclusion.**

Since $\chi_0^2 = 27.0 > \chi^2_{0.05,4} = 9.49$, reject H_0 at $\alpha = 0.05$. It is concluded that students in different class standings have significantly different opinions on the curriculum change at $\alpha = 0.05$.

10 Statistical Inference for Two Samples

OUTLINE

10-2 Inference for a Difference in Means of Two Normal Distributions, Variances Known

Learning Goals

- ❑ Test a hypothesis on $\mu_1 - \mu_2$ when σ_1^2 and σ_2^2 are known (z-test).
- ❑ Determine the sample size of a z-test for statistical inference on $\mu_1 - \mu_2$ by using an appropriate sample size formula and operating characteristic (OC) curve.
- ❑ Establish a $100(1 - \alpha)\%$ confidence interval (CI) on $\mu_1 - \mu_2$ when σ_1^2 and σ_2^2 are known.
- ❑ Determine the sample size of a z-test to satisfy a preselected level of error (E) in estimating $\mu_1 - \mu_2$.

Inference Context

Parameter of interest: $\mu_1 - \mu_2$

Point estimator of $\mu_1 - \mu_2$: $\overline{X}_1 - \overline{X}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$

(Note) $\overline{X}_1 \sim N(\mu_1, \frac{\sigma_1^2}{n_1})$ and $\overline{X}_2 \sim N(\mu_2, \frac{\sigma_2^2}{n_2})$, σ_1^2 and σ_2^2 known;

$\overline{X}_1$ and $\overline{X}_2$ are independent

Test statistic of $\mu_1 - \mu_2$: $$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Sampling Distribution of $\overline{X}_1 - \overline{X}_2$

The sampling distribution of $\overline{X}_1 - \overline{X}_2$ is

$$\overline{X}_1 - \overline{X}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$$

where: $\overline{X}_1 \sim N(\mu_1, \frac{\sigma_1^2}{n_1})$, $X_{11}, X_{12}, \ldots, X_{1n} \sim i.i.d.\ N(\mu_1, \sigma_1^2)$;

$\overline{X}_2 \sim N(\mu_2, \frac{\sigma_2^2}{n_2})$, $X_{21}, X_{22}, \ldots, X_{2n} \sim i.i.d.\ N(\mu_2, \sigma_2^2)$;

$\overline{X}_1$ and $\overline{X}_2$ are independent.

(Derivation) $\overline{X}_1 - \overline{X}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$

Since $\overline{X}_1$ and $\overline{X}_2$ are independent and normal with means and variances $E(\overline{X}_1) = \mu_1$, $E(\overline{X}_2) = \mu_2$, $V(\overline{X}_1) = \sigma_1^2 / n_1$, and $V(\overline{X}_2) = \sigma_2^2 / n_2$, respectively, $\overline{X}_1 - \overline{X}_2$ is normal with mean and variance

$$E(\overline{X}_1 - \overline{X}_2) = E(\overline{X}_1) - E(\overline{X}_2) = \mu_1 - \mu_2$$

$$V(\overline{X}_1 - \overline{X}_2) = V(\overline{X}_1) + V(\overline{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Test Procedure (z-test)

Step 1: State H_0 and H_1.

H_0: $\mu_1 - \mu_2 = \delta_0$

H_1: $\mu_1 - \mu_2 \neq \delta_0$ for two-sided test,

$\mu_1 - \mu_2 > \delta_0$ or $\mu_1 - \mu_2 < \delta_0$ for one-sided test.

Step 2: Determine a **test statistic and its value.**

$$Z_0 = \frac{(\overline{X}_1 - \overline{X}_2) - \delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Step 3: Determine a **critical value(s) for α.**

$z_{\alpha/2}$ for two-sided test; z_α for one-sided test

Step 4: Make a **conclusion**. Reject H_0 if

$|z_0| > z_{\alpha/2}$ for two-sided test

$z_0 > z_\alpha$ for upper-sided test

$z_0 < -z_\alpha$ for lower-sided test

Sample Size Formula

For a z- test on two samples, the following formulas are applied to obtain a predefined power of the test:

$$n = \frac{(z_{\alpha/2} + z_\beta)^2 (\sigma_1^2 + \sigma_2^2)}{(\Delta - \Delta_0)^2} \quad \text{for two-sided test}$$

Sample Size Formula (*cont.*)

$$n = \frac{(z_\alpha + z_\beta)^2(\sigma_1^2 + \sigma_2^2)}{(\Delta - \Delta_0)^2} \quad \text{for one-sided test}$$

where: $\Delta = \mu_1 - \mu_2$ (true mean difference),

$\Delta \neq \Delta_0$ (hypothesized mean difference), and

$n_1 = n_2 = n$

Operating Characteristic (OC) Curve

Table 10-1 displays a list of OC charts in MR and a formula of the OC parameter d for a z-test on μ_1 - μ_2. By using the table, the appropriate OC chart for a particular z-test is chosen (e.g., for a one-sided z-test at $\alpha = 0.05$, chart VI*c* is selected).

Table 10-1 Operating Characteristic Charts for z-test – Two Samples

Test		α	Chart VI (Appendix A in MR)	OC parameter
z-test	Two-sided	0.05	(a)	$d = \dfrac{\lvert\Delta - \Delta_0\rvert}{\sqrt{\sigma_1^2 + \sigma_2^2}}$
		0.01	(b)	
	One-sided	0.05	(c)	
		0.01	(d)	

Confidence Interval Formula

A 100(1 - α)% CI on $\mu_1 - \mu_2$ when σ_1^2 and σ_2^2 are known is

$$\bar{X}_1 - \bar{X}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \le \bar{X}_1 - \bar{X}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad \text{for two-sided CI}$$

$$\bar{X}_1 - \bar{X}_2 - z_{\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \quad \text{for lower-confidence bound}$$

$$\mu_1 - \mu_2 \le \bar{X}_1 - \bar{X}_2 + z_{\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad \text{for upper-confidence bound}$$

(Derivation) Two-sided confidence interval* on $\mu_1 - \mu_2$, σ_1^2 and σ_2^2 known

By using the test statistic $Z = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim N(0,1)$,

$$P(L \le \mu \le U) = 1 - \alpha$$

$$\Rightarrow P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) = 1 - \alpha$$

$$\Rightarrow P\left(-z_{\alpha/2} \le \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \le z_{\alpha/2}\right) = 1 - \alpha$$

$$\Rightarrow P\left(\bar{X}_1 - \bar{X}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \le \bar{X}_1 - \bar{X}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = 1 - \alpha$$

Therefore,

$$L = \bar{X}_1 - \bar{X}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad \text{and} \quad U = \bar{X}_1 - \bar{X}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Sample Size Formula for Predefined Error

As an extension of the sample size formula for estimation on μ with a preselected level of error (E) (described in Section 8-1), the following formula is used for estimation on $\mu_1 - \mu_2$:

$$n = \left(\frac{z_{\alpha/2}}{E}\right)^2 (\sigma_1^2 + \sigma_2^2), \text{ where } n_1 = n_2 = n$$

Example 10.1

The life lengths of INFINITY (X_1; unit: hour) and FOREVER (X_2; unit: hour) light bulbs are under study. Suppose that X_1 and X_2 are normally distributed with $\sigma_1^2 = 40^2$ and $\sigma_2^2 = 30^2$, respectively. A random sample of INFINITY light bulbs is presented in Example 8.1 and a random sample of FOREVER light bulbs is shown below:

No	Life Length	No	Life Length	No	Life Length
1	789	11	755	21	837
2	835	12	813	22	798
3	765	13	828	23	837
4	796	14	771	24	841
5	797	15	829	25	766
6	776	16	756		
7	769	17	787		
8	836	18	788		
9	847	19	794		
10	769	20	822		

The two random samples are summarized as follows:

Brand of Light Bulb	Sample Size	Sample Mean	Variance
INFINITY (1)	$n_1 = 30$	$\bar{x}_1 = 780$ hrs	$\sigma_1^2 = 40^2$
FOREVER (2)	$n_2 = 25$	$\bar{x}_2 = 800$ hrs	$\sigma_2^2 = 30^2$

1. (**Hypothesis Test on $\mu_1 - \mu_2$, σ_1^2 and σ_2^2 Known; Two-Sided Test**) Test if the mean life length of an INFINITY light bulb is different from that of a FOREVER light bulb at $\alpha = 0.05$.

Step 1: State H_0 and H_1.

H_0: $\mu_1 - \mu_2 = 0$

H_1: $\mu_1 - \mu_2 \neq 0$

Step 2: Determine a **test statistic and its value.**

$$z_0 = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} = \frac{(780 - 800) - 0}{\sqrt{\dfrac{40^2}{30} + \dfrac{30^2}{25}}} = -2.12$$

Step 3: Determine a **critical value(s) for α.**

$$z_{\alpha/2} = z_{0.05/2} = z_{0.025} = 1.96$$

Step 4: Make a **conclusion.**

Since $|z_0| = 2.12 > z_{0.025} = 1.96$, reject H_0 at $\alpha = 0.05$.

Example 10.1 (*cont.*)

2. (**Sample Size Determination for Predefined Power of Test**) Determine the sample size n ($= n_1 = n_2$) required for this two-sided z-test to detect the true difference in mean life length as high as 20 hours with 0.8 of power. Apply an appropriate sample size formula and OC curve.

(1) Sample Size Formula

power $= P(\text{reject } H_0 \mid H_0 \text{ is false}) = 1 - \beta = 0.8 \Rightarrow \beta = 0.2$

$$n = \frac{(z_{\alpha/2} + z_\beta)^2(\sigma_1^2 + \sigma_2^2)}{(\Delta - \Delta_0)^2} = \frac{(z_{0.05/2} + z_{0.2})^2(40^2 + 30^2)}{(20 - 0)^2}$$

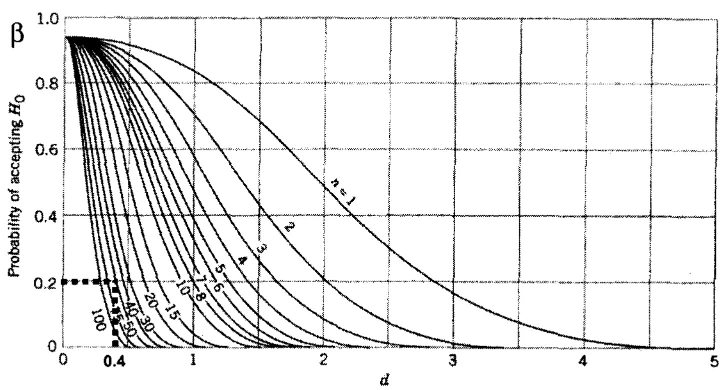
$$= \frac{(1.96 + 0.84)^2 \times (40^2 + 30^2)}{20^2} \cong 50$$

(2) OC Curve

To design a two-sided z-test at $\alpha = 0.05$ for two samples, OC Chart VI*a* is applicable with the parameter

$$d = \frac{|\Delta - \Delta_0|}{\sqrt{\sigma_1^2 + \sigma_2^2}} = \frac{|20 - 0|}{\sqrt{40^2 + 30^2}} = 0.4$$

By using $d = 0.4$ and $\beta = 0.2$, the required sample size is determined as $n = 50$ as displayed below.



(*a*) O.C. curves for different values of n for the two-sided normal test for a level of significance $\alpha = 0.05$.

3. (**Confidence Interval; Two-Sided CI**) Construct a 95% two-sided confidence interval on the mean difference in life length ($\mu_1 - \mu_2$). Based on this 95% two-sided CI on $\mu_1 - \mu_2$, test H_0: $\mu_1 - \mu_2 = 0$ vs. H_1: $\mu_1 - \mu_2 \neq 0$ at $\alpha = 0.05$.

$1 - \alpha = 0.95 \quad \Rightarrow \quad \alpha = 0.05$

95% two-sided CI on $\mu_1 - \mu_2$:

$$\bar{x}_1 - \bar{x}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \le \bar{x}_1 - \bar{x}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$\Rightarrow \quad 780 - 800 - z_{0.05/2}\sqrt{\frac{40^2}{30} + \frac{30^2}{25}} \le \mu_1 - \mu_2 \le 780 - 800 + z_{0.05/2}\sqrt{\frac{40^2}{30} + \frac{30^2}{25}}$$

$$\Rightarrow \quad -20 - 1.96 \times \sqrt{\frac{40^2}{30} + \frac{30^2}{25}} \le \mu_1 - \mu_2 \le -20 + 1.96 \times \sqrt{\frac{40^2}{30} + \frac{30^2}{25}}$$

$$\Rightarrow \quad -38.5 \le \mu_1 - \mu_2 \le -1.5$$

Example 10.1 (*cont.*)

Since this 95% two-sided CI on $\mu_1 - \mu_2$ does not include the hypothesized value zero, reject H_0 at $\alpha = 0.05$.

4. (**Sample Size Determination for Predefined Error**) Find the sample size n (= $n_1 = n_2$) to construct a two-sided confidence interval on $\mu_1 - \mu_2$ within 20 hours of error at $\alpha = 0.05$.

$$n = \left(\frac{z_{\alpha/2}}{E}\right)^2 (\sigma_1^2 + \sigma_2^2) = \left(\frac{z_{0.05/2}}{20}\right)^2 (40^2 + 30^2) = \left(\frac{1.96}{20}\right)^2 \times 50^2 \cong 25$$

Exercise 10.1 (MR 10-2)

Two types of plastic (say, plastic 1 and plastic 2) are used for an electronics component. It is known that the breaking strengths (unit: psi) of plastic 1 (X_1) and plastic 2 (X_2) are normal with $\sigma_1^2 = \sigma_2^2 = 1.0$. From two random samples of n_1 = 10 and n_2 = 12, we obtain $\bar{x}_1 = 162.5$ and $\bar{x}_2 = 155.0$.

1. The company will not adopt plastic 1 unless its breaking strength exceeds that of plastic 2 by at least 10 psi. Based on the sample test results, should they use plastic 1? Use $\alpha = 0.05$ in reaching a decision.
2. Determine the sample size n (= $n_1 = n_2$) required for this one-sided z-test to detect the true difference in mean breaking strength as high as 11.5 psi with 0.9 of power. Apply an appropriate sample size formula and OC curve.
3. Construct a 95% upper-sided confidence interval on the mean difference in breaking strength ($\mu_1 - \mu_2$).

10-3 Inference for a Difference in Means of Two Normal Distributions, Variances Unknown

Learning Goals

- ❑ Test a hypothesis on $\mu_1 - \mu_2$ when σ_1^2 and σ_2^2 are unknown (t-test).
- ❑ Determine the sample size of a z-test for statistical inference on $\mu_1 - \mu_2$ by using an appropriate operating characteristic (OC) curve.
- ❑ Establish a 100(1 - α)% confidence interval (CI) on $\mu_1 - \mu_2$ when σ_1^2 and σ_2^2 are unknown.

Inference Context

Parameter of interest: $\mu_1 - \mu_2$

Point estimator of $\mu_1 - \mu_2$: $\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$

(Note) $\bar{X}_1 \sim N(\mu_1, \frac{\sigma_1^2}{n_1})$ and $\bar{X}_2 \sim N(\mu_2, \frac{\sigma_2^2}{n_2})$, σ_1^2 and σ_2^2 unknown; $\bar{X}_1$ and $\bar{X}_2$ are independent

Inference Context (*cont.*)

Test statistic of $\mu_1 - \mu_2$: Different test statistics of $\mu_1 - \mu_2$ are used depending on the equality of σ_1^2 and σ_2^2 as follows:

(1) **Case 1: Equal variances** ($\sigma_1^2 = \sigma_2^2 = \sigma^2$)

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim t(v), \quad v = n_1 + n_2 - 2$$

where: $S_p^2 = \dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$ (pooled estimator of σ^2)

(2) **Case 2: Unequal variances** ($\sigma_1^2 \neq \sigma_2^2$)

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}} \sim t(v), \quad v = \frac{(S_1^2/n_1 + S_2^2/n_2)^2}{\dfrac{(S_1^2/n_1)^2}{n_1 + 1} + \dfrac{(S_2^2/n_2)^2}{n_2 + 1}} - 2$$

(Note) The equality of two variances ($\sigma_1^2 = \sigma_2^2$) can be checked by using an F test described in Section 10-5.

Test Procedure (t-test)

Step 1: State H_0 and H_1.

H_0: $\mu_1 - \mu_2 = \delta_0$

H_1: $\mu_1 - \mu_2 \neq \delta_0$ for two-sided test

$\mu_1 - \mu_2 > \delta_0$ or $\mu_1 - \mu_2 < \delta_0$ for one-sided test

Step 2: Determine a **test statistic and its value.**

(1) **Case 1: Equal Variances** ($\sigma_1^2 = \sigma_2^2 = \sigma^2$)

$$T_0 = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim t(v), \quad v = n_1 + n_2 - 2$$

where: $S_p^2 = \dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$ (estimator of σ^2)

(2) **Case 2: Unequal Variances** ($\sigma_1^2 \neq \sigma_2^2$)

$$T_0 = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}} \sim t(v), \quad v = \frac{(S_1^2/n_1 + S_2^2/n_2)^2}{\dfrac{(S_1^2/n_1)^2}{n_1 + 1} + \dfrac{(S_2^2/n_2)^2}{n_2 + 1}} - 2$$

Step 3: Determine a **critical value(s) for α.**

$t_{\alpha/2,v}$ for two-sided test; $t_{\alpha,v}$ for one-sided test

Step 4: Make a **conclusion**. Reject H_0 if

$|t_0| > t_{\alpha/2,v}$ for two-sided test

$t_0 > t_{\alpha,v}$ for upper-sided test

$t_0 < -t_{\alpha,v}$ for lower-sided test

Operating Characteristic (OC) Curve

Table 10-2 displays a list of OC charts in MR and a formula of the OC parameter d for a t-test on $\mu_1 - \mu_2$ where $\sigma_1^2 = \sigma_2^2 = \sigma^2$ and $n_1 = n_2 = n$. Note that OC curves are unavailable for a t-test when $\sigma_1^2 \neq \sigma_2^2$ because the corresponding t distribution is unknown. By using the table, the appropriate OC chart for a particular t-test is chosen (e.g., for a two-sided t-test at $\alpha = 0.01$, chart VI*f* is selected). The sample size n^* obtained from an OC curve is used to determine the sample size n ($= n_1 = n_2$) as follows:

$$n = \frac{n^* + 1}{2}, \quad \text{where } n^* \text{ from an OC curve}$$

Table 10-2 Operating Characteristic Charts for t-test – Two Samples

Test		α	Chart VI (Appendix A in MR)	OC parameter
t-test	Two-sided	0.05	(e)	$d = \dfrac{\lvert\Delta - \Delta_0\rvert}{2\hat{\sigma}}$
		0.01	(f)	
	One-sided	0.05	(g)	
		0.01	(h)	

(Note) For $\hat{\sigma}$, use s_p (pooled estimate of common standard deviation) or a subjective estimate.

Confidence Interval Formula

A 100(1 - α)% CI on $\mu_1 - \mu_2$ when σ_1^2 and σ_2^2 are unknown depends on the equality of σ_1^2 and σ_2^2 as follows:

(1) **Case 1: Equal variances** ($\sigma_1^2 = \sigma_2^2 = \sigma^2$)

$$\bar{X}_1 - \bar{X}_2 - t_{\alpha/2,v} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \le \mu_1 - \mu_2 \le \bar{X}_1 - \bar{X}_2 + t_{\alpha/2,v} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad \text{for two-sided CI}$$

$$\bar{X}_1 - \bar{X}_2 - t_{\alpha/2,v} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \le \mu_1 - \mu_2 \quad \text{for lower-confidence bound}$$

$$\mu_1 - \mu_2 \le \bar{X}_1 - \bar{X}_2 + t_{\alpha/2,v} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad \text{for upper-confidence bound}$$

(2) **Case 2: Unequal variances** ($\sigma_1^2 \neq \sigma_2^2$)

$$\bar{X}_1 - \bar{X}_2 - t_{\alpha/2,v} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \le \mu_1 - \mu_2 \le \bar{X}_1 - \bar{X}_2 + t_{\alpha/2,v} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \quad \text{for two-sided CI}$$

$$\bar{X}_1 - \bar{X}_2 - t_{\alpha/2,v} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \le \mu_1 - \mu_2 \quad \text{for lower-confidence bound}$$

$$\mu_1 - \mu_2 \le \bar{X}_1 - \bar{X}_2 + t_{\alpha/2,v} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \quad \text{for upper-confidence bound}$$

CI and Hypothesis Test for Large Sample

If the sample sizes are large (n_1 and $n_2 \ge 30$), the z-based CI formulas and test procedure in Section 10-2 can be applied to inference on $\mu_1 - \mu_2$ regardless of whether the underlying populations are normal or non-normal according to the central limit theorem (described in Section 7-5).

Example 10.2

For the light bulb life length data in Example 10.1, the following results have been obtained:

Brand of Light Bulb	Sample Size	Sample Mean	Sample Variance
INFINITY (1)	$n_1 = 30$	$\bar{x}_1 = 780$ hrs	$s_1^2 = 40^2$
FOREVER (2)	$n_2 = 25$	$\bar{x}_2 = 800$ hrs	$s_2^2 = 30^2$

1. (**Hypothesis Test on $\mu_1 - \mu_2$, σ_1^2 and σ_2^2 Unknown and Unequal; Two-Sided Test**) Assuming $\sigma_1^2 \neq \sigma_2^2$, test if the mean life length of an INFINITY light bulb is different from that of a FOREVER light bulb at $\alpha = 0.05$.

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $\mu_1 - \mu_2 = 0$

H_1: $\mu_1 - \mu_2 \neq 0$

Step 2: Determine a **test statistic and its value.**

$$t_0 = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} = \frac{(780 - 800) - 0}{\sqrt{\dfrac{40^2}{30} + \dfrac{30^2}{25}}} = -2.12$$

$$\nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\dfrac{(s_1^2/n_1)^2}{n_1+1} + \dfrac{(s_2^2/n_2)^2}{n_2+1}} - 2 = \frac{(40^2/30 + 30^2/25)^2}{\dfrac{(40^2/30)^2}{30+1} + \dfrac{(30^2/25)^2}{25+1}} - 2 \cong 54$$

Step 3: Determine a **critical value(s) for α.**

$$t_{\alpha/2,\nu} = t_{0.025,54} = 2.00$$

Step 4: Make a **conclusion.**

Since $|t_0| = 2.12 > t_{0.025,54} = 2.00$, reject H_0 at $\alpha = 0.05$.

2. (**Sample Size Determination**) Assuming $\sigma_1^2 = \sigma_2^2$, determine the sample size n ($= n_1 = n_2$) required for this two-sided t-test to detect the true difference in mean life length as high as 20 hours with 0.8 of power. Apply an appropriate OC curve.

To design a two-sided t-test at $\alpha = 0.05$ for two samples, OC Chart VI*e* is applicable with the parameter

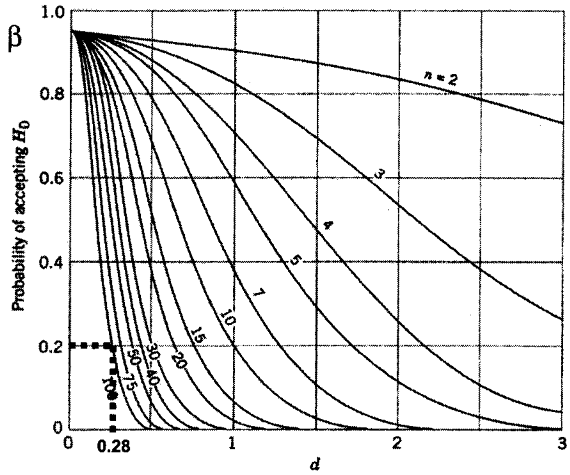
$$d = \frac{|\Delta - \Delta_0|}{2\hat{\sigma}} = \frac{|\Delta - \Delta_0|}{2s_p} = \frac{|20 - 0|}{2 \times 35.8} = 0.28$$

(Note) $$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$= \frac{(30-1) \times 40^2 + (25-1) \times 30^2}{30 + 25 - 2} = 1283.0 = 35.8^2$$

Example 10.2 (*cont.*)

For $d = 0.28$ and $\beta = 0.2$, $n^* = 100$ as displayed below.



(*e*) *O.C.* curves for different values of *n* for the two-sided *t*-test for a level of significance $\alpha = 0.05$.

Thus, the required sample size n is

$$n = \frac{n^*+1}{2} = \frac{100+1}{2} = 50.5 \cong 51$$

3. **(Confidence Interval on $\mu_1 - \mu_2$, σ_1^2 and σ_2^2 Unknown but Equal; Two-Sided CI)** Assuming $\sigma_1^2 = \sigma_2^2$, construct a 95% two-sided confidence interval on the difference in mean life length ($\mu_1 - \mu_2$).

$1-\alpha = 0.95 \Rightarrow \alpha = 0.05;\ \ \nu = n_1 + n_2 - 2 = 30 + 25 - 2 = 53;\ \ s_p^2 = 35.8^2$

95% two-sided CI on $\mu_1 - \mu_2$:

$$\bar{x}_1 - \bar{x}_2 - t_{\alpha/2,\nu} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \le \mu_1 - \mu_2 \le \bar{x}_1 - \bar{x}_2 + t_{\alpha/2,\nu} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$\Rightarrow 780 - 800 - t_{0.05/2,53} \times 35.8\sqrt{\frac{1}{30} + \frac{1}{25}} \le \mu_1 - \mu_2 \le$$

$$780 - 800 + t_{0.05/2,53} \times 35.8\sqrt{\frac{1}{30} + \frac{1}{25}}$$

$$\Rightarrow -20 - 2.01 \times 9.69 \le \mu_1 - \mu_2 \le -20 + 2.01 \times 9.69$$

$$\Rightarrow -39.5 \le \mu_1 - \mu_2 \le -0.6$$

(Note) This t-based CI is <u>wider</u> than the corresponding z-based CI $-38.5 \le \mu_1 - \mu_2 \le -1.5$ in Example 10-1.

4. **(Confidence Interval on $\mu_1 - \mu_2$, σ_1^2 and σ_2^2 Unknown and Unequal; Two-Sided CI)** Assuming $\sigma_1^2 \ne \sigma_2^2$, construct a 95% two-sided confidence interval on the difference in mean life length ($\mu_1 - \mu_2$). Based on this 95% two-sided CI on μ_1 - μ_2, test H_0: $\mu_1 - \mu_2 = 0$ vs. H_1: $\mu_1 - \mu_2 \ne 0$ at $\alpha = 0.05$

Example 10.2 (*cont.*)

95% two-sided CI on $\mu_1 - \mu_2$:

$$\bar{x}_1 - \bar{x}_2 - t_{\alpha/2,v}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} \le \mu_1 - \mu_2 \le \bar{x}_1 - \bar{x}_2 + t_{\alpha/2,v}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

$$\Rightarrow\ 780-800-t_{0.05/2,54}\sqrt{\frac{40^2}{30}+\frac{30^2}{25}} \le \mu_1 - \mu_2 \le 780-800+t_{0.05/2,54}\sqrt{\frac{40^2}{30}+\frac{30^2}{25}}$$

$$\Rightarrow\ -20-2.00\times 9.45 \le \mu_1 - \mu_2 \le -20+2.00\times 9.45$$

$$\Rightarrow\ -39.0 \le \mu_1 - \mu_2 \le -1.1$$

Since this 95% two-sided CI on μ_1 - μ_2 when σ_1^2 and σ_2^2 are unknown and unequal does not include the hypothesized value zero, reject H_0 at $\alpha = 0.05$.

Exercise 10.2 (MR 10-22)

Two suppliers manufacture a plastic gear used in a laser printer. The impact strength of these gears measured in foot-pounds is an important characteristic. A random sample of $n_1 = 10$ gears from supplier 1 results in $\bar{x}_1 = 290$ and $s_1 = 12$, while another random sample of $n_2 = 16$ gears from supplier 2 results in $\bar{x}_2 = 321$ and $s_2 = 22$. Assume that X_1 and X_2 are normally distributed.

1. Assuming $\sigma_1^2 = \sigma_2^2$, test if supplier 2 provides gears with <u>higher</u> mean strength at $\alpha = 0.05$.
2. Assuming $\sigma_1^2 = \sigma_2^2$, determine the sample size n ($= n_1 = n_2$) required for this one-sided t-test to detect the true mean difference as high as 25 foot-pounds with 0.9 of power. Apply an appropriate OC curve.
3. Assuming $\sigma_1^2 = \sigma_2^2$, construct a 95% upper-confidence bound on $\mu_1 - \mu_2$.
4. Assuming $\sigma_1^2 \neq \sigma_2^2$, construct a 95% upper-confidence bound on $\mu_1 - \mu_2$.

10-4 Paired *t*-Test

Learning Goals

- ❑ Explain a paired experiment and its purpose.
- ❑ Test a hypothesis on μ_D for paired observations when σ_D^2 is unknown (paired t-test).
- ❑ Establish a 100(1 - α)% confidence interval (CI) on μ_D for paired observations when σ_D^2 is unknown.

Paired Experiment

A paired experiment collects a pair of observations (X_1 and X_2) for each specimen (experimental unit) and analyzes their differences (instead of the original data). This paired experiment is used when **heterogeneity** exists between specimens and this heterogeneity can significantly affect X_1 and X_2; in other words, X_1 and X_2 are <u>not</u> independent.

For instance, in Table 10-3, the effect of a diet program on the change in weight is under study with n participants, who are heterogeneous. This heterogeneity is likely to confound the effect of the diet program on the weight change; in other words, the weight before (X_1) and the weight after (X_2) the diet program are not independent. To block the effect of the heterogeneity on the weight change, it is proper to analyze the differences (D) of the paired observations.

Paired Experiment (*cont.*)

Table 10-3 A Diet Program Study

Participant (Specimen)	Weight Before (X_1)	Weight After (X_2)	$D = X_1 - X_2$
1			
2			
⋮			
n			

Inference Context

Parameter of interest: μ_D

Point estimator of μ_D: $\overline{D} = \overline{X_1 - X_2} \sim N(\mu_D, \frac{\sigma_D^2}{n})$, σ_D^2 unknown;

X_1 and X_2 are not independent.

Test statistic of μ_D: $T = \dfrac{\overline{D} - \mu_D}{S_D / \sqrt{n}} \sim t(\nu)$, $\nu = n - 1$

Test Procedure (paired *t*-test)

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $\mu_D = \delta_0$

H_1: $\mu_D \neq \delta_0$ for two-sided test,

$\mu_D > \delta_0$ or $\mu_D < \delta_0$ for one-sided test.

Step 2: Determine a **test statistic and its value**.

$$T_0 = \frac{\overline{D} - \delta_0}{S_D / \sqrt{n}} \sim t(n-1)$$

Step 3: Determine a **critical value(s) for α**.

$t_{\alpha/2,n-1}$ for two-sided test; $t_{\alpha,n-1}$ for one-sided test

Step 4: Make a **conclusion**. Reject H_0 if

$|t_0| > t_{\alpha/2,n-1}$ for two-sided test

$t_0 > t_{\alpha,n-1}$ for upper-sided test

$t_0 < -t_{\alpha,n-1}$ for lower-sided test

Confidence Interval Formula

$$\overline{D} - t_{\alpha/2,\nu} \frac{S_D}{\sqrt{n}} \le \mu_D \le \overline{D} + t_{\alpha/2,\nu} \frac{S_D}{\sqrt{n}} \quad \text{for two-sided CI}$$

$$\overline{D} - t_{\alpha,\nu} \frac{S_D}{\sqrt{n}} \le \mu_D \quad \text{for lower-confidence bound}$$

$$\mu_D \le \overline{D} + t_{\alpha,\nu} \frac{S_D}{\sqrt{n}} \quad \text{for upper-confidence bound}$$

CI and Hypothesis Test for Large Sample

If the sample size is large ($n \ge 30$), the z-based CI formulas and test procedure in Section 9-2 can be applied to inference on μ_D according to the central limit theorem.

Example 10.3 The weights (unit: lbs) before and after a diet program for 30 participants are measured below.

No	Before	After	No	Before	After	No	Before	After
1	160	153	11	158	140	21	170	154
2	172	160	12	205	196	22	218	213
3	154	136	13	164	158	23	147	134
4	210	198	14	225	207	24	173	158
5	173	166	15	186	182	25	195	187
6	145	136	16	155	149	26	220	209
7	198	182	17	184	176	27	215	206
8	165	160	18	173	156	28	206	201
9	180	178	19	179	167	29	165	156
10	172	171	20	168	152	30	170	154

The summary of the weight data is as follows:

Sample Size (no. participants)	Sample Mean (weight loss)	Sample Variance
$n = 30$	$\bar{d} = 10$ lbs	$s_D^2 = 5^2$

1. (**Hypothesis Test on μ_D, σ_D^2 Unknown; Two-Sided CI**) Test if there is a significant effect of the diet program on weight loss. Use $\alpha = 0.05$.

Step 1: State H_0 and H_1.

H_0: $\mu_D = 0$

H_1: $\mu_D \neq 0$

Step 2: Determine a **test statistic and its value.**

$$t_0 = \frac{\bar{d} - \delta_0}{s_D / \sqrt{n}} = \frac{10 - 0}{5 / \sqrt{30}} = 10.95$$

Step 3: Determine a **critical value(s) for** α.

$$t_{\alpha/2, n-1} = t_{0.05/2, 30-1} = t_{0.025, 29} = 2.045$$

Step 4: Make a **conclusion.**

Since $|t_0| = 10.95 > t_{0.025,24} = 2.045$, reject H_0 at $\alpha = 0.05$.

2. (**Confidence Interval on μ_D, σ_D^2 Unknown; Two-Sided CI**) Construct a 95% two-sided confidence interval on the mean weight loss (μ_D) due to the diet program. Based on this 95% two-sided CI on μ_D, test H_0: $\mu_D = 0$ vs. H_1: $\mu_D \neq 0$ at $\alpha = 0.05$.

$1 - \alpha = 0.95 \Rightarrow \alpha = 0.05$; $\nu = n - 1 = 30 - 1 = 29$

95% two-sided CI on μ_D:

$$\bar{d} - t_{\alpha/2, \nu} \frac{s_D}{\sqrt{n}} \le \mu_D \le \bar{d} + t_{\alpha/2, \nu} \frac{s_D}{\sqrt{n}}$$

$$\Rightarrow 10 - t_{0.05/2, 29} \frac{5}{\sqrt{30}} \le \mu_D \le 10 + t_{0.05/2, 29} \frac{5}{\sqrt{30}}$$

$$\Rightarrow 10 - 2.045 \times 0.91 \le \mu_D \le 10 + 2.045 \times 0.91$$

Example 10.3 (*cont.*)

$$\Rightarrow\ 8.1 \le \mu_D \le 11.9$$

Since this 95% two-sided CI on μ_D does not include the hypothesized value zero, reject H_0 at $\alpha = 0.05$.

Exercise 10.3 (MR 10-38)

A computer scientist is investigating the usefulness of two different design languages in improving programming tasks. Twelve expert programmers, familiar with both languages, are asked to code a standard function in both languages, and their coding times in minutes are recorded as follows:

	Time (unit: min.)		
Programmer	Design Language 1 (X_1)	Design Language 2 (X_2)	$D = X_1 - X_2$
1	17	18	-1
2	16	14	2
3	21	19	2
4	14	11	3
5	18	23	-5
6	24	21	3
7	16	10	6
8	14	13	1
9	21	19	2
10	23	24	-1
11	13	15	-2
12	18	20	-2

The summary quantities of the coding time data are

$n = 12,\quad \bar{d} = 0.67,\quad \text{and}\quad s_D^2 = 2.96^2$

1. Does the data suggest that there is no difference in mean coding time for the two languages? Use $\alpha = 0.05$ in drawing a conclusion.
2. Construct a 95% two-sided confidence interval on the difference between the two languages in coding time (μ_D). Based on this 95% CI on μ_D, test H_0: $\mu_D = 0$ vs. H_1: $\mu_D \ne 0$ at $\alpha = 0.05$.

10-5 Inference on the Variances of Two Normal Populations

Learning Goals

- ☐ Explain the characteristics of an F distribution.
- ☐ Read the F table.
- ☐ Test a hypothesis on the ratio of two variances (σ_1^2/σ_2^2) (F-test).
- ☐ Determine the sample size of an F-test for statistical inference on σ_1^2/σ_2^2 by using an appropriate operating characteristic (OC) curve.
- ☐ Establish a 100(1 - α)% confidence interval (CI) on σ_1^2/σ_2^2.

F Distribution

The probability density function of an F distribution with u and v degrees of freedom is

$$f(x) = \frac{\Gamma[(u+v)/2](u/v)^{u/2}}{\Gamma(u/2)\Gamma(v/2)} \frac{x^{u/2-1}}{[1+(u/v)x]^{(u+v)/2}}, x > 0$$

The mean and variance of the F distribution are

$$E(X) = v/(v-2) \text{ for } v > 2$$

$$V(X) = \frac{2v^2(u+v-2)}{u(v-2)^2(v-4)} \text{ for } v > 4$$

An F distribution is unimodal and skewed to the right (see Figure 10-4 in MR) like a χ^2 distribution. However, an F distribution is more flexible in shape by having an additional parameter of degrees of freedom.

F Table

The F table (see Appendix Table V in MR) provides 100α **upper-tail percentage points** ($f_{\alpha,u,v}$) of F distributions with various values of u and v, i.e.,

$$P(F_{u,v} > f_{\alpha,u,v})$$

The 100α **lower-tail percentage points** ($f_{1-\alpha,u,v}$) can be found as follows:

$$f_{1-\alpha,u,v} = \frac{1}{f_{\alpha,v,u}}$$

(e.g.) Reading the F table

(1) $P(F_{7,15} > f) = 0.05: \ f = f_{0.05,7,15} = 2.71$

(2) $P(F_{7,15} > f) = 0.95: \ f = f_{0.95,7,15} = 1/f_{0.05,15,7} = 1/3.51 = 0.28$

Inference Context

Parameter of interest: $\dfrac{\sigma_1^2}{\sigma_2^2}$

Point estimator of $\dfrac{\sigma_1^2}{\sigma_2^2}$: $\dfrac{S_1^2}{S_2^2}$

(Note) $S_1^2 = \sum_{i=1}^{n_1} (X_{1i} - \overline{X}_1)^2 / n_1 - 1$ and $S_2^2 = \sum_{j=1}^{n_2} (X_{2i} - \overline{X}_2)^2 / n_2 - 1$;

$X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$;

X_1 and X_2 are independent

Test statistic of $\dfrac{\sigma_1^2}{\sigma_2^2}$: $F = \dfrac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$

Test Procedure (*F*-test)

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

$$H_0:\ \frac{\sigma_1^2}{\sigma_2^2} = \frac{\sigma_{1,0}^2}{\sigma_{2,0}^2}$$

$$H_1:\ \frac{\sigma_1^2}{\sigma_2^2} \neq \frac{\sigma_{1,0}^2}{\sigma_{2,0}^2} \quad \text{for two-sided test,}$$

$$\frac{\sigma_1^2}{\sigma_2^2} > \frac{\sigma_{1,0}^2}{\sigma_{2,0}^2} \quad \text{or} \quad \frac{\sigma_1^2}{\sigma_2^2} < \frac{\sigma_{1,0}^2}{\sigma_{2,0}^2} \quad \text{for one-sided test.}$$

Step 2: Determine a **test statistic and its value.**

$$F_0 = \frac{S_1^2/\sigma_{1,0}^2}{S_2^2/\sigma_{2,0}^2} = \frac{S_1^2}{S_2^2}\frac{\sigma_{2,0}^2}{\sigma_{1,0}^2} \sim f(n_1-1, n_2-1)$$

Step 3: Determine a **critical value(s) for α.**

$$f_{\alpha/2,n_1-1,n_2-1} \text{ and } f_{1-\alpha/2,n_1-1,n_2-1}\left(=\frac{1}{f_{\alpha/2,n_2-1,n_1-1}}\right) \quad \text{for two-sided test}$$

$$f_{\alpha,n_1-1,n_2-1} \quad \text{for upper-sided test}$$

$$f_{1-\alpha,n_1-1,n_2-1}\left(=\frac{1}{f_{\alpha,n_2-1,n_1-1}}\right) \quad \text{for lower-sided test}$$

Step 4: Make a **conclusion**. Reject H_0 if

$$f_0 > f_{\alpha/2,n_1-1,n_2-1} \text{ or } f_0 < f_{1-\alpha/2,n_1-1,n_2-1} \quad \text{for two-sided test}$$

$$f_0 > f_{\alpha,n_1-1,n_2-1} \quad \text{for upper-sided test}$$

$$f_0 < f_{1-\alpha,n_1-1,n_2-1} \quad \text{for lower-sided test}$$

Operating Characteristic (OC) Curve

Table 10-4 displays a list of OC charts in MR and a formula of the OC parameter λ for an F-test on σ_1^2/σ_2^2 where $n_1 = n_2 = n$. By using the table, the appropriate OC chart for a particular F-test is chosen.

Table 10-4 Operating Characteristic Charts for F-test

Test		α	Chart VI (Appendix A in MR)	OC parameter
F-test	Two-sided	0.05	(o)	$\lambda = \frac{\sigma_1}{\sigma_2}$
		0.01	(p)	
	One-sided	0.05	(q)	
		0.01	(r)	

Confidence Interval Formula

$$\frac{s_1^2}{s_2^2} f_{1-\alpha/2,n_2-1,n_1-1} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{s_1^2}{s_2^2} f_{\alpha/2,n_2-1,n_1-1} \quad \text{for two-sided CI}$$

$$\frac{s_1^2}{s_2^2} f_{1-\alpha,n_2-1,n_1-1} \le \frac{\sigma_1^2}{\sigma_2^2} \quad \text{for lower-confidence bound}$$

$$\frac{\sigma_1^2}{\sigma_2^2} \le \frac{s_1^2}{s_2^2} f_{\alpha,n_2-1,n_1-1} \quad \text{for upper-confidence bound}$$

(Derivation) Two-sided confidence interval* on σ_1^2/σ_2^2

By using the test statistic $F = \dfrac{S_2^2/\sigma_2^2}{S_1^2/\sigma_1^2} \sim F(n_2-1, n_1-1)$,

$$P\left(L \le \frac{\sigma_1^2}{\sigma_2^2} \le U\right) = 1-\alpha$$

$$\Rightarrow P\left(f_{1-\alpha/2,n_2-1,n_1-1} \le F \le f_{\alpha/2,n_2-1,n_1-1}\right) = 1-\alpha$$

$$\Rightarrow P\left(f_{1-\alpha/2,n_2-1,n_1-1} \le \frac{S_2^2/\sigma_2^2}{S_1^2/\sigma_1^2} \le f_{\alpha/2,n_2-1,n_1-1}\right) = 1-\alpha$$

$$\Rightarrow P\left(\frac{S_1^2}{S_2^2} f_{1-\alpha/2,n_2-1,n_1-1} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{S_1^2}{S_2^2} f_{\alpha/2,n_2-1,n_1-1}\right) = 1-\alpha$$

Therefore,

$$L = \frac{S_1^2}{S_2^2} f_{1-\alpha/2,n_2-1,n_1-1} \quad \text{and} \quad U = \frac{S_1^2}{S_2^2} f_{\alpha/2,n_2-1,n_1-1}$$

* For a one-sided CI, use f_α instead of $f_{\alpha/2}$ to derive the corresponding limit.

Example 10.4

For the light bulb life length data in Example 10.1, the following results have been obtained:

Brand of Light Bulb	Sample Size	Sample Mean	Sample Variance
INFINITY (1)	$n_1 = 30$	$\bar{x}_1 = 780$ hrs	$s_1^2 = 40^2$
FOREVER (2)	$n_2 = 25$	$\bar{x}_2 = 800$ hrs	$s_2^2 = 30^2$

1. **(Hypothesis Test on σ_1^2/σ_2^2; Two-Sided Test)** Test H_0: $\sigma_1^2/\sigma_2^2 = 1$ vs. H_1: $\sigma_1^2/\sigma_2^2 \ne 1$ at $\alpha = 0.05$.

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

$$H_0: \ \sigma_1^2/\sigma_2^2 = 1$$

$$H_1: \ \sigma_1^2/\sigma_2^2 \ne 1$$

Step 2: Determine a **test statistic and its value.**

$$f_0 = \frac{s_1^2}{s_2^2}\frac{\sigma_{2,0}^2}{\sigma_{1,0}^2} = \frac{40^2}{30^2} \times 1 = 1.78$$

Example 10.4 (*cont.*)

Step 3: Determine a **critical value(s) for α.**

$$f_{\alpha/2,n_1-1,n_2-1} = f_{0.05/2,30-1,25-1} = f_{0.025,29,24} = 2.22 \text{ and}$$

$$f_{1-\alpha/2,n_1-1,n_2-1} = f_{0.975,29,24} = \frac{1}{f_{0.025,24,29}} = \frac{1}{2.15} = 0.46$$

Step 4: Make a **conclusion.**

Since, $f_0 = 1.78 \not> f_{0.025,29,24} = 2.22$ and

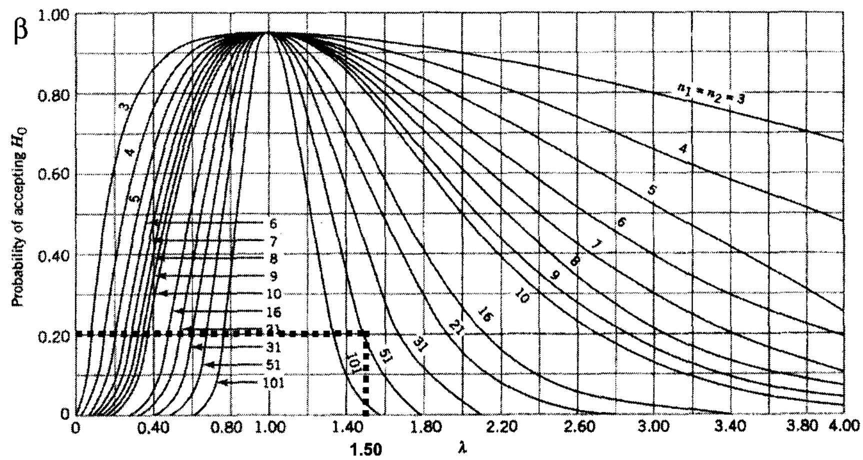
$f_0 = 1.78 \not< f_{0.975,29,24} = 0.46$, fail to reject H_0 at $\alpha = 0.05$.

2. (**Sample Size Determination**) Determine the sample size n ($= n_1 = n_2$) required for this two-sided F-test to detect the ratio of σ_1 to σ_2 as high as 1.5 with 0.8 of power. Apply an appropriate OC curve.

To design a two-sided F-test at $\alpha = 0.05$, OC Chart VI*o* is applicable with the parameter

$$\lambda = \frac{\sigma_1}{\sigma_2} = 1.5$$

By using $\lambda = 1.5$ and $\beta = 0.2$ (because power $= 1 - \beta = 0.8$), the sample size required is determined as n ($= n_1 = n_2$) $= 50$ as displayed below.



(*o*) *O.C.* curves for different values of *n* for the two-sided *F*-test for a level of significance α = 0.05.

3. (**Confidence Interval on σ_1^2/σ_2^2; Two-Sided CI**) Construct a 95% two-sided confidence interval on σ_1^2/σ_2^2. Based on this 95% CI on σ_1^2/σ_2^2, test H_0: $\sigma_1^2/\sigma_2^2 = 1$ vs. H_1: $\sigma_1^2/\sigma_2^2 \neq 1$ at $\alpha = 0.05$.

$1 - \alpha = 0.95 \quad \Rightarrow \quad \alpha = 0.05$

95% two-sided CI on σ_1^2/σ_2^2:

$$\frac{s_1^2}{s_2^2} f_{1-\alpha/2,n_2-1,n_1-1} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{s_1^2}{s_2^2} f_{\alpha/2,n_2-1,n_1-1}$$

$$\Rightarrow \frac{40^2}{30^2} f_{1-0.05/2,24,29} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{40^2}{30^2} f_{0.05/2,24,29}$$

Example 10.4 (*cont.*)

$$\Rightarrow\ 1.78\frac{1}{f_{0.025,29,24}} \le \frac{\sigma_1^2}{\sigma_2^2} \le 1.78 f_{0.025,24,29}$$

$$\Rightarrow\ 1.78\times\frac{1}{2.22} \le \frac{\sigma_1^2}{\sigma_2^2} \le 1.78\times 2.15$$

$$\Rightarrow\ 0.80 \le \frac{\sigma_1^2}{\sigma_2^2} \le 3.83$$

Since this 95% two-sided CI on σ_1^2/σ_2^2 includes the hypothesized value unity, fail to reject H_0 at $\alpha = 0.05$.

Exercise 10.4 (MR 10-49)

In semiconductor manufacturing, wet chemical etching is often used to remove silicon from the backs of wafers prior to metalization. The etching rate is an important characteristic in this process. Two different etching solutions have been compared, using two random samples of 10 wafers for each solution. The observed etching rates (unit: mils/min) are as follows:

Solution 1		Solution 2	
9.9	10.6	10.2	10.0
9.4	10.3	10.6	10.2
9.3	10.0	10.7	10.7
9.6	10.3	10.4	10.4
10.2	10.1	10.5	10.3

The sample variances of the etching rates are $s_1^2 = 0.42^2$ and $s_2^2 = 0.23^2$. Assume that the etching rates are normally distributed.

1. Test H_0: $\sigma_1^2/\sigma_2^2 = 1$ vs. H_1: $\sigma_1^2/\sigma_2^2 \neq 1$ at $\alpha = 0.05$.
2. Determine the sample size n ($= n_1 = n_2$) required for this two-sided F-test to detect the ratio of σ_1 to σ_2 as high as 2 with 0.9 of power. Apply an appropriate OC curve.
3. Construct a 95% two-sided confidence interval on σ_1^2/σ_2^2. Based on this 95% two-sided CI on σ_1^2/σ_2^2, test H_0: $\sigma_1^2/\sigma_2^2 = 1$ vs. H_1: $\sigma_1^2/\sigma_2^2 \neq 1$ at $\alpha = 0.05$.

10-6 Inference on Two Population Proportions

Learning Goals

- ❑ Test a hypothesis on $p_1 - p_2$ (z-test).
- ❑ Determine the sample size of a z-test for statistical inference on $p_1 - p_2$ by using an appropriate sample size formula.
- ❑ Establish a 100(1 - α)% confidence interval (CI) on $p_1 - p_2$.

Inference Context

Parameter of interest: $p_1 - p_2$

Point estimator of $p_1 - p_2$: $\hat{P}_1 - \hat{P}_2 \sim N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$

Inference Context (*cont.*)

(Note) $X_1 \sim B(n_1, p_1)$, $X_2 \sim B(n_2, p_2)$;
$n_1 p_1$, $n_1(1-p_1)$, $n_2 p_2$, and $n_2(1-p_2)$ are greater than 5;
X_1 and X_2 are independent;

$$\hat{P}_1 = \frac{X_1}{n_1} \sim N\left(p_1, \frac{p_1(1-p_1)}{n_1}\right) \text{ and } \hat{P}_2 = \frac{X_2}{n_2} \sim N\left(p_2, \frac{p_2(1-p_2)}{n_2}\right)$$

Test statistic of p_1 - p_2: The test statistic of p_1 - p_2 depends on the equality of p_1 and p_2 as follows:

(1) **Case 1: Unequal proportions** ($p_1 \neq p_2$)

$$Z = \frac{(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}} \sim N(0,1)$$

(2) **Case 2: Equal proportions** ($p_1 = p_2 = p$)

$$Z = \frac{(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}} = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{p(1-p)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$$

$$= \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{P}(1-\hat{P})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \sim N(0,1), \text{ where } \hat{P} = \frac{X_1 + X_2}{n_1 + n_2} \text{ (estimator of } p)$$

Test Procedure (*z*-test)

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $p_1 - p_2 = \delta_0$
H_1: $p_1 - p_2 \neq \delta_0$ for two-sided test,
$p_1 - p_2 > \delta_0$ or $p_1 - p_2 < \delta_0$ for one-sided test.

Step 2: Determine a **test statistic and its value**.

(1) **Case 1: Unequal proportions** ($p_1 \neq p_2$)

$$Z_0 = \frac{\hat{P}_1 - \hat{P}_2 - \delta_0}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}} = \frac{\hat{P}_1 - \hat{P}_2 - \delta_0}{\sqrt{\dfrac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \dfrac{\hat{P}_2(1-\hat{P}_2)}{n_2}}}$$

(2) **Case 2: Equal proportions** ($p_1 = p_2 = p$)

$$Z_0 = \frac{\hat{P}_1 - \hat{P}_2 - 0}{\sqrt{p(1-p)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{P}(1-\hat{P})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}, \quad \hat{P} = \frac{X_1 + X_2}{n_1 + n_2}$$

Step 3: Determine a **critical value(s) for** $\boldsymbol{\alpha}$.

$z_{\alpha/2}$ for two-sided test; z_α for one-sided test

Step 4: Make a **conclusion**. Reject H_0 if

$|z_0| > z_{\alpha/2}$ for two-sided test
$z_0 > z_\alpha$ for upper-sided test
$z_0 < -z_\alpha$ for lower-sided test

Sample Size Formula

For a hypothesis test on $p_1 - p_2$, the following formulas are applied to determine the sample size:

$$n = \left(\frac{z_{\alpha/2}\sqrt{(p_1+p_2)(q_1+q_2)/2} + z_\beta\sqrt{p_1q_1+p_2q_2}}{p_1-p_2} \right)^2 \quad \text{for two-sided test}$$

$$= \left(\frac{z_{\alpha}\sqrt{(p_1+p_2)(q_1+q_2)/2} + z_\beta\sqrt{p_1q_1+p_2q_2}}{p_1-p_2} \right)^2 \quad \text{for one-sided test}$$

where: $n_1 = n_2 = n$, $q_1 = 1 - p_1$, and $q_2 = 1 - p_2$

Confidence Interval Formula

Like the test statistic on $p_1 - p_2$, a $100(1 - \alpha)\%$ CI on $p_1 - p_2$ depends on the equality of p_1 and p_2 as follows:

1. **Case 1: Unequal proportions** ($p_1 \neq p_2$)

$$\hat{P}_1 - \hat{P}_2 - z_{\alpha/2}\sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{n_2}}$$
$$\le p_1 - p_2 \le \hat{P}_1 - \hat{P}_2 + z_{\alpha/2}\sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{n_2}} \quad \text{for two-sided CI}$$

$$\hat{P}_1 - \hat{P}_2 - z_{\alpha/2}\sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{n_2}} \le p_1 - p_2 \quad \text{for lower-confidence bound}$$

$$p_1 - p_2 \le \hat{P}_1 - \hat{P}_2 + z_{\alpha/2}\sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{n_2}} \quad \text{for upper-confidence bound}$$

2. **Case 2: Equal proportions** ($p_1 = p_2 = p$)

$$\hat{P}_1 - \hat{P}_2 - z_{\alpha/2}\sqrt{\hat{P}(1-\hat{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$
$$\le p_1 - p_2 \le \hat{P}_1 - \hat{P}_2 + z_{\alpha/2}\sqrt{\hat{P}(1-\hat{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \quad \text{for two-sided CI}$$

$$\hat{P}_1 - \hat{P}_2 - z_{\alpha/2}\sqrt{\hat{P}(1-\hat{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \le p_1 - p_2 \quad \text{for lower-confidence bound}$$

$$p_1 - p_2 \le \hat{P}_1 - \hat{P}_2 + z_{\alpha/2}\sqrt{\hat{P}(1-\hat{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \quad \text{for upper-confidence bound}$$

Example 10.5

Random samples of bridges are tested for metal corrosion in the HAPPY and GREAT counties, resulting in the following:

County	Sample Size (no. bridges)	X (no. corroded bridges)	Sample Proportion
HAPPY (1)	$n_1 = 40$	$x_1 = 28$	$\hat{p}_1 = 0.7$
GREAT (2)	$n_2 = 30$	$x_2 = 15$	$\hat{p}_2 = 0.5$

Example 10.5 (*cont.*)

1. (**Hypothesis Test on $p_1 - p_2$; Unequal Proportions; Upper-Sided Test**) Assuming $p_1 \neq p_2$, test if the proportion of corroded bridges of HAPPY County exceeds that of GREAT County by at least 0.1. Use $\alpha = 0.05$.

Since $n_1\hat{p}_1 = 40 \times 0.7 = 28$, $n_1(1-\hat{p}_1) = 40 \times 0.3 = 12$, $n_2\hat{p}_2 = 30 \times 0.5 = 15$, and $n_2(1-\hat{p}_2) = 30 \times 0.5 = 15$ are greater than five, the sampling distributions of $\hat{P}_1$ and $\hat{P}_2$ are approximately normal.

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $p_1 - p_2 = 0.1$
H_1: $p_1 - p_2 > 0.1$

Step 2: Determine a **test statistic and its value.**

$$z_0 = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} = \frac{0.7 - 0.5 - 0.1}{\sqrt{\frac{0.7 \times (1-0.7)}{40} + \frac{0.5 \times (1-0.5)}{30}}} = 0.86$$

Step 3: Determine a **critical value(s) for α.**

$$z_\alpha = z_{0.05} = 1.64$$

Step 4: Make a **conclusion.**

Since $|z_0| = 0.86 \not> z_{0.05} = 1.64$, fail to reject H_0 at $\alpha = 0.05$.

2. (**Sample Size Determination**) Suppose that $p_1 = 0.7$ and $p_2 = 0.5$. Determine the sample size n ($= n_1 = n_2$) required for this two-sided z-test to detect the difference of the two proportions with power of 0.9.

power = P(reject H_0 | H_0 is false) = $1 - \beta = 0.9 \Rightarrow \beta = 0.1$

$$n = \left(\frac{z_\alpha\sqrt{(p_1+p_2)(q_1+q_2)/2} + z_\beta\sqrt{p_1q_1+p_2q_2}}{p_1 - p_2}\right)^2$$

$$= \left(\frac{z_{0.05}\sqrt{(0.7+0.5)(0.3+05)/2} + z_{0.01}\sqrt{0.7\times0.3+0.5\times0.5}}{0.7-0.5}\right)^2$$

$$= \left(\frac{1.64\times0.69 + 1.28\times0.68}{0.2}\right)^2 \cong 101$$

3. (**Confidence Interval on $p_1 - p_2$; Unequal Proportions; Upper-Confidence Bound**) Assuming $p_1 \neq p_2$, construct a 95% upper-confidence bound on the difference of the two corroded bridge proportions ($p_1 - p_2$).

95% two-sided CI on $p_1 - p_2$:

$$p_1 - p_2 \le \hat{p}_1 - \hat{p}_2 + z_\alpha\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$\Rightarrow\ p_1 - p_2 \le 0.7 - 0.5 + z_{0.05}\sqrt{\frac{0.7(1-0.7)}{40} + \frac{0.5(1-0.5)}{30}}$$

Example 10.5 (*cont.*)

$$\Rightarrow\ p_1 - p_2 \le 0.2 + 1.64 \times 0.117$$
$$\Rightarrow\ p_1 - p_2 \le 0.39$$

Exercise 10.5 (MR 10-65)

A random sample of $n_1 = 500$ adult residents in MARICOPA County found that $x_1 = 385$ were in favor of increasing the highway speed limit to 70 mph, while another random sample of $n_2 = 400$ adult residents in PIMA County found that $x_2 = 267$ were in favor of the increased speed limit.

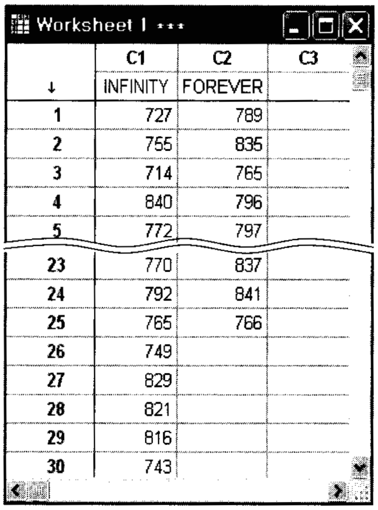
1. Does the survey data indicate that there is a difference between the residents of the two counties in support of increasing the speed limit? Assume $p_1 - p_2 = 0$ and use $\alpha = 0.05$.
2. Suppose that $p_1 = 0.75$ and $p_2 = 0.65$. Determine the sample size n ($= n_1 = n_2$) required for this two-sided z-test to detect the difference of the two proportions with power of 0.8.
3. Assuming $p_1 = p_2$, construct a 95% two-sided confidence interval on the difference between the favor proportions of the two counties for the speed limit increase ($p_1 - p_2$). Based on this 95% two-sided CI on $p_1 - p_2$, test H_0: $p_1 - p_2 = 0$ vs. H_1: $p_1 - p_2 \neq 0$ at $\alpha = 0.05$.

MINITAB Applications

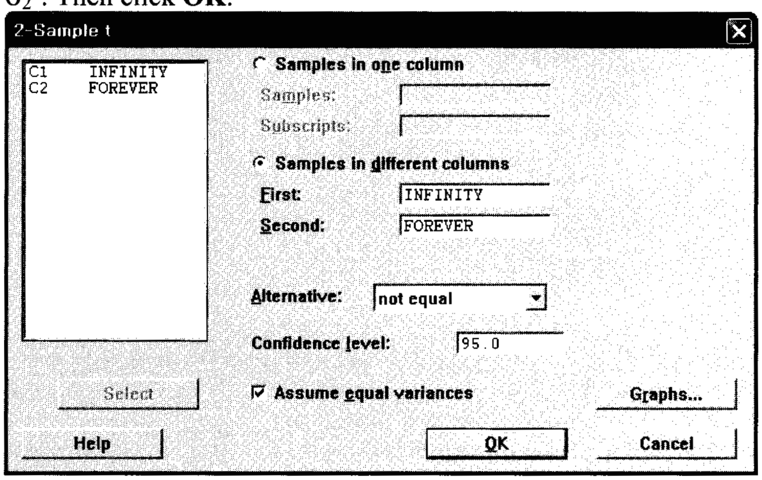
Example 10.2 **(Inference on $\mu_1 - \mu_2$, σ_1^2 and σ_2^2 Unknown)**

(1) Choose **File ➤ New**, click **Minitab Project**, and click **OK**.

(2) Enter the life length data of the INFINITY and FOREVER light bulbs on the worksheet.

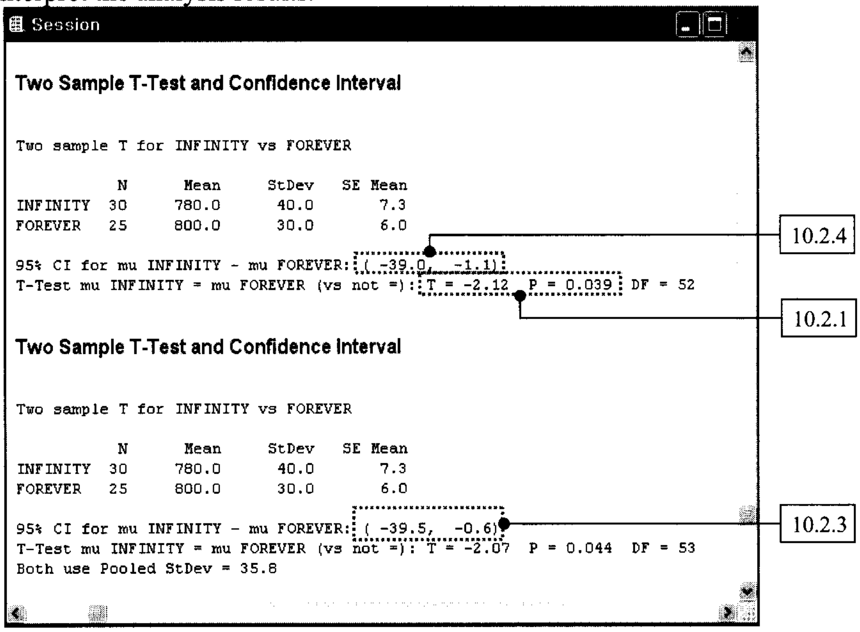


(3) Choose **Stat ➤ Basic Statistics ➤ 2-Sample t**. Click **Samples in different columns** and select INFINITY and FOREVER in **First** and **Second**, respectively. Select **not equal** in **Alternative** and type the level of confidence in **Confidence level**. Check **Assume equal variances** if $\sigma_1^2 = \sigma_2^2$. Then click **OK**.



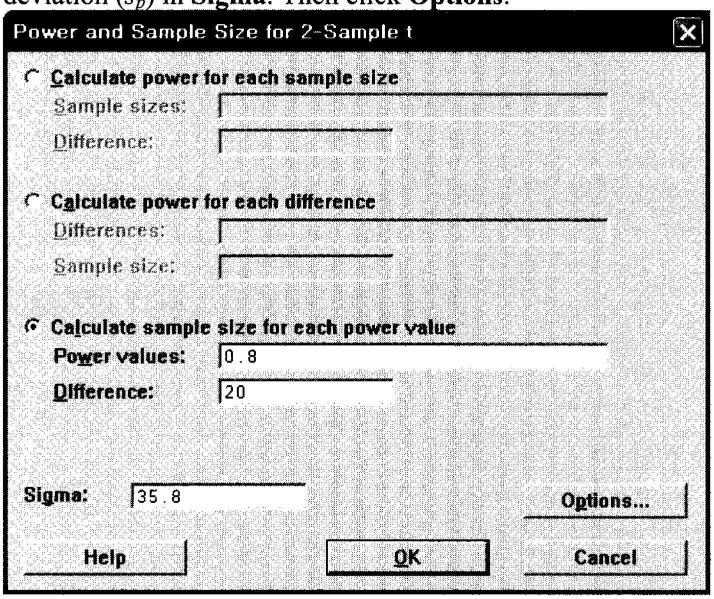
Example 10.2
(cont.)

(4) Interpret the analysis results.



```
Session

Two Sample T-Test and Confidence Interval

Two sample T for INFINITY vs FOREVER

           N      Mean     StDev   SE Mean
INFINITY  30     780.0      40.0       7.3
FOREVER   25     800.0      30.0       6.0

95% CI for mu INFINITY - mu FOREVER: ( -39.0,  -1.1)
T-Test mu INFINITY = mu FOREVER (vs not =): T = -2.12  P = 0.039  DF = 52

Two Sample T-Test and Confidence Interval

Two sample T for INFINITY vs FOREVER

           N      Mean     StDev   SE Mean
INFINITY  30     780.0      40.0       7.3
FOREVER   25     800.0      30.0       6.0

95% CI for mu INFINITY - mu FOREVER: ( -39.5,  -0.6)
T-Test mu INFINITY = mu FOREVER (vs not =): T = -2.07  P = 0.044  DF = 53
Both use Pooled StDev = 35.8
```

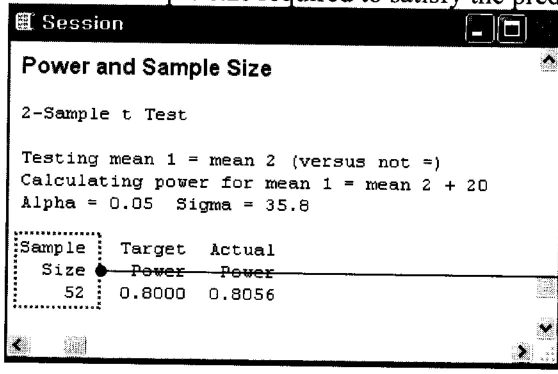

(5) Choose **Stat ➤ Power and Sample Size ➤ 2-Sample Z**. Click **Calculate sample size for each power value**. Enter the power of the test predefined in **Power values,** the difference between the true mean and hypothesized mean to be detected in **Difference**, and the pooled estimate of standard deviation (s_p) in **Sigma**. Then click **Options**.



(6) Under **Alternative Hypothesis** select the type of alternative hypothesis, and in **Significance level** enter the probability of type I error (α).Then click **OK** twice.

Example 10.2 (*cont.*)

(7) Obtain the sample size required to satisfy the predefined test condition.



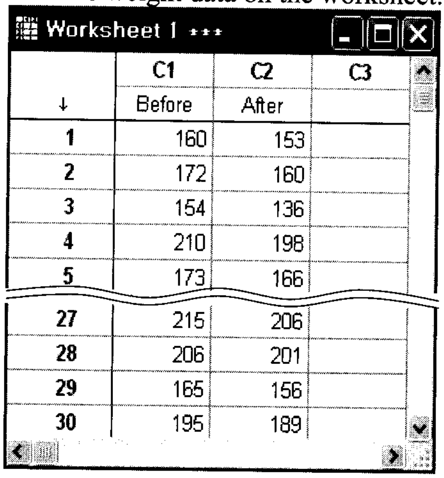
10.2.2

Example 10.3

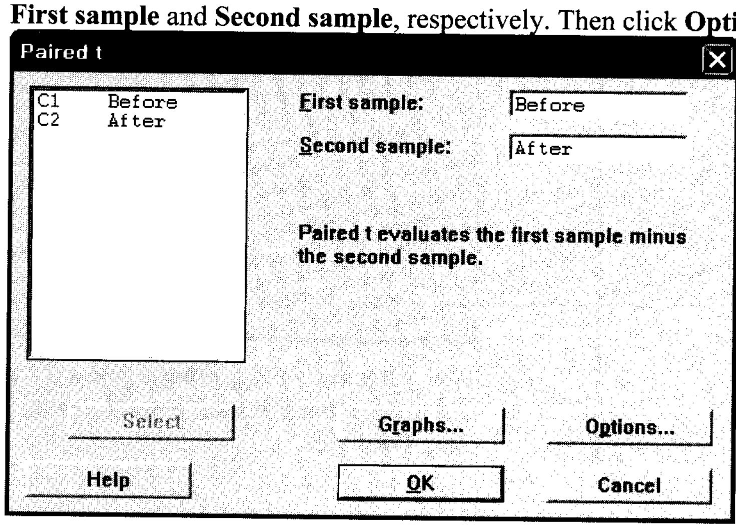
(Inference on μ_D, σ_D^2 Unknown)

(1) Choose **File ➤ New**, click **Minitab Project**, and click **OK**.

(2) Enter the weight data on the worksheet.

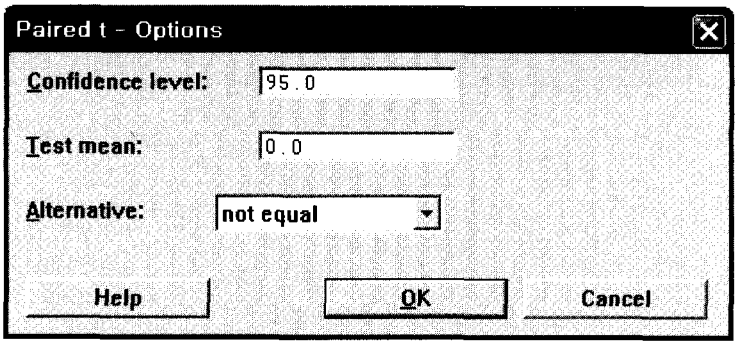


(3) Choose **Stat ➤ Basic Statistics ➤ Paired t**. Select BEFORE and AFTER in **First sample** and **Second sample**, respectively. Then click **Options**.

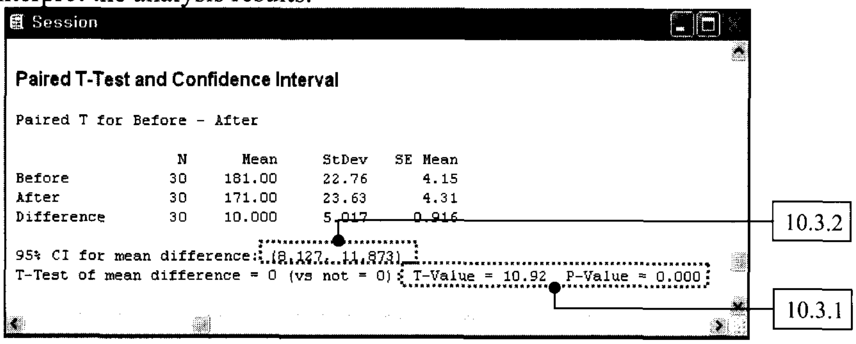


Example 10.3 (*cont.*)

(4) Enter the level of confidence in **Confidence level**, the hypothesized difference in mean in **Test mean**, and **not equal** in **Alternative**. Then click **OK** twice.

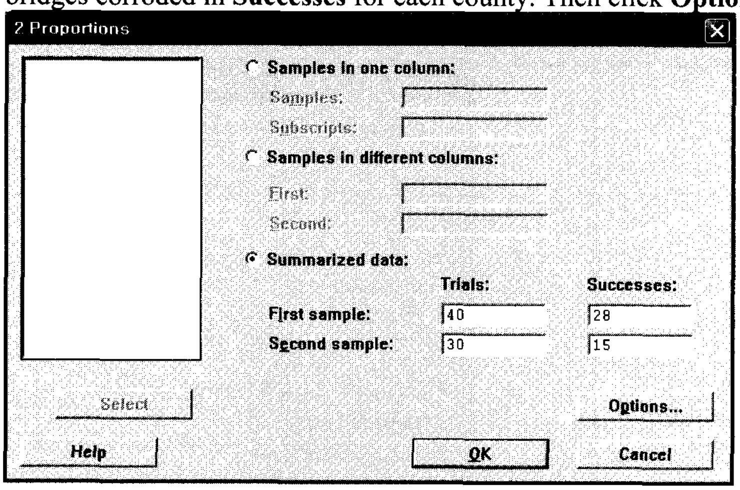


(5) Interpret the analysis results.



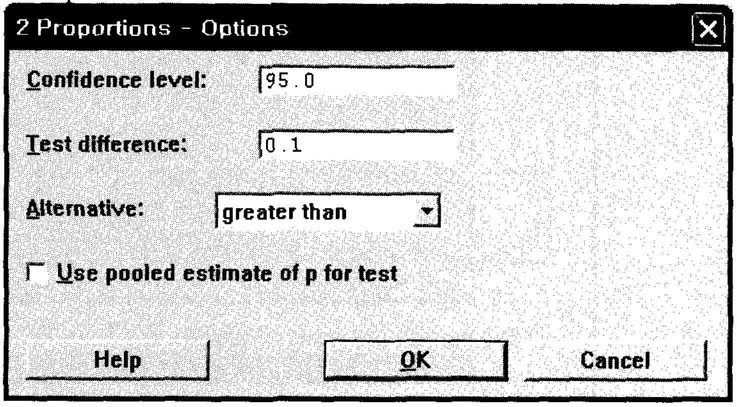
Example 10.5 **(Inference on $p_1 - p_2$)**

(1) Choose **Stat ➤ Basic Statistics ➤ 2 Proportions**. Click **Summarized data** and enter the number of bridges surveyed in **Trials** and the number of bridges corroded in **Successes** for each county. Then click **Options**.

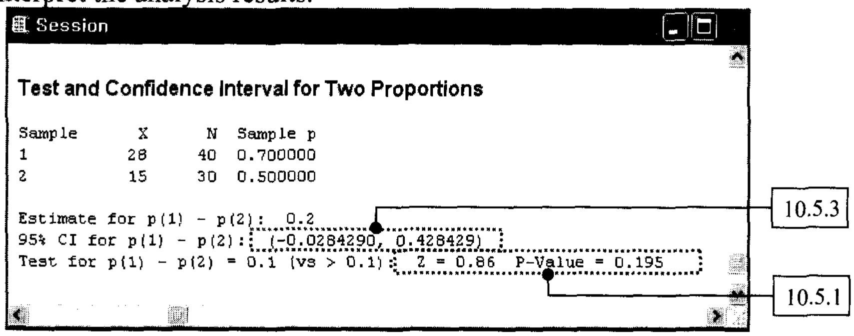


Example 10.5 *(cont.)*

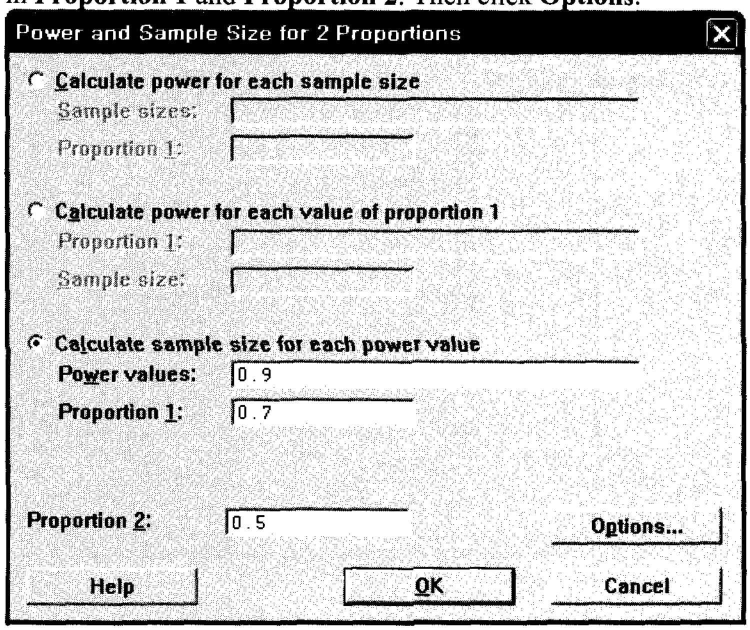
(2) Enter the level of confidence in **Confidence level**, the hypothesized difference in proportion in **Test difference**, and **greater than** in **Alternative**. Check **Use pooled estimate of p for test** if the proportions are equal. Then click **OK** twice.



(3) Interpret the analysis results.

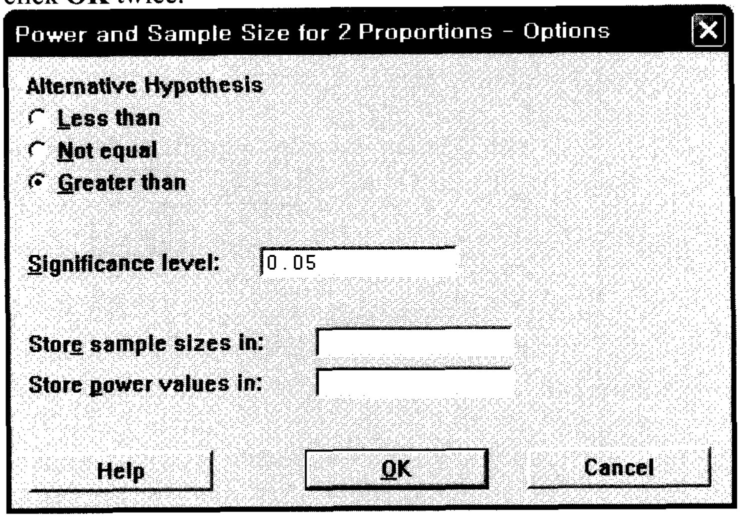


(4) Choose **Stat ➤ Power and Sample Size ➤ 2 Proportions**. Click **Calculate sample size for each power value**. Enter the power of the test predefined in **Power values** and the true proportions of two populations in **Proportion 1** and **Proportion 2**. Then click **Options**.

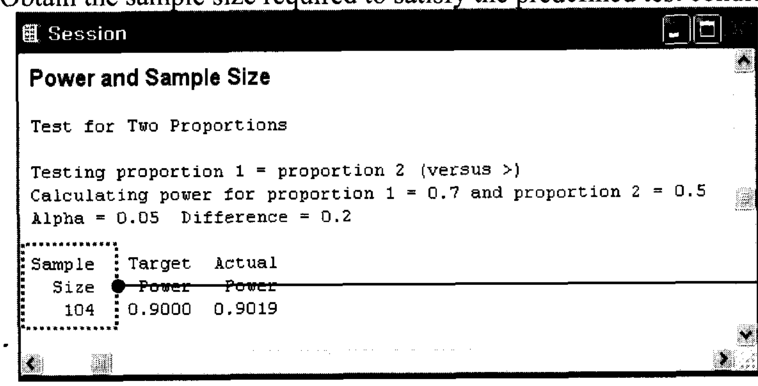


Example 10.5
(cont.)

(5) Under **Alternative Hypothesis** select the type of alternative hypothesis, and in **Significance level** enter the probability of type I error (α). Then click **OK** twice.



(6) Obtain the sample size required to satisfy the predefined test condition.



10.5.2

Answers to Exercises

Exercise 10.1

1. **(Hypothesis Test on $\mu_1 - \mu_2$, σ^2 Known; Upper-Sided Test)**

   Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

   H_0: $\mu_1 - \mu_2 = 10$

   H_1: $\mu_1 - \mu_2 > 10$

   Step 2: Determine a **test statistic and its value.**

   $$z_0 = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(162.5 - 155.0) - 10}{\sqrt{\frac{1}{10} + \frac{1}{12}}} = -5.84$$

   Step 3: Determine a **critical value(s) for α.**

   $$z_\alpha = z_{0.05} = 1.65$$

   Step 4: Make a **conclusion.**

   Since $z_0 = -5.84 \not> z_{0.05} = 1.65$, fail to reject H_0 at $\alpha = 0.05$. We do not have significant evidence to support use of plastic 1 at $\alpha = 0.05$.

2. **(Sample Size Determination)**

   (1) Sample Size Formula

   $$n = \frac{(z_\alpha + z_\beta)^2(\sigma_1^2 + \sigma_2^2)}{(\Delta - \Delta_0)^2} = \frac{(z_{0.05} + z_{0.1})^2(1^2 + 1^2)}{(11.5 - 10)^2}$$

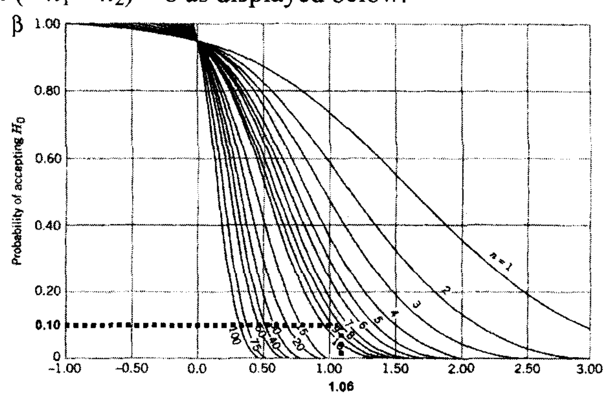
   $$= \frac{(1.64 + 1.28)^2 \times (1^2 + 1^2)}{1.5^2} \cong 8$$

   (2) OC Curve

   To design a one-sided z-test at $\alpha = 0.05$ for two samples, OC Chart VI*c* is applicable with the parameter

   $$d = \frac{|\Delta - \Delta_0|}{\sqrt{\sigma_1^2 + \sigma_2^2}} = \frac{|11.5 - 10|}{\sqrt{1^2 + 1^2}} = 1.06$$

   By using $d = 1.06$ and $\beta = 0.1$, the required sample size is determined as $n\ (= n_1 = n_2) = 8$ as displayed below.



(c) O.C. curves for different values of n for the one-sided normal test for a level of significance $\alpha = 0.05$.

Exercise 10.1 (*cont.*)

3. **(Confidence Interval; Upper-Confidence Bound)**

   $1 - \alpha = 0.95 \Rightarrow \alpha = 0.05$

   95% upper-confidence bound on $\mu_1 - \mu_2$:

$$\mu_1 - \mu_2 \le \bar{x}_1 - \bar{x}_2 + z_\alpha \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$\Rightarrow \mu_1 - \mu_2 \le 162.5 - 155.0 + z_{0.05}\sqrt{\frac{1}{10} + \frac{1}{12}}$$

$$\Rightarrow \mu_1 - \mu_2 \le 7.5 + 1.65 \times \sqrt{\frac{1}{10} + \frac{1}{12}}$$

$$\Rightarrow \mu_1 - \mu_2 \le 8.2$$

Exercise 10.2

1. **(Hypothesis Test on $\mu_1 - \mu_2$, σ_1^2 and σ_2^2 Unknown but Equal; Lower-Sided Test)**

   Step 1: State H_0 and H_1.

   H_0: $\mu_1 - \mu_2 = 0$

   H_1: $\mu_1 - \mu_2 < 0$

   Step 2: Determine a **test statistic and its value.**

$$t_0 = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(290 - 321) - 0}{18.9 \times \sqrt{\frac{1}{10} + \frac{1}{16}}} = -4.07$$

(Note) $$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$= \frac{(10-1)\times 12^2 + (16-1)\times 22^2}{10 + 16 - 2} = 356.5 = 18.9^2$$

$$\nu = n_1 + n_2 - 2 = 10 + 16 - 2 = 24$$

   Step 3: Determine a **critical value(s) for α.**

$$t_{\alpha,\nu} = t_{0.05,24} = 1.711$$

   Step 4: Make a **conclusion.**

   Since $t_0 = -4.07 < -t_{0.05,24} = -1.711$, reject H_0 at $\alpha = 0.05$.

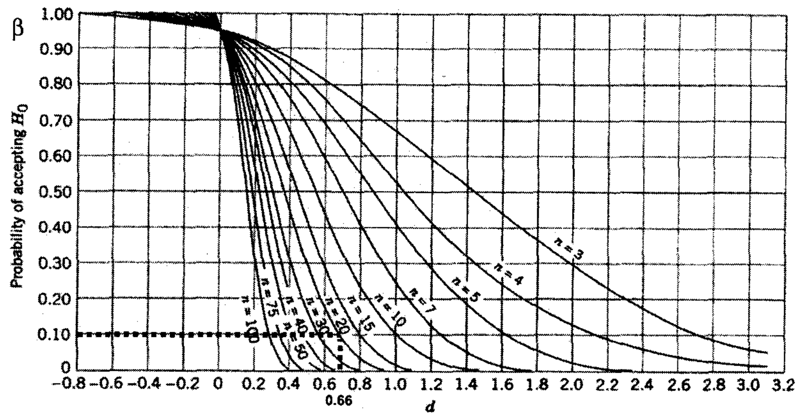
2. **(Sample Size Determination)**

   To design a one-sided t-test at $\alpha = 0.05$ for two samples, OC Chart VIg is applicable with the parameter

$$d = \frac{|\Delta - \Delta_0|}{2\hat{\sigma}} = \frac{|\Delta - \Delta_0|}{2s_p} = \frac{|25 - 0|}{2 \times 18.9} = 0.66 \quad (s_p = 18.9)$$

Exercise 10.2 (*cont.*)

When $d = 0.66$ and $\beta = 0.1$, $n^* = 20$ as displayed below.



(g) *O.C.* curves for different values of *n* for the one-sided *t*-test for a level of significance $\alpha = 0.05$.

Thus, the required sample size n is

$$n = \frac{n^* + 1}{2} = \frac{20 + 1}{2} = 10.5 \cong 11$$

3. **(Confidence Interval on $\mu_1 - \mu_2$, σ_1^2 and σ_2^2 Unknown but Equal; Upper-Confidence Bound)**

$1 - \alpha = 0.95 \Rightarrow \alpha = 0.05;\ \ \nu = n_1 + n_2 - 2 = 10 + 16 - 2 = 24;\ \ s_p^2 = 18.9^2$

95% upper-confidence bound on $\mu_1 - \mu_2$:

$$\mu_1 - \mu_2 \le \bar{x}_1 - \bar{x}_2 + t_{\alpha,\nu} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$\Rightarrow\ \mu_1 - \mu_2 \le 290 - 321 + t_{0.05,24} \times 18.9 \times \sqrt{\frac{1}{15} + \frac{1}{10}}$$

$$\Rightarrow\ \mu_1 - \mu_2 \le -31 + 1.711 \times 7.61$$

$$\Rightarrow\ \mu_1 - \mu_2 \le -18.0$$

4. **(Confidence Interval on $\mu_1 - \mu_2$, σ_1^2 and σ_2^2 Unknown and Unequal; Upper-Confidence Bound)**

$1 - \alpha = 0.95 \ \Rightarrow\ \alpha = 0.05$

$$\nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\dfrac{(s_1^2/n_1)^2}{n_1 + 1} + \dfrac{(s_2^2/n_2)^2}{n_2 + 1}} - 2 = \frac{(12^2/10 + 22^2/16)^2}{\dfrac{(12^2/10)^2}{10 + 1} + \dfrac{(22^2/16)^2}{16 + 1}} - 2 \cong 25$$

95% upper-confidence bound on $\mu_1 - \mu_2$:

$$\mu_1 - \mu_2 \le \bar{x}_1 - \bar{x}_2 + t_{\alpha,\nu} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$\Rightarrow\ \mu_1 - \mu_2 \le 290 - 321 + t_{0.05,25} \sqrt{\frac{12^2}{10} + \frac{22^2}{16}}$$

$$\Rightarrow\ \mu_1 - \mu_2 \le -31 + 1.708 \times 6.68$$

$$\Rightarrow\ \mu_1 - \mu_2 \le -19.59$$

Exercise 10.3

1. **(Hypothesis Test on μ_D, σ_D^2 Unknown; Two-Sided Test)**

Step 1: State H_0 and H_1.

H_0: $\mu_D = 0$
H_1: $\mu_D \neq 0$

Step 2: Determine a **test statistic and its value.**

$$t_0 = \frac{\bar{d} - \delta_0}{s_D / \sqrt{n}} = \frac{0.67 - 0}{2.96 / \sqrt{12}} = 0.783$$

Step 3: Determine a **critical value(s) for α.**

$$t_{\alpha/2,n-1} = t_{0.05/2,12-1} = t_{0.025,11} = 2.201$$

Step 4: Make a **conclusion.**

Since $|t_0| = 0.783 \ngtr t_{0.025,11} = 2.201$, fail to reject H_0 at $\alpha = 0.05$. Thus, we conclude that the two design languages are not significantly different in mean coding time at $\alpha = 0.05$.

2. **(Confidence Interval on μ_D, σ_D^2 Unknown; Two-Sided CI)**

$1 - \alpha = 0.95 \quad \Rightarrow \quad \alpha = 0.05; \quad \nu = n - 1 = 12 - 1 = 11$

95% two-sided CI on μ_D:

$$\bar{d} - t_{\alpha/2,\nu}\frac{s_D}{\sqrt{n}} \le \mu_D \le \bar{d} + t_{\alpha/2,\nu}\frac{s_D}{\sqrt{n}}$$

$$\Rightarrow \quad 0.67 - t_{0.05/2,11}\frac{2.96}{\sqrt{12}} \le \mu_D \le 0.67 + t_{0.05/2,11}\frac{2.96}{\sqrt{12}}$$

$$\Rightarrow \quad 0.67 - 2.201 \times 0.85 \le \mu_D \le 0.67 + 2.201 \times 0.85$$

$$\Rightarrow \quad -1.21 \le \mu_D \le 2.55$$

Since this 95% two-sided CI on μ_D includes the hypothesized value zero, fail to reject H_0 at $\alpha = 0.05$.

Exercise 10.4

1. **(Hypothesis Test on σ_1^2/σ_2^2; Two-Sided Test)**

Step 1: State H_0 and H_1.

H_0: $\sigma_1^2 = \sigma_2^2 \Rightarrow \sigma_1^2 / \sigma_2^2 = 1$

H_1: $\sigma_1^2 \neq \sigma_2^2 \Rightarrow \sigma_1^2 / \sigma_2^2 \neq 1$

Step 2: Determine a **test statistic and its value.**

$$f_0 = \frac{s_1^2}{s_2^2}\frac{\sigma_{2,0}^2}{\sigma_{1,0}^2} = \frac{0.42^2}{0.23^2} \times 1 = 3.33$$

Step 3: Determine a **critical value(s) for α.**

$$f_{\alpha/2,n_1-1,n_2-1} = f_{0.05/2,10-1,10-1} = f_{0.025,9,9} = 4.03 \text{ and}$$

$$f_{1-\alpha/2,n_1-1,n_2-1} = f_{1-0.05/2,9,9} = f_{0.975,9,9} = \frac{1}{f_{0.025,9,9}} = \frac{1}{4.03} = 0.25$$

Exercise 10.4 (*cont.*)

Step 4: Make a **conclusion.**

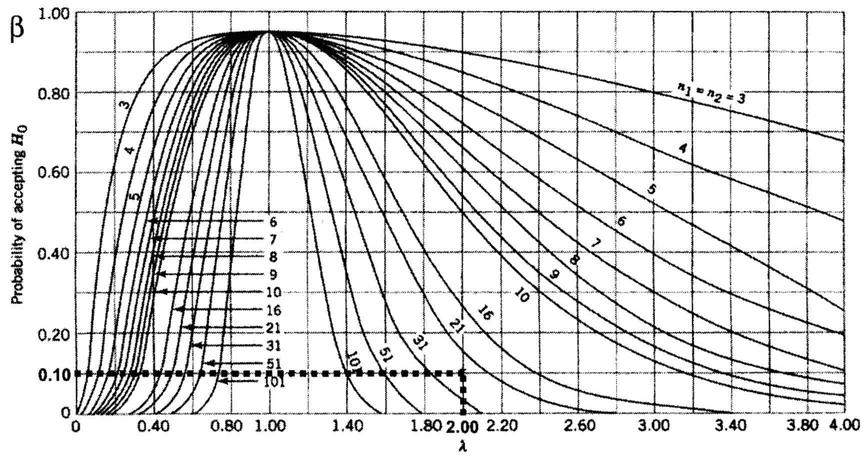
Since $f_0 = 3.33 \ngtr f_{0.025,9,9} = 4.03$ and $f_0 = 3.33 \nless f_{0.975,9,9} = 0.25$, fail to reject H_0 at $\alpha = 0.05$.

2. (**Sample Size Determination**)

To design a two-sided F-test at $\alpha = 0.05$, OC Chart VI*o* is applicable with the parameter

$$\lambda = \frac{\sigma_1}{\sigma_2} = 2.0$$

When $\lambda = 2.0$ and $\beta = 0.1$, the required sample size is determined as n $(= n_1 = n_2) = 25$ as displayed below.



(*o*) *O.C.* curves for different values of *n* for the two-sided *F*-test for a level of significance α = 0.05.

3. (**Confidence Interval on σ_1^2/σ_2^2; Two-Sided CI**)

$1 - \alpha = 0.95 \quad \Rightarrow \quad \alpha = 0.05$

95% two-sided CI on σ_1^2/σ_2^2:

$$\frac{s_1^2}{s_2^2} f_{1-\alpha/2, n_2-1, n_1-1} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{s_1^2}{s_2^2} f_{\alpha/2, n_2-1, n_1-1}$$

$$\Rightarrow \frac{0.42^2}{0.23^2} f_{1-0.05/2,9,9} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{0.42^2}{0.23^2} f_{0.05/2,9,9}$$

$$\Rightarrow 3.33 \frac{1}{f_{0.025,9,9}} \le \frac{\sigma_1^2}{\sigma_2^2} \le 3.33 f_{0.025,9,9}$$

$$\Rightarrow 3.33 \times \frac{1}{4.03} \le \frac{\sigma_1^2}{\sigma_2^2} \le 3.33 \times 4.03$$

$$\Rightarrow 0.83 \le \frac{\sigma_1^2}{\sigma_2^2} \le 13.43$$

Since this 95% two-sided CI on σ_1^2/σ_2^2 includes the hypothesized value unity, fail to reject H_0 at $\alpha = 0.05$.

Exercise 10.5

1. **(Hypothesis Test on $p_1 - p_2$, Equal Proportions; Two-Sided Test)**

The survey data is summarized as follows:

County	Sample Size	X	Sample Proportion
MARICOPA (1)	$n_1 = 500$	$x_1 = 385$	$\hat{p}_1 = 0.77$
PIMA (2)	$n_2 = 400$	$x_2 = 267$	$\hat{p}_2 = 0.67$

Since $n_1\hat{p}_1 = 500 \times 0.77 = 385$, $n_1(1-\hat{p}_1) = 500 \times 0.23 = 115$, $n_2\hat{p}_2 = 400 \times 0.67 = 267$, and $n_2(1-\hat{p}_2) = 400 \times 0.33 = 133$ are greater than five, the sampling distributions of $\hat{P}_1$ and $\hat{P}_2$ are approximately normal.

Step 1: State H_0 and H_1.

H_0: $p_1 - p_2 = 0$

H_1: $p_1 - p_2 \neq 0$

Step 2: Determine a **test statistic and its value.**

$$z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.77 - 0.67}{\sqrt{0.72 \times (1-0.72) \times \left(\frac{1}{500} + \frac{1}{400}\right)}} = 3.32$$

$$\text{(Note)}\ \hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{385 + 267}{500 + 400} = 0.72$$

Step 3: Determine a **critical value(s) for** α.

$$z_{\alpha/2} = z_{0.05/2} = z_{0.025} = 1.96$$

Step 4: Make a **conclusion.**

Since $|z_0| = 3.32 > z_{0.025} = 1.96$, reject H_0 at $\alpha = 0.05$.

2. **(Sample Size Determination)**

power = P(reject H_0 | H_0 is false) = 1 - β = 0.8 ⇒ β = 0.2

$$n = \left(\frac{z_{\alpha/2}\sqrt{(p_1 + p_2)(q_1 + q_2)/2} + z_\beta\sqrt{p_1q_1 + p_2q_2}}{p_1 - p_2}\right)^2$$

$$= \left(\frac{z_{0.05/2}\sqrt{(0.75 + 0.65)(0.25 + 0.35)/2} + z_{0.02}\sqrt{0.75 \times 0.25 + 0.65 \times 0.35}}{0.75 - 0.65}\right)^2$$

$$= \left(\frac{1.96 \times 0.65 + 0.84 \times 0.64}{0.1}\right)^2 \cong 329$$

Exercise 10.5 (*cont.*)

3. **(Confidence Interval on p_1 - p_2; Equal Proportions; Two-Sided CI)**

$$1 - \alpha = 0.95 \quad \Rightarrow \quad \alpha = 0.05,\ \hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{385 + 267}{500 + 400} = 0.72$$

95% two-sided CI on p_1 - p_2:

$$\hat{p}_1 - \hat{p}_2 - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$\le p_1 - p_2 \le \hat{p}_1 - \hat{p}_2 + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$\Rightarrow \quad 0.77 - 0.67 - z_{0.05/2}\sqrt{0.72(1-0.72)\left(\frac{1}{500} + \frac{1}{400}\right)}$$

$$\le p_1 - p_2 \le 0.77 - 0.67 + z_{0.05/2}\sqrt{0.72(1-0.72)\left(\frac{1}{500} + \frac{1}{400}\right)}$$

$$\Rightarrow \quad 0.1 - 1.96 \times 0.03 \le p_1 - p_2 \le 0.1 + 1.96 \times 0.03$$

$$\Rightarrow \quad 0.04 \le p_1 - p_2 \le 0.16$$

Since this 95% two-sided CI on p_1 - p_2 does not include the hypothesized value zero, reject H_0 at $\alpha = 0.05$.

11 Simple Linear Regression and Correlation

OUTLINE

11-1 Empirical Models

Learning Goals

- ❑ Describe the use of regression analysis.
- ❑ Explain the terms *intercept, slope, independent variable, dependent variable,* and *random error.*
- ❑ Explain the assumptions of a simple linear model.

Regression Analysis Regression analysis is a statistical technique to model the relationship between two or more variables so that we can predict the response of a variable at a given condition of the other variable(s).

Regression Analysis (*cont.*)

For example, in Figure 11-1, the scatter diagram of price versus sales for a plywood product indicates that the volume of sales linearly decreases as the price of the product increases. By modeling this linear relationship, we may want to know how many pieces of plywood would be sold if the price is set at \$8.50. It may not be easy to answer the question because there is no simple, linear curve which passes through all the points in the diagram—the data randomly scatter along a straight line.

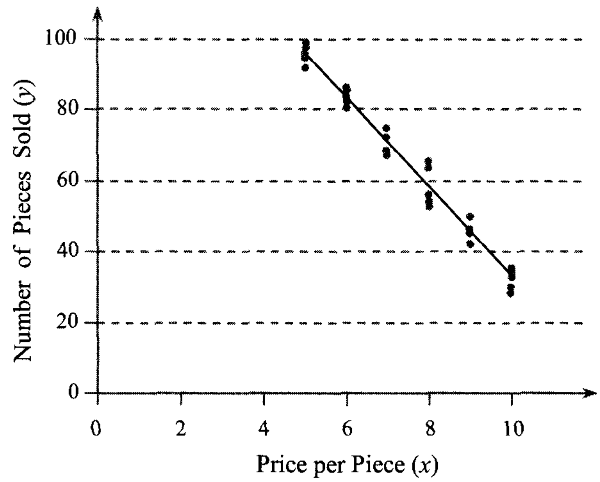


Figure 11-1 Scatter diagram of plywood price versus sales.

Simple Linear Model

A simple linear model which explains the linear relationship between two variables (x and Y) is

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

where: (1) **Regression coefficients (β_0 and β_1)**: The coefficients β_0 and β_1 denote the intercept and slope of the regression line, respectively. The slope β_1 measures the expected change in Y for a unit change in x.

(2) **Independent variable (x; regressor, predictor)**: Since the independent variable x is under control with negligible error in an experiment, x represents controlled constants (not random outcomes).

(3) **Random error (ε)**: The variable ε represents the random variation of Y around the straight regression line $\beta_0 + \beta_1 x$. It is assumed that ε is normally distributed with mean 0 and constant variance σ^2, i.e.,

$$\varepsilon \sim N(0, \sigma^2)$$

(4) **Dependent (response) variable (Y)**: Since outcomes in Y at the same value of x can vary randomly, Y is a random variable with the following distribution:

$$Y \sim N(\beta_0 + \beta_1 x, \sigma^2)$$

In summary, four assumptions are made for the simple linear regression model:

1. **Linear relationship between x and Y**: The variables x and Y are linearly related.
2. **Randomness of error**: The errors (spreads of Y along the regression line) are random (not following any particular pattern).

Simple Linear Model (*cont.*)

3. **Constant σ^2**: The variance of the random errors is constant as σ^2. In other words, the spreads of Y along the regression line are the same for different values of x (see Figure 11-2).
4. **Normality of error**: The random errors are normally distributed.

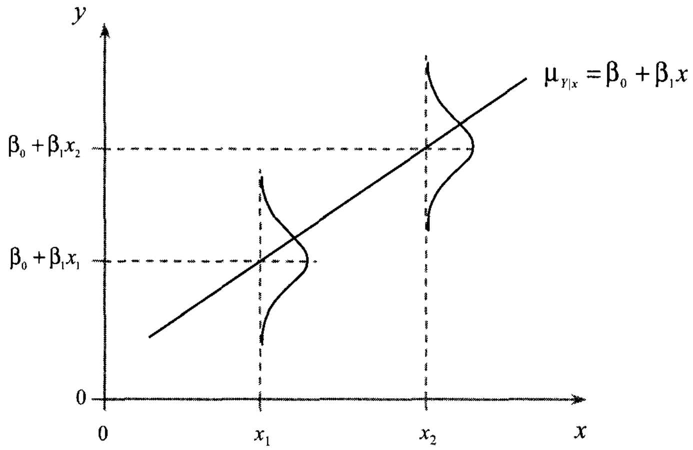


Figure 11-2 The variability of Y over x.

The analyst should check if the assumptions of the linear model are met for proper regression analysis, which is called **assessment of model adequacy** (presented in Section 11-8).

11-2 Simple Linear Regression

Learning Goals

- ❑ Explain the method of least squares.
- ❑ Estimate regression coefficients by using the least squares method.
- ❑ Calculate mean responses and residuals by using a fitted regression line.
- ❑ Estimate error variance σ^2.

Least Squares Method

In the 1800s Karl Gauss proposed 'the method of least squares' to estimate the regression coefficients of a linear model. Suppose that we have n pairs of observations (x_i, y_i), $i = 1, 2, \ldots, n$, and the n observations are modeled by a simple linear model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

The least squares method finds the estimates of β_0 and β_1 which minimize the **sum of the squares of the errors** (denoted by SS_E; called **error sum of squares**; sum of the squares of the vertical deviations of data from the regression line in Figure 11-3)

$$SS_E = L = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Least Squares Method (*cont.*)

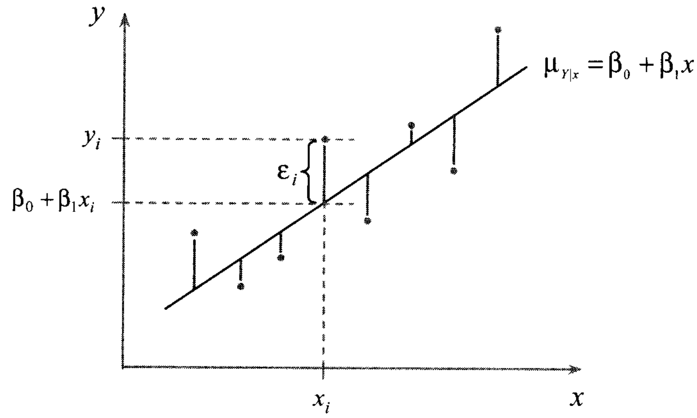


Figure 11-3 Deviations (ε_i) of data from a regression line.

The least squares estimators of β_0 and β_1 are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{\sum_{i=1}^{n} y_i (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x}$$

(Derivation) Least square estimators $\hat{\beta}_0$ and $\hat{\beta}_1$

Recall that, in calculus, when we want to find the value of a variable which minimizes (or maximizes) a certain function, the following procedures are used:

(1) Take the derivative of the function with respect to the variable designated,
(2) Set the derivative as equal to zero, and
(3) Solve the equation.

Since the function SS_E (or L) has two unknowns (β_0 and β_1), we need to take a partial derivative with respect to each regression coefficient as follows:

$$\begin{cases} \dfrac{\partial L}{\partial \beta_0} = -2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \Rightarrow n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \dfrac{\partial L}{\partial \beta_1} = -2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \Rightarrow \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i \end{cases} \tag{11-1}$$

Least Squares Method (*cont.*)

By multiplying the first equation above by $\sum_{i=1}^{n} x_i$ and the last equation by n each,

$$\begin{cases} n\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \left(\sum_{i=1}^{n} x_i\right)^2 = \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) \\ n\hat{\beta}_0 \sum_{i=1}^{n} x_i + n\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = n\sum_{i=1}^{n} y_i x_i \end{cases} \qquad (11\text{-}2)$$

By subtracting the last equation above from the first equation above,

$$\hat{\beta}_1 \left[\left(\sum_{i=1}^{n} x_i\right)^2 - n\sum_{i=1}^{n} x_i^2\right] = \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) - n\sum_{i=1}^{n} y_i x_i \qquad (11\text{-}3)$$

Thus,

$$\hat{\beta}_1 = \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) - n\sum_{i=1}^{n} y_i x_i}{\left(\sum_{i=1}^{n} x_i\right)^2 - n\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{S_{xy}}{S_{xx}}$$

where: $$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$S_{xy} = \sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n} = \sum_{i=1}^{n} y_i (x_i - \bar{x})^2$$

Lastly, from Equation (11-2),

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x}$$

Fitted Regression Model

By using the least squares estimates of β_0 and β_1, the fitted (estimated) regression line is determined as follows:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

By applying the fitted regression line, the mean response at x_i is estimated. Note that an actual response y_i can be different from the corresponding predicted response $\hat{y}_i$. The difference between the actual and predicted responses is called the **residual** (denoted as e_i), which is an estimate of the random error ε_i:

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \ldots, n$$

Estimation of Error Variance (σ^2)

Since the error variance σ^2 is unknown in most cases, estimation on σ^2 is needed. The point estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2} = \frac{SS_E}{n-2} = \frac{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2 - \hat{\beta}_1 S_{xy}}{n-2}$$

(Derivation) $SS_E = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 - \hat{\beta}_1 S_{xy}$

$$SS_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right]^2 \quad \text{because } \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$= \sum_{i=1}^{n} \left[y_i - (\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i)\right]^2 \quad \text{because } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$= \sum_{i=1}^{n} \left[(y_i - \bar{y}) - \hat{\beta}_1 (\bar{x} - x_i)\right]^2$$

$$= \sum_{i=1}^{n} \left[(y_i - \bar{y})^2 - 2\hat{\beta}_1 (y_i - \bar{y})(\bar{x} - x_i) + \hat{\beta}_1^2 (\bar{x} - x_i)^2\right]$$

$$= \sum_{i=1}^{n} (y_i - \bar{y})^2 - 2\hat{\beta}_1 \sum_{i=1}^{n} (y_i - \bar{y})(\bar{x} - x_i) + \hat{\beta}_1^2 \sum_{i=1}^{n} (\bar{x} - x_i)^2$$

$$= SS_T - 2\hat{\beta}_1 \sum_{i=1}^{n} y_i (\bar{x} - x_i) + 2\hat{\beta}_1 \bar{y} \sum_{i=1}^{n} (\bar{x} - x_i) + \hat{\beta}_1^2 S_{xx}$$

(Notes) $SS_T = S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$ (called **total sum of squares**)

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

$$= SS_T - 2\hat{\beta}_1 S_{xy} + \hat{\beta}_1^2 S_{xx} \quad \text{because } S_{xy} = \sum_{i=1}^{n} y_i (\bar{x} - x_i) \text{ and } \sum_{i=1}^{n} (\bar{x} - x_i) = 0$$

$$= SS_T - 2\hat{\beta}_1 S_{xy} + \hat{\beta}_1 \frac{S_{xy}}{S_{xx}} S_{xx} \quad \text{because } \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

$$= SS_T - 2\hat{\beta}_1 S_{xy} + \hat{\beta}_1 S_{xy}$$

$$= SS_T - \hat{\beta}_1 S_{xy}$$

$$= \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 - \hat{\beta}_1 S_{xy} \quad \text{because } SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

Example 11.1

The sales' volumes (Y) of a plywood product at various prices (x) are surveyed as follows (sorted by price):

No.	Price (x; unit: \$)	Sales Volume (Y; unit: ea.)	No.	Price (x; unit: \$)	Sales Volume (Y; unit: ea.)
1	5	95	16	8	63
2	5	98	17	8	65
3	5	96	18	8	54
4	5	92	19	8	53
5	5	99	20	8	56
6	6	80	21	9	45
7	6	84	22	9	50
8	6	85	23	9	46
9	6	82	24	9	45
10	6	86	25	9	42
11	7	67	26	10	30
12	7	68	27	10	28
13	7	72	28	10	35
14	7	75	29	10	33
15	7	72	30	10	34

The summary quantities of the plywood price-sales data are

$n = 30$, $\sum y_i = 1{,}930$, $\sum y_i^2 = 138{,}656$, $\sum x_i = 225$, $\sum x_i^2 = 1{,}775$, and $\sum x_i y_i = 13{,}360$

Assume that x and Y are linearly related.

1. (**Estimation of Regression Coefficients**) Calculate the least squares estimates of the intercept and slope of the linear model for x and Y.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)/n}{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2/n} = \frac{S_{xy}}{S_{xx}}$$

$$= \frac{13{,}360 - 1{,}930 \times 225/30}{1{,}775 - 225^2/30} = \frac{-1{,}115}{87.5} = -12.7$$

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i}{n} = \frac{1{,}930 - (-12.7) \times 225}{30} = 159.9$$

The fitted regression line of the plywood price-sales data is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 159.9 - 12.7x$$

2. (**Calculation of Residual**) Calculate the residual of $y = 84$ at $x = \$6$.

The predicted sales at $x = \$6$ is

$$\hat{y}_{x=6} = \hat{\beta}_0 + \hat{\beta}_1 x = 159.9 - 12.7 \times 6 = 83.4$$

Thus,

$$e_{x=6} = y_{x=6} - \hat{y}_{x=6} = 84 - 83.4 = 0.6 \text{ (underestimate)}$$

Example 11.1 ***(cont.)***

3. (**Estimation of σ^2**) Estimate the error variance σ^2.

$$SS_E = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 - \hat{\beta}_1 S_{xy} = 138{,}656 - 30 \times \left(\frac{1{,}930}{30}\right)^2 - (-12.7) \times (-1{,}115)$$
$$= 284.4$$

$$\hat{\sigma}^2 = \frac{SS_E}{n-2} = \frac{284.4}{30-2} = 10.2 = 3.2^2$$

Exercise 11.1 **(MR 11-9)**

An article in the *Journal of Sound and Vibration* (Vol. 151, 1991, pp. 383-394) described a study investigating the relationship between noise exposure (x) and hypertension (Y). The following data (sorted by sound pressure level) are representative of those reported in the article:

No.	Sound Pressure Level (x; unit: dB)	Blood Pressure Rise (Y; unit: mmHg)	No.	Sound Pressure Level (x; unit: dB)	Blood Pressure Rise (Y; unit: mmHg)
1	60	1	11	85	5
2	63	0	12	89	4
3	65	1	13	90	6
4	70	2	14	90	8
5	70	5	15	90	4
6	70	1	16	90	5
7	80	4	17	94	7
8	80	6	18	100	9
9	80	2	19	100	7
10	80	3	20	100	6

The summary quantities of the noise-hypertension data are

$n = 20$, $\sum y_i = 86$, $\sum y_i^2 = 494$, $\sum x_i = 1{,}646$, $\sum x_i^2 = 138{,}476$, and $\sum x_i y_i = 7{,}594$

Assume that x and Y are linearly related.

1. Calculate the least squares estimates of the intercept and slope of the linear model for x and Y.
2. Calculate the residual of y = 5 mmHg at x = 85 dB.
3. Estimate the error variance σ^2.

11-3 Properties of the Least Squares Estimators

Learning Goals

- ❑ Identify the sampling distributions of the least squares slope and intercept estimators.
- ❑ Check if the least square estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators.

Sampling Distribution of Slope Estimator ($\hat{\beta}_1$)

The sampling distribution of the least squares slope estimator $\hat{\beta}_1$ is

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

(Note) **Estimated standard error** of $\hat{\beta}_1 = \sqrt{\dfrac{\hat{\sigma}^2}{S_{xx}}}$

(Derivation) Sampling distribution of $\hat{\beta}_1$

Recall that the least squares estimator of β_1 is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

Also recall that $Y = \beta_0 + \beta_1 x + \varepsilon \sim N(\beta_0 + \beta_1 x, \sigma^2)$, where $\varepsilon \sim N(0, \sigma^2)$, and x represents constants. Since $\hat{\beta}_1$ is a linear combination of the normal random variables y_i's, $\hat{\beta}_1$ is <u>normally</u> distributed with the following mean and variance:

$$\begin{aligned}
E(\hat{\beta}_1) &= E\left(\frac{S_{xy}}{S_{xx}}\right) = \frac{1}{S_{xx}}E(S_{xy}) = \frac{1}{S_{xx}}E\left[\sum_{i=1}^{n} y_i(x_i - \bar{x})\right] \\
&= \frac{1}{S_{xx}}E\left[\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i + \varepsilon_i)(x_i - \bar{x})\right] \\
&= \frac{1}{S_{xx}}\left\{E\left[\sum_{i=1}^{n}\beta_0(x_i - \bar{x})\right] + E\left[\sum_{i=1}^{n}\beta_1 x_i(x_i - \bar{x})\right] + E\left[\sum_{i=1}^{n}\varepsilon_i(x_i - \bar{x})\right]\right\} \\
&= \frac{1}{S_{xx}}\left\{\beta_0\sum_{i=1}^{n}(x_i - \bar{x}) + \beta_1\sum_{i=1}^{n}x_i(x_i - \bar{x}) + \sum_{i=1}^{n}(x_i - \bar{x})E(\varepsilon_i)\right\} \\
&= \frac{1}{S_{xx}}\left\{\beta_0\sum_{i=1}^{n}(x_i - \bar{x}) + \beta_1\sum_{i=1}^{n}x_i(x_i - \bar{x})\right\} \quad \text{because } E(\varepsilon_i) = 0 \\
&= \frac{1}{S_{xx}}\beta_1\sum_{i=1}^{n}x_i(x_i - \bar{x}) \quad \text{because } \sum_{i=1}^{n}(x_i - \bar{x}) = 0 \\
&= \frac{1}{S_{xx}}\beta_1 S_{xx} = \beta_1
\end{aligned}$$

$$\begin{aligned}
V(\hat{\beta}_1) &= V\left(\frac{S_{xy}}{S_{xx}}\right) = \frac{1}{S_{xx}^2}V(S_{xy}) = \frac{1}{S_{xx}^2}V\left[\sum_{i=1}^{n} y_i(x_i - \bar{x})\right] \\
&= \frac{1}{S_{xx}^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 V(y_i) = \frac{1}{S_{xx}^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\sigma^2 = \frac{1}{S_{xx}^2}S_{xx}\sigma^2 = \frac{\sigma^2}{S_{xx}}
\end{aligned}$$

(Note) Since $E(\hat{\beta}_1) = \beta_1$, $\hat{\beta}_1$ is an unbiased estimator of β_1.

Sampling Distribution of Intercept Estimator ($\hat{\beta}_0$)

The sampling distribution of the least squares intercept estimator $\hat{\beta}_0$ is

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]\right)$$

(Note) **Estimated standard error** of $\hat{\beta}_0 = \sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]}$

(Derivation) Sampling distribution of $\hat{\beta}_0$

Recall that the least squares estimator of β_0 is

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x}$$

Also recall that $Y \sim N(\beta_0 + \beta_1 x, \sigma^2)$, $\hat{\beta}_1 \sim N(\beta_1, \sigma^2 / S_{xx})$, and x denotes constants. Since $\hat{\beta}_0$ is a linear combination of the normal random variables y_i's and $\hat{\beta}_1$, $\hat{\beta}_0$ is normally distributed with the following mean and variance:

$$E(\hat{\beta}_1) = E(\bar{y} - \hat{\beta}_1\bar{x}) = E(\bar{y}) - \bar{x}E(\hat{\beta}_1) = E(\beta_0 + \beta_1\bar{x}) - \beta_1\bar{x}$$
$$= \beta_0 + \beta_1\bar{x} - \beta_1\bar{x} = \beta_0$$

$$V(\hat{\beta}_0) = V(\bar{y} - \hat{\beta}_1\bar{x}) = V(\bar{y}) + \bar{x}^2 V(\hat{\beta}_1) - 2\bar{x}\,\mathrm{cov}(\bar{y}, \hat{\beta}_1) = V(\bar{y}) + \bar{x}^2 V(\hat{\beta}_1)$$

(Note) $\mathrm{cov}(\bar{y}, \hat{\beta}_1) = 0$ because $\bar{y}$ and $\hat{\beta}_1$ are independent.

$$= \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{S_{xx}} = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]$$

(Note) Since $E(\hat{\beta}_0) = \beta_0$, $\hat{\beta}_0$ is an unbiased estimator of β_0.

11-4 Some Comments on Uses of Regression

Learning Goals

- ❑ Explain common misuses of regression.

Common Misuses of Regression

Care should be taken on the following to avoid technical misuse of regression:

1. **Practical validity of regressor**: A statistical significance can be found on the relationship between independent (x) and dependent (Y) variables, although they are not related from a practical sense. Therefore, the practical significance of the relationship among the variables should be checked separately.

Common Misuses of Regression (*cont.*)

2. **Empirical/theoretical validity of regressor**: The direction and/or magnitude of a regression coefficient in the model can contradict common understanding of the relationship between the variables. Therefore, it should be checked if the sign and magnitude of each regression coefficient are reasonable as compared with previous findings.

3. **Association of regressor with predictor**: A significant relationship between the regressor and predictor in the model does not necessarily indicate a cause-effect relationship between the variables. Designed experiments are the only way to determine cause-effect relationships between variables.

4. **Limited generalizability of the model**: The regression relationship established between the variables may be valid only over the range(s) of the regressor(s) examined in the study. To apply the regression model beyond the original range(s) of the regressor(s), a new study should be designed for the extrapolation purpose.

11-5 Hypothesis Tests in Simple Linear Regression

11-5.1 Use of *t*-Tests

Learning Goals

- ☐ Conduct hypothesis tests on β_1 and β_0 each.

Inference Context on Slope (β_1)

Parameter of interest: β_1

Point estimator of β_1: $\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}} \sim N\left(\beta_1, \dfrac{\sigma^2}{S_{xx}}\right)$, σ^2 is unknown.

Test statistic of β_1: $T_0 = \dfrac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2 / S_{xx}}} \sim t(n-2)$

Hypothesis Test on β_1 (*t*-test)

Step 1: State H_0 and H_1.

H_0: $\beta_1 = \beta_{1,0}$

H_1: $\beta_1 \neq \beta_{1,0}$

Step 2: Determine a **test statistic and its value.**

$$T_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2 / S_{xx}}} \sim t(n-2)$$

Step 3: Determine a **critical value(s) for α.**

$$t_{\alpha/2,n-2}$$

Step 4: Make a **conclusion**. Reject H_0 if

$$|t_0| > t_{\alpha/2,n-2}$$

The Significance of Regression

It is typical to test if the regressor x is significant to explain the variability in Y by using H_0: $\beta_1 = 0$, which is called **testing the significance of regression.**

Note that failure to reject H_0: $\beta_1 = 0$ may imply that

(1) The regressor x is of little value in explaining the variability in Y, or
(2) The true relationship between x and Y is not linear.

On the other hand, rejection of H_0: $\beta_1 = 0$ may imply that

(1) The regressor x is of value in explaining the variability in Y,
(2) The simple linear model is adequate, or
(3) A better model could be obtained by adding a higher order term(s) of x while a linear effect of x on Y remains significant.

Inference Context on Intercept (β_0)

Parameter of interest: β_0

Point estimator of β_0: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \sim N\left(\beta_0, \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]\right)$, σ^2 is unknown.

Test statistic of β_0: $T_0 = \dfrac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}} \sim t(n-2)$

Hypothesis Test on β_0 (t-test)

Step 1: State H_0 and H_1.

H_0: $\beta_0 = \beta_{0,0}$
H_1: $\beta_0 \neq \beta_{0,0}$

Step 2: Determine a **test statistic and its value.**

$$T_0 = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}} \sim t(n-2)$$

Step 3: Determine a **critical value(s) for α.**

$t_{\alpha/2,n-2}$

Step 4: Make a **conclusion.** Reject H_0 if

$|t_0| > t_{\alpha/2,n-2}$

Example 11.2

For the plywood price-sales data in Example 11-1, the following summary quantities have been calculated:

$n = 30$, $\hat{\sigma}^2 = 3.2^2$, $\sum x_i = 225$, $S_{xx} = 87.5$, $\hat{\beta}_1 = -12.7$, and $\hat{\beta}_0 = 159.9$

1. (**Hypothesis Test on β_1**) Test H_0: $\beta_1 = 0$ at $\alpha = 0.05$.

Example 11.2 (*cont.*)

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $\beta_1 = 0$ $\qquad$ H_1: $\beta_1 \neq 0$

Step 2: Determine a **test statistic and its value.**

$$t_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2 / S_{xx}}} = \frac{-12.7 - 0}{\sqrt{3.2^2 / 87.5}} = -37.4$$

Step 3: Determine a **critical value(s) for α.**

$$t_{\alpha/2,n-2} = t_{0.025,28} = 2.05$$

Step 4: Make a **conclusion.**

Since $|t_0| = 37.4 > t_{\alpha/2,n-2} = 2.05$, reject H_0 at $\alpha = 0.05$.

2. (**Hypothesis Test on $\boldsymbol{\beta_0}$**) Test H_0: $\beta_0 = 0$ at $\alpha = 0.05$.

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $\beta_0 = 0$

H_1: $\beta_0 \neq 0$

Step 2: Determine a **test statistic and its value.**

$$t_0 = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]}} = \frac{159.9 - 0}{\sqrt{3.2^2 \left[\frac{1}{30} + \frac{(225/30)^2}{87.5} \right]}} = 61.0$$

Step 3: Determine a **critical value(s) for α.**

$$t_{\alpha/2,n-2} = t_{0.025,28} = 2.05$$

Step 4: Make a **conclusion.**

Since $|t_0| = 61.0 > t_{\alpha/2,n-2} = 2.05$, reject H_0 at $\alpha = 0.05$.

Exercise 11.2 (MR 11-25)

For the noise-hypertension data in Exercise 11-1, the following summary quantities have been calculated:

$$n = 20,\ \hat{\sigma}^2 = 1.4^2,\ \sum x_i = 1{,}646,\ S_{xx} = 3{,}010.2,\ \hat{\beta}_1 = 0.17, \text{ and } \hat{\beta}_0 = -9.81$$

1. Test H_0: $\beta_1 = 0$ at $\alpha = 0.05$.
2. Test H_0: $\beta_0 = 0$ at $\alpha = 0.05$.

11-5.2 Analysis of Variance Approach to Test Significance of Regression

Learning Goals

- ❑ Explain the analysis of variance (ANOVA) technique.
- ❑ Describe the segmentation of the total variability of a response variable (SS_T) into the regression sum of squares (SS_R) and error sum of squares (SS_E).
- ❑ Establish an ANOVA table and conduct a hypothesis test on β_1.

Analysis of Variance (ANOVA)

The analysis of variance (ANOVA) technique provides a statistical procedure to evaluate the contribution of a variable (X_i) to the variability in the response variable (Y). As illustrated in Figure 11-4, the ANOVA method divides the total variance of Y (SS_T) into the segments (SS_{X_i}'s) explained by variables X_i's and the segment unexplained (SS_E), implying that the larger the size of the segment SS_{X_i}, the higher the contribution of X_i to the total variability in Y.

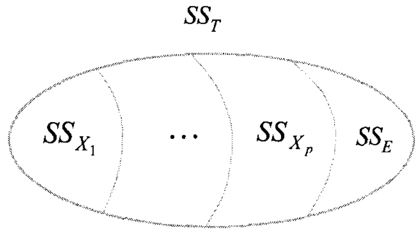


Figure 11-4 Segmentation of the total variability in Y (SS_T).

ANOVA for Test on β_1

The ANOVA method can be used to test the significance of regression (H_0: $\beta_1 = 0$). For a simple linear model, as shown in Table 11-1, the total variability in Y (**SS_T or S_{yy}, total sum of squares**; degrees of freedom = $n-1$) is partitioned into two components:

(1) Variability explained by the regression model (**SS_R, regression sum of squares**; degrees of freedom =1)

(2) Variability due to random error (**SS_E, error sum of squares**; degrees of freedom = $n-2$)

Table 11-1 Partitioning of the Total Variability in Y (SS_T)

	SS_T (Total SS)		SS_R (Regression SS)		SS_E (Error SS)
	$\sum_{i=1}^{n}(y_i-\bar{y})^2$	=	$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$	+	$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$
DF	$n-1$	=	1	+	$n-2$

(Note) SS: Sum of Squares; DF: Degrees of Freedom

A graphical interpretation of the ANOVA quantities is provided in Figure 11-5. The deviation of an observation y_i from the corresponding mean $\bar{y}$ is divided into the y_i-deviation explained by the regression line and the y_i-deviation due to error. These deviation are squared and then added to those of the other observations to calculate SS_T, SS_R, and SS_E as follows:

$$SS_T = \sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

$$SS_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 = SS_T - \hat{\beta}_1 S_{xy} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 - \hat{\beta}_1 S_{xy}$$

$$SS_R = SS_T - SS_E = SS_T - (SS_T - \hat{\beta}_1 S_{xy}) = \hat{\beta}_1 S_{xy}$$

ANOVA for Test on β_1 (*cont.*)

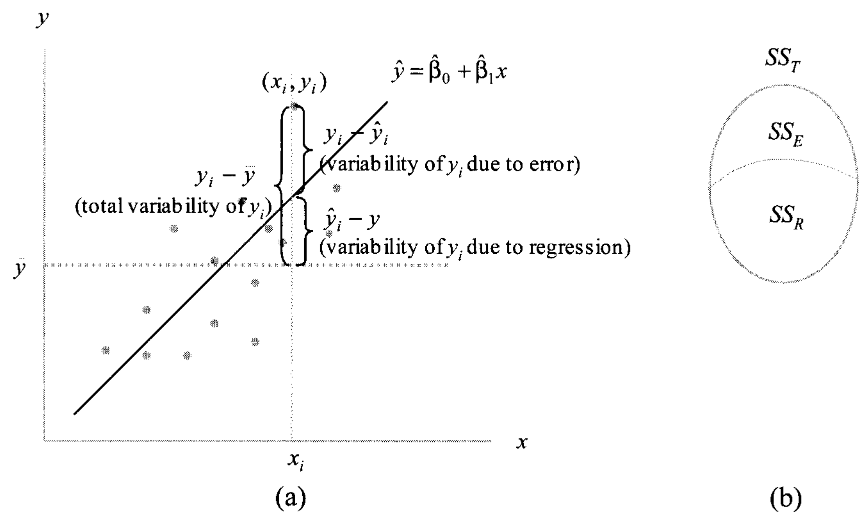


Figure 11-5 Partitioning of the total variability of y.

The ratio of $SS_R/1$ to $SS_E/(n-2)$ follows an F distribution:

$$F = \frac{SS_R/1}{SS_E/(n-2)} = \frac{MS_R}{MS_E} \sim F(1, n-2)$$

This statistic F is used to test H_0: $\beta_1 = 0$, because F becomes small as $\beta_1 = 0$ is true. The quantities MS_R **(regression mean square)** and MS_E **(error mean square)** are adjusted SS_R and SS_E by their degrees of freedom, respectively. Note that

$$MS_E = \frac{SS_E}{n-2} = \hat{\sigma}^2$$

ANOVA Table

The variation quantities from the ANOVA analysis are summarized in Table 11-2 to test the significance of regression.

Table 11-2 ANOVA Table for Testing the Significance of Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	$SS_R = \hat{\beta}_1 S_{xy}$	1	MS_R	MS_R / MS_E
Error	$SS_E = SS_T - SS_R$	$n-2$	MS_E	
Total	$SS_T = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$	$n-1$		

Hypothesis Test on β_1 (*F*-test)

Step 1: State H_0 and H_1.

H_0: $\beta_1 = 0$
H_1: $\beta_1 \neq 0$

Step 2: Determine a **test statistic and its value.**

$$F_0 = \frac{SS_R/1}{SS_E/(n-2)} = \frac{MS_R}{MS_E} \sim F(1, n-2)$$

Hypothesis Test on β_1 (*F*-test) (*cont.*)

Step 3: Determine a **critical value(s) for α.**

$$f_{\alpha,1,n-2}$$

Step 4: Make a **conclusion**. Reject H_0 if

$$f_0 > f_{\alpha,1,n-2}$$

Relationship between *t*- and *F*-tests

In testing the significance of regression, the ANOVA test procedure is equivalent to the t-test procedure in Section 11-5.1, because the square of the t-test statistic becomes the F-test statistic when $\beta_1 = 0$. Therefore, use of either test procedure will lead to the same conclusion on β_1.

(Derivation) $T^2 = F$ when $\beta_1 = 0$

$$T^2 = \frac{\hat{\beta}_1^2 - \beta_1}{\hat{\sigma}^2 / S_{xx}} = \frac{\hat{\beta}_1^2 S_{xx}}{\hat{\sigma}^2} = \frac{\hat{\beta}_1 \cdot \hat{\beta}_1 S_{xx}}{\hat{\sigma}^2}$$

$$= \frac{\hat{\beta}_1 S_{xy}}{\hat{\sigma}^2} \quad \text{because } S_{xy} = \hat{\beta}_1 S_{xx} \ (\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}})$$

$$= \frac{MS_R}{MS_E} \quad \text{because } \hat{\beta}_1 S_{xy} = SS_R = MS_R \text{ and } \hat{\sigma}^2 = MS_E$$

$$= F$$

Note that the t-test is more <u>flexible</u> than the F-test in testing a hypothesis on β_1, because the t-test can have either a one- or two-sided alternative hypothesis whereas the F-test can have only a two-sided alternative hypothesis. However, the F-test is used to evaluate the <u>relative contributions</u> of multiple regressors to the variability in Y (see Section 12-2).

Example 11.3

(ANOVA for Hypothesis Test on β_1) For the plywood price-sales data in Example 11-1, the following summary quantities have been calculated:

$$n = 30,\ \sum y_i = 1{,}930,\ \sum y_i^2 = 138{,}656,\ S_{xy} = -1{,}115, \text{ and } \hat{\beta}_1 = -12.7$$

Establish an ANOVA table and conduct a hypothesis test on β_1 at $\alpha = 0.05$.

$$SS_T = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = 138{,}656 - 30 \times \left(\frac{1{,}930}{30}\right)^2 = 14{,}492.7$$

$$SS_R = \hat{\beta}_1 S_{xy} = -12.7 \times (-1{,}115) = 14{,}208.3$$

$$SS_E = SS_T - SS_R = 14{,}492.7 - 14{,}208.3 = 284.4$$

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	f_0
Regression	14,208.3	1	14,208.3	1398.9
Error	284.4	28	10.2	
Total	14,492.7	29		

(Note) $t_0^2 = (-37.4)^2 = 1398.9 = f_0$

Example 11.3 (*cont.*)

Step 1: State H_0 and H_1.

H_0: $\beta_1 = 0$
H_1: $\beta_1 \neq 0$

Step 2: Determine a **test statistic and its value.**

$$f_0 = \frac{SS_R/1}{SS_E/(n-2)} = \frac{MS_R}{MS_E} = \frac{14{,}208.3}{10.2} = 1{,}398.9$$

Step 3: Determine a **critical value(s) for α.**

$$f_{\alpha,1,n-2} = f_{0.05,1,30-2} = 4.20$$

Step 4: Make a **conclusion.**

Since $f_0 = 1{,}398.9 > f_{\alpha,1,n-2} = 4.20$, reject H_0 at $\alpha = 0.05$.

Exercise 11.3 (MR 11-25)

For the noise-hypertension data in Exercise 11-1, the following summary quantities have been calculated:

$$n = 20,\ \sum y_i = 86,\ \sum y_i^2 = 494,\ S_{xy} = 516.2, \text{ and } \hat{\beta}_1 = 0.17$$

Establish an ANOVA table and conduct a hypothesis test on β_1 at $\alpha = 0.05$.

11-6 Confidence Intervals

11-6.1 Confidence Intervals on the Slope and Intercept

Learning Goals

- ☐ Establish confidence intervals on β_1 and β_0 each.

Inference Context on Slope (β_1)

Parameter of interest: β_1

Point estimator of β_1: $\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}} \sim N\left(\beta_1, \dfrac{\sigma^2}{S_{xx}}\right)$, σ^2 is unknown.

(Note) **Estimated standard error** of $\hat{\beta}_1 = \sqrt{\dfrac{\hat{\sigma}^2}{S_{xx}}}$

Confidence Interval on β_1

A 100(1 - α)% confidence interval on β_1 is

$$\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2/S_{xx}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2/S_{xx}}$$

Inference Context on Intercept (β_0)

Parameter of interest: β_0

Point estimator of β_0: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \sim N\left(\beta_0, \sigma^2\left[\dfrac{1}{n} + \dfrac{\bar{x}^2}{S_{xx}}\right]\right)$, σ^2 is unknown.

(Note) **Estimated standard error** of $\hat{\beta}_0 = \sqrt{\hat{\sigma}^2\left[\dfrac{1}{n} + \dfrac{\bar{x}^2}{S_{xx}}\right]}$

Confidence Interval on β_0

A 100(1 - α)% confidence interval on β_0 is

$$\hat{\beta}_0 - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]} \le \beta_0 \le \hat{\beta}_0 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]}$$

Example 11.4

For the plywood price-sales data in Example 11-1, the following summary quantities have been calculated:

$$n = 30,\ \sum x_i = 225,\ \hat{\sigma}^2 = 3.2^2,\ S_{xx} = 87.5,\ \hat{\beta}_1 = -12.7, \text{ and } \hat{\beta}_0 = 159.9$$

1. (**Confidence Interval on β_1**) Establish a 95% two-sided confidence interval on the slope.

$1-\alpha = 0.95 \Rightarrow \alpha = 0.05$

95% two-sided CI on β_1:

$$\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2 / S_{xx}} \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2 / S_{xx}}$$

$$\Rightarrow -12.7 - t_{0.05/2,30-2}\sqrt{3.2^2/87.5} \le \beta_1 \le -12.7 + t_{0.05/2,30-2}\sqrt{3.2^2/87.5}$$

$$\Rightarrow -12.7 - 2.05\times 0.34 \le \beta_1 \le -12.7 + 2.05\times 0.34$$

$$\Rightarrow -13.4 \le \beta_1 \le -12.0$$

2. (**Confidence Interval on β_0**) Establish a 95% two-sided confidence interval on the intercept.

$1-\alpha = 0.95 \Rightarrow \alpha = 0.05$

95% two-sided CI on β_0:

$$\hat{\beta}_0 - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]} \le \beta_0 \le \hat{\beta}_0 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]}$$

$$\Rightarrow 159.9 - t_{0.05/2,30-2}\sqrt{3.2^2\left[\frac{1}{30}+\frac{(225/30)^2}{87.5}\right]} \le \beta_0 \le$$

$$159.9 + t_{0.05/2,30-2}\sqrt{3.2^2\left[\frac{1}{30}+\frac{(225/30)^2}{87.5}\right]}$$

$$\Rightarrow 159.9 - 2.05\times 2.62 \le \beta_0 \le 159.9 + 2.05\times 2.62$$

$$\Rightarrow 154.5 \le \beta_0 \le 165.3$$

Exercise 11.4 (MR 11-38)

For the noise-hypertension data in Exercise 11-1, the following summary quantities have been calculated:

$$n = 20,\ \sum x_i = 1{,}646,\ \hat{\sigma}^2 = 1.4^2,\ S_{xx} = 3{,}010.2,\ S_{xx} = 3{,}010.2,\ \hat{\beta}_1 = 0.17,$$

and $\hat{\beta}_0 = -9.81$

1. Establish a 95% two-sided confidence interval on the slope.
2. Establish a 95% two-sided confidence interval on the intercept.

11-6.2 Confidence Interval on the Mean Response

Learning Goals

- ☐ Identify the sampling distribution of the mean response estimator at $x = x_0$.
- ☐ Establish a confidence interval on the mean response at $x = x_0$.

Sampling Distribution of Mean Response Estimator at $x = x_0$ ($\hat{\mu}_{Y|x=x_0}$)

The mean response at $x = x_0$ is

$$\mu_{Y|x_0} = E(Y \mid x_0) = E[(\beta_0 + \beta_1 x + \varepsilon) \mid x_0] = \beta_0 + \beta_1 x_0$$

The point estimator of $\mu_{Y|x_0}$ is

$$\hat{\mu}_{Y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

The sampling distribution of $\hat{\mu}_{Y|x_0}$ is

$$\hat{\mu}_{Y|x_0} \sim N\left(\mu_{Y|x_0}, \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]\right)$$

(Derivation) Sampling distribution of $\hat{\mu}_{Y|x_0}$

Recall that

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]\right) \quad \text{and} \quad \hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

Since $\hat{\mu}_{Y|x_0}$ is a linear combination of $\hat{\beta}_0$ and $\hat{\beta}_1$, $\hat{\mu}_{Y|x_0}$ has a normal distribution with the following mean and variance:

$$E(\hat{\mu}_{Y|x_0}) = E(\hat{\beta}_0 + \hat{\beta}_1 x_0) = E(\hat{\beta}_0) + x_0 E(\hat{\beta}_1) = \beta_0 + x_0 \beta_1 = \mu_{Y|x_0}$$

$$V(\hat{\mu}_{Y|x_0}) = V(\hat{\beta}_0 + \hat{\beta}_1 x_0) = V(\hat{\beta}_0) + x_0^2 V(\hat{\beta}_1) + 2\,\mathrm{cov}(\hat{\beta}_0, \hat{\beta}_1)$$

$$= \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right] + x_0^2 \frac{\sigma^2}{S_{xx}} + 2\frac{-\bar{x}x_0\sigma^2}{S_{xx}} = \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]$$

(Note) Since $E(\hat{\mu}_{Y|x_0}) = \mu_{Y|x_0}$, $\hat{\mu}_{Y|x_0}$ is an unbiased estimator of $\mu_{Y|x_0}$.

Inference Context on Mean Response at $x = x_0$ ($\mu_{Y|x_0}$)

Parameter of interest: $\mu_{Y|x_0}$

Point estimator of $\mu_{Y|x_0}$: $\hat{\mu}_{Y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0 \sim N\left(\mu_{Y|x_0}, \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]\right)$,

σ^2 is unknown.

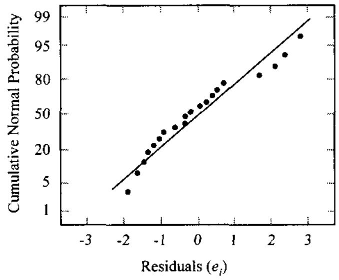
Example 11.7 (*cont.*)

The normal probability plot displays that the residuals lie closely along a straight line. In addition, the standardized residual plot shows more than 95% of the standardized residuals (29/30 = 96.7%) fall in the interval (–2, 2). Therefore, it is concluded that the residuals are normally distributed. (Notice that there is no indication of an outlier in the plywood price-sales data because all the standardized residuals are within the interval (-3, 3).)

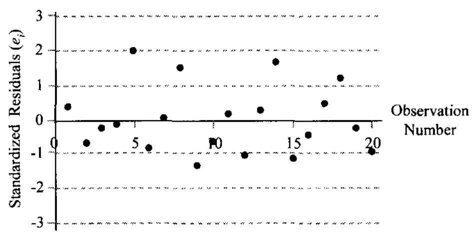
Next, the plots e_i vs. $\hat{y}_i$ and e_i vs. x_i indicate that the residuals are random and have a constant variance. The variability of the residuals is slightly increased in the mid ranges of y and x; however, this deviation from the constant variance is not serious enough to reject the adequacy of the regression model.

Exercise 11.7 (MR 11-47)

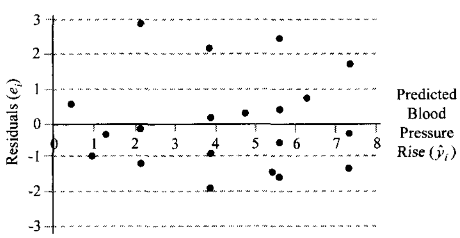
In the noise-hypertension data in Exercise 11-1, the following residual plots are obtained for the fitted regression model $\hat{y} = -9.81 + 0.17x$:



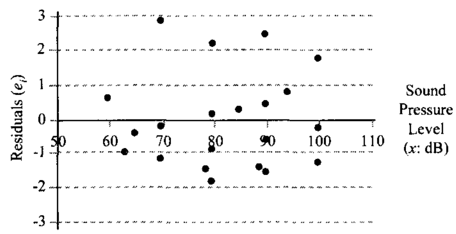
(a) normal probability plot



(b) standardized residual plot



(c) e_i vs. $\hat{y}_i$



(d) e_i vs. x_i

Discuss the adequacy of the regression model by using the residual plots.

11-8.2 Coefficient of Determination (R^2)

Learning Goals

- ❑ Explain the term *coefficient of determination.*
- ❑ Calculate the coefficient of determination of a regression model.

Coefficient of Determination (R^2)

The coefficient of determination (denoted as R^2) indicates the amount of variability of the data explained by the regression model. In other words, R^2 is the proportion of the total variability in Y which is accounted for by the regression model:

Coefficient of Determination (R^2) (*cont.*)

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}, \quad 0 \le R^2 \le 1$$

(Caution) Use of R^2
A large value of R^2 does not necessarily indicate the adequacy of the regression model. By adding new regressors or high-order terms of the regressors to an existing model, R^2 will always increase even if the new model is less desirable than the existing model in terms of simplicity, stability, and statistical significance. For instance, n data points of x and y can be 'perfectly' fit by a polynomial of degree $n - 1$ (e.g., four data points can be perfectly fit by a polynomial with a degree of three). Although the R^2-value of the perfect-fit model is one, the high-order polynomial is complex to use and often unstable when a different sample of data is used. Lastly, unless the error sum of the squares (SS_E) of the new model is reduced by the error mean square of the existing model, the new model will have a larger error mean square (MS_E), resulting in a decreased level of statistical significance.

Example 11.8

(Coefficient of Determination) For the plywood price-sales data in Example 11-1, the following summary quantities have been calculated:

$SS_T = 14{,}492.7$, $SS_R = 14{,}208.3$, and $SS_E = 284.4$

Calculate the coefficient of determination (R^2) of the fitted regression model $\hat{y} = 159.9 - 12.7x$.

$$R^2 = \frac{SS_R}{SS_T} = \frac{14{,}208.3}{14{,}492.7} = 98.0\%$$

The fitted regression model explains 98.0% of the variability in the sales of the plywood product.

Exercise 11.8 (MR 11-47)

For the noise-hypertension data in Exercise 11-1, the following summary quantities have been calculated:

$SS_T = 124.2$, $SS_R = 88.5$, and $SS_E = 35.7$

Calculate the coefficient of determination (R^2) of the fitted regression model $\hat{y} = -9.81 + 0.17x$.

11-8.3 Lack-of-Fit Test

Learning Goals

- ☐ Describe the purpose of the lack-of-fit test on a regression model.
- ☐ Conduct a lack-fit-test on a fitted regression model.

Lack-of-Fit Test

The lack-of-fit (goodness-of-fit) test on a regression model identifies if the order of the model is appropriate. For a lack-of-fit test, repeated observations of Y are needed for at least one level of x.

ANOVA for Lack-of-Fit Test

The ANOVA method underlies the lack-of-fit test on a regression model. Suppose that the observations for m levels of x are as follows:

X	No. repeated observations	Y
x_1	n_1	$y_{11}, y_{12}, \ldots, y_{1n_1}$
x_2	n_2	$y_{21}, y_{22}, \ldots, y_{2n_2}$
$\vdots$	$\vdots$	$\vdots$
x_m	n_m	$y_{m1}, y_{m2}, \ldots, y_{mn_1}$

By applying the ANOVA technique, as shown in Table 11-3, the error sum of squares (SS_E; degrees of freedom = $n - 2$; see Section 11-5.2) is further partitioned into two components:

(1) Variability due to pure error (***SS_{PE}*, pure error sum of squares**; degrees of freedom = $n - m$)

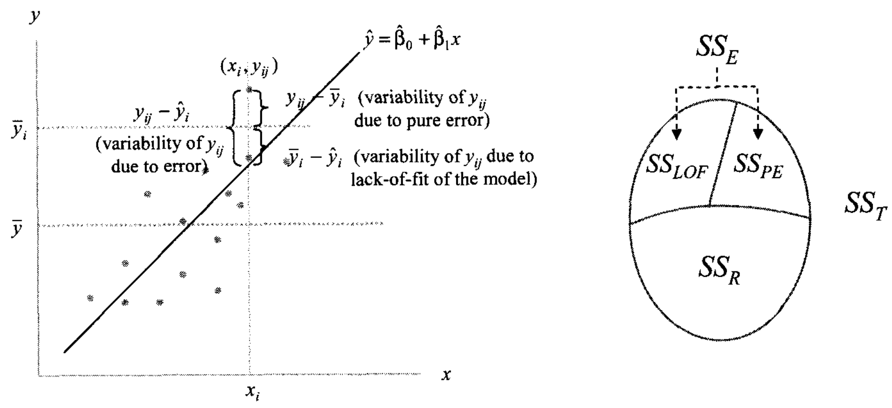
(2) Variability due to the lack-of-fit of the model (***SS_{LOF}*, lack-of-fit sum of squares**; degrees of freedom = m - 2)

Table 11-3 Partitioning of the Error Sum of Squares (SS_E)

	SS_E (Error SS)		SS_{PE} (Pure error SS)		SS_{LOF} (Lack-of-fit SS)
	$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \hat{y}_i)^2$	=	$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$	+	$\sum_{i=1}^{m}(\bar{y}_i - \hat{y}_i)^2$
DF	$n-2$	=	$n-m$	+	$m-2$

(Note) $n = \sum_{i=1}^{m} n_i$; SS: Sum of Squares; DF: Degrees of Freedom

A graphical interpretation of SS_E, SS_{PE}, and SS_{LOF} is provided in Figure 11-7. The deviation of an observation y_{ij} from the corresponding estimate $\hat{y}_i$ at x_i is divided into the deviation of y_{ij} from the corresponding mean $\bar{y}_i$ due to pure error and the deviation of $\bar{y}_i$ from the corresponding estimate $\hat{y}_i$ due to the lack-of-fit of the



(a) (b)

Figure 11-7 Partitioning of the variability of y due to error.

ANOVA for Lack-of-Fit Test (*cont.*)

model. These three deviations are squared and then added to those of the other observations to calculate SS_E, SS_{PE}, and SS_{LOF}. In Figure 11-7, the error due to the lack-of-fit of the model could have been zero if a higher order model (dotted line) was employed.

The ratio of $SS_{LOF}/(m - 2)$ to $SS_{PE}/(n - m)$ follows an F distribution:

$$F = \frac{SS_{LOF}/(m-2)}{SS_{PE}/(n-m)} = \frac{MS_{LOF}}{MS_{PE}} \sim F(m-2, n-m)$$

Since the statistic F becomes small as the order of the model is appropriate, F is used to test H_0: The order of the regression model is correct. Note that the quantities MS_{LOF} (**lack-of-fit mean square**) and MS_{PE} (**pure-error mean square**) are adjusted SS_{LOF} and SS_{PE} by their degrees of freedom, respectively.

ANOVA Table and Lack-of-Fit Test

The lack-of-fit and pure-error quantities from the ANOVA analysis are summarized in Table 11-4 to test for the adequacy of model.

Table 11-4 ANOVA Table for the Lack-of-Fit Test

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Lack-of-Fit	SS_{LOF}	$m-2$	MS_{LOF}	MS_{LOF} / MS_{PE}
Pure Error	SS_{PE}	$n-m$	MS_{PE}	
Error	SS_E	$n-2$	MS_E	

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: The order of the model is correct.
H_1: The order of the model is not correct.

Step 2: Determine a **test statistic and its value.**

$$F_0 = \frac{SS_{LOF}/(m-2)}{SS_{PE}/(n-m)} = \frac{MS_{LOF}}{MS_{PE}} \sim F(m-2, n-m)$$

Step 3: Determine a **critical value(s) for α.**

$f_{\alpha,m-2,n-m}$

Step 4: Make a **conclusion**. Reject H_0 if

$f_0 > f_{\alpha,m-2,n-m}$

Example 11.9

(Lack-of-Fit Test) For the plywood price-sales data in Example 11-1, the following summary quantities have been calculated:

$$n = 30,\ m = 6,\ SS_E = 284.4,\ SS_{LOF} = \sum_{i=1}^{m}(\bar{y}_i - \hat{y}_i)^2 = 2.4, \text{ and}$$

$$SS_{PE} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 = 282.0$$

Conduct a lack-of-fit test on the regression model $\hat{y} = 159.9 - 12.7x$ at $\alpha = 0.05$.

Example 11.9 (*cont.*)

An ANOVA table to test the lack-of-fit of the regression model is as follows:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Lack-of-Fit	2.4	4	0.6	0.05
Pure Error	282.0	24	11.8	
Error	284.4	28		

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: The order of the model is correct.

H_1: The order of the model is not correct.

Step 2: Determine a **test statistic and its value.**

$$f_0 = \frac{MS_{LOF}}{MS_{PE}} = \frac{0.6}{11.8} = 0.05$$

Step 3: Determine a **critical value(s) for α.**

$$f_{\alpha,m-2,n-m} = f_{0.05,4,24} = 2.78$$

Step 4: Make a **conclusion.**

Since $f_0 = 0.05 \not> f_{\alpha,m-2,n-m} = 2.78$, fail to reject H_0 at $\alpha = 0.05$. It is concluded that the order of the model is adequate.

Exercise 11.9

For the noise-hypertension data in Exercise 11-1, the following summary quantities have been calculated:

$$n = 20,\ m = 10,\ SS_E = 35.7,\ SS_{LOF} = \sum_{i=1}^{m} (\bar{y}_i - \hat{y}_i)^2 = 4.8, \text{ and}$$

$$SS_{PE} = \sum_{i=1}^{m}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 = 30.8$$

Conduct a lack-of-fit test on the regression model $\hat{y} = -9.81 + 0.17x$ at $\alpha = 0.05$.

11-9 Transformations to a Straight Line

Learning Goals

- ❑ Explain the relationship between linear and non-linear models.

Transformation of Nonlinear Model

A non-linear relationship between independent (x) and dependent (Y) variables can be transformed into a linear model by mathematical operations. Examples are presented in Table 11-5.

Transformation of Nonlinear Model (*cont.*)

Table 11-5 Transformation of Nonlinear Models (illustrated)

Nonlinear Model		Linear Model
$Y = \beta_0 e^{\beta_1 x}\varepsilon$	$\Rightarrow$	$\ln Y = \ln\beta_0 + \beta_1 x + \varepsilon$
$Y = \beta_0 + \beta_1 x^3 + \varepsilon$	$\Rightarrow$	$Y = \beta_0 + \beta_1 z + \varepsilon$, where $z = x^3$
$Y = \beta_0 + \beta_1 \dfrac{1}{x} + \varepsilon$	$\Rightarrow$	$Y = \beta_0 + \beta_1 z + \varepsilon$, where $z = \dfrac{1}{x}$
$Y = \dfrac{1}{\exp(\beta_0 + \beta_1 x + \varepsilon)}$	$\Rightarrow$	$Z = \beta_0 + \beta_1 x + \varepsilon$, where $Z = -\ln Y$

11-11 Correlation

Learning Goals

- ❑ Calculate the sample correlation coefficient (R) between random variables X and Y.
- ❑ Construct a confidence interval on the population correlation coefficient (ρ) between X and Y.
- ❑ Conduct a hypothesis test on ρ.

Correlation vs. Regression

A close relationship exists between the correlation of random variables X and Y and the regression model of independent and dependent variables x and Y. Recall that in the regression model the regressor x is assumed to be a mathematical variable, not a random variable (see Section 11-1).

Suppose that X and Y are random variables and the observations (X_i, Y_i), $i = 1, 2, \ldots, n$, are obtained from a bivariate normal distribution. The **population correlation coefficient** ρ_{XY} (see Section 5-5) is

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}, \quad -1 \le \rho_{XY} \le 1$$

The conditional probability distribution of Y given $X = x$ is a normal distribution with mean and variance (see Section 5-6)

$$E(Y \mid x) = \mu_Y + \rho_{XY}\frac{\sigma_Y}{\sigma_X}(x - \mu_X) = \mu_Y - \mu_X \rho_{XY}\frac{\sigma_Y}{\sigma_X} + \rho_{XY}\frac{\sigma_Y}{\sigma_X}x$$

$$= \beta_0 + \beta_1 x, \quad \text{where } \beta_0 = \mu_Y - \mu_X \rho_{XY}\frac{\sigma_Y}{\sigma_X} \text{ and } \beta_1 = \rho_{XY}\frac{\sigma_Y}{\sigma_X}$$

$$V(Y \mid x) = \sigma_Y^2(1 - \rho_{XY})$$

Therefore, it is identified that the conditional probability distribution of Y given $X = x$ is equivalent to the probability distribution of a simple linear regression model with mean $\beta_0 + \beta_1 x$ and variance σ^2.

Estimation of Correlation Coefficient (ρ_{XY})

The estimator of ρ_{XY}, called **sample correlation coefficient** (denoted as R), is

$$R = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X \hat{\sigma}_Y} = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}, \quad -1 \le R \le 1$$

The square of R becomes the coefficient of determination R^2 of the regression model for Y and X having a bivariate normal distribution:

$$R^2 = \frac{S_{XY}^2}{S_{XX} S_{YY}} = \frac{SS_R}{SS_T}$$

(Derivation) $R^2 = \dfrac{SS_R}{SS_T}$

$$R = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}}$$

$$\Rightarrow R^2 = \frac{S_{XY}^2}{S_{XX} S_{YY}} = \frac{S_{XY}^2}{S_{XX} SS_T} \quad \text{because } SS_T = S_{YY}$$

$$= \frac{\hat{\beta}_1 S_{XY}}{SS_T} \quad \text{because } \hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}$$

$$= \frac{SS_R}{SS_T} \quad \text{because } SS_R = \hat{\beta}_1 S_{XY}$$

Inference Context on ρ_{XY}

Parameter of interest: ρ_{XY}

Point estimator of ρ_{XY}: $R = \dfrac{S_{XY}}{S_X S_Y}$

Test statistic of ρ_{XY}

$$T_0 = \frac{R - \rho_0}{\sqrt{(1 - R^2)/(n-2)}} \sim t(n-2) \quad \text{for testing } H_0\text{: } \rho = 0$$

$$Z_0 = (\operatorname{arctanh} R - \operatorname{arctanh} \rho_0)\sqrt{n-3} \quad \text{for testing } H_0\text{: } \rho = \rho_0 \ (\neq 0),\ n \ge 25$$

Confidence Interval on ρ_{XY}

An <u>approximate</u> $100(1 - \alpha)\%$ confidence interval on ρ is

$$\tanh\left(\operatorname{arctanh} r - \frac{z_{\alpha/2}}{\sqrt{n-3}}\right) \le \rho \le \tanh\left(\operatorname{arctanh} r + \frac{z_{\alpha/2}}{\sqrt{n-3}}\right)$$

Hypothesis Test on β_1 (*t*-test)

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $\rho = \rho_0$
H_1: $\rho \neq \rho_0$

Step 2: Determine a **test statistic and its value.**

$$T_0 = \frac{R - \rho_0}{\sqrt{(1 - R^2)/(n-2)}} \sim t(n-2) \quad \text{for testing } H_0\text{: } \rho = 0$$

$$Z_0 = (\operatorname{arctanh} R - \operatorname{arctanh} \rho_0)\sqrt{n-3} \quad \text{for testing } H_0\text{: } \rho = \rho_0 \ (\neq 0),\ n \ge 25$$

Hypothesis Test on β_1 (t-test) (*cont.*)

Step 3: Determine a **critical value(s) for α.**

$t_{\alpha/2,n-2}$ for testing H_0: $\rho = 0$

$z_{\alpha/2}$ for testing H_0: $\rho = \rho_0$ ($\neq 0$), $n \geq 25$

Step 4: Make a **conclusion**. Reject H_0 if

$|t_0| > t_{\alpha/2,n-2}$ for testing H_0: $\rho = 0$

$|z_0| > z_{\alpha/2}$ for testing H_0: $\rho = \rho_0$ ($\neq 0$), $n \geq 25$

Example 11.10

The relationship between ergonomics grade (X) and statistics grade (Y) is under study. A random sample of $n = 50$ students is selected who have taken both ergonomics and statistics courses. Their final scores in ergonomics and statistics are summarized as follows:

$n = 50$, $S_{yy} = 6{,}652.5$, $S_{xx} = 3{,}999.0$, and $S_{xy} = 3{,}303.1$

Assume that ergonomics grade and statistics grade have a bivariate normal distribution.

1. **(Sample Correlation Coefficient)** Calculate the sample correlation coefficient (R) between ergonomics grade (X) and statistics grade (Y).

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{3{,}303.1}{\sqrt{3{,}999.0 \times 6{,}652.5}} = 0.64$$

2. **(Confidence Interval on ρ_{XY})** Construct a 95% confidence interval on the population correlation coefficient (ρ) between X and Y.

$$\tanh\left(\operatorname{arctanh} r - \frac{z_{\alpha/2}}{\sqrt{n-3}}\right) \leq \rho \leq \tanh\left(\operatorname{arctanh} r + \frac{z_{\alpha/2}}{\sqrt{n-3}}\right)$$

$$\Rightarrow \tanh\left(\operatorname{arctanh} 0.64 - \frac{z_{0.05/2}}{\sqrt{50-3}}\right) \leq \rho \leq \tanh\left(\operatorname{arctanh} 0.64 + \frac{z_{0.05/2}}{\sqrt{50-3}}\right)$$

$$\Rightarrow 0.44 \leq \rho \leq 0.78$$

3. **(Hypothesis Test on ρ_{XY})** Test if $\rho_{XY} \neq 0$ at $\alpha = 0.05$.

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $\rho = 0$

H_1: $\rho \neq 0$

Step 2: Determine a **test statistic and its value.**

$$t_0 = \frac{r - \rho_0}{\sqrt{(1-r^2)/(n-2)}} = \frac{0.64 - 0}{\sqrt{(1-0.64^2)/(50-2)}} = 5.78$$

Step 3: Determine a **critical value(s) for α.**

$$t_{\alpha/2,n-2} = t_{0.05/2,50-2} = t_{0.025,48} = 2.01$$

Example 11.10
(cont.)

Step 4: Make a **conclusion**.

Since $|t_0| = 5.78 > t_{\alpha/2,n-2} = 2.01$, reject H_0. It is concluded that ergonomics grade (X) and statistics grade (Y) are significantly related at $\alpha = 0.05$.

Exercise 11.10
(MR 11-56)

The measurements of weight (X) and systolic blood pressure (Y) for $n = 26$ randomly selected males in the age group 25 to 30 are collected, resulting in the following summary quantities:

$$n = 26\,,\ S_{yy} = 4{,}502.2\,,\ S_{xx} = 15{,}312.3\,,\ \text{and}\ S_{xy} = 6{,}422.2$$

Assume that weight and systolic blood pressure are jointly normally distributed.

1. Calculate the sample correlation coefficient (R) between X and Y.
2. Construct a 95% confidence interval on the population correlation coefficient (ρ) between X and Y.
3. Test if $\rho_{XY} \neq 0$ at $\alpha = 0.05$.

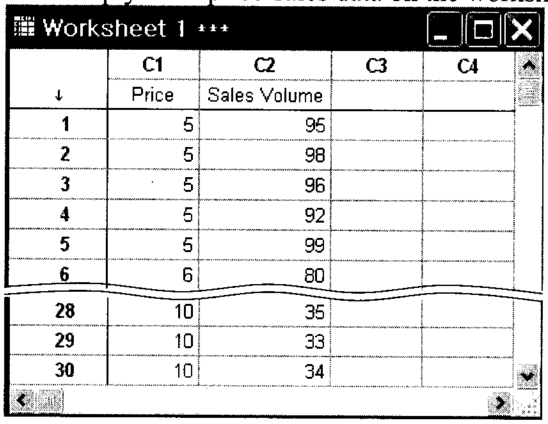
MINITAB Applications

Examples 11.1-3, 5, 6, 8

(Simple Linear Regression)

(1) Choose **File ➤ New**, click **Minitab Project**, and click **OK**.

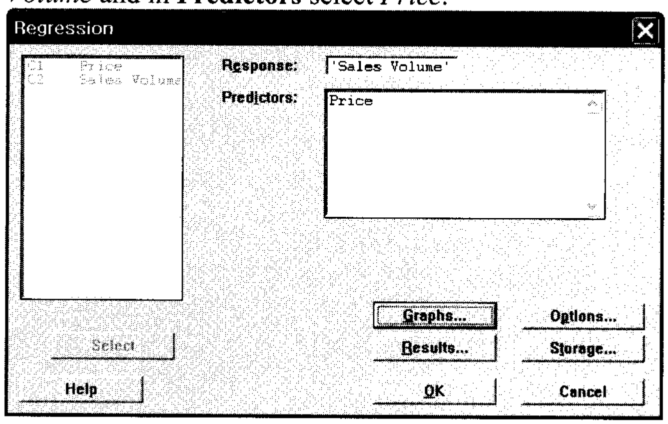
(2) Enter the plywood price-sales data on the worksheet.



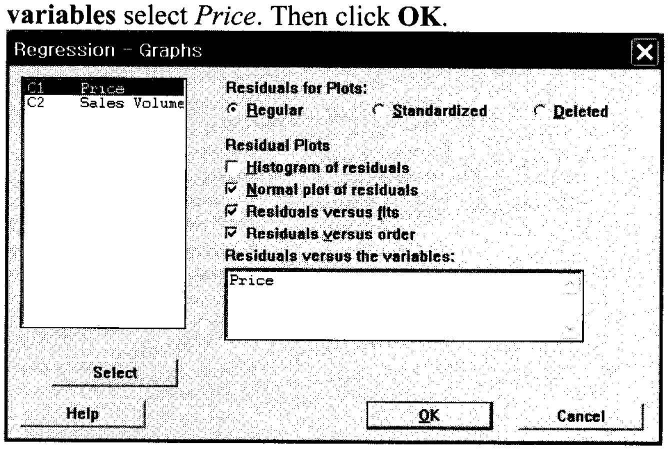
Worksheet 1 ***

↓	C1 Price	C2 Sales Volume	C3	C4
1	5	95		
2	5	98		
3	5	96		
4	5	92		
5	5	99		
6	6	80		
28	10	35		
29	10	33		
30	10	34		

(3) Choose **Stat ➤ Regression ➤ Regression**. In **Response** select *Sales Volume* and in **Predictors** select *Price*.

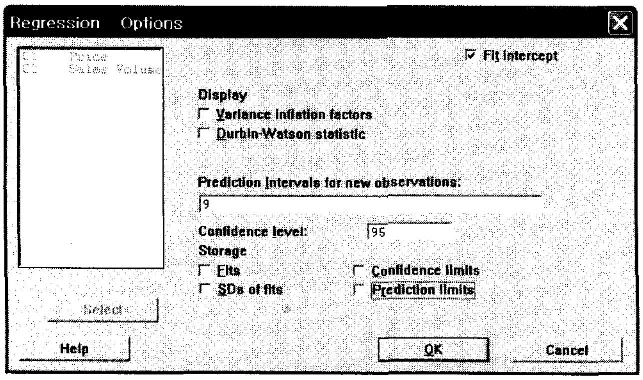


(4) Click **Graphs**. For **Residuals for Plots**, select **Regular** or **Standardized**. Under **Residual Plots**, check **Normal plot of residuals**, **Residuals versus fits**, and/or **Residuals versus order**. In **Residuals versus the variables** select *Price*. Then click **OK**.

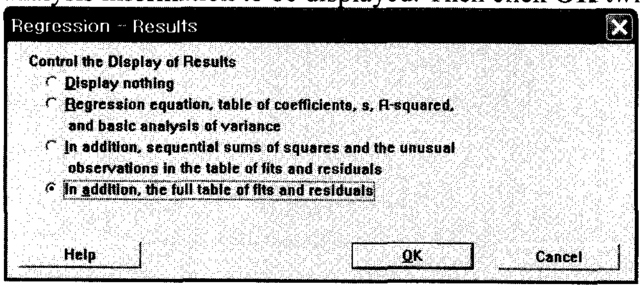


Examples 11.1-3, 5, 6, 8 (*cont.*)

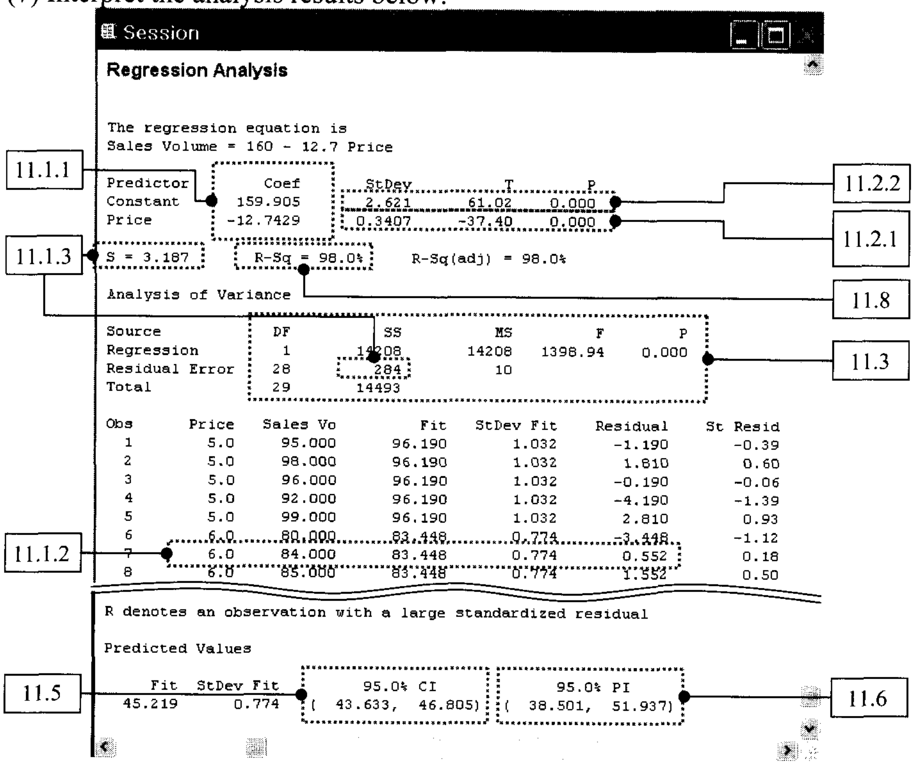
(5) Click **Options**. Check **Fit intercept**. In **Prediction Intervals for new observations**, enter the value of price of interest for the confidence interval on the mean response and/or prediction interval of a future observation. In **Confidence level**, type the level of confidence. Then click **OK**.



(6) Click **Results**. Under **Control the Display of Results**, click the level of analysis information to be displayed. Then click **OK** twice.

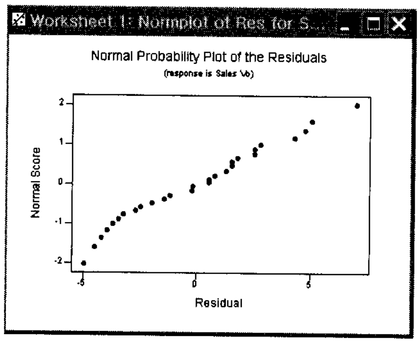


(7) Interpret the analysis results below.

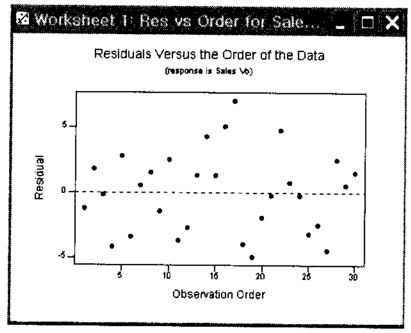


Example 11.7 **(Residual Analysis)**

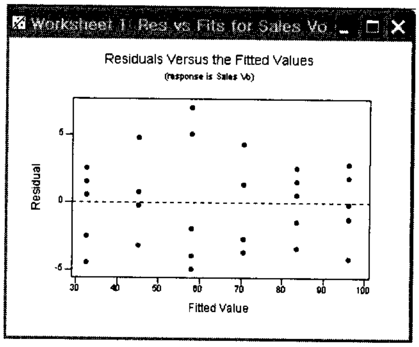
(1) Normal probability plot



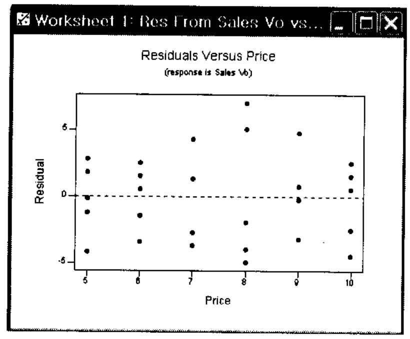
(2) Standardized residual plot



(3) e_i vs. $\hat{y}_i$



(4) e_i vs. x_i



Answers to Exercises

Exercise 11.1

1. **(Estimation of Regression Coefficients)**

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)/n}{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 / n} = \frac{S_{xy}}{S_{xx}}$$

$$= \frac{7{,}594 - 86 \times 1{,}646/20}{138{,}476 - 1{,}646^2/20} = \frac{516.2}{3{,}010.2} = 0.17$$

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i}{n} = \frac{86 - 0.17 \times 1{,}646}{20} = -9.81$$

The fitted regression line of the noise-hypertension data is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = -9.81 + 0.17x$$

2. **(Calculation of Residual)**

$$\hat{y}_{x=85} = \hat{\beta}_0 + \hat{\beta}_1 x = -9.81 + 0.17 \times 85 = 4.76$$

$$e_{x=85} = y_{x=85} - \hat{y}_{x=85} = 5 - 4.76 = 0.24$$

3. **(Estimation of σ^2)**

$$SS_E = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 - \hat{\beta}_1 S_{xy} = 494 - 20 \times (86/20)^2 - 0.17 \times 516.2 = 35.7$$

$$\hat{\sigma}^2 = \frac{SS_E}{n-2} = \frac{35.7}{20-2} = 1.98 = 1.4^2$$

Exercise 11.2

1. **(Hypothesis Test on β_1)**

Step 1: State H_0 and H_1.

H_0: $\beta_1 = 0$ $\qquad$ H_1: $\beta_1 \neq 0$

Step 2: Determine a **test statistic and its value.**

$$t_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2 / S_{xx}}} = \frac{0.17 - 0}{\sqrt{1.4^2 / 3{,}010.2}} = 6.68$$

Step 3: Determine a **critical value(s) for α.**

$$t_{\alpha/2, n-2} = t_{0.025,18} = 2.10$$

Step 4: Make a **conclusion.**

Since $|t_0| = 6.68 > t_{\alpha/2,n-2} = 2.10$, reject H_0 at $\alpha = 0.05$.

Exercise 11.2 (*cont.*)

2. **(Hypothesis Test on β_0)**

Step 1: State H_0 and H_1.

H_0: $\beta_0 = 0$

H_1: $\beta_0 \neq 0$

Step 2: Determine a **test statistic and its value.**

$$t_0 = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}} = \frac{-9.81 - 0}{\sqrt{1.4^2\left[\frac{1}{20} + \frac{(1{,}646/20)^2}{3{,}010.2}\right]}} = -4.6$$

Step 3: Determine a **critical value(s) for α.**

$$t_{\alpha/2,n-2} = t_{0.025,18} = 2.10$$

Step 4: Make a **conclusion.**

Since $|t_0| = 4.6 > t_{\alpha/2,n-2} = 2.10$, reject H_0 at $\alpha = 0.05$.

Exercise 11.3

(ANOVA for Hypothesis Test on β_1)

$$SS_T = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = 494 - 20 \times \left(\frac{86}{20}\right)^2 = 124.2$$

$$SS_R = \hat{\beta}_1 S_{xy} = 0.17 \times 516.2 = 88.5$$

$$SS_E = SS_T - SS_R = 124.2 - 88.5 = 35.7$$

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	f_0
Regression	88.5	1	88.5	44.7
Error	35.7	18	2.0	
Total	124.2	19		

(Note) $t_0^2 = 6.7^2 = 44.7 = f_0$

Step 1: State H_0 and H_1.

H_0: $\beta_1 = 0$ $\quad$ H_1: $\beta_1 \neq 0$

Step 2: Determine a **test statistic and its value.**

$$f_0 = \frac{SS_R/1}{SS_E/(n-2)} = \frac{MS_R}{MS_E} = \frac{88.5}{2.0} = 44.7$$

Step 3: Determine a **critical value(s) for α.**

$$f_{\alpha,1,n-2} = f_{0.05,1,20-2} = 4.41$$

Step 4: Make a **conclusion.**

Since $f_0 = 44.7 > f_{\alpha,1,n-2} = 4.41$, reject H_0 at $\alpha = 0.05$.

Exercise 11.4

1. **(Confidence Interval on β_1)**

$1-\alpha = 0.95 \quad \Rightarrow \quad \alpha = 0.05$

95% two-sided CI on β_1:

$$\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2 / S_{xx}} \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2 / S_{xx}}$$

$$\Rightarrow \; 0.17 - t_{0.05/2,20-2}\sqrt{1.4^2 / 3{,}010.2} \le \beta_1 \le 0.17 + t_{0.05/2,20-2}\sqrt{1.4^2 / 3{,}010.2}$$

$$\Rightarrow \; 0.17 - 2.10 \times 0.03 \le \beta_1 \le 0.17 + 2.10 \times 0.03$$

$$\Rightarrow \; 0.12 \le \beta_1 \le 0.23$$

2. **(Confidence Interval on β_0)**

$1-\alpha = 0.95 \quad \Rightarrow \quad \alpha = 0.05$

95% two-sided CI on β_0:

$$\hat{\beta}_0 - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]} \le \beta_0 \le \hat{\beta}_0 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]}$$

$$\Rightarrow \; -9.81 - t_{0.05/2,20-2}\sqrt{1.4^2\left[\frac{1}{20}+\frac{(1{,}646/20)^2}{3{,}010.2}\right]} \le \beta_0 \le$$

$$-9.81 + t_{0.05/2,20-2}\sqrt{1.4^2\left[\frac{1}{20}+\frac{(1{,}646/20)^2}{3{,}010.2}\right]}$$

$$\Rightarrow \; -9.81 - 2.10 \times 2.14 \le \beta_0 \le -9.81 + 2.10 \times 2.14$$

$$\Rightarrow \; -14.3 \le \beta_0 \le -5.3$$

Exercise 11.5

(Confidence Interval on $\mu_{Y|x_0}$)

$$\hat{\mu}_{Y|85} = \hat{\beta}_0 + \hat{\beta}_1 x_0 = -9.81 + 0.17 \times 85 = 4.7$$

$1-\alpha = 0.95 \quad \Rightarrow \quad \alpha = 0.05$

95% two-sided CI on $\mu_{Y|85}$:

$$\hat{\mu}_{Y|85} - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]} \le \mu_{Y|85} \le$$

$$\hat{\mu}_{Y|85} + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]}$$

$$\Rightarrow \; 4.7 - t_{0.05/2,20-2}\sqrt{1.4^2\left[\frac{1}{20}+\frac{(85-1{,}646/20)^2}{87.5}\right]} \le \mu_{Y|85} \le$$

$$4.7 + t_{0.05/2,20-2}\sqrt{1.4^2\left[\frac{1}{20}+\frac{(85-1{,}646/20)^2}{87.5}\right]}$$

$$\Rightarrow \; 4.7 - 2.10 \times 0.32 \le \mu_{Y|85} \le 4.7 + 2.10 \times 0.32$$

$$\Rightarrow \; 4.1 \le \mu_{Y|85} \le 5.4$$

Exercise 11.6

(Prediction Interval on y_0)

$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = -9.81 + 0.17 \times 85 = 4.7$

$1 - \alpha = 0.95 \quad \Rightarrow \quad \alpha = 0.05$

The 95% two-sided prediction interval on y_0 at $x = 85$ dB is

$$\hat{y}_0 - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]} \le y_0 \le \hat{y}_0 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}$$

$$\Rightarrow \; 4.7 - t_{0.05/2,20-2}\sqrt{1.4^2\left[1 + \frac{1}{20} + \frac{(85 - 1{,}646/20)^2}{87.5}\right]} \le y_0 \le 4.7 + t_{0.05/2,20-2}\sqrt{1.4^2\left[1 + \frac{1}{20} + \frac{(85 - 1{,}646/20)^2}{87.5}\right]}$$

$$\Rightarrow \; 4.7 - 2.10 \times 1.44 \le y_0 \le 4.7 + 2.10 \times 1.44$$

$$\Rightarrow \; 1.7 \le y_0 \le 7.8$$

(Note) The prediction interval $1.7 \le y_0 \le 7.8$ is wider than the corresponding confidence interval on the mean response $4.1 \le \mu_{Y|85} \le 5.4$ in Exercise 11-5.

Exercise 11.7

(Residual Analysis)

The normal probability plot displays that the residuals lie closely along a straight line. In addition, the standardized residual plot shows that all the standardized residuals fall in the interval (–2, 2). Therefore, it is concluded that the residuals are normally distributed.

Next, the plots e_i vs. $\hat{y}_i$ and e_i vs. x_i indicate that the residuals are random and have a constant variance. The variability of the residuals is slightly reduced in the low ranges of y and x; however, this deviation from the constant variance is not serious enough to reject the adequacy of the regression model.

Exercise 11.8

(Coefficient of Determination)

$$R^2 = \frac{SS_R}{SS_T} = \frac{88.5}{124.2} = 71.3\%$$

The fitted regression model explains 71.3% of the variability in blood pressure rise.

Exercise 11.9 **(Lack-of-Fit Test)**

An ANOVA table to test the lack-of-fit of the regression model is as follows:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Lack-of-Fit	4.8	8	0.6	0.20
Pure Error	30.8	10	3.1	
Error	35.7	18		

Step 1: State H_0 and H_1.

H_0: The order of the model is correct.

H_1: The order of the model is not correct.

Step 2: Determine a **test statistic and its value.**

$$f_0 = \frac{MS_{LOF}}{MS_{PE}} = \frac{0.6}{3.1} = 0.20$$

Step 3: Determine a **critical value(s) for α.**

$$f_{\alpha,m-2,n-m} = f_{0.05,8,10} = 3.07$$

Step 4: Make a **conclusion.**

Since $f_0 = 0.20 \not> f_{\alpha,m-2,n-m} = 3.07$, fail to reject H_0 at $\alpha = 0.05$. It is concluded that the order of the model is adequate.

Exercise 11.10 1. **(Sample Correlation Coefficient)**

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{6{,}422.2}{\sqrt{15{,}312.3 \times 4{,}502.2}} = 0.77$$

2. **(Confidence Interval on ρ_{XY})**

$$\tanh\left(\operatorname{arctanh} r - \frac{z_{\alpha/2}}{\sqrt{n-3}}\right) \le \rho \le \tanh\left(\operatorname{arctanh} r + \frac{z_{\alpha/2}}{\sqrt{n-3}}\right)$$

$$\Rightarrow \tanh\left(\operatorname{arctanh} 0.77 - \frac{z_{0.05/2}}{\sqrt{26-3}}\right) \le \rho \le \tanh\left(\operatorname{arctanh} 0.77 + \frac{z_{0.05/2}}{\sqrt{26-3}}\right)$$

$$\Rightarrow 0.55 \le \rho \le 0.89$$

3. **(Hypothesis Test on ρ_{XY})**

Step 1: State H_0 and H_1.

H_0: $\rho = 0$

H_1: $\rho \ne 0$

Step 2: Determine a **test statistic and its value.**

$$t_0 = \frac{r - \rho_0}{\sqrt{(1-r^2)/(n-2)}} = \frac{0.77 - 0}{\sqrt{(1-0.77^2)/(26-2)}} = 5.91$$

Exercise 11.10
(*cont.*)

Step 3: Determine a **critical value(s) for α.**

$$t_{\alpha/2,n-2} = t_{0.05/2,26-2} = t_{0.025,24} = 2.06$$

Step 4: Make a **conclusion**.

Since $|t_0| = 5.91 > t_{\alpha/2,n-2} = 2.06$, reject H_0. It is concluded that weight (X) and systolic blood pressure (Y) are significantly related at $\alpha = 0.05$.

12 Multiple Linear Regression

OUTLINE

12-1 Multiple Linear Regression Model

Learning Goals

- ❑ Explain the assumptions of a multiple linear regression model.
- ❑ Transform a non-linear model into a linear model.
- ❑ Calculate residuals.
- ❑ Estimate the error variance σ^2.

Multiple Linear Model As an extension of a simple linear model, a multiple linear model explains the linear relationship between the response variable (Y) and k (> 1) multiple regressors (x_j's):

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

where: (1) **Regression coefficients ($\beta_0, \beta_1, \ldots, \beta_k$)**: The coefficient β_0 denotes

Multiple Linear Model (*cont.*)

the intercept of the model and the coefficients β_j's, $j = 1$ to k, denote the slopes of the regressors. Each slope coefficient β_j measures the expected change in Y per unit change in x_j when the other regressors are held constant (this is why β_j's are sometimes called **partial regression coefficients**).

(2) **Independent variables ($x_1, x_2, \ldots, x_k$; regressors, predictors)**: Since the regressors x_j's will be controlled with negligible error, x_j's represent controlled constants (not random outcomes).

(3) **Random error (ε)**: The variable ε represents the random variation of Y around the linear line $\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$. It is assumed that ε is normally distributed with mean 0 and constant variance σ^2, i.e.,

$$\varepsilon \sim N(0, \sigma^2)$$

(4) **Dependent (response) variable (Y)**: Since outcomes in Y at the same values of $x_1, x_2, \ldots, x_k$ can vary randomly, Y is a random variable with the normal distribution

$$Y \sim N(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k, \sigma^2)$$

Like the assumptions of a simple linear model (see Section 11-1), the multiple linear model has four assumptions:

(1) Linear relationship between x_j's and Y
(2) Randomness of error
(3) Constant σ^2
(4) Normality of error

The analyst should check if these four assumptions are met for proper regression analysis (see the assessment of model adequacy in Section 12-5).

Note that computers are mostly used in multiple regression analysis so that the detailed calculation process is not presented in this chapter.

Transformation of Nonlinear Model

Like the transformation of a non-linear model in simple linear regression analysis (see Section 11-9), a non-linear relationship between x_j's and Y can be transformed into a linear model by applying appropriate mathematical operations. Therefore, any regression models with regressors linearly combined can be treated as a linear model.

(e.g.) Transformation of a non-linear model

(1) $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon$

$\Rightarrow Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$, where $\beta_3 = \beta_{12}$ and $x_3 = x_1 x_2$

(2) $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \dfrac{1}{x_1} + \beta_4 x_2^3 + \varepsilon$

$\Rightarrow Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$, where $x_3 = \dfrac{1}{x_1}$ and $x_4 = x_2^3$

Matrix Notation

Matrix notations and operations are useful in multiple regression. Suppose that n observations consisting of k regressors and the response variable, $(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i)$, $i = 1, 2, \ldots, n$, are modeled by a multiple linear model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

Matrix Notation (*cont.*)

The system of n equations is expressed in matrix notation as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} \quad \text{and} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Model Building

The least squares method finds the estimates of regression coefficients $\boldsymbol{\beta}$ which minimize the sum of the squares of the errors

$$SS_E = L = \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}$$

The **least squares estimators of $\boldsymbol{\beta}$** are

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Thus, the **fitted regression** model is

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

Residual and Error Variance

The **residuals** of the fitted model are

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$$

The **estimate of the error variance σ^2** is

$$\hat{\sigma}^2 = \frac{SS_E}{n-p} = \frac{\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}}{n-p}, \quad \text{where } p = \text{number of regression coefficients}$$
$$= k + 1 \; (k = \text{number of regressors})$$

Standard Error of Regression Coefficients $\hat{\beta}_j$

The **covariances of $\hat{\boldsymbol{\beta}}$** are

$$\text{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2\mathbf{C}, \quad \text{where } \mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$$

This covariance matrix is a $(p \times p)$ symmetric matrix whose jj^{th} component is the variance of $\hat{\beta}_j$ and whose ij^{th} component is the covariance of $\hat{\beta}_i$ and $\hat{\beta}_j$, where i and $j = 0$ to k, $i \neq j$.

Then, the **estimated standard error of $\hat{\beta}_j$** is

$$se(\hat{\beta}_j) = \hat{\sigma}\sqrt{C_{jj}}$$

Standard Error of $\hat{\beta}_j$ (*cont.*)

The standard error of a regression coefficient is a useful measure to represent the precision of estimation for the regression coefficient: the smaller the standard error, the better the estimation precision.

Example 12.1

The electric power (y) consumed each month at a chemical plant is under study. It is assumed that the monthly electric consumption is related to four factors: average ambient temperature (x_1), number of days in the month (x_2), average product purity (x_3), and volume of production (x_4). The corresponding records at the plant in the past year are as follows:

No.	Power consumption (y; unit: kW)	Temperature (x_1; unit: °F)	Number of days (x_2)	Product purity (x_3; unit: %)	Production volume (x_4; unit: ton)
1	240	25	24	91	100
2	236	31	21	90	95
3	290	45	24	88	110
4	274	60	25	87	88
5	301	65	25	91	94
6	316	72	26	94	99
7	300	80	25	87	97
8	296	84	25	86	96
9	267	75	24	88	110
10	276	60	25	91	105
11	288	50	25	90	100
12	261	38	23	89	98

By using a statistical software program, a fitted regression model of the power consumption data including all the four regressors (x_1, x_2, x_3, and x_4) and the corresponding error sum of squares (SS_E) are obtained as follows:

$$\hat{y} = -102.71 + 0.61x_1 + 8.92x_2 + 1.44x_3 + 0.01x_4 \quad \text{and} \quad SS_E = 1{,}699.0$$

1. (**Calculation of Residual**) Calculate the residual of $y = 301$ when $x_1 = 65$°F, $x_2 = 25$ days, $x_3 = 91$%, and $x_4 = 94$ tons.

The estimate of the power consumption given that $x_1 = 65$, $x_2 = 25$, $x_3 = 91$, and $x_4 = 94$ is

$$\hat{y} = -102.71 + 0.61 \times 65 + 8.92 \times 25 + 1.44 \times 91 + 0.01 \times 94 = 291.8$$

Thus,

$$e = y - \hat{y} = 301 - 291.8 = 9.2 \quad \text{(underestimate)}$$

2. (**Estimation of σ^2**) Estimate the error variance σ^2.

$n = 12$ and p (number of regression coefficients) $= 5$

$$\hat{\sigma}^2 = \frac{SS_E}{n-p} = \frac{1{,}699.0}{12-5} = 242.7 = 15.6^2$$

Exercise 12.1 (MR 12-8)

The pull strength of a wire bond is under study. Information of pull strength (y), die height (x_1), post height (x_2), loop height (x_3), wire length (x_4), bond width on the die (x_5), and bond width on the post (x_6) is collected as follows:

No.	pull strength (y)	die height (x_1)	post height (x_2)	loop height (x_3)	wire length (x_4)	bond width on the die (x_5)	bond width on the post (x_6)
1	8.0	5.2	19.6	29.6	94.9	2.1	2.3
2	8.3	5.2	19.8	32.4	89.7	2.1	1.8
3	8.5	5.8	19.6	31.0	96.2	2.0	2.0
4	8.8	6.4	19.4	32.4	95.6	2.2	2.1
5	9.0	5.8	18.6	28.6	86.5	2.0	1.8
6	9.3	5.2	18.8	30.6	84.5	2.1	2.1
7	9.3	5.6	20.4	32.4	88.8	2.2	1.9
8	9.5	6.0	19.0	32.6	85.7	2.1	1.9
9	9.8	5.2	20.8	32.2	93.6	2.3	2.1
10	10.0	5.8	19.9	31.8	86.0	2.1	1.8
11	10.3	6.4	18.0	32.6	87.1	2.0	1.6
12	10.5	6.0	20.6	33.4	93.1	2.1	2.1
13	10.8	6.2	20.2	31.8	83.4	2.2	2.1
14	11.0	6.2	20.2	32.4	94.5	2.1	1.9
15	11.3	6.2	19.2	31.4	83.4	1.9	1.8
16	11.5	5.6	17.0	33.2	85.2	2.1	2.1
17	11.8	6.0	19.8	35.4	84.1	2.0	1.8
18	12.3	5.8	18.8	34.0	86.9	2.1	1.8
19	12.5	5.6	18.6	34.2	83.0	1.9	2.0

By using a statistical software program, a fitted regression model of the pull strength data including four regressors (x_2, x_3, x_4, and x_5) and the corresponding error sum of squares (SS_E) are obtained as follows:

$$\hat{y} = 7.46 - 0.03x_2 + 0.52x_3 - 0.10x_4 - 2.16x_5 \quad \text{and} \quad SS_E = 10.91$$

1. Calculate the residual of $y = 9.0$ when $x_2 = 18.6$, $x_3 = 28.6$, $x_4 = 86.5$, and $x_5 = 2.0$.

2. Estimate the error variance σ^2.

12-2 Hypothesis Tests in Multiple Linear Regression

12-2.1 Test on Significance of Regression

Learning Goals

- ❑ Test the significance of multiple regression.
- ❑ Distinguish between the coefficient of multiple determination (R^2) and adjusted R^2.
- ❑ Calculate the R^2 and adjusted R^2 of a regression model.

Significance of Regression

The test for significance of regression determines if there is at least one x_j, out of k regressors, that has a significant linear relationship with the response variable Y, i.e.,

H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ ($\Rightarrow$ $\boldsymbol{\beta} = \mathbf{0}$)

H_1: $\beta_j \neq 0$ for at least one $j = 1$ to k

Like simple regression, the ANOVA method is used to test the significance of regression. For the multiple linear model, in Table 12-1, the total variability in Y (**SS_T or S_{yy}, total sum of squares**; degrees of freedom = $n - 1$) is partitioned into two components:

(1) Variability explained by the regression model (**SS_R, regression sum of squares**; degrees of freedom = k)

(2) Variability due to random error (**SS_E, error sum of squares**; degrees of freedom = $n - k - 1 = n - (k + 1) = n - p$)

Table 12-1 Partitioning of the Total Variability in Y (SS_T)

	SS_T (Total SS)		SS_R (Regression SS)		SS_E (Error SS)
	$\sum_{i=1}^{n}(y_i - \bar{y})^2$	=	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	+	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
$\Rightarrow$	$\mathbf{y}'\mathbf{y} - \dfrac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$	=	$\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \dfrac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$	+	$\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}$
DF	$n - 1$	=	k	+	$n - p$

(Note) SS: Sum of Squares; DF: Degrees of Freedom

The ratio of SS_R/k to $SS_E/(n - p)$ follows an F distribution:

$$F = \frac{SS_R / k}{SS_E / (n - p)} = \frac{MS_R}{MS_E} \sim F(k, n - p)$$

Since F becomes small as H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ is true, it is used to test the significance of regression. Note that **MS_R (regression mean square)** and **MS_E (error mean square)** are adjusted SS_R and SS_E by their corresponding degrees of freedom, respectively. Also note that MS_E is an estimate of the error variance:

$$MS_E = \frac{SS_E}{n - 2} = \hat{\sigma}^2$$

Test Procedure (F-test)

The variation quantities from the ANOVA analysis are summarized in Table 12-2 to test the significance of regression.

Table 12-2 ANOVA Table for Testing the Significance of Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-value
Regression	SS_R	k	MS_R	MS_R / MS_E	
Error	SS_E	$n - p$	MS_E		
Total	SS_T	$n - 1$			

Test Procedure (*F*-test) (*cont.*)

Step 1: State H_0 and H_1.

H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$

H_1: $\beta_j \neq 0$ for at least one j, $j = 1$ to k

Step 2: Determine a **test statistic and its value.**

$$F_0 = \frac{SS_R / k}{SS_E / (n-p)} = \frac{MS_R}{MS_E} \sim F(k, n-p)$$

Step 3: Determine a **critical value(s) for α.**

$$f_{\alpha,k,n-p}$$

Step 4: Make a **conclusion**. Reject H_0 if

$$f_0 > f_{\alpha,k,n-p}$$

R^2 vs. Adjusted R^2

The coefficient of determination (denoted as R^2) indicates the amount of variability in Y explained by the regressors $x_1, x_2, \ldots, x_k$:

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}, \quad 0 \leq R^2 \leq 1$$

As discussed in Section 11-8.2, a large value of R^2 does not necessarily imply that the regression model is adequate. As a new regressor is added to the existing model, R^2 will always increase regardless of whether the additional regressor is statistically significant or not; thus, a regression model having a large value of R^2 may unsatisfactorily estimate mean responses or predict new observations.

In contrast, the measure adjusted R^2 considers the number of regressors (p) included in the model in the computation:

$$\text{adjusted } R^2 = 1 - \frac{SS_E / (n-p)}{SS_T / (n-1)} = 1 - \frac{n-1}{n-p}(1 - R^2), \quad 0 \leq \text{adjusted } R^2 \leq 1$$

Adjusted R^2 is preferred to R^2 when comparison of several regression models having different numbers of regressors is needed. While R^2 simply increases as an additional regressor is included in the existing model, adjusted R^2 may decease if the reduction in SS_E by addition of a new regressor to the existing model is smaller than the MS_E of the existing model.

Example 12.2

For the power consumption data in Example 12-1, the following ANOVA table is obtained:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	f_0	*P*-value
Regression	4,957.2	4	1,239.3	5.1	0.03
Error	1,699.0	7	242.7		
Total	6,656.3	11			

1. **(Test on the Significance of Regression)** Test the significance of regression at $\alpha = 0.05$.

Example 12.2 (*cont.*)

Step 1: State H_0 and H_1.

H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

H_1: $\beta_j \neq 0$ for at least one j, j = 1 to 4

Step 2: Determine a **test statistic and its value.**

$$f_0 = \frac{SS_R / k}{SS_E /(n-p)} = \frac{MS_R}{MS_E} = \frac{1{,}239.3}{242.7} = 5.1$$

Step 3: Determine a **critical value(s) for** α.

$$f_{\alpha,k,n-p} = f_{0.05,4,7} = 4.12$$

Step 4: Make a **conclusion.**

Since $f_0 = 5.1 > f_{\alpha,k,n-p} = 4.12$, reject H_0 at $\alpha = 0.01$. This indicates that at least one of the four regressors has a significant linear relationship with the monthly power consumption at $\alpha = 0.05$.

2. (**R^2 and Adjusted R^2**) Calculate the coefficient of determination (R^2) and adjusted R^2 of the regression model.

$$R^2 = \frac{SS_R}{SS_T} = \frac{4{,}957.2}{6{,}656.3} = 74.5\%$$

The four regressors (x_1, x_2, x_3, and x_4) explain 74.5% of the variability in the power consumption Y.

$$\text{Adjusted } R^2 = 1 - \frac{n-1}{n-p}(1-R^2) = 1 - \frac{12-1}{12-5}(1-0.745) = 0.599$$

Exercise 12.2 (MR 12-22)

For the pull strength data in Exercise 12-1, the following ANOVA table is obtained by multiple regression including the four regressors (x_2, x_3, x_4, and x_5):

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	f_0	*P*-value
Regression	22.31	4	5.58	7.16	< 0.01
Error	10.91	14	0.78		
Total	33.22	18			

1. Test the significance of regression at $\alpha = 0.05$.
2. Calculate the coefficient of determination (R^2) and adjusted R^2 of the regression model.

12-2.2 Tests on Individual Regression Coefficients and Subsets of Coefficients

Learning Goals

- ❑ Test the significance of an individual regressor.
- ❑ Test the significance of a subset of regressors.

Significance of Individual Regressor

Like the testing on the significance of a regression coefficient in simple regression (see Section 11-5.1), the significance of any individual regressor x_j is evaluated by t test (called **partial** or **marginal test**).

Inference Context on β_j

Parameter of interest: β_j

Point estimator of β_j: $\hat{\beta}_j \sim N(\beta_j, \sigma^2 C_{jj})$, σ^2 is unknown.

Test statistic of β_j: $T = \dfrac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} = \dfrac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \sim t(n-p)$

(Note) **Estimated standard error** of $\hat{\beta}_j$: $se(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 C_{jj}}$, where $\hat{\sigma}^2 = MS_E$

Hypothesis Test on β_j (t-test)

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $\beta_j = \beta_{j,0}$
H_1: $\beta_j \neq \beta_{j,0}$

Step 2: Determine a **test statistic and its value.**

$$T_0 = \frac{\hat{\beta}_j - \beta_{j,0}}{\sqrt{\hat{\sigma}^2 C_{jj}}} \sim t(n-p)$$

Step 3: Determine a **critical value(s) for α.**

$t_{\alpha/2,n-p}$

Step 4: Make a **conclusion**. Reject H_0 if

$|t_0| > t_{\alpha/2,n-p}$

Example 12.3

(**Hypothesis Test on β_j**) Consider the power consumption data in Example 12-1. The following quantities are obtained for production volume (x_4) in the regression analysis:

$\hat{\beta}_4 = 0.014$, $se(\hat{\beta}_4) = 0.734$, $n = 12$, and $p = k + 1 = 5$

Test the significance of x_4 at $\alpha = 0.05$.

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $\beta_4 = 0$
H_1: $\beta_4 \neq 0$

Step 2: Determine a **test statistic and its value.**

$$t_0 = \frac{\hat{\beta}_4 - \beta_{4,0}}{se(\hat{\beta}_4)} = \frac{0.014 - 0}{0.734} = 0.02$$

Step 3: Determine a **critical value(s) for α.**

$t_{\alpha/2,n-p} = t_{0.05/2,12-5} = t_{0.025,7} = 2.365$

Step 4: Make a **conclusion.**

Since $|t_0| = 0.02 \not> t_{\alpha/2,n-p} = 2.365$, fail to reject H_0 at $\alpha = 0.05$.

Exercise 12.3

Consider the wire bond pull strength data in Exercise 12-1. The following quantities are obtained for bond width on the die (x_5) in the regression analysis:

$$\hat{\beta}_5 = -2.16,\ se(\hat{\beta}_5) = 2.39,\ n = 19, \text{ and } p = k + 1 = 5$$

Test the significance of x_5 at $\alpha = 0.05$.

Significance of the Subset of Regressors

The significance of a subset of regressors, $x_1, x_2, \ldots, x_r$ ($r < k$), is tested by the **extra sum of squares method (partial *F*-test; general regression significance test)**, which is an extension of the ANOVA method explained in Section 12-2.1. Like the test for significance of regression, this test determines if there is at least one x_j, out of r regressors, that has a significant linear relationship with Y, i.e.,

$$H_0:\ \beta_1 = \beta_2 = \cdots = \beta_r = 0$$
$$H_1:\ \beta_j \neq 0 \quad \text{for at least one } j, j = 1 \text{ to } r$$

To test the significance of the subset of regressors, let the vector of regression coefficients ($\boldsymbol{\beta}$) be partitioned into $\boldsymbol{\beta}_1$ for the regressors $x_1, x_2, \ldots, x_r$ and $\boldsymbol{\beta}_2$ for the other regressors $x_{r+1}, x_{r+2}, \ldots, x_k$:

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}$$

Also let $\mathbf{X}_1$ and $\mathbf{X}_2$ denote the columns of the regressors $\mathbf{X}$ related to $\boldsymbol{\beta}_1$ and those related to $\boldsymbol{\beta}_2$, respectively. Then, the multiple linear model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$$

In Section 12-2.1, the regression sum of squares (SS_R) and mean error square (MS_E) are

$$SS_R(\boldsymbol{\beta}) = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \quad \text{and} \quad MS_E = \frac{\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}}{n-p}$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Note that $SS_R(\boldsymbol{\beta})$ indicates the contribution of the regressors $\mathbf{X}$ to the variability in Y.

Now, to identify the contribution of the regressors $\mathbf{X}_1$, fit a reduced model for $\mathbf{X}_2$, i.e.,

$$\mathbf{y} = \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$$

The least squares estimate of $\boldsymbol{\beta}_2$ is $\hat{\boldsymbol{\beta}}_2 = (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{y}$ and the corresponding regression sum of squares is

$$SS_R(\boldsymbol{\beta}_2) = \hat{\boldsymbol{\beta}}_2'\mathbf{X}_2'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

Significance of the Subset of Regressors (*cont.*)

Then, the regression sum of squares due to $\boldsymbol{\beta}_1$ given that $\boldsymbol{\beta}_2$ is already in the model is

$$SS_R(\boldsymbol{\beta}_1 \mid \boldsymbol{\beta}_2) = SS_R(\boldsymbol{\beta}) - SS_R(\boldsymbol{\beta}_2) = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \hat{\boldsymbol{\beta}}_2'\mathbf{X}_2'\mathbf{y}$$

In other words, $SS_R(\boldsymbol{\beta}_1 \mid \boldsymbol{\beta}_2)$ indicates an increase in regression sum of squares by adding $\mathbf{X}_1$ to the reduced model, which is called extra sum of squares due to $\boldsymbol{\beta}_1$. The ratio of $SS_R(\boldsymbol{\beta}_1 \mid \boldsymbol{\beta}_2)/r$ to $SS_E/(n-p)$ follows an F distribution:

$$F = \frac{SS_R(\boldsymbol{\beta}_1 \mid \boldsymbol{\beta}_2)/r}{SS_E/(n-p)} = \frac{MS_R(\boldsymbol{\beta}_1 \mid \boldsymbol{\beta}_2)}{MS_E} \sim F(r, n-p)$$

Like the full model, the statistic F becomes <u>small</u> as H_0: $\beta_1 = \beta_2 = \cdots = \beta_r = 0$ is true.

Test Procedure (*F*-test)

The variation quantities of the ANOVA analysis for testing the significance of the subset of regressors $\boldsymbol{\beta}_1$ are summarized in Table 12-3.

Table 12-3 ANOVA Table for Testing the Significance of a Subset of Regressors $\boldsymbol{\beta}_1$

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression ($\boldsymbol{\beta}$)	SS_R	k	MS_R	MS_R / MS_E
$\boldsymbol{\beta}_1$	$SS_R(\boldsymbol{\beta}_1 \mid \boldsymbol{\beta}_2)$	r	$MS_R(\boldsymbol{\beta}_1 \mid \boldsymbol{\beta}_2)$	$\dfrac{MS_R(\boldsymbol{\beta}_1 \mid \boldsymbol{\beta}_2)}{MS_E}$
$\boldsymbol{\beta}_2$	$SS_R(\boldsymbol{\beta}_2)$	$k-r$	$MS_R(\boldsymbol{\beta}_2)$	
Error	SS_E	$n-p$	MS_E	
Total	SS_T	$n-1$		

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $\boldsymbol{\beta}_1 = \mathbf{0}$ $(\beta_1 = \beta_2 = \cdots = \beta_r = 0)$

H_1: $\boldsymbol{\beta}_1 \neq \mathbf{0}$ $(\beta_j \neq 0$ for at least one j, $j = 1$ to $r)$

Step 2: Determine a **test statistic and its value.**

$$F_0 = \frac{SS_R(\boldsymbol{\beta}_1 \mid \boldsymbol{\beta}_2)/r}{SS_E/(n-p)} = \frac{MS_R(\boldsymbol{\beta}_1 \mid \boldsymbol{\beta}_2)}{MS_E} \sim F(r, n-p)$$

Step 3: Determine a **critical value(s) for** $\boldsymbol{\alpha}$.

$$f_{\alpha,r,n-p}$$

Step 4: Make a **conclusion**. Reject H_0 if

$$f_0 > f_{\alpha,r,n-p}$$

Example 12.4

(**Test on the Subset of Regressors**) Consider the power consumption data in Example 12-1. Suppose that the four regressors ($\boldsymbol{\beta}$) are grouped into two categories:

(1) Non-production regressor ($\boldsymbol{\beta}_1$): temperature (x_1)

(2) Production regressors ($\boldsymbol{\beta}_2$): number of days (x_2), product purity (x_3) and production volume (x_4)

Multiple regression is performed separately with the four regressors ($\boldsymbol{\beta}$) and the non-production regressor ($\boldsymbol{\beta}_1$), yielding the ANOVA results below.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	f_0
Regression ($\boldsymbol{\beta}$)	4,957.2	4	1,239.3	5.1
Error	1,699.0	7	242.7	
Total	6,656.3	11		

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	f_0
Regression ($\boldsymbol{\beta}_1$)	3,758.9	1	3,758.9	13.0
Error	2,897.3	10	289.7	
Total	6,656.3	11		

Test the significance of the production regressors ($\boldsymbol{\beta}_2$) at $\alpha = 0.05$.

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $\boldsymbol{\beta}_2 = \mathbf{0}$ $(\beta_2 = \beta_3 = \beta_4 = 0)$

H_1: $\boldsymbol{\beta}_2 \neq \mathbf{0}$ ($\beta_j \neq 0$ for at least one j, $j = 2$ to 4)

Step 2: Determine a **test statistic and its value.**

$$SS_R(\boldsymbol{\beta}_2 \mid \boldsymbol{\beta}_1) = SS_R(\boldsymbol{\beta}) - SS_R(\boldsymbol{\beta}_1) = 4{,}957.2 - 3{,}758.9 = 1{,}198.3$$

$$MS_R(\boldsymbol{\beta}_2 \mid \boldsymbol{\beta}_1) = \frac{SS_R(\boldsymbol{\beta}_2 \mid \boldsymbol{\beta}_1)}{r} = \frac{1{,}198.3}{3} = 399.4$$

$$f_0 = \frac{SS_R(\boldsymbol{\beta}_2 \mid \boldsymbol{\beta}_1)/r}{SS_E/(n-p)} = \frac{MS_R(\boldsymbol{\beta}_2 \mid \boldsymbol{\beta}_1)}{MS_E} = \frac{399.4}{242.7} = 1.65$$

Step 3: Determine a **critical value(s) for α.**

$$f_{\alpha,r,n-p} = f_{0.05,3,7} = 4.35$$

Step 4: Make a **conclusion.**

Since $f_0 = 1.65 \ngtr f_{\alpha,r,n-p} = 4.35$, fail to reject H_0 at $\alpha = 0.05$. This indicates that none of the production regressors have a significant linear relationship with the power consumption at $\alpha = 0.05$.

Exercise 12.4

Consider the wire bond pull strength data in Exercise 12-1. Suppose that the six regressors ($\boldsymbol{\beta}$) are grouped into two categories:

(1) Process regressors ($\boldsymbol{\beta}_1$): die height (x_1), post height (x_2), and loop height (x_3)

(2) Wire-bond regressors ($\boldsymbol{\beta}_2$): wire length (x_4), bond width on the die (x_5), and bond width on the post (x_6)

Exercise 12.4 (*cont.*)

Multiple regression is performed separately with the six regressors (β) and the wire-bond regressors (β_2), yielding the ANOVA results below.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	f_0
Regression (β)	23.63	6	3.94	4.93
Error	9.59	12	0.80	
Total	33.22	18		

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	f_0
Regression (β_2)	10.81	3	3.61	2.41
Error	22.41	15	1.49	
Total	33.22	18		

Test the significance of the process regressors (β_1) at $\alpha = 0.05$.

12-3 Confidence Intervals in Multiple Regression

12-3.1 Confidence Intervals on Individual Regression Coefficients

Learning Goals

- ❑ Establish a confidence interval on β_j.

Inference Context on β_j

Test statistic of β_j: $T = \dfrac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} = \dfrac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \sim t(n-p)$

(Note) **Estimated standard error** of $\hat{\beta}_j$: $se(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 C_{jj}}$, where $\hat{\sigma}^2 = MS_E$

Confidence Interval on β_j

A 100(1 - α)% confidence interval on β_j is

$$\hat{\beta}_j - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}} \le \beta_j \le \hat{\beta}_j + t_{\alpha/2,n-j}\sqrt{\hat{\sigma}^2 C_{jj}}$$

Example 12.5

(**Confidence Interval on β_j**) Consider the power consumption data in Example 12-1. The following quantities are obtained for production volume (x_4) in the regression analysis:

$$\hat{\beta}_4 = 0.014,\ se(\hat{\beta}_4) = 0.734,\ n = 12, \text{ and } p = k + 1 = 5$$

Establish a 95% two-sided confidence interval on β_4.

$1 - \alpha = 0.95 \Rightarrow \alpha = 0.05$

95% two-sided CI on β_4:

$$\hat{\beta}_4 - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}} \le \beta_4 \le \hat{\beta}_4 + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}}$$

$$\Rightarrow 0.014 - t_{0.05/2,12-5} \times 0.734 \le \beta_4 \le 0.014 + t_{0.05/2,12-5} \times 0.734$$

Example 12.5 (*cont.*)

$\Rightarrow\ 0.014 - 2.365 \times 0.734 \le \beta_4 \le 0.014 + 2.365 \times 0.734$

$\Rightarrow\ -1.72 \le \beta_4 \le 1.75$

Exercise 12.5 (MR 12-31)

Consider the wire bond pull strength data in Exercise 12-1. The following quantities are obtained for bond width on the die (x_5) in the regression analysis:

$$\hat{\beta}_5 = \text{-}2.16,\ se(\hat{\beta}_5) = 2.39,\ n = 19, \text{ and } p = k + 1 = 5$$

Establish a 95% two-sided confidence interval on β_5.

12-3.2 Confidence Interval on the Mean Response

Learning Goals

- ☐ Establish a confidence interval on the mean response at $\mathbf{x} = \mathbf{x_0}$.

Sampling Distribution of Mean Response Estimator at $\mathbf{x} = \mathbf{x}_0$ ($\hat{\mu}_{Y|\mathbf{x}=\mathbf{x}_0}$)

Suppose that the vector $\mathbf{x_0}$ represents a particular point of the regressors $x_1, x_2, \ldots, x_k$:

$$\mathbf{x}_0 = \begin{bmatrix} 1 \\ x_{01} \\ x_{02} \\ \vdots \\ x_{0k} \end{bmatrix}$$

The mean response at $\mathbf{x} = \mathbf{x_0}$ is

$$\mu_{Y|\mathbf{x}_0} = E(Y \mid \mathbf{x}_0) = E(\mathbf{X\beta} + \varepsilon \mid \mathbf{x}_0) = \mathbf{x}_0'\boldsymbol{\beta}$$

Then, the point estimator of $\mu_{Y|\mathbf{x}_0}$ is

$$\hat{\mu}_{Y|\mathbf{x}_0} = \mathbf{x}_0'\hat{\boldsymbol{\beta}}$$

The sampling distribution of $\hat{\mu}_{Y|\mathbf{x}_0}$ is

$$\hat{\mu}_{Y|\mathbf{x}_0} \sim N(\mu_{Y|\mathbf{x}_0}, \sigma^2 \mathbf{x}_0'(\mathbf{X'X})^{-1}\mathbf{x}_0)$$

Inference Context on Mean Response at $\mathbf{x} = \mathbf{x}_0$ ($\mu_{Y|\mathbf{x}_0}$)

Parameter of interest: $\mu_{Y|\mathbf{x}_0}$

Point estimator of $\mu_{Y|\mathbf{x}_0}$: $\hat{\mu}_{Y|\mathbf{x}_0} = \mathbf{x}_0'\hat{\boldsymbol{\beta}} \sim N(\mu_{Y|\mathbf{x}_0}, \sigma^2 \mathbf{x}_0'(\mathbf{X'X})^{-1}\mathbf{x}_0)$, σ^2 is unknown.

Test statistic of $\mu_{Y|\mathbf{x}_0}$: $T = \dfrac{\hat{\mu}_{Y|\mathbf{x}_0} - \mu_{Y|\mathbf{x}_0}}{\sqrt{\hat{\sigma}^2 \mathbf{x}_0'(\mathbf{X'X})^{-1}\mathbf{x}_0}} \sim t(n-p)$

(Note) **Estimated standard error** of $\hat{\mu}_{Y|\mathbf{x}_0}$: $se(\hat{\mu}_{Y|\mathbf{x}_0}) = \sqrt{\hat{\sigma}^2 \mathbf{x}_0'(\mathbf{X'X})^{-1}\mathbf{x}_0}$, where $\hat{\sigma}^2 = MS_E$.

Confidence Interval on $\mu_{Y|\mathbf{x}_0}$

A 100(1 - α)% confidence interval on $\mu_{Y|\mathbf{x}_0}$ is

$$\hat{\mu}_{Y|\mathbf{x}_0} - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} \le \mu_{Y|\mathbf{x}_0} \le \hat{\mu}_{Y|\mathbf{x}_0} + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}$$

Example 12.6

(Confidence Interval on $\mu_{Y|\mathbf{x}_0}$) Consider the power consumption data in Example 12-1. The following quantities are obtained for the mean power consumption when $x_1 = 65°\text{F}$, $x_2 = 25$ days, $x_3 = 91\%$, and $x_4 = 94$ tons in the regression analysis:

$$\hat{\mu}_{Y|\mathbf{x}_0} = 291.81,\ se(\hat{\mu}_{Y|\mathbf{x}_0}) = 7.68,\ n = 12, \text{ and } p = 5$$

Establish a 95% two-sided confidence interval on the mean response.

$1-\alpha = 0.95 \quad \Rightarrow \quad \alpha = 0.05$

95% two-sided CI on $\mu_{Y|\mathbf{x}_0}$:

$$\hat{\mu}_{Y|\mathbf{x}_0} - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} \le \mu_{Y|\mathbf{x}_0} \le \hat{\mu}_{Y|\mathbf{x}_0} + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}$$

$$\Rightarrow\ 291.81 - t_{0.05/2,12-5} \times 7.68 \le \mu_{Y|\mathbf{x}_0} \le 291.81 + t_{0.05/2,12-5} \times 7.68$$

$$\Rightarrow\ 291.81 - 2.365 \times 7.68 \le \mu_{Y|\mathbf{x}_0} \le 291.81 + 2.365 \times 7.68$$

$$\Rightarrow\ 273.65 \le \mu_{Y|\mathbf{x}_0} \le 309.97$$

Exercise 12.6 (MR 12-31)

Consider the wire bond pull strength data in Exercise 12-1. The following quantities are obtained for the mean pull strength when $x_2 = 18.6$, $x_3 = 28.6$, $x_4 = 86.5$, and $x_5 = 2.0$ in the regression analysis:

$$\hat{\mu}_{Y|\mathbf{x}_0} = 8.66,\ se(\hat{\mu}_{Y|\mathbf{x}_0}) = 0.57,\ n = 19, \text{ and } p = 5$$

Establish a 95% two-sided confidence interval on the mean response.

12-4 Prediction of New Observations

Learning Goals

- ❑ Establish a confidence interval on the mean response at $\mathbf{x} = \mathbf{x}_0$.

Prediction Interval on Future Observation (y_0)

Suppose that the vector $\mathbf{x}_0' = [1, x_{01}, x_{02}, \cdots, x_{0k}]$ denote a particular point of the regressors $x_1, x_2, \ldots, x_k$. Then, a point estimate of a future observation y_0 at $\mathbf{x} = \mathbf{x}_0'$ is

$$\hat{y}_0 = \mathbf{x}_0'\hat{\boldsymbol{\beta}}$$

A 100(1 - α)% prediction interval on y_0 at $\mathbf{x}_0$ is

$$\hat{y}_0 - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2(1+\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)} \le y_0 \le \hat{y}_0 + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2(1+\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)}$$

As discussed in prediction of new observations in simple regression (Section 11-7), a prediction interval (PI) is always wider than the corresponding confidence interval (CI) on the mean response because the PI depends on both the error from the fitted regression model and the inherent variability in the random variable Y, while the CI depends on only the error from the fitted regression model.

Example 12.7

(Prediction Interval on y_0) Consider the power consumption data in Example 12-1. The following quantities are obtained in the regression analysis for a future power consumption observation when $x_2 = 18.6$, $x_3 = 28.6$, $x_4 = 86.5$, and $x_5 = 2.0$:

$$\hat{y}_0 = 291.81,\ \hat{\sigma}^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 = 7.68^2,\ \hat{\sigma}^2 = 15.58^2,\ n = 12, \text{ and } p = 5$$

Establish a 95% two-sided prediction interval on the power consumption.

$1 - \alpha = 0.95 \Rightarrow \alpha = 0.05$

95% two-sided CI on y_0:

$$\hat{y}_0 - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2(1+\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)} \le y_0 \le \hat{y}_0 + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2(1+\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)}$$

$$\Rightarrow 291.81 - t_{0.05/2,12-5} \times \sqrt{7.68^2 + 15.58^2} \le y_0 \le$$

$$291.81 + t_{0.05/2,12-5} \times \sqrt{7.68^2 + 15.58^2}$$

$$\Rightarrow 291.81 - 2.365 \times 17.37 \le y_0 \le 291.81 + 2.365 \times 17.37$$

$$\Rightarrow 250.74 \le y_0 \le 332.88$$

Exercise 12.7 (MR 12-31)

Consider the wire bond pull strength data in Exercise 12-1. The following quantities are obtained in the regression analysis for a future pull strength observation when $x_1 = 65°F$, $x_2 = 25$ days, $x_3 = 91\%$, and $x_4 = 94$ tons:

$$\hat{y}_0 = 8.66,\ \hat{\sigma}^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 = 0.57^2,\ \hat{\sigma}^2 = 0.88^2,\ n = 19, \text{ and } p = 5$$

Establish a 95% two-sided prediction interval on the pull strength.

12-5 Model Adequacy Checking

12-5.1 Residual Analysis

Learning Goals

- ❑ Describe the purpose of residual analysis.

Residual Analysis

Recall the four assumptions of a multiple linear model (see Section 12-1): (1) linearity between x_j's and Y, (2) randomness of error, (3) constant variance of error, and (4) normality of error. These four assumptions can be checked by analyzing residuals ($e_i = y_i - \hat{y}_i$), which is called the **assessment of model adequacy**.

The first three assumptions (linearity, randomness, and constant variance) can be checked by examining the following residual plots:

(1) e_i vs. $\hat{y}_i$

(2) e_i vs. x_{ij}, $j = 1, 2, \ldots, r$

(3) e_i vs. t (if time sequence is known)

Next, the normality of error is checked by one of the following:

(1) Frequency histogram of residuals

Residual Analysis (*cont.*)

(2) Normal probability plot of residuals

(3) Standardization of residuals: If the residuals are normally distributed, about 95% of the standardized residuals

$$d_i = \frac{e_i}{\sqrt{\hat{\sigma}^2}}, \; i = 1, 2, \ldots, n$$

are in the interval (-2, 2). Residuals whose standardized values are beyond ±3 may be outliers.

An undesirable residual pattern (see Section 11-8.1) may be due to non-linearity or insignificance of the relationship between the corresponding regressor and response variable (y). To correct an undesirable residual pattern, the appropriate method out of the following can be applied:

(1) Transform x_j or y into another form (such as $1/y$, $\sqrt{y}$, y^2, and $\ln y$).

(2) Add a high order term(s) of x_j to the model.

(3) Include a new regressor(s) in the model.

12-5.2 Influential Observations

Learning Goals

- ❑ Identify influential observations by using the measure Cook's distance.

Cook's Distance (D_i)

The measure Cook's distance (denoted by D_i) is the squared distance between the least squares estimate $\hat{\boldsymbol{\beta}}$ based on all n observations and $\hat{\boldsymbol{\beta}}_{(i)}$ based on all the data except the i^{th} observation:

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{p\hat{\sigma}^2}, \quad i = 1, 2, \ldots, n$$

As the i^{th} observation is unusually influential in determining the estimates of regression coefficients, D_i becomes large—a value of $\boldsymbol{D_i > 1}$ would indicate that the point is influential. If an influential observation is due to error during the experiment, this observation should be excluded from the analysis.

Example 12.8

(Influential Observation) Consider the power consumption data in Example 12-1. The Cook's distances of the observations are calculated by using software:

No.	Power consumption (y; unit: kW)	Cook's Distance (D_i)	No.	Power consumption (y; unit: kW)	Cook's Distance (D_i)
1	240	0.43	7	300	0.01
2	236	0.09	8	296	< 0.01
3	290	0.82	9	267	0.65
4	274	0.16	10	276	0.05
5	301	0.03	11	288	0.01
6	316	0.20	12	261	0.01

Check if any influential observations exist in the data.

No influential observations whose $D_i > 1$ are found in the data.

Exercise 12.8 (MR 12-40)

Consider the wire bond pull strength data in Exercise 12-1. The Cook's distances of the observations are calculated by using software as follows:

No.	Pull strength (y)	Cook's Distance (D_i)	No.	Pull strength (y)	Cook's Distance (D_i)
1	8.0	< 0.01	11	10.3	0.01
2	8.3	0.06	12	10.5	0.01
3	8.5	0.04	13	10.8	0.09
4	8.8	0.03	14	11.0	0.14
5	9.0	0.04	15	11.3	0.08
6	9.3	0.01	16	11.5	0.17
7	9.3	0.02	17	11.8	0.09
8	9.5	0.03	18	12.3	0.07
9	9.8	0.09	19	12.5	0.02
10	10.0	< 0.01			

Check if any influential observations exist in the data.

12-6 Aspects of Multiple Regression Modeling

12-6.1 Polynomial Regression Models

Learning Goals

- ❑ Establish a polynomial regression model.
- ❑ Test the significance of a high order term.

Polynomial Regression Model

When the response is curvilinear, a polynomial (a linear combination of powers in one or several variables) can be fit by using a multiple linear regression model. For example, a second-order polynomial in one variable can be transformed to a linear regression model as follows:

$$Y = \beta_0 + \beta_1 x + \beta_{11} x^2 + \varepsilon$$

$$\Rightarrow Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon, \text{ where } x_1 = x,\ x_2 = x^2, \text{ and } \beta_2 = \beta_{11},$$

When fitting a polynomial model, the lowest-order model is preferred for simplicity reasons (see Section 12-6.3). The significance of a higher-order term(s) is tested by t test or the extra sum of squares method.

Example 12.9

(Polynomial Model) Consider the power consumption data in Example 12-1. By including only temperature (x_1), the following second-order regression model is developed:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_1^2 = 162.76 + 3.50 x_1 - 0.02 x_1^2$$

The corresponding ANOVA table is as follows:

Example 12.9 (*cont.*)

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	f_0
Regression (x_1, x_1^2)	4,362.8	2	2,181.4	8.6
$SS_R(\beta_1 \mid \beta_0)$	3,758.9	1	3,758.9	13.0
$SS_R(\beta_2 \mid \beta_1, \beta_0)$	603.8	1	603.8	2.4
Error	2,293.5	9	254.8	
Total	6,656.3	11		

Test if the quadratic term of temperature (x_1^2) is significant at $\alpha = 0.05$.

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $\beta_2 = 0$
H_1: $\beta_2 \neq 0$

Step 2: Determine a **test statistic and its value.**

$$f_0 = \frac{MS_R(\beta_2 \mid \beta_1, \beta_0)}{MS_E} = \frac{603.8}{254.8} = 2.4$$

Step 3: Determine a **critical value(s) for** $\boldsymbol{\alpha}$.

$$f_{\alpha,r,n-p} = f_{0.05,1,9} = 5.12$$

Step 4: Make a **conclusion**.

Since $f_0 = 2.4 \not> f_{\alpha,r,n-p} = 5.12$, fail to reject H_0 at $\alpha = 0.05$. This indicates that the quadratic term does not significantly contribute to the model.

Exercise 12.9 (MR 12-50)

The following data are collected during an experiment to determine the change in thrust efficiency (y, in percent) as the divergence angle of a rocket nozzle (x) changes:

y	24.60	24.71	23.90	39.50	39.60	57.12
x	4.0	4.0	4.0	5.0	5.0	6.0

y	67.11	67.24	67.15	77.87	80.11	84.67
x	6.5	6.5	6.75	7.0	7.1	7.3

A polynomial model is developed for the thrust efficiency data:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 = -4.46 + 1.38x + 1.47x^2$$

The corresponding ANOVA table is as follows:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	f_0
Regression (x_1, x_1^2)	5,740.6	2	2,870.3	1,045.0
$SS_R(\beta_1 \mid \beta_0)$	5,716.3	1	5,716.3	1,167.0
$SS_R(\beta_2 \mid \beta_1, \beta_0)$	24.3	1	24.3	8.8
Error	24.7	9	2.7	
Total	5,765.3	11		

Test if the quadratic term of the nozzle divergence angle (x^2) is significant at $\alpha = 0.05$.

12-6.2 Categorical Regressors and Indicator Variables

Learning Goals

- ☐ Explain use of indicator variables in regression analysis.

Indicator Variables

The regression models presented in previous sections are based on continuous (quantitative) variables. However, categorical (qualitative) variables such as gender and education level are sometimes considered in regression analysis. To incorporate categorical variables in regression, indicator (dummy) variables are used. A categorical variable with c levels can be modeled by $c - 1$ indicator variables.

(e.g.) Indicator variables

(1) gender = {male, female}

$$x = \begin{cases} 0 & \text{for male} \\ 1 & \text{for female} \end{cases}$$

(2) education = {high school, college, and postgraduate}

x_1	x_2	
0	0	for high school
1	0	for college
0	1	for postgraduate

Another method of analyzing data including categorical information is to fit a separate regression model to the data of each category. Compared to this approach, the indicator variable approach has the following advantages:

1. **One regression model**: One regression model integrates several regression models for different categories.
2. **Increased statistical power**: By pooling the data of different categories, the error term will have a smaller mean square error due to a larger number of degrees of freedom ($n - p$), which results in increased power in statistical testing.

Example 12.10

(Indicator Variable) A relationship between liver weight (y; unit: g) and body surface area (x_1; unit: m^2) is under study. The following measurements are obtained from 20 cadavers (10 males and 10 females; gender (x_2) = 0 for male and 1 for female):

No.	Liver weight (y)	Body surface area (x_1)	Gender (x_2)	No.	Liver weight (y)	Body surface area (x_1)	Gender (x_2)
1	1690.2	1.6	0	11	1201.7	1.8	1
2	1314.1	1.5	0	12	1978.4	1.7	1
3	1530.0	1.6	0	13	1616.9	1.8	1
4	1270.1	1.8	0	14	1388.3	1.4	1
5	1924.7	1.8	0	15	1768.4	1.6	1
6	1016.1	1.5	0	16	1133.3	1.4	1
7	1323.8	1.6	0	17	1453.8	1.5	1
8	1314.1	1.7	0	18	1582.7	1.7	1
9	1760.6	1.7	0	19	923.3	1.5	1
10	1792.8	1.6	0	20	1583.7	1.7	1

Example 12.10 (*cont.*)

By assuming that the interaction between x_1 and x_2 is negligible, a fitted regression model is obtained for the liver weight data:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 = -108.7 + 977.0 x_1 - 1.3 x_2$$

Predictor	Coefficient	Standard error	t_0	*P*-value
Constant	-108.7	829.2	-0.13	0.897
BSA (x_1)	977.0	502.7	1.94	0.069
Gender (x_2)	-1.3	126.7	-0.01	0.992

Discuss the effect of gender (x_1) on liver weight (y) at $\alpha = 0.05$.

Since the *P*-value for β_2 is greater than 0.05, gender (x_1) does not have a significant effect on liver weight (y) at $\alpha = 0.05$.

Exercise 12.10 (MR 12-53)

A mechanical engineer is investigating the surface finish (y; unit: μm) of metal parts produced on a lathe and its relationship to the speed (x_1; unit: RPM) of the lathe. The following measurements are obtained by using two different types of cutting tools ($x_2 = 0$ for tool type 302 and 1 for tool type 416):

No.	Surface finish (y)	Lathe speed (x_1)	Type of cutting tool (x_2)	No.	Surface finish (Y)	Lathe speed (x_1)	Type of cutting tool (x_2)
1	45.44	225	0	11	33.50	224	1
2	42.03	200	0	12	31.23	212	1
3	50.10	250	0	13	37.52	248	1
4	48.75	245	0	14	37.13	260	1
5	47.92	235	0	15	34.70	243	1
6	47.79	237	0	16	33.92	238	1
7	52.26	265	0	17	32.13	224	1
8	50.52	259	0	18	35.47	251	1
9	45.58	221	0	19	33.49	232	1
10	44.78	218	0	20	32.29	216	1

The following regression model is fit to the surface finish data by using software:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2 = 11.50 + 0.15 x_1 - 6.09 x_2 - 0.03 x_1 x_2$$

Predictor	Coefficient	Standard error	t_0	*P*-value
Constant	11.50	2.50	4.59	<0.01
Lathe speed (x_1)	0.15	0.01	14.43	<0.01
Tool type (x_2)	-6.09	4.02	-1.51	0.15
$x_1 \times x_2$	-0.03	0.02	-1.79	0.09

Test if two regression models (with two different slopes and intercepts) are required to adequately model the tool life data.

12-6.3 Selection of Variables and Model Building

Learning Goals

- ❑ Explain the features of a satisfactory regression model.
- ❑ Compare variable selection techniques with each other: (1) all-possible regression evaluation; (2) forced selection/elimination; (3) forward selection; (4) backward elimination; and (5) stepwise regression.
- ❑ Determine the best set of regressors by the all-possible regression evaluation technique.
- ❑ Select a subset of regressors for model building by the stepwise regression technique.

Variable Selection for Good Model

In regression analysis, the features of a 'good' model include

1. **Validity**: The model includes regressors having empirical (engineering or practical) significance as well as statistical significance.
2. **Satisfactory performance**: The model provides estimates of the response (given conditions of the regressors) within an acceptable accuracy.
3. **Simplicity**: The model includes as few regressors as possible for ease of use.
4. **Stability**: The model does not change sensitively by use of a different set of observations.

These four criteria often conflict in many applications, thus a compromise is necessary to select appropriate regressors in a model. In this variable selection process, experience, judgment, and discussion among the analysts and practitioners involved in the study are important to establish a satisfactory model.

Variable Selection Techniques

No 'best' technique exists to select the best subset of regressors out of k candidate regressors. Available selection techniques include (1) all-possible regression evaluation; (2) forced selection/elimination; (3) forward selection; (4) backward elimination; and (5) stepwise regression. Each selection technique will show its own trade-off between computational efficiency and model performance in a specific application.

All-Possible Regression Evaluation

The all-possible regression evaluation technique generates all possible regression equations from the combinations of k regressors and then evaluates the equations by using selected criteria (which will be presented in the following paragraphs). This approach finds the best regression equation through comprehensive search and evaluation, but the number of equations to be examined increases geometrically as the number of candidate regressors increases (e.g., if $k = 3$, there are $2^3 = 8$ possible equations; if $k = 10$, there are $2^{10} = 1{,}024$ possible equations).

To determine a subset of regressors for the best model, four statistical criteria (R_p^2, adjusted R_p^2, $MS_E(p)$, and C_p; p = number of regression coefficients in the model, $p \le k + 1$) are available to use (see Figure 12-1). First, the best set of regressors is found where an increment in $\boldsymbol{R_p^2}$ (coefficient of determination) begins smaller than a value predetermined by the analyst. Note that R_p^2 ever increases as p increases (see Section 12-2.1).

All-Possible Regression Evaluation (*cont.*)

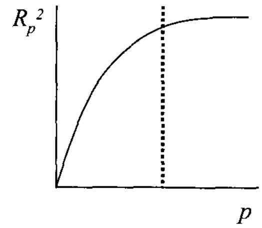
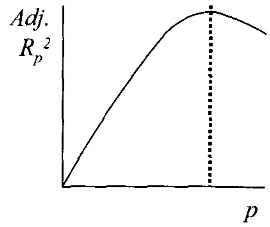
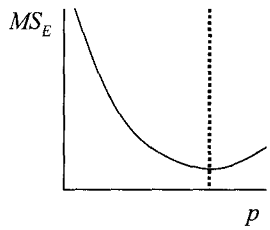
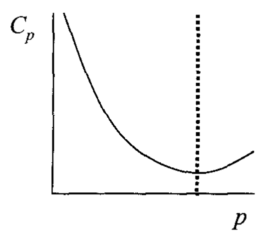
Criteria	Plot against p ($\le k + 1$)	Selection condition of p
R_p^2		Increase in R_p^2 smaller than the value predetermined
Adjusted R_p^2		Maximum adjusted R_p^2
$MS_E(p)$		Minimum $MS_E(p)$
C_p		Minimum C_p or $C_p \cong p$

Figure 12-1 Plots of variable selection criteria against p (number of regression coefficients).

Second, the best set of p regressors is determined at which **adjusted R_p^2** is at maximum. Note that adjusted R_p^2 may decrease as p increases unless the reduction in error sum of squares (SS_E) is greater than the error mean square (MS_E) of the existing model (see Section 12-2.1).

Third, the best set of p regressors is determined where $\boldsymbol{MS_E(p)}$ is at minimum. This MS_E criterion is equivalent to the adjusted R_p^2 criterion in the aspect that $MS_E(p)$ will increase if the reduction in SS_E is smaller than $MS_E(p - 1)$ due to the loss of one degree of freedom in the error term by the addition of a new regressor.

Lastly, the best set of p regressors is determined where $\boldsymbol{C_p}$, which is a measure of the total mean square error, is at minimum or close to p:

$$C_p = \frac{SS_E(p)}{\hat{\sigma}^2} - n + 2p, \quad \hat{\sigma}^2 = MS_E(k+1)$$

If the p-term regression model has negligible bias, the value of C_p is close to p; if the model is significantly biased, the value of C_p is significantly greater than p.

Example 12.11

(All-Possible Regression Evaluation) Consider the power consumption data in Example 12-1. By using software, the values of R_p^2, adjusted R_p^2, $MS_E(p)$, and C_p are obtained as follows:

No. regressors in the model	p	Regressors in the model	R_p^2 (%)	Adjusted R_p^2 (%)	$MS_E(p)$	C_p
1	2	x_2	64.5	60.9	15.4	1.7
1	2	x_1	56.5	52.1	17.0	3.9
1	2	x_3	0.2	<0.1	25.8	19.4
1	2	x_4	<0.1	<0.1	25.8	19.4
2	3	x_1, x_2	73.1	67.2	14.1	1.4
2	3	x_2, x_3	64.6	56.8	16.2	3.7
2	3	x_2, x_4	64.5	56.6	16.2	3.7
2	3	x_1, x_3	64.1	56.2	16.3	3.8
2	3	x_1, x_4	56.5	46.8	17.9	5.9
3	4	x_1, x_2, x_3	74.5	64.9	14.6	3.0
3	4	x_1, x_2, x_4	73.2	63.1	14.9	3.4
3	4	x_2, x_3, x_4	64.7	51.4	17.1	5.7
3	4	x_1, x_3, x_4	64.1	50.7	17.3	5.8
4	5	x_1, x_2, x_3, x_4	74.5	59.9	15.6	5.0

Discuss which model is preferable for the power consumption data.

A two-regressor model including x_1 and x_2 is recommended for it has the maximum value of adjusted R_p^2 and minimum values of $MS_E(p)$ and C_p.

Exercise 12.11 (MR 12-57)

(All-Possible Regression Evaluation) Consider the wire bond pull strength data in Exercise 12-1. By using software, the values of R_p^2, adjusted R_p^2, $MS_E(p)$, and C_p are obtained as follows:

No. regressors in the model	p	Regressors in the model	R_p^2 (%)	Adjusted R_p^2 (%)	MS_E (p)	C_p
1	2	x_3	47.7	44.6	1.01	6.7
1	2	x_4	31.1	27.0	1.16	13.6
1	2	x_5	11.2	6.0	1.32	21.9
2	2	x_3, x_4	64.8	60.4	0.86	1.6
2	3	x_3, x_5	57.2	51.8	0.94	4.8
2	3	x_2, x_3	54.1	48.3	0.98	6.1
3	3	x_1, x_3, x_4	67.3	60.7	0.85	2.6
3	3	x_3, x_4, x_5	67.1	60.6	0.85	2.7
3	3	x_2, x_3, x_4	65.3	58.3	0.88	3.4
4	4	x_1, x_3, x_4, x_5	68.7	59.8	0.86	4.0
4	4	x_1, x_3, x_4, x_6	68.6	59.6	0.86	4.1
4	4	x_3, x_4, x_5, x_6	68.2	59.2	0.87	4.2
5	5	x_1, x_3, x_4, x_5, x_6	71.1	60.0	0.86	5.0
5	5	x_1, x_2, x_3, x_4, x_6	69.0	57.0	0.89	5.9
5	5	x_1, x_2, x_3, x_4, x_5	68.8	56.8	0.89	6.0
6	6	$x_1, x_2, x_3, x_4, x_5, x_6$	71.1	56.7	0.89	7.0

Discuss which model is preferable for the pull strength data.

Forced Selection/ Elimination

The forced selection/elimination technique includes/excludes regressors designated by the analyst in/from the model regardless of their statistical significance.

Forward Selection

The forward selection technique adds to the existing model the regressor x_i out of candidate regressors (not in the model) that has the largest partial F-value (f_i) if f_i exceeds the entry criterion f_{in} (corresponding probability $p_{\text{in}} = 0.10$ or 0.15; preselected by the analyst):

$$f_i = \frac{SS_R(\beta_i \mid \boldsymbol{\beta}_{\text{model}})}{MS_E(\boldsymbol{\beta}_{\text{model}}, \beta_i)}$$

In other words, regressors are added to the model one at a time as long as their partial F-values are <u>greater</u> than f_{in}.

Backward Elimination

The backward elimination technique includes all candidate regressors in the model and then eliminates from the model the regressor x_j that has the smallest partial F-value (f_j) if f_j is less than the removal criterion f_{out} (corresponding probability $p_{\text{out}} = 0.15$ or 0.20; predefined by the analyst):

$$f_j = \frac{SS_R(\beta_j \mid \boldsymbol{\beta}_{\text{model}} \text{ without } \beta_j)}{MS_E(\boldsymbol{\beta}_{\text{model}})}$$

In other words, regressors are removed from the model one at a time as long as their partial F-values are <u>less</u> than f_{out}.

Stepwise Regression

The stepwise regression technique is the mix of the forward selection and backward elimination techniques: forward selection followed by backward elimination. This mixed variable selection technique is widely used. The procedure of stepwise regression is as follows.

Step 1: Out of candidate regressors, include x_i in the model having the largest F-value if $f_i > f_{\text{in}}$ (corresponding $p_{\text{in}} = 0.10$ or 0.15).

Step 2: Among the regressors in the model, remove x_j from the model having the smallest F-value if $f_j > f_{\text{out}}$ (corresponding $p_{\text{out}} = 0.15$ or 0.20; $p_{\text{in}} \le p_{\text{out}}$).

Step 3: Repeat steps 1 and 2 until no other regressors are added to or removed from the model.

Example 12.12

(Stepwise Regression) Consider the power consumption data in Example 12-1. Out of the four regressors (x_1, x_2, x_3, and x_4), determine an appropriate set of regressors by the stepwise regression technique to build a regression model. Use software with $p_{\text{in}} = p_{\text{out}} = 0.15$ (equivalent to $f_{\text{in}} = f_{\text{out}} = f_{1,\,12-p} = 2.5$).

Two regressors, x_1 and x_2, are selected for model building.

Exercise 12.12 (MR 12-57)

Consider the wire bond pull strength data in Exercise 12-1. Out of the six regressors (x_1, x_2, x_3, x_4, x_5, and x_6), determine an appropriate set of regressors by the stepwise regression technique to build a regression model. Use software with $p_{\text{in}} = p_{\text{out}} = 0.15$ (equivalent to $f_{\text{in}} = f_{\text{out}} = f_{1,\,19-p} = 2.3$).

12-6.4 Multicollinearity

Learning Goals

- ☐ Explain the terms *multicollinearity* and *variance inflation factor* (VIF).
- ☐ Describe undesirable effects due to multicollinearity in multiple regression analysis.
- ☐ Identify regressors having multicollinearity based on VIF.
- ☐ Describe remedial measures to resolve the problem of multicollinearity.

Multicollinearity

When there are strong dependencies among the regressors x_j's, it is said that multicollinearity exists. The following undesirable effects can be observed due to multicollinearity:

1. **Unstable estimation of regression coefficients**: The estimates of regression coefficients, $\hat{\beta}_j$'s, are sensitive to a small change in data.
2. **High standard errors of regression coefficient estimates**: The standard errors of regression coefficient estimates, $se(\hat{\beta}_j)$'s, are unreasonably large.
3. **Contradicting test results**: Although the full regression model is significant by F-test, no individual regressors are found significant by t-test.

A regression model with the problem of multicollinearity will produce unsatisfactory estimations of mean responses or predictions of new observations beyond the region of the x-space surveyed (where extrapolation is needed). However, this regression model may produce satisfactory estimations on the response within the region of the x-space surveyed (where interpolation is performed).

Variance Inflation Factor (VIF)

The measure *variance inflation factor* (VIF) indicates the extent of multicollinearity:

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R_j^2}, \quad j = 1, 2, \ldots, n$$

where R_j^2 is the coefficient of determination between x_j and the other $k - 1$ regressors. The stronger the multicollinearity, the larger the value of R_j^2. If a VIF exceeds 10 (some researchers suggest 4 or 5), the problem of multicollinearity exists.

Remedial Measures

Remedial measures against multicollinearity include

1. **Inclusion of additional data**: Collect additional data to explore the possibility to break the existing linear dependencies between the regressors.
2. **Removal of regressors with large values of VIF**: Exclude regressors having large values of VIF from the model.

Example 12.13 **(Multicollinearity)** Consider the power consumption data in Example 12-1. The variance inflation factors of the four regressors (x_1, x_2, x_3, and x_4) are obtained by using software as follows:

Regressor	Variance inflation factor
Ambient temperature (x_1)	2.3
Number of days in the month (x_2)	2.2
Product purity (x_3)	1.3
Volume of production (x_4)	1.0

Interpret the VIF values.

Since all the VIF values are less than 10, the regression model does not have the problem of multicollinearity.

Exercise 12.13 Consider the wire bond pull strength data in Exercise 12-1. The variance inflation factors of the four regressors (x_2, x_3, x_4, and x_5) are obtained by using software as follows:

Regressor	Variance inflation factor
Post height (x_2)	1.44
Loop height (x_3)	1.07
Wire length (x_4)	1.41
Bond width on the die (x_5)	1.36

Interpret the VIF values.

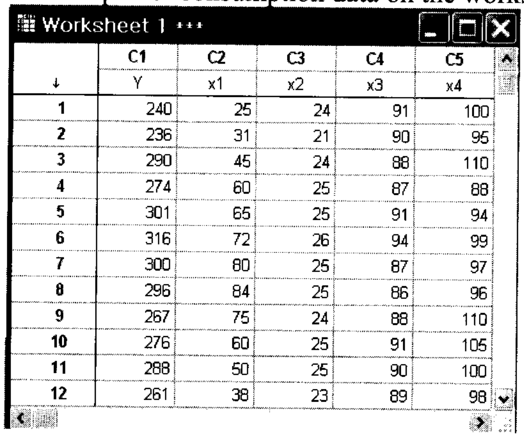
MINITAB Applications

Examples
12.1-3, 5-8, 13

(Multiple Linear Regression)

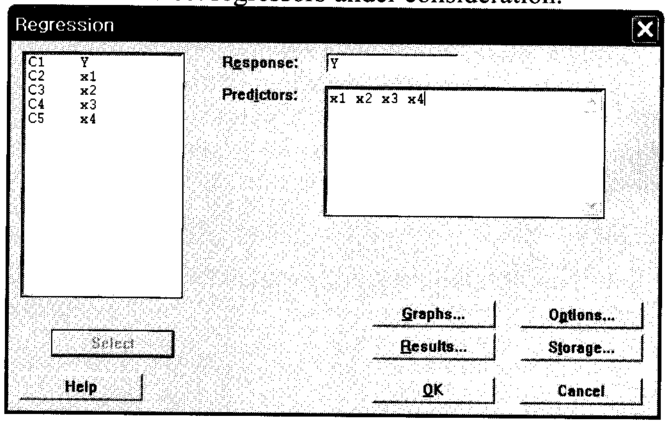
(1) Choose **File ➤ New**, click **Minitab Project**, and click **OK**.

(2) Enter the power consumption data on the worksheet.

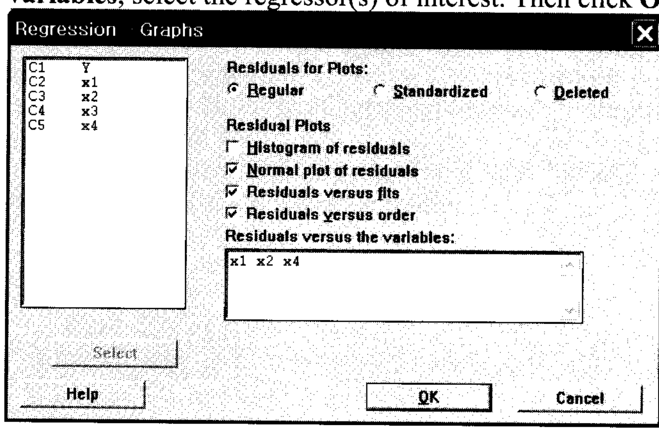

Worksheet 1 ***

	C1	C2	C3	C4	C5
↓	Y	x1	x2	x3	x4
1	240	25	24	91	100
2	236	31	21	90	95
3	290	45	24	88	110
4	274	60	25	87	88
5	301	65	25	91	94
6	316	72	26	94	99
7	300	80	25	87	97
8	296	84	25	86	96
9	267	75	24	88	110
10	276	60	25	91	105
11	288	50	25	90	100
12	261	38	23	89	98

(3) Choose **Stat ➤ Regression ➤ Regression**. In **Response** select Y and in **Predictors** select regressors under consideration.

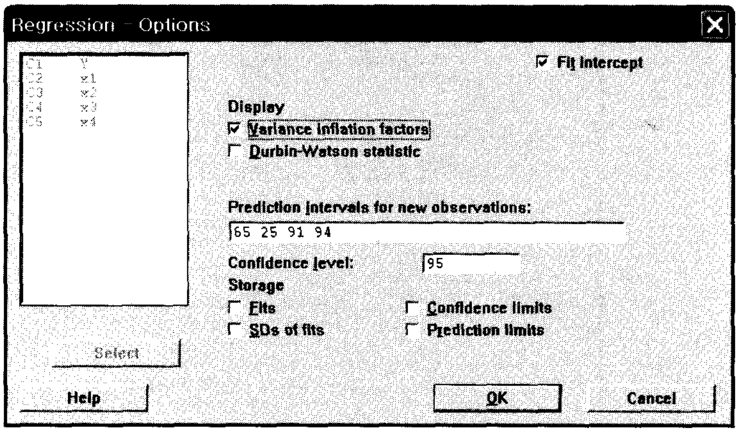


(4) Click **Graphs**. For **Residuals for Plots**, select **Regular** or **Standardized**. Under **Residual Plots**, check **Normal plot of residuals**, **Residuals versus fits**, and/or **Residuals versus order**. In **Residuals versus the variables**, select the regressor(s) of interest. Then click **OK**.

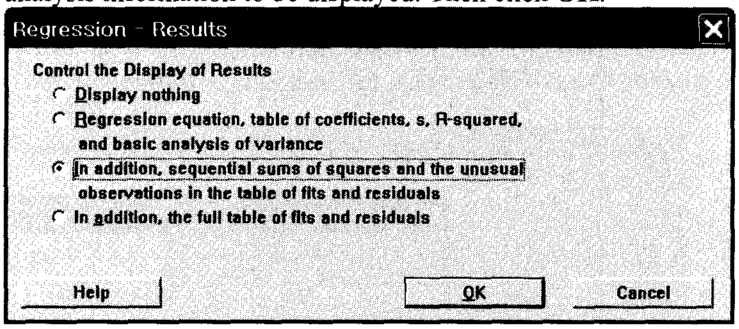


Examples 12.1-3, 5-8, 13 (*cont.*)

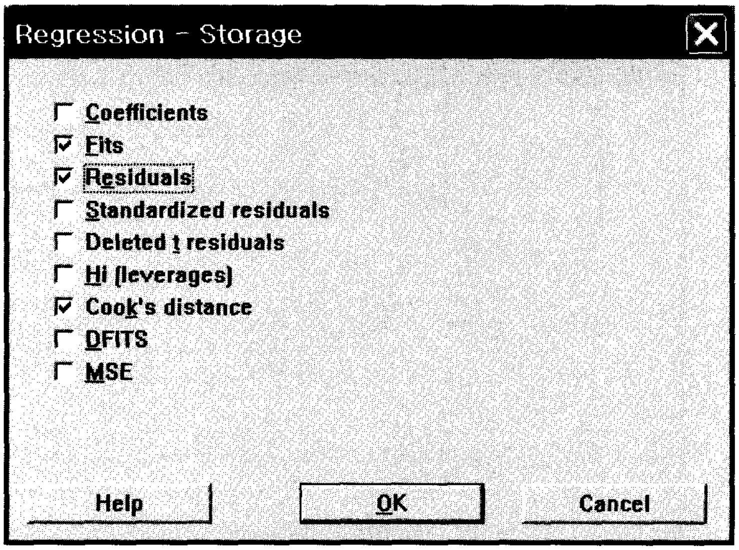
(5) Click **Options**. Check **Fit intercept**. Under **Display**, check **Variance inflation factors**. In **Prediction intervals for new observations**, enter the values of the regressors under consideration for the confidence interval on the mean response and/or prediction interval of a future observation. In **Confidence level**, type the level of confidence. Then click **OK**.



(6) Click **Results**. Under **Control the Display of Results**, click the level of analysis information to be displayed. Then click **OK**.



(7) Click **Storage**. Check **Fits, Residuals**, and **Cook's distance**. Then click **OK** twice.



Examples 12.1-3, 5-8, 13 (*cont.*)

(8) Interpret the analysis results.

```
Session

Regression Analysis

The regression equation is
Y = - 103 + 0.605 x1 + 8.92 x2 + 1.44 x3 + 0.014 x4

Predictor        Coef       StDev         T        P      VIF
Constant       -102.7       207.9     -0.49    0.636
x1             0.6054      0.3689      1.64    0.145      2.3      12.13
x2              8.924       5.301      1.68    0.136      2.2
x3              1.437       2.392      0.60    0.567      1.3
12.3, 5   x4   0.0136      0.7338      0.02    0.986      1.0

S = 15.58      R-Sq = 74.5%      R-Sq(adj) = 59.9%          12.2.2

Analysis of Variance

Source           DF          SS          MS        F        P
Regression        4      4957.2      1239.3     5.11    0.030    12.2.1
12.1.2  Residual Error    7      1699.8       242.7
Total            11      6656.3

Source       DF      Seq SS
x1            1      3758.9
x2            1      1109.4
x3            1        88.9
x4            1         0.1

Unusual Observations
Obs        x1          Y        Fit    StDev Fit    Residual    St Resid
  3      45.0     290.00     266.69        10.85       23.31       2.09R

R denotes an observation with a large standardized residual

Predicted Values

12.6     Fit   StDev Fit         95.0% CI              95.0% PI       12.7
      291.81        7.68   ( 273.63,  310.00)   ( 250.71,  332.92)
```

Worksheet 1 ***

	C1	C2	C3	C4	C5	C6	C7	C8
↓	Y	x1	x2	x3	x4	FITS1	RESI1	COOK1
1	240	25	24	91	100	258.758	-18.7580	0.430586
2	236	31	21	90	95	234.114	1.8862	0.092382
3	290	45	24	88	110	266.689	23.3109	0.820139
4	274	60	25	87	88	282.956	-8.9565	0.159823
5	301	65	25	91	94	291.815	9.1852	0.029524
6	316	72	26	94	99	309.356	6.6436	0.200280
7	300	80	25	87	97	295.186	4.8136	0.009353
8	296	84	25	86	96	296.157	-0.1568	0.000021
9	267	75	24	88	110	284.850	-17.8502	0.646149
10	276	60	25	91	105	288.938	-12.9376	0.047782
11	288	50	25	90	100	281.378	6.6218	0.010787
12	261	38	23	89	98	254.802	6.1980	0.010213

12.1.1

12.8

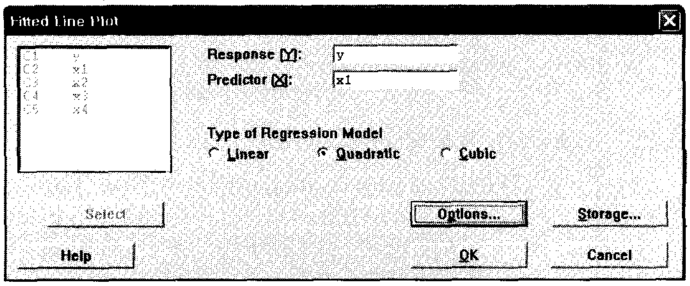
Example 12.4 **(Test on the Subset of Regressors)**

The regression model including x_1 is as follows:

```
Session

Regression Analysis

The regression equation is
Y = 224 + 0.953 x1

Predictor        Coef       StDev          T        P
Constant       224.38       15.88      14.13    0.000
x1             0.9525      0.2645       3.60    0.005

S = 17.02       R-Sq = 56.5%      R-Sq(adj) = 52.1%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      3758.9      3758.9     12.97    0.005
Residual Error    10      2897.3       289.7
Total             11      6656.3
```

Example 12.9 **(Polynomial Model)**

(1) Choose **Stat** ➤ **Regression** ➤ **Fitted Line Plot**. In **Response** select Y, in **Predictor** select the regressor under consideration, and under **Type of Regression Model** click the type of the polynomial to be developed. Then click **OK**.



(2) Interpret the analysis results.

```
Session

Polynomial Regression

Y = 162.755 + 3.49685X - 2.32E-02X**2
R-Sq = 65.5 %

Analysis of Variance

SOURCE           DF        SS         MS          F          P
Regression        2   4362.75    2181.38    8.56002   8.27E-03
Error             9   2293.50     254.83
Total            11   6656.25


SOURCE      DF     Seq SS          F          P
Linear       1    3758.92    12.9737   4.83E-03
Quadratic    1     603.83    2.36953   0.158109
```

Example 12.10 **(Indicator Variables)**

(1) Choose **File ➤ New**, click **Minitab Project**, and click **OK**.

(2) Enter the liver weight data on the worksheet.

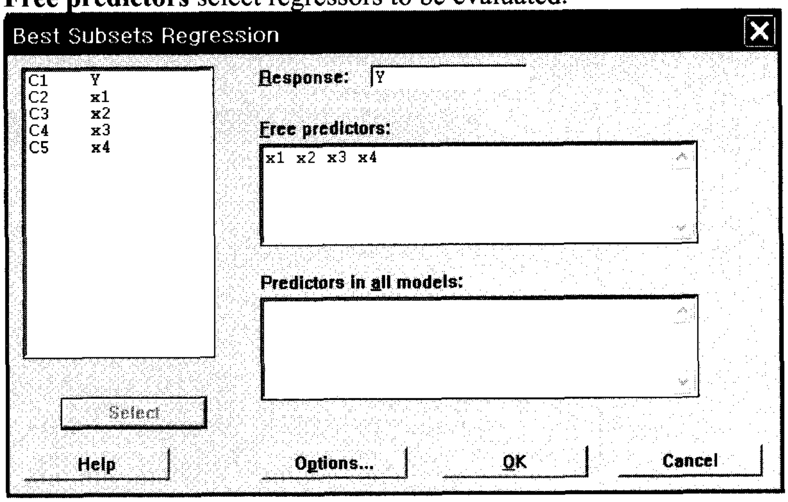
(3) Choose **Stat ➤ Regression ➤ Regression**. In **Response** select Y and in **Predictors** select regressors under consideration.

(4) Interpret the analysis results.

```
Session

Regression Analysis

The regression equation is
y = - 109 + 977 x1 - 1 x2

Predictor        Coef      StDev        T        P
Constant       -108.7      829.2    -0.13    0.897
x1              977.0      502.7     1.94    0.069
x2               -1.3      126.7    -0.01    0.992

S = 281.2       R-Sq = 18.4%     R-Sq(adj) = 8.8%

Analysis of Variance

Source            DF          SS         MS        F        P
Regression         2      303477     151738     1.92    0.177
Residual Error    17     1344478      79087
Total             19     1647955

Source        DF       Seq SS
x1             1       303469
x2             1            8
```

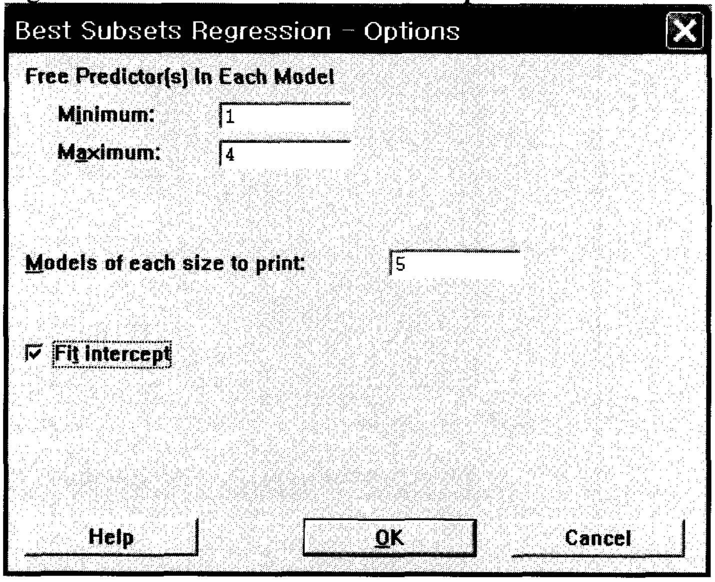
Example 12.11 **(All Possible Regression Evaluation)**

(1) Choose **Stat ➤ Regression ➤ Best Subsets**. In **Response** select Y and in **Free predictors** select regressors to be evaluated.



Example 12.11 (*cont.*)

(2) Click **Options**. In **Minimum** and **Maximum** enter the minimum and maximum numbers of regressors that will be included in regression models, respectively. In **Models of each size to print**, specify the number (1 to 5) of best regression models to be printed for each number of regressors in a model. Check **Fit intercept** and then click **OK** twice.



(3) Interpret the analysis results..

Session

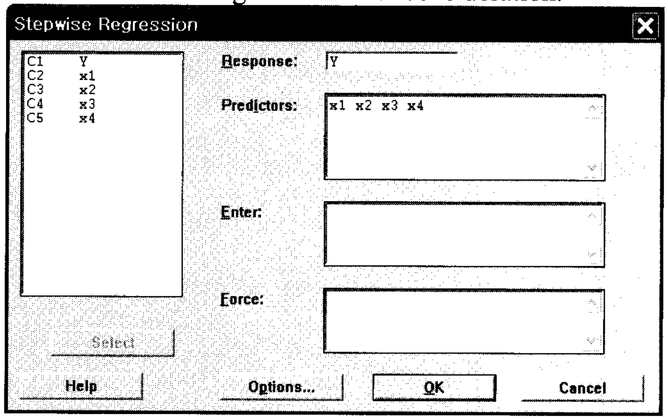
Best Subsets Regression

Response is y

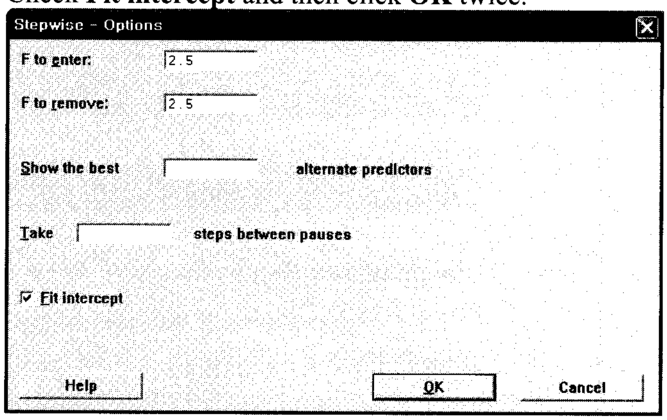
Vars	R-Sq	Adj. R-Sq	C-p	$\sqrt{MS_E} = s$	x 1	x 2	x 3	x 4
1	64.5	60.9	1.7	15.381		X		
1	56.5	52.1	3.9	17.022	X			
1	0.2	0.0	19.4	25.769			X	
1	0.0	0.0	19.4	25.799				X
2	73.1	67.2	1.4	14.095	X	X		
2	64.6	56.8	3.7	16.173		X	X	
2	64.5	56.6	3.7	16.210		X		X
2	64.1	56.2	3.8	16.289	X		X	
2	56.5	46.8	5.9	17.941	X			X
3	74.5	64.9	3.0	14.573	X	X	X	
3	73.2	63.1	3.4	14.944	X	X		X
3	64.7	51.4	5.7	17.149		X	X	X
3	64.1	50.7	5.8	17.273	X		X	X
4	74.5	59.9	5.0	15.579	X	X	X	X

Example 12.12 **(Stepwise Regression)**

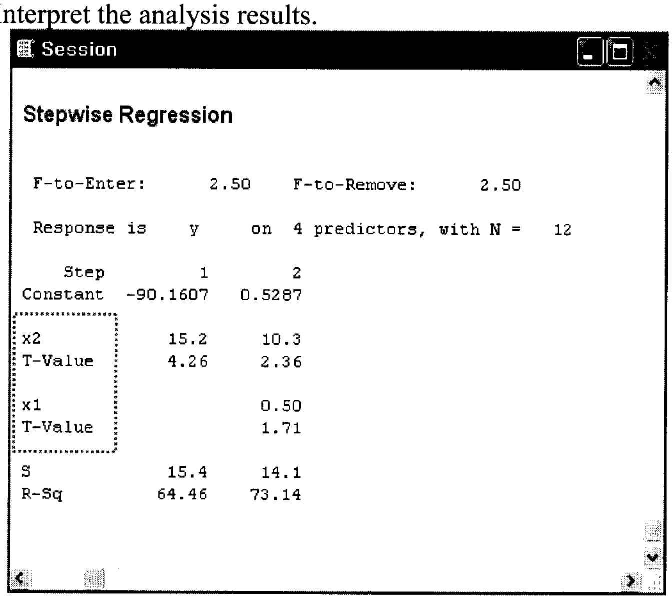
(1) Choose **Stat ➤ Regression ➤ Stepwise**. In **Response** select Y and in **Predictors** select regressors under consideration.



(2) Click **Options**. In **F to enter** and **F to enter** enter f_{in} and f_{out}, respectively. Check **Fit intercept** and then click **OK** twice.



(3) Interpret the analysis results.



Answers to Exercises

Exercise 12.1

(Calculation of Residual)

The estimate of the pull strength when $x_2 = 18.6$, $x_3 = 28.6$, $x_4 = 86.5$, and $x_5 = 2.0$ is

$$\hat{y} = 7.46 - 0.03\times 18.6 + 0.52\times 28.6 - 0.10\times 86.5 - 2.16\times 2.0 = 8.66$$

Thus,

$$e = y - \hat{y} = 9.0 - 8.66 = 0.34 \qquad \text{(underestimate)}$$

2. **(Estimation of σ^2)**

$n = 19$ and p (number of regression coefficients) = 5

$$\hat{\sigma}^2 = \frac{SS_E}{n-p} = \frac{10.91}{19-5} = 0.78 = 0.88^2$$

Exercise 12.2

1. **(Test on the Significance of Regression)**

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $\beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

H_1: $\beta_j \neq 0$ for at least one j, $j = 1$ to 4

Step 2: Determine a **test statistic and its value.**

$$f_0 = \frac{SS_R / k}{SS_E /(n-p)} = \frac{MS_R}{MS_E} = \frac{5.58}{0.78} = 7.16$$

Step 3: Determine a **critical value(s) for α.**

$$f_{\alpha,k,n-p} = f_{0.05,4,14} = 3.11$$

Step 4: Make a **conclusion.**

Since $f_0 = 7.16 > f_{\alpha,k,n-p} = 3.11$, reject H_0 at $\alpha = 0.05$. This indicates that at least one of the four regressors in the model has a significant linear relationship with the pull strength of a wire bond at $\alpha = 0.05$.

2. **(R^2 and Adjusted R^2)**

$$R^2 = \frac{SS_R}{SS_T} = \frac{22.31}{33.22} = 67.2\%$$

The four regressors (x_2, x_3, x_4, and x_5) explain 67.2% of the variability in the pull strength Y.

$$\text{Adjusted } R^2 = 1 - \frac{n-1}{n-p}(1 - R^2) = 1 - \frac{19-1}{19-5}(1 - 0.672) = 0.578$$

Exercise 12.3

(Hypothesis Test on β_j)

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $\beta_5 = 0$

H_1: $\beta_5 \neq 0$

Step 2: Determine a **test statistic and its value.**

$$t_0 = \frac{\hat{\beta}_5 - \beta_{5,0}}{se(\hat{\beta}_5)} = \frac{-2.16-0}{2.39} = -0.90$$

Step 3: Determine a **critical value(s) for α.**

$$t_{\alpha/2,n-p} = t_{0.05/2,19-5} = t_{0.025,14} = 2.145$$

Step 4: Make a **conclusion.**

Since $|t_0| = 0.90 \not> t_{\alpha/2,n-p} = 2.145$, fail to reject H_0 at $\alpha = 0.05$. This indicates that production volume (x_4) does not have a significant linear relationship with the pull strength at $\alpha = 0.05$.

Exercise 12.4

(Test on the Subset of Regressors)

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $\boldsymbol{\beta_1} = \mathbf{0}$ $(\beta_1 = \beta_2 = \beta_3 = 0)$

H_1: $\boldsymbol{\beta_1} \neq \mathbf{0}$ ($\beta_j \neq 0$ for at least one j, $j = 1$ to 3)

Step 2: Determine a **test statistic and its value.**

$$SS_R(\boldsymbol{\beta}_1 \mid \boldsymbol{\beta}_2) = SS_R(\boldsymbol{\beta}) - SS_R(\boldsymbol{\beta}_2) = 22.31 - 10.81 = 11.50$$

$$MS_R(\boldsymbol{\beta}_1 \mid \boldsymbol{\beta}_2) = \frac{SS_R(\boldsymbol{\beta}_1 \mid \boldsymbol{\beta}_2)}{r} = \frac{11.50}{3} = 3.83$$

$$f_0 = \frac{SS_R(\boldsymbol{\beta}_1 \mid \boldsymbol{\beta}_2)/r}{SS_E/(n-p)} = \frac{MS_R(\boldsymbol{\beta}_1 \mid \boldsymbol{\beta}_2)}{MS_E} = \frac{3.83}{0.80} = 4.79$$

Step 3: Determine a **critical value(s) for α.**

$$f_{\alpha,r,n-p} = f_{0.05,3,12} = 3.49$$

Step 4: Make a **conclusion.**

Since $f_0 = 4.79 > f_{\alpha,k,n-p} = 3.49$, reject H_0 at $\alpha = 0.05$. This indicates that at least one of the process regressors has a significant linear relationship with the pull strength at $\alpha = 0.05$.

Exercise 12.5

(Confidence Interval on β_j)

$1 - \alpha = 0.95 \quad \Rightarrow \quad \alpha = 0.05$

95% two-sided CI on β_5:

$$\hat{\beta}_5 - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}} \le \beta_5 \le \hat{\beta}_5 + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}}$$

$$\Rightarrow \; -2.16 - t_{0.05/2,19-5} \times 2.39 \le \beta_5 \le -2.16 + t_{0.05/2,19-5} \times 2.39$$

$$\Rightarrow \; -2.16 - 2.145 \times 2.39 \le \beta_5 \le -2.16 + 2.145 \times 2.39$$

$$\Rightarrow \; -7.29 \le \beta_5 \le 2.97$$

Exercise 12.6

(Confidence Interval on $\mu_{Y|\mathbf{x}_0}$)

$1-\alpha = 0.95 \quad \Rightarrow \quad \alpha = 0.05$

95% two-sided CI on $\mu_{Y|\mathbf{x}_0}$:

$$\hat{\mu}_{Y|\mathbf{x}_0} - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} \le \mu_{Y|\mathbf{x}_0} \le \hat{\mu}_{Y|\mathbf{x}_0} + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}$$

$$\Rightarrow \; 8.66 - t_{0.05/2,19-5} \times 0.57 \le \mu_{Y|\mathbf{x}_0} \le 8.66 + t_{0.05/2,19-5} \times 0.57$$

$$\Rightarrow \; 8.66 - 2.145 \times 0.57 \le \mu_{Y|\mathbf{x}_0} \le 8.66 + 2.145 \times 0.57$$

$$\Rightarrow \; 7.44 \le \mu_{Y|\mathbf{x}_0} \le 9.88$$

Exercise 12.7

(Prediction Interval on y_0)

$1-\alpha = 0.95 \quad \Rightarrow \quad \alpha = 0.05$

95% two-sided CI on y_0:

$$\hat{y}_0 - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2(1+\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)} \le y_0 \le \hat{y}_0 + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2(1+\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)}$$

$$\Rightarrow \; 8.66 - t_{0.05/2,19-5} \times \sqrt{0.57^2 + 0.88^2} \le y_0 \le$$

$$8.66 + t_{0.05/2,19-5} \times \sqrt{0.57^2 + 0.88^2}$$

$$\Rightarrow \; 8.66 - 2.145 \times 1.05 \le y_0 \le 8.66 + 2.145 \times 1.05$$

$$\Rightarrow \; 6.41 \le y_0 \le 10.91$$

Exercise 12.8

(Influential Observation)

Since all the D_i's are less than one, no influential observations are found in the data.

Exercise 12.9

(Polynomial Model)

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $\beta_2 = 0$

H_1: $\beta_2 \ne 0$

Step 2: Determine a **test statistic and its value.**

$$f_0 = \frac{MS_R(\beta_2 \mid \beta_1, \beta_0)}{MS_E} = \frac{24.3}{2.7} = 9.0$$

Step 3: Determine a **critical value(s) for $\boldsymbol{\alpha}$.**

$$f_{\alpha,r,n-p} = f_{0.05,1,9} = 5.12$$

Step 4: Make a **conclusion.**

Since $f_0 = 9.0 > f_{\alpha,r,n-p} = 5.12$, reject H_0 at $\alpha = 0.05$. This indicates that the quadratic term significantly contributes to the model.

Exercise 12.10

(Indicator Variable)

For tool type 302 ($x_2 = 0$), the fitted model is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 = 11.50 + 0.15 x_1$$

In contrast, for tool type 416 ($x_2 = 1$), the fitted model is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 + \hat{\beta}_3 x_1 = (\hat{\beta}_0 + \hat{\beta}_2) + (\hat{\beta}_1 + \hat{\beta}_3) x_1$$
$$= (11.50 - 6.09) + (0.15 - 0.03) x_1 = 5.41 + 0.12 x_1$$

However, since the P-values for β_2 and β_3 are greater than 0.05, it is concluded that two regression equations are not needed to model the tool life data.

Exercise 12.11

(All-Possible Regression Evaluation)

Either a two-regressor model including x_3 and x_4 or three-regressor model including x_1, x_3 and x_4 is recommended. The two-regressor model has the minimum value of C_p and the three-regressor model has the maximum value of adjusted R_p^2 and minimum value of $MS_E(p)$.

Exercise 12.12

(Stepwise Regression)

Two regressors, x_3 and x_4, are selected for model building.

Exercise 12.13

(Multicollinearity)

Since all the VIF values are less than 10, the regression model does not have the problem of multicollinearity.

13-2 The Completely Randomized Single-Factor Experiment

13-2.1 An Example

Learning Goals

- ☐ Identify dependent, independent, and extraneous variables in an experiment.
- ☐ Explain the terms *treatment* and *replicate.*
- ☐ Compare the following pairs of concepts each: controllable vs. uncontrollable variables; fixed-effects vs. random-effects factors; balanced vs. unbalanced designs; complete vs. restricted randomizations.
- ☐ Explain the purpose of randomization.

Tensile Strength Experiment (example)

A tensile strength experiment is introduced to explain fundamental terminologies of experimental design. Suppose that a manufacturer of paper is interested in improving the tensile strength of the product. The manufacturer wishes to identify if the tensile strength is affected by the hardwood concentration in the pulp used for the paper. The range of hardwood concentration of interest is between 5 and 20%; and, within the range, four levels ($a = 4$) of hardwood concentration are examined: 5, 10, 15, and 20%. At each concentration level, three specimens ($n_i = 3$, $i = 1$ to a) are tested in a laboratory. The measurements ($y_{ij}, j = 1$ to n_i, $i = 1$ to a) of this experiment can be summarized in a table format like Table 13-1.

Table 13-1 Data Structure of a Single Factor Experiment

	Factor Levels						
Replicate	1	2	...	i	...	a	
1	y_{11}	y_{21}	...	y_{i1}	...	y_{a1}	
2	y_{12}	y_{22}	...	y_{i2}	...	y_{a2}	
⋮	⋮	⋮	...	⋮	...	⋮	
j	y_{1j}	y_{2j}	...	y_{ij}	...	y_{aj}	
⋮	⋮	⋮	...	⋮	...	⋮	
n_i	y_{1n_1}	y_{2n_2}	...	y_{in_i}	...	y_{an_a}	
Totals	$y_{1.}$	$y_{2.}$	...	$y_{i.}$	...	$y_{a.}$	$y_{..}$
Averages	$\bar{y}_{1.}$	$\bar{y}_{2.}$	...	$\bar{y}_{i.}$	...	$\bar{y}_{a.}$	$\bar{y}_{..}$

Types of Variables

Three categories of variables exist in an experiment:

1. **Response (dependent) variable**: The outcome (response) of interest (e.g., tensile strength)
2. **Independent variable (factor)**: The variable whose effect on the response variable is under study (e.g., hardwood concentration)
3. **Extraneous variable (nuisance factor)**: The variable whose effect on the response is not under consideration but may affect the response variable (e.g., difference between specimens, temperature, etc.)

Treatment (Factor Level)

The term **treatment (factor level)** of a factor refers to an individual condition of the factor. For example, the tensile strength experiment has four factor levels: 5, 10, 15, and 20%.

Fixed-Effects vs. Random-Effects Factors

Two types of factors are defined depending on the selection method of factor levels: fixed-effects and random-effects factors. While the levels of a fixed-effects factor are specifically selected by the experimenter, those of a random-effects factor are chosen by random mechanism from a population of factor levels. The conclusion of a fixed-effects factor experiment cannot be extended beyond the factor levels selected, while that of a random-effects factor experiment can be extended to all the factor levels. For example, in the tensile strength experiment hardwood concentration is a fixed-effects factor because the four factor levels are specifically (not randomly) selected by the experimenter; thus, the corresponding conclusion is restricted to the four hardwood concentrations examined.

Controllable vs. Uncontrollable Variables

An extraneous variable can be classified into either a controllable or uncontrollable variable depending on if the condition of the variable can be controlled by the experimenter with an acceptable variation. For example, if the tensile strength experiment is conducted in a laboratory where its temperature is under control, temperature is a controllable variable.

Replicate

The term **replicate** refers to a repeated measurement under the same experimental condition. As an example, the tensile strength experiment has six replicates for each treatment.

Balanced vs. Unbalanced Designs

A design of experiment is classified into either a balanced or unbalanced design depending on the equality of replicates. A balanced design has the same number of replicates between the treatments, while an unbalanced design does not. For example, the tensile strength experiment is a balanced design.

Block

Recall that the paired t-test is used to block out the effect of heterogeneity between specimens (experimental units) (see Section 10-4). This heterogeneity between experimental units becomes a nuisance factor; and each experimental unit is called block.

Complete vs. Restricted Randomization

Randomization means that the order of trials is determined by random mechanism (such as random number table and random number generator) and two types of randomization are defined according to restriction in randomization:

(1) Complete randomization: All trials are randomized (see Table 13-2(b)).
(2) Restricted randomization: Trials are randomized within each block (see Table 13-2(c)).

Randomization balances out the effects of extraneous variables on the response. For example, in the tensile strength experiment, suppose that the tensile strength measurement device has a warm-up effect—an increase of two units in error per measurement. Table 13-2 shows errors in measurement due to the warm-up effect for different types of running trials; the warm-up effect becomes balanced out by randomization (from the range [30, 48] in sequential experiment to [34, 44] in complete randomization and [34, 42] in restricted randomization).

Complete vs. Restricted Randomization (*cont.*)

Table 13-2 Errors Due to Warm-Up Effects

(a) Sequential Order

Replicate	Hardwood Concentration (%) (order of trial) 5	10	15	20
1	2 (1)	4 (2)	6 (3)	8 (4)
2	10 (5)	12 (6)	14 (7)	16 (8)
3	18 (9)	20 (10)	22 (11)	24 (12)
Totals	30	36	42	48

(b) Complete Randomization

Replicate	Hardwood Concentration (%) (order of trial) 5	10	15	20
1	20 (10)	4 (2)	2 (1)	18 (9)
2	6 (3)	24 (12)	10 (5)	14 (7)
3	12 (6)	16 (8)	22 (11)	8 (4)
Totals	38	44	34	40

(c) Restricted Randomization

Block	Hardwood Concentration (%) (order of trial) 5	10	15	20
1	12 (6)	10 (5)	16 (8)	14 (7)
2	2 (1)	6 (3)	8 (4)	4 (2)
3	20 (10)	24 (12)	18 (9)	22 (11)
Totals	34	40	42	40

13-2.2 The Analysis of Variance

Learning Goals

- ❑ Test the significance of a fixed-effects factor in a completely randomized design.
- ❑ Establish a $100(1 - \alpha)\%$ confidence interval on μ_i and $\mu_i - \mu_j$ each..

Context

1. **Factor**: Fixed-effects factor with a treatments
2. **Blocking**: No blocks defined
3. **Repetition**: n replicates for each factor level (balanced design)
   (Note) N (total number of observations) $= n \times a$
4. **Randomization**: Complete randomization

Model

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} = \mu_i + \varepsilon_{ij} \quad \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, n \end{cases}$$

where: μ = overall mean

$\tau_i = i^{\text{th}}$ treatment effect, $\sum_{i=1}^{a} \tau_i = 0$

$\mu_i = i^{\text{th}}$ treatment mean

ε_{ij} = random error ~ *i.i.d.* $N(0, \sigma^2)$

Model (*cont.*)

(Notes) 1. $Y_{ij} \sim N(\mu_i, \sigma^2)$

2. In the fixed-effects model, the treatment effects (τ_i) are defined as deviations from the overall mean; thus, the sum of τ_i's is zero.

Least Square Estimators

The least squares method finds the estimates of μ and τ_i's (i = 1 to a) which minimize the sum of the squares of the errors (SS_E)

$$SS_E = L = \sum_{i=1}^{a}\sum_{j=1}^{n} \varepsilon_{ij}^2 = \sum_{i=1}^{a}\sum_{j=1}^{n} \left[y_{ij} - (\mu + \tau_i)\right]^2$$

By solving the partial derivatives of L with respect to μ and τ_i's, the least square estimators of μ, τ_i's, and μ_i's are determined:

$$\hat{\mu} = \bar{y}_{..}$$

$$\hat{\tau}_i = \bar{y}_{i.} - \bar{y}_{..}, \quad i = 1, 2, \ldots, a$$

$$\hat{\mu}_i = \hat{\mu} + \hat{\tau}_i = \bar{y}_{i.}, \quad i = 1, 2, \ldots, a$$

ANOVA

The ANOVA method is used to test the significance of a factor of interest. Like the simple regression model, the total variability in Y (***SS_T*, total sum of squares**; degrees of freedom = $N - 1$) is partitioned into two components (see Table 13-3):

(1) Variability explained by the factor (***$SS_{\text{Treatments}}$*, treatment sum of squares**; degrees of freedom = $a - 1$)

(2) Variability due to random error (***SS_E*, error sum of squares**; degrees of freedom = $N - a$)

Table 13-3 Partitioning of the Total Variability in Y (SS_T)

	SS_T (Total SS)		$SS_{\text{Treatments}}$ (Treatment SS)		SS_E (Error SS)
	$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2$	=	$\sum_{i=1}^{a}\sum_{j=1}^{n}(\bar{y}_{i.} - \bar{y}_{..})^2$	+	$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2$
$\Rightarrow$	$\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - \frac{1}{N}y_{..}^2$	=	$\frac{1}{n}\sum_{i=1}^{a} y_{i.}^2 - \frac{1}{N}y_{..}^2$	+	$\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - \frac{1}{n}\sum_{i=1}^{a} y_{i.}^2$
DF	$N-1$	=	$a-1$	+	$N-a$

(Note) SS: Sum of Squares; DF: Degrees of Freedom

The ratio of $SS_{\text{Treatments}}/(a - 1)$ to $SS_E/(N - a)$ follows an F distribution:

$$F = \frac{SS_{\text{Treatments}}/(a-1)}{SS_E/(N-a)} = \frac{MS_{\text{Treatments}}}{MS_E} \sim F(a-1, N-a)$$

This statistic F is used to test H_0: $\tau_1 = \tau_2 = \cdots = \tau_a = 0$, because F becomes small as the effect of the factor is insignificant. The quantities ***$MS_{\text{Treatments}}$* (treatment mean square)** and ***MS_E* (error mean square)** are adjusted $SS_{\text{Treatments}}$ and SS_E by their degrees of freedom, respectively. Note that

$$MS_E = \frac{SS_E}{N-a} = \hat{\sigma}^2$$

ANOVA (cont.)

The variation quantities from the ANOVA analysis are summarized in Table 13-4 to test the significance of the factor under consideration.

Table 13-4 ANOVA Table for Testing the Significance of a Factor: Completely Randomized Design

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Treatments	$SS_{\text{Treatments}}$	$a-1$	$MS_{\text{Treatments}}$	$MS_{\text{Treatments}} / MS_E$
Error	SS_E	$N-a$	MS_E	
Total	SS_T	$N-1$		

Hypothesis Test (*F*-test)

Step 1: State H_0 and H_1.

H_0: $\tau_1 = \tau_2 = \cdots = \tau_a = 0$

H_1: $\tau_i \neq 0$ for at least one i, $i = 1, 2, \ldots, a$

Step 2: Determine a **test statistic and its value.**

$$F_0 = \frac{SS_{\text{Treatments}} /(a-1)}{SS_E /(N-a)} = \frac{MS_{\text{Treatments}}}{MS_E} \sim F(a-1, N-a)$$

Step 3: Determine a **critical value(s) for** α.

$$f_{\alpha,a-1,N-a}$$

Step 4: Make a **conclusion**. Reject H_0 if

$$f_0 > f_{\alpha,a-1,N-a}$$

Unbalanced Design

The same ANOVA method of the balanced design is applicable to an unbalanced design with slight modifications. If n_i denotes the number of observations under the i^{th} treatment, the total number of observations is

$$N = \sum_{i=1}^{a} n_i$$

Then, the sums of squares for the unbalanced design are

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{a}\sum_{j=1}^{n_i} y_{ij}^{\,2} - \frac{1}{N} y_{..}^2$$

$$SS_{\text{Treatments}} = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(\bar{y}_{i.} - \bar{y}_{..})^2 = \sum_{i=1}^{a} \frac{y_{i.}^{\,2}}{n_i} - \frac{1}{N} y_{..}^2$$

$$SS_E = SS_T - SS_{\text{Treatments}}$$

A balanced design has two advantages compared to an unbalanced design:

(1) **Robustness**: The test procedure of the balanced design is less sensitive to a small departure from the assumption of equal error variance (σ^2).

(2) **Statistical power**: The power of the test of the balanced design is higher when the sample sizes of the two designs are the same.

Confidence Interval on μ_i

A 100(1 - α)% confidence interval on μ_i is

$$\bar{y}_{i.} - t_{\alpha/2,N-a}\sqrt{\frac{MS_E}{n}} \le \mu_i \le \bar{y}_{i.} + t_{\alpha/2,N-a}\sqrt{\frac{MS_E}{n}}$$

Confidence Interval on $\mu_i - \mu_j$

A 100(1 - α)% confidence interval on $\mu_i - \mu_j$ is

$$\bar{y}_{i.} - \bar{y}_{j.} - t_{\alpha/2,N-a}\sqrt{\frac{2MS_E}{n}} \le \mu_i - \mu_j \le \bar{y}_{i.} - \bar{y}_{j.} + t_{\alpha/2,N-a}\sqrt{\frac{2MS_E}{n}}$$

Example 13.1

A paper manufacturer is examining if the tensile strength of a paper product is affected by the hardwood concentration of the pulp used for the product. Four hardwood concentrations (5, 10, 15, and 20%; $a = 4$) are selected by the analyst and five specimens ($n = 5$) are tested at each concentration for tensile strength, resulting in the following:

Replicate	Hardwood Concentration (%; *i*) 5 (1)	10 (2)	15 (3)	20 (4)	
1	7	12	14	19	
2	8	17	18	25	
3	15	13	19	22	
4	11	18	17	23	
5	9	19	16	18	
Totals	50	79	84	107	320
Averages	10.0	15.8	16.8	21.4	16.0

The following summary quantities are also provided for the tensile strength data:

$$N = 20,\ \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 = 5{,}576,\ \text{and}\ \sum_{i=1}^{a} y_{i.}^2 = 27{,}246$$

(Note) The subscripts 1, 2, 3, and 4 correspond to 5, 10, 15, and 20% of hardwood concentration, respectively.

1. (**Test on the Significance of a Fixed-Effects Factor; Completely Randomized Design**) Perform the analysis of variance to test if there is any significant difference in tensile strength due to hardwood concentration at $\alpha = 0.05$.

The analysis of variance of the tensile strength data is as follows:

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - \frac{1}{N}y_{..}^2 = 5{,}576 - \frac{320^2}{20} = 456.0$$

$$SS_{\text{Treatments}} = \frac{1}{n}\sum_{i=1}^{a} y_{i.}^2 - \frac{1}{N}y_{..}^2 = 5{,}449.2 - \frac{320^2}{20} = 329.2$$

$$SS_E = SS_T - SS_{\text{Treatments}} = 456.0 - 329.2 = 126.8$$

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	f_0
Hardwood concentration	329.2	3	109.7	13.8
Error	126.8	16	7.9	
Total	456.0	19		

Example 13.1 (*cont.*)

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $\tau_1 = \tau_2 = \cdots = \tau_a = 0$

H_1: $\tau_i \neq 0$ for at least one i, $i = 1$ to 4

Step 2: Determine a **test statistic and its value.**

$$f_0 = \frac{SS_{\text{Treatments}}/(a-1)}{SS_E/(N-a)} = \frac{MS_{\text{Treatments}}}{MS_E} = \frac{109.7}{7.9} = 13.8$$

Step 3: Determine a **critical value(s) for α.**

$$f_{\alpha,a-1,N-a} = f_{0.05,4-1,20-4} = f_{0.05,3,16} = 3.24$$

Step 4: Make a **conclusion.**

Since $f_0 = 13.8 > f_{\alpha,a-1,N-a} = 3.24$, reject H_0 at $\alpha = 0.05$.

2. (**Confidence Interval on μ_i**) Establish a 95% two-sided confidence interval on μ_2 (mean of tensile strength at 10% of hardwood concentration).

$1 - \alpha = 0.95 \Rightarrow \alpha = 0.05$

95% two-sided CI on μ_2:

$$\bar{y}_{2.} - t_{\alpha/2,N-a}\sqrt{\frac{MS_E}{n}} \le \mu_2 \le \bar{y}_{2.} + t_{\alpha/2,N-a}\sqrt{\frac{MS_E}{n}}$$

$$\Rightarrow 15.8 - t_{0.05/2,20-4}\sqrt{\frac{7.9}{5}} \le \mu_2 \le 15.8 + t_{0.05/2,20-4}\sqrt{\frac{7.9}{5}}$$

$$\Rightarrow 15.8 - 2.120 \times 1.257 \le \mu_2 \le 15.8 + 2.120 \times 1.257$$

$$\Rightarrow 13.1 \le \mu_2 \le 18.5$$

3. (**Confidence Interval on μ_i - μ_j**) Establish a 95% two-sided confidence interval on μ_1 - μ_2.

95% two-sided CI on μ_1 - μ_2:

$$\bar{y}_{1.} - \bar{y}_{2.} - t_{\alpha/2,N-a}\sqrt{\frac{2MS_E}{n}} \le \mu_1 - \mu_2 \le \bar{y}_{1.} - \bar{y}_{2.} + t_{\alpha/2,N-a}\sqrt{\frac{2MS_E}{n}}$$

$$\Rightarrow 10.0 - 15.8 - t_{0.05/2,20-4}\sqrt{\frac{2\times 7.9}{5}} \le \mu_1 - \mu_2 \le$$

$$10.0 - 15.8 + t_{0.05/2,20-4}\sqrt{\frac{2\times 7.9}{5}}$$

$$\Rightarrow -5.8 - 2.120 \times 1.778 \le \mu_1 - \mu_2 \le -5.8 + 2.120 \times 1.778$$

$$\Rightarrow -9.6 \le \mu_1 - \mu_2 \le -2.0$$

Exercise 13.1 (MR 13-5)

The effect of cathode ray tube coating type on the tube conductivity is under study. Five different types of coating ($a = 5$) are selected by the analyst and four replicates ($n = 4$) are run for each coating type, resulting in the data as follows:

Exercise 13.1 (*cont.*)

	Coating Type					
Replicate	1	2	3	4	5	
1	143	152	134	129	147	
2	141	149	133	127	148	
3	150	137	132	132	144	
4	146	143	127	129	142	
Totals	580	581	526	517	581	2,785
Averages	145.0	145.3	131.5	129.3	145.3	139.3

The following summary quantities are also provided for the tensile strength data:

$$N = 20,\ \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^{2} = 389{,}115,\ \text{and}\ \sum_{i=1}^{a} y_{i.}^{2} = 1{,}555{,}487$$

1. Perform the analysis of variance to test if there is any significant difference in conductivity due to coating type at $\alpha = 0.05$.
2. Establish a 95% two-sided confidence interval on μ_1 (mean of tube conductivity for coating type 1).
3. Establish a 95% two-sided confidence interval on $\mu_2 - \mu_3$.

13-2.3 Multiple Comparisons Following the ANOVA

Learning Goals

❑ Perform multiple comparisons by Fisher's least significant difference (LSD) method.

Multiple Comparison Methods

Several methods are available to identify which treatment means are different from each other:

1. Scheffe's orthogonal contrast
2. Fisher's least significant difference (LSD)
3. Duncan's multiple range test
4. Tukey-Kramer's honestly significant difference (HSD) test

In this book, only the LSD method is presented.

Least Significant Difference (LSD) Method

Fisher's least significant difference (LSD) method is used to compare all possible pairs of treatment means, having the following null hypotheses

H_0: $\mu_i = \mu_j$ for all $i \neq j$, i and $j = 1, 2, \ldots, a$

Each null hypothesis H_0: $\mu_i = \mu_j$ will be rejected if

$$|t_0| > t_{\alpha/2, N-a}$$

In other words,

$$|t_0| = \frac{|\bar{y}_{i.} - \bar{y}_{j.}|}{\sqrt{2\hat{\sigma}^2 / n}} = \frac{|\bar{y}_{i.} - \bar{y}_{j.}|}{\sqrt{2MS_E / n}} > t_{\alpha/2, N-a}$$

$$\Rightarrow \quad |\bar{y}_{i.} - \bar{y}_{j.}| > t_{\alpha/2, N-a}\sqrt{\frac{2MS_E}{n}} = LSD$$

LSD Method (*cont.*)

As shown above, the pair of μ_i and μ_j will be concluded significantly different if the absolute difference of the corresponding sample means is greater than the LSD. Note that for an unbalanced design the LSD is defined as

$$LSD = t_{\alpha/2,N-a}\sqrt{MS_E\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

The procedure of the LSD method is as follows:

Step 1: Determine the **LSD** for α.
Step 2: Arrange the means of a treatments in **descending** order.
Step 3: Compare the difference between the **largest** and **smallest** means with the LSD. Continue this comparison with the **next smallest** mean as long as the mean difference is greater than the LSD.
Step 4: Continue Step 3 for the **next largest** mean until this iteration reaches the **second smallest** mean.
Step 5: Summarize the pairwise comparison results by arranging the a treatments in order of mean and then **underlining** treatments whose means are **not significantly different**.

Example 13.2

(Least Significant Difference Method) Consider the tensile strength data in Example 13-1. The following summary quantities are obtained from the analysis of variance:

$a = 4$, $n = 5$, $N = 20$, $MS_E = 7.9$

$\bar{y}_{1.} = 10.0$, $\bar{y}_{2.} = 15.8$, $\bar{y}_{3.} = 16.8$, and $\bar{y}_{4.} = 21.4$

Compare the means of the four treatments by using Fisher's LSD method. Use $\alpha = 0.05$.

Step 1: Determine the **LSD** for α.

$$LSD = t_{\alpha/2,N-a}\sqrt{\frac{2MS_E}{n}} = t_{0.05/2,20-4}\sqrt{\frac{2\times 7.9}{5}} = 2.120\times 1.778 = 3.8$$

Step 2: Arrange the means of a treatments in **descending** order.

$\bar{y}_{4.} = 21.4$, $\bar{y}_{3.} = 16.8$, $\bar{y}_{2.} = 15.8$, and $\bar{y}_{1.} = 10.0$

Step 3: Compare the difference between the **largest** and **smallest** means with the LSD. Continue this comparison with the **next smallest** mean as long as the mean difference is greater than the LSD.

4 vs. 1 = 21.4 − 10.0 = 11.4 > 3.8
4 vs. 2 = 21.4 − 15.8 = 5.6 > 3.8
4 vs. 3 = 21.4 − 16.8 = 4.6 > 3.8

Step 4: Continue Step 3 for the **next largest** mean until this iteration reaches the **second smallest** mean.

3 vs. 1 = 16.8 − 10.0 = 6.8 > 3.8
3 vs. 2 = 16.8 − 15.8 = 1.0 < 3.8
2 vs. 1 = 15.8 − 10.0 = 5.8 > 3.8

Step 5: Summarize the pairwise comparison results by arranging the a treatments in order of mean and then **underlining** treatments whose means are **not significantly different**.

$4 \quad \underline{\ 3 \quad 2\ } \quad 1$

Exercise 13.2 (MR 13-14)

Consider the tube conductivity data in Exercise 13-1. The following summary quantities are obtained from the analysis of variance:

$a = 5$, $n = 4$, $N = 20$, $MS_E = 16.2$

$\bar{y}_{1.} = 145.0$, $\bar{y}_{2.} = 145.3$, $\bar{y}_{3.} = 131.5$, $\bar{y}_{4.} = 129.3$, and $\bar{y}_{5.} = 145.3$

Compare the means of the five treatments by using Fisher's LSD method. Use $\alpha = 0.05$.

13-2.5 Residual Analysis and Model Checking

Learning Goals

- ❑ Describe the purpose of residual analysis in the analysis of variance.
- ❑ Calculate residuals.

Residual Analysis

In Section 13-2.2, the single-factor ANOVA model has three assumptions on the error term: (1) normality, (2) constant variance, and (3) randomness. These assumptions can be checked by analyzing residuals ($e_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_{i.}$), which is called the **assessment of model adequacy**.

First, the normality of error is checked by one of the following:

(1) Frequency histogram of residuals

(2) Normal probability plot of residuals

(3) Standardization of residuals: If the residuals are normally distributed, about 95% of the standardized residuals

$$d_{ij} = \frac{e_{ij}}{\sqrt{\hat{\sigma}^2}} = \frac{e_{ij}}{\sqrt{MS_E}}, \quad i = 1, 2, \ldots, a \text{ and } j = 1, 2, \ldots, n$$

are in the interval (-2, 2). Residuals whose standardized values are beyond ±3 may be outliers.

Next, the constant variance and randomness of error can be checked by examining the following residual plots:

(1) e_{ij} vs. factor level i, $i = 1, 2, \ldots, a$

(2) e_{ij} vs. $\bar{y}_{i.}$

(3) e_{ij} vs. t (if time sequence is known)

An undesirable residual pattern (see Section 11-8.1) in a residual plot may be corrected by transformation of the response variable (y) into another form such as $\sqrt{y}$ or $\log y$.

Example 13.3

(Calculation of Residual) Consider the tensile strength data in Example 13-1. The following summary quantities are obtained from the analysis of variance:

$\bar{y}_{1.} = 10.0$, $\bar{y}_{2.} = 15.8$, $\bar{y}_{3.} = 16.8$, and $\bar{y}_{4.} = 21.4$

Calculate the residual of $y_{12} = 8$.

$e_{12} = y_{12} - \hat{y}_{12} = y_{12} - \bar{y}_{1.} = 8 - 10 = -2$ (overestimate)

Exercise 13.3 (MR 13-5)

Consider the tube conductivity data in Exercise 13-1. The following summary quantities are obtained from the analysis of variance:

$\bar{y}_{1.} = 145.0$, $\bar{y}_{2.} = 145.3$, $\bar{y}_{3.} = 131.5$, $\bar{y}_{4.} = 129.3$, and $\bar{y}_{5.} = 145.3$

Calculate the residual of $y_{24} = 143$.

13-2.6 Determining Sample Size

Learning Goals

- ❑ Determine the sample size of a single-factor experiment by using an appropriate operating characteristic (OC) curve.

Operating Characteristic (OC) Curve

Operating characteristic (OC) curves in Figure 13-6 of MR plot required sample sizes n for a fixed-effects factor experiment at different levels of the parameters α, β, Φ, ν_1, and ν_2, i.e.,

$$n = f(\alpha, \beta, \Phi, \nu_1, \text{ and } \nu_2)$$

where: $$\Phi = \sqrt{\frac{n\sum_{i=1}^{a}\tau_i^2}{a\sigma^2}}$$

$$\nu_1 = a - 1$$

$$\nu_2 = a(n-1)$$

(Notes)

1. If the error variance σ^2 is unknown, an estimate of σ^2 (such as MS_E, conventional value, and subjective estimate) can be used.
2. The analyst can define the ratio $\sum_{i=1}^{a}\tau_i^2/\sigma^2$ (or σ_τ^2/σ^2) in sample size determination.
3. The sample size required increases as α, β, and $\sum_{i=1}^{a}\tau_i^2$ (or σ_τ^2) decrease and σ increases.

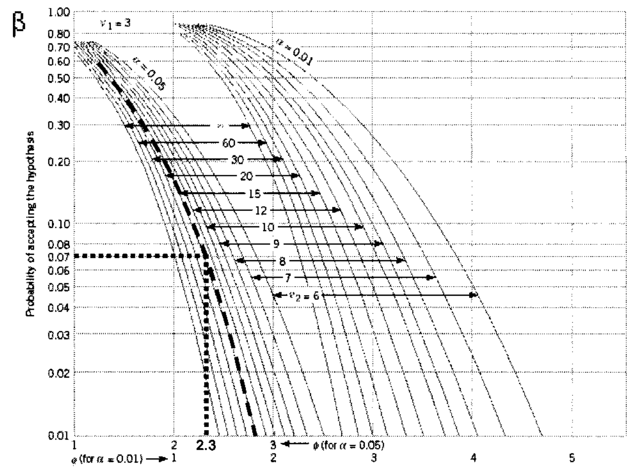
Example 13.4

(**Sample Size Determination; Fixed-Effects Factor**) Consider the tensile strength experiment in Example 13-1. Four hardwood concentrations ($a = 4$) are selected and their tensile strength means tested at $\alpha = 0.05$. Suppose that the four normal populations have common variance $\sigma^2 = 3^2$ and means $\mu_1 = 12$, $\mu_2 = 15$, $\mu_3 = 18$, and $\mu_4 = 20$ (i.e., $\sum_{i=1}^{a}\tau_i^2/\sigma^2 = 4.1$). Determine the number of replicates within each treatment to reject the hypothesis of equality of means with power of at least 0.90 (= 1 - β).

Use the operating characteristic curve of Figure 13-6(a) in MR for a fixed-effects factor with $\nu_1 = 3$ (= $a - 1 = 4 - 1$). From the OC curve, the power of the test according to n is found as follows:

n	Φ	$\nu_2 = a(n-1)$	β	Power (1 - β)
4	2.0	12	0.17	0.83
5	2.3	16	0.07	0.93

Example 13.4 (*cont.*)



Thus, at least $n = 5$ replicates are needed to achieve the designated power of the test.

Exercise 13.4 (MR 13-20)

Suppose that five normal populations ($a = 5$) have common variance $\sigma^2 = 10^2$ and means $\mu_1 = 175$, $\mu_2 = 190$, $\mu_3 = 160$, $\mu_4 = 200$, and $\mu_5 = 215$ (i.e., $\sum_{i=1}^{a} \tau_i^2 / \sigma^2 = 18.3$). How many observations per population must be taken so that the probability of rejecting the hypothesis of equality of means is at least 0.95? Use $\alpha = 0.01$.

13-3 The Random Effects Factor

Learning Goals

- ❑ Test the significance of a random-effects factor in a completely randomized design.
- ❑ Estimate the variances of error (σ^2), treatment effect (σ_τ^2), and response variable ($V(Y_{ij})$).

Context

1. **Factor**: Random-effects factor with a treatments
2. **Blocking**: No blocks defined
3. **Repetition**: n for each factor level (balanced design)
   (Note) N (total number of observations) $= n \times a$
4. **Randomization**: Complete randomization

Model

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} = \mu_i + \varepsilon_{ij} \quad \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, n \end{cases}$$

where: μ = overall mean

τ_i = i^{th} treatment effect ~ *i.i.d.* $N(0, \sigma_\tau^2)$

μ_i = i^{th} treatment mean

ε_{ij} = random error ~ *i.i.d.* $N(0, \sigma^2)$

Model (*cont.*)

(Notes) 1. The random variables τ_i and ε_{ij} are assumed independent of each other.

2. $Y_{ij} \sim N(\mu_i, \sigma_\tau^2 + \sigma^2)$

3. In the random-effects model, testing the individual treatment effects (τ_i) are meaningless because the treatments are selected at random.

ANOVA

The analysis of variance procedure of the fixed-effects model is valid for the random-effects model. The computation procedure for an ANOVA table in the fixed-effects model (see Section 13-2.2) is identical to that of the random-effects model.

However, the conclusion of ANOVA for the random-effects model can be extended to the entire population of treatments. This generalization of conclusion is due to random selection of treatments in the experiment.

The variance components (σ_τ^2 and σ^2) of the random-effects model can be estimated by using $MS_{\text{Treatnebts}}$ and MS_E:

$$\hat{\sigma}^2 = MS_E = \frac{SS_E}{N-a} \quad \text{and} \quad \hat{\sigma}_\tau^2 = \frac{MS_{\text{Treatments}} - MS_E}{n}$$

Therefore, the estimate of the variance of Y_{ij} is

$$\hat{V}(Y_{ij}) = \hat{\sigma}_\tau^2 + \hat{\sigma}^2$$

Hypothesis Test (*F*-test)

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

$$H_0: \ \sigma_\tau^2 = 0$$

$$H_1: \ \sigma_\tau^2 \neq 0$$

(Note) If all treatment effects are identical, $\sigma_\tau^2 = 0$; but if there is variability between treatment effects, $\sigma_\tau^2 \neq 0$.

Step 2: Determine a **test statistic and its value**.

$$F_0 = \frac{SS_{\text{Treatments}}/(a-1)}{SS_E/(N-a)} = \frac{MS_{\text{Treatments}}}{MS_E} \sim F(a-1, N-a)$$

Step 3: Determine a **critical value(s) for $\boldsymbol{\alpha}$**.

$$f_{\alpha,a-1,N-a}$$

Step 4: Make a **conclusion**. Reject H_0 if

$$f_0 > f_{\alpha,a-1,N-a}$$

Example 13.5

(**Estimation of Variance Components**) Consider the tensile strength data in Example 13-1. Assume that the four levels of hardwood concentration are randomly selected. The following summary quantities are obtained from the analysis of variance:

$$n = 5, \ MS_{\text{Treatments}} = 109.7, \text{ and } MS_E = 7.9$$

Estimate the (1) variability in tensile strength due to random error (σ^2), (2) variability due to hardwood concentration (σ_τ^2), and (3) total variability in tensile strength ($V(Y_{ij})$). Also interpret these variability estimation results.

Example 13.5 (*cont.*)

(1) $\hat{\sigma}^2 = MS_E = 7.9$

(2) $\hat{\sigma}_\tau^2 = \dfrac{MS_{\text{Treatments}} - MS_E}{n} = \dfrac{109.7 - 7.9}{5} = 20.4$

(3) $\hat{V}(Y_{ij}) = \hat{\sigma}_\tau^2 + \hat{\sigma}^2 = 20.4 + 7.9 = 28.3$

These variance estimates indicate that about 72% (= 20.4/28.3) of the total variability in tensile strength is attributable to hardwood concentration.

Exercise 13.5

Consider the tube conductivity data in Exercise 13-1. Assume that the five coating types are chosen at random. The following summary quantities are obtained from the analysis of variance:

$n = 4$, $MS_{\text{Treatments}} = 265.1$, and $MS_E = 16.2$

Estimate the (1) variability in tube conductivity due to random error (σ^2), (2) variability due to coating type (σ_τ^2), and (3) total variability in tube conductivity ($V(Y_{ij})$). Also interpret these variability estimation results.

13-4 Randomized Complete Block Design

Learning Goals

- ❑ Distinguish between complete block design and incomplete block design.
- ❑ Explain the effects of blocking on the error term.
- ❑ Test the significance of a fixed-effects factor in a randomized complete block design.
- ❑ Perform multiple comparisons by Fisher's least significant difference (LSD) method.

Complete vs. Incomplete Block Designs

A complete block design contains all the treatments in each block (see Table 13-5), while an incomplete block design may omit one or more treatments in each block. Recall that blocking is used to balance out the effect of a nuisance factor (heterogeneity between experimental units) on the response variable.

Table 13-5 Data Structure of a Complete Block Design

	Treatments							
Blocks	1	2	⋯	i	⋯	a	Totals	Means
1	y_{11}	y_{21}	⋯	y_{i1}	⋯	y_{a1}	$y_{.1}$	$\bar{y}_{.1}$
2	y_{12}	y_{22}	⋯	y_{i2}	⋯	y_{a2}	$y_{.2}$	$\bar{y}_{.2}$
⋮	⋮	⋮	⋯	⋮	⋯	⋮	⋮	⋮
j	y_{1j}	y_{2j}	⋯	y_{ij}	⋯	y_{aj}	$y_{.j}$	$\bar{y}_{.j}$
⋮	⋮	⋮	⋯	⋮	⋯	⋮	⋮	⋮
b	y_{1b}	y_{2b}	⋯	y_{ib}	⋯	y_{ab}	$y_{.b}$	$\bar{y}_{.b}$
Totals	$y_{1.}$	$y_{2.}$	⋯	$y_{i.}$	⋯	$y_{a.}$	$y_{..}$	
Means	$\bar{y}_{1.}$	$\bar{y}_{2.}$	⋯	$\bar{y}_{i.}$	⋯	$\bar{y}_{a.}$		$\bar{y}_{..}$

Context

1. **Factor**: Fixed-effects factor with a treatments
2. **Blocking**: Complete block design with b fixed-effect blocks
3. **Repetition**: One observation for each combination of factor levels and blocks (Note) N (total number observations) = $a \times b \times 1$
4. **Randomization**: Restricted randomization (randomization within each block)

Model

$$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij} = \mu_i + \beta_j + \varepsilon_{ij} \quad \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, b \end{cases}$$

where: μ = overall mean

τ_i = i^{th} treatment effect, $\sum_{i=1}^{a} \tau_i = 0$

β_j = j^{th} block effect, $\sum_{j=1}^{b} \beta_j = 0$

μ_i = i^{th} treatment mean

ε_{ij} = random error ~ *i.i.d.* $N(0, \sigma^2)$

(Notes) 1. $Y_{ij} \sim N(\mu_i, \sigma^2)$

2. The treatments and blocks are assumed to be independent of (not interacting with) each other.
3. Since the treatments and blocks are assumed to be fixed-effects factors in the model, the treatment effects (τ_i) and block effects (β_j) are defined as deviations from the overall mean; thus, the sums of τ_i's and β_j's are zero.

ANOVA

The analysis of variance method is used to test the significance of a factor in a randomized complete block design. The total variability in Y (**SS_T, total sum of squares**; degrees of freedom = $N - 1$) is partitioned into three components:

(1) Variability explained by the treatments (**$SS_{\text{Treatments}}$, treatment sum of squares**; degrees of freedom = a - 1)

(2) Variability explained by the blocks (**SS_{Blocks}, block sum of squares**; degrees of freedom =b - 1)

(3) Variability due to random error (**SS_E, error sum of squares**; degrees of freedom = $(a-1)(b-1)$)

In other words,

$$SS_T = SS_{\text{Treatments}} + SS_{\text{Blocks}} + SS_E$$

$$\Rightarrow \sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{..})^2 = b\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2 + a\sum_{j=1}^{b}(\bar{y}_{.j} - \bar{y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2$$

$$DF \quad N - 1 = a - 1 + b - 1 + (a-1)(b-1)$$

where: $$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij}^2 - \frac{y_{..}^2}{ab}$$

$$SS_{\text{Treatments}} = b\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2 = \frac{1}{b}\sum_{i=1}^{a} y_{i.}^2 - \frac{y_{..}^2}{ab}$$

ANOVA (*cont.*)

$$SS_{\text{Blocks}} = a\sum_{j=1}^{b}(\bar{y}_{.j} - \bar{y}_{..})^2 = \frac{1}{a}\sum_{j=1}^{b} y_{.j}{}^2 - \frac{y_{..}^2}{ab}$$

$$SS_E = SS_T - SS_{\text{Treatments}} - SS_{\text{Blocks}}$$

The ratio of $SS_{\text{Treatments}}/(a-1)$ to $SS_E/(a-1)(b-1)$ follows an F distribution:

$$F = \frac{SS_{\text{Treatments}}/(a-1)}{SS_E/(a-1)(b-1)} = \frac{MS_{\text{Treatments}}}{MS_E} \sim F(a-1,(a-1)(b-1))$$

This statistic F is used to test H_0: $\tau_1 = \tau_2 = \cdots = \tau_a = 0$, because F becomes small as the effect of the factor is insignificant. The quantities $\boldsymbol{MS_{\text{Treatments}}}$ **(treatment mean square)**, $\boldsymbol{MS_{\text{Blocks}}}$ **(block mean square)**, and $\boldsymbol{MS_E}$ **(error mean square)** are adjusted $SS_{\text{Treatments}}$, SS_{Blocks}, and SS_E by their corresponding degrees of freedom, respectively. Note that

$$MS_E = \frac{SS_E}{(a-1)(b-1)} = \hat{\sigma}^2$$

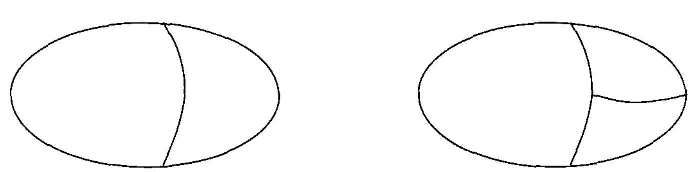
The variation quantities from the ANOVA analysis are summarized in Table 13-6 to test the significance of a factor under consideration.

Table 13-6 ANOVA Table for Testing the Significance of a Factor: Randomized Complete Block Design

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Treatments	$SS_{\text{Treatments}}$	$a-1$	$MS_{\text{Treatments}}$	$MS_{\text{Treatments}}/MS_E$
Blocks	SS_{Blocks}	$b-1$	MS_{Blocks}	
Error	SS_E	$(a-1)(b-1)$	MS_E	
Total	SS_T	$N-1$		

Effectiveness of Blocking

Compared to a completely randomized design (presented in Section 13.-2.2), a randomized complete block design reduces the sum of squares and degrees of freedom for error. In the block design the error sum of squares (see Figure 13-1) and corresponding degrees of freedom are decreased by SS_{Blocks} and $b-1$ (= $(N-a)-(a-1)(b-1)$; $N = ab$), respectively.



(a) (b)

Figure 13-1 The effect of blocking on error sum of squares (SS_E): (a) completely randomized design; (b) randomized complete block design.

Effectiveness of Blocking (*cont.*)

The effectiveness of blocking depends on the significance of block effects. If the effect of blocks is significant, a complete block design will result in a smaller error mean square (MS_E) than a completely randomized design, which makes the test more sensitive to detect the effect of treatments.

In general, when the significance of block effects is uncertain, apply blocking and check if the block effects exist. There would be a slight loss in the degrees of freedom for error, which may be negligible in the test unless the number of trials is very small.

Hypothesis Test (*F*-test)

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $\tau_1 = \tau_2 = \cdots = \tau_a = 0$

H_1: $\tau_i \neq 0$ for at least one i, $i = 1, 2, \ldots, a$

Step 2: Determine a **test statistic and its value**.

$$F_0 = \frac{SS_{\text{Treatments}}/(a-1)}{SS_E/(a-1)(b-1)} = \frac{MS_{\text{Treatments}}}{MS_E} \sim F(a-1, (a-1)(b-1))$$

Step 3: Determine a **critical value(s) for** $\boldsymbol{\alpha}$.

$$f_{\alpha, a-1, (a-1)(b-1)}$$

Step 4: Make a **conclusion**. Reject H_0 if

$$f_0 > f_{\alpha, a-1, (a-1)(b-1)}$$

Multiple Comparison Methods

The multiple comparison methods for the completely randomized design (see Section 13-2.3) are applicable to the complete block design. As an example, the value of Fisher's LSD is

$$LSD = t_{\alpha/2, (a-1)(b-1)} \sqrt{\frac{2MS_E}{b}}$$

Residual Analysis

The complete block design model has three assumptions on the error term: (1) normality, (2) constant variance, and (3) randomness. These assumptions can be checked by analyzing residuals

$$e_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - (\bar{y}_{i.} + \bar{y}_{.j} - \bar{y}_{..})$$

These residuals are analyzed by the model adequacy assessment methods explained in Section 13-2.5.

Example 13.6

The effect of grip angle on maximum grip strength (unit: lb.) is under study. Three grip angles ($a = 3$; 20°, 50°, and 80° from the horizontal) are selected by the experimenter. Five participants are recruited in the experiment; and each participant exerts his/her maximum strength at the three different grip angles presented at random. The grip strength measurements are as follows:

Example 13.6 (*cont.*)

Participant	Grip Angle 1 (20°)	2 (50°)	3 (80°)	Totals	Averages
1	45	53	50	148	49.3
2	64	75	71	210	70.0
3	54	59	52	165	55.0
4	72	80	79	231	77.0
5	62	71	65	198	66.0
Totals	297	338	317	952	
Averages	59.4	67.6	63.4		63.5

The following summary quantities are also provided for the grip strength data:

$$N = 15,\ \sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij}^{\ 2} = 62{,}132,\ \sum_{i=1}^{a} y_{i.}^{\ 2} = 302{,}942,\ \text{and}\ \sum_{j=1}^{b} y_{.j}^{\ 2} = 185{,}794$$

1. (**Test on the Significance of a Fixed-Effects Factor; Complete Block Design**) Perform the analysis of variance to test if there is any significant difference in maximum grip strength due to grip angle at $\alpha = 0.05$.

The analysis of variance of the grip strength data are as follows:

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij}^{\ 2} - \frac{y_{..}^2}{ab} = 62{,}132 - \frac{952^2}{15} = 1{,}711.7$$

$$SS_{\text{Treatments}} = \frac{1}{b}\sum_{i=1}^{a} y_{i.}^{\ 2} - \frac{y_{..}^2}{ab} = \frac{302{,}942}{5} - \frac{952^2}{15} = 168.1$$

$$SS_{\text{Blocks}} = \frac{1}{a}\sum_{j=1}^{b} y_{.j}^{\ 2} - \frac{y_{..}^2}{ab} = \frac{185{,}794}{3} - \frac{952^2}{15} = 1{,}511.1$$

$$SS_E = SS_T - SS_{\text{Treatments}} - SS_{\text{Blocks}} = 1{,}711.7 - 168.1 - 1{,}511.1 = 32.5$$

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Grip Angle	168.1	2	84.1	20.7
Participant	1,511.1	4	377.8	
Error	32.5	8	4.1	
Total	1,711.7	14		

Step 1: State H_0 and H_1.

H_0: $\tau_1 = \tau_2 = \cdots = \tau_a = 0$

H_1: $\tau_i \neq 0$ for at least one i, $i = 1$ to 3

Step 2: Determine a **test statistic and its value.**

$$f_0 = \frac{SS_{\text{Treatments}}/(a-1)}{SS_E/(a-1)(b-1)} = \frac{MS_{\text{Treatments}}}{MS_E} = \frac{84.1}{4.1} = 20.7$$

Step 3: Determine a **critical value(s) for** α.

$$f_{\alpha,a-1,(a-1)(b-1)} = f_{0.05,3-1,(3-1)(5-1)} = f_{0.05,2,8} = 4.46$$

Step 4: Make a **conclusion**.

Since $f_0 = 20.7 > f_{\alpha,a-1,(a-1)(b-1)} = 4.46$, reject H_0 at $\alpha = 0.05$.

Example 13.6 (*cont.*)

2. (**Least Significant Difference Method**) Compare these three treatment means by using Fisher's LSD method at $\alpha = 0.05$.

Step 1: Determine the **LSD** for α.

$$LSD = t_{\alpha/2,(a-1)(b-1)}\sqrt{\frac{2MS_E}{b}} = t_{0.05/2,8}\sqrt{\frac{2\times 4.1}{5}} = 2.306\times 1.281 = 3.0$$

Step 2: Arrange the means of a treatments in **descending** order.

$\bar{y}_{2.} = 67.6$, $\bar{y}_{3.} = 63.4$, and $\bar{y}_{1.} = 59.4$

Step 3: Compare the difference between the **largest** and **smallest** means with the LSD. Continue this comparison with the **next smallest** mean as long as the mean difference is greater than the LSD.

2 vs. 1 = 67.6 – 59.4 = 8.2 > 3.0
2 vs. 3 = 67.6 – 63.4 = 4.2 > 3.0

Step 4: Continue Step 3 for the **next largest** mean until this iteration reaches the **second smallest** mean.

3 vs. 1 = 63.4 – 59.4 = 4.0 > 3.0

Step 5: Summarize the pairwise comparison results.

2 3 1 (significantly different from each other)

3. (**Calculation of Residual**) Calculate the residual of $y_{34} = 79$.

$e_{34} = y_{34} - \hat{y}_{34} = y_{34} - (\bar{y}_{3.} + \bar{y}_{.4} - \bar{y}_{..}) = 79 - (63.4 + 77.0 - 63.5) = 2.1$

Exercise 13.6 (MR 13-29)

An experiment is conducted to investigate leaking current in a near-micron SOS MOSFETS device. This experiment investigates how leakage current varies as the channel length changes. Four channel lengths ($a = 4$) are selected by the experimenter; and for each channel length five different channel widths ($b = 5$) are used. Suppose that channel width is considered a nuisance factor. The data are as follows:

Channel Width	Channel Length 1	2	3	4	Totals	Means
1	0.7	0.8	0.9	1.0	3.4	0.9
2	0.8	0.8	1.0	1.5	4.1	1.0
3	0.8	0.9	1.7	2.0	5.4	1.4
4	0.9	0.9	2.0	3.0	6.8	1.7
5	1.0	1.0	4.0	2.0	8.0	2.0
Totals	4.2	4.4	9.6	9.5	27.7	
Means	0.8	0.9	1.9	1.9		1.4

The following summary quantities are also provided for the leaking current data:

$$N = 20,\ \sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij}^2 = 52.4,\ \sum_{i=1}^{a} y_{i.}^2 = 219.4, \text{ and } \sum_{j=1}^{b} y_{.j}^2 = 167.8$$

1. Perform the analysis of variance to test if there is any significant difference in leaking current due to the difference in channel length at $\alpha = 0.05$.
2. Compare these four treatment means by using Fisher's LSD method at $\alpha = 0.05$.
3. Calculate the residual of $y_{21} = 0.8$.

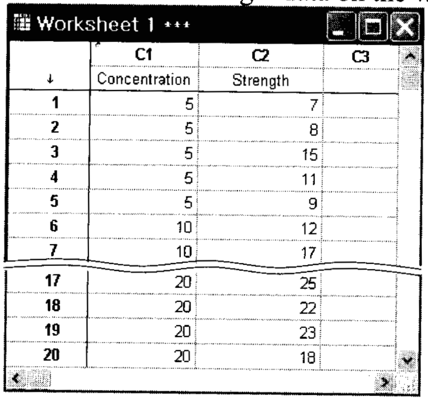
MINITAB Applications

Examples 13.1 - 3

(Fixed-Effects Factor; Completely Randomized Design)

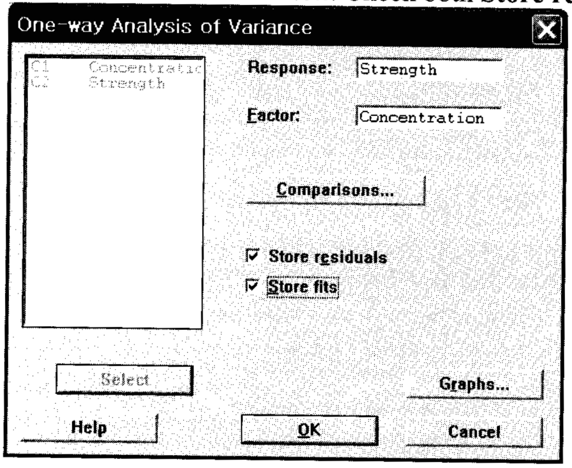
(1) Choose **File ➤ New**, click **Minitab Project**, and click **OK**.

(2) Enter the tensile strength data on the worksheet.

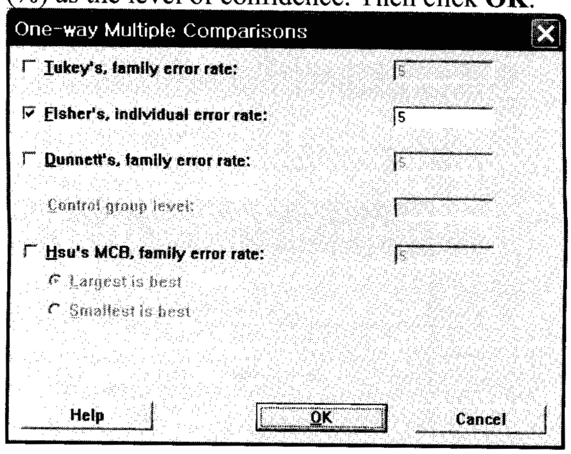


	C1	C2	C3
↓	Concentration	Strength	
1	5	7	
2	5	8	
3	5	15	
4	5	11	
5	5	9	
6	10	12	
7	10	17	
17	20	25	
18	20	22	
19	20	23	
20	20	18	

(3) Choose **Stat ➤ ANOVA ➤ One-way**. In **Response** select *Strength* and in **Factor** select *Concentration*. Check both **Store residuals** and **Store fits**.

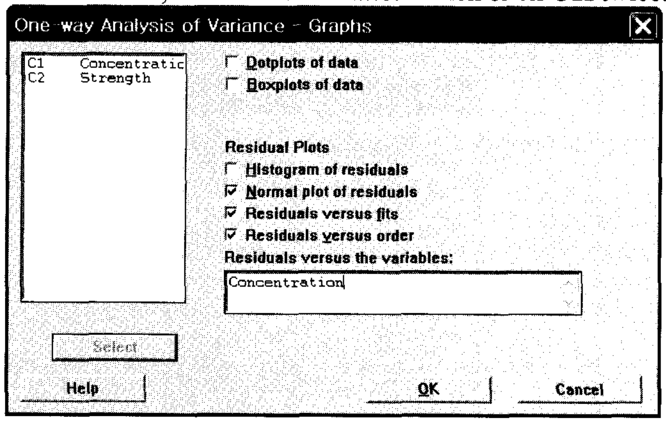


(4) Click **Comparisons**. Check **Fisher's, individual error rate** and type '5' (%) as the level of confidence. Then click **OK**.



Examples 13.1 - 3 (*cont.*)

(5) Click **Graphs**. Under **Residual Plots**, check **Normal plot of residuals, Residuals versus fits,** and **Residuals versus order**. In **Residuals versus the variables**, select *Concentration*. Then click **OK** twice.



(6) Interpret the analysis results.

Session

One-way Analysis of Variance

```
Analysis of Variance for Strength
Source      DF        SS        MS        F        P
Concentr     3    329.20    109.73    13.85    0.000      13.1.1
Error       16    126.80      7.93
Total       19    456.00
                                    Individual 95% CIs For Mean
                                    Based on Pooled StDev
Level        N      Mean     StDev  ------+---------+---------+---------+
 5           5    10.000     3.162  (----*----)
10           5    15.800     3.114             (-----*----)               13.1.2
15           5    16.800     1.924               (-----*----)
20           5    21.400     2.881                        (-----*----)
                                    ------+---------+---------+---------+
Pooled StDev =    2.815                 10.0      15.0      20.0      25.0

Fisher's pairwise comparisons

    Family error rate = 0.189
Individual error rate = 0.0500

Critical value = 2.120

Intervals for (column level mean) - (row level mean)                  13.2

                   5          10          15

      10      -9.575                                                  13.1.3
              -2.025

      15     -10.575      -4.775
              -3.025       2.775

      20     -15.175      -9.375      -8.375
              -7.625      -1.825      -0.825
```

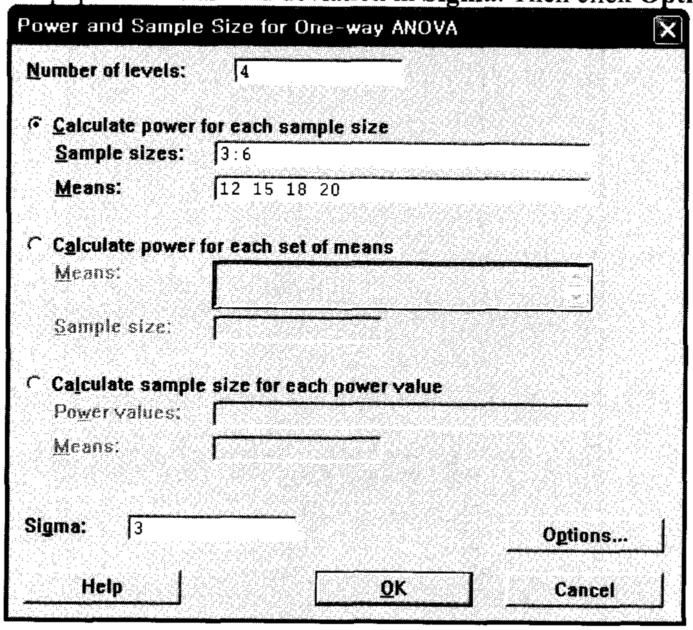
Worksheet 1 ***

	C1	C2	C3	C4
↓	Concentration	Strength	RESI1	FITS1
1	5	7	-3.0	10.0
2	5	8	-2.0	10.0
3	5	15	5.0	10.0

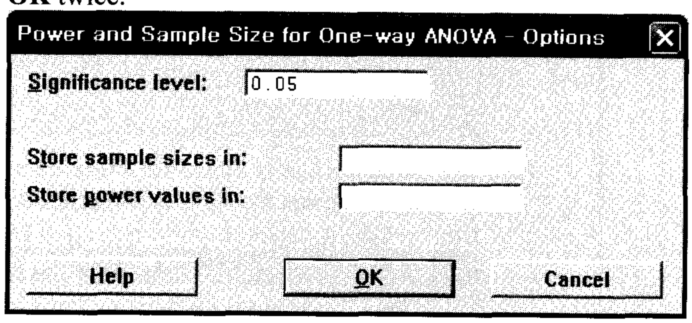
13.3

Example 13.4 **(Sample Size Determination; Fixed-Effects Factor)**

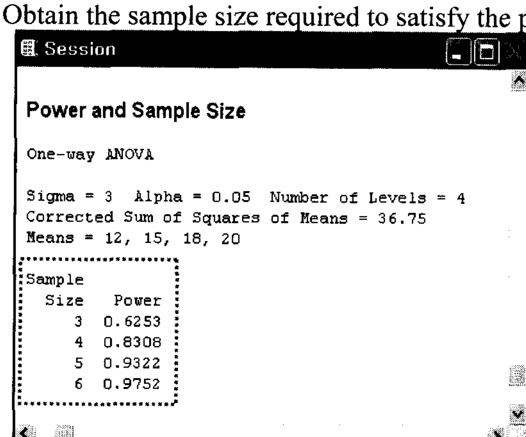
(1) Choose **Stat ➤ Power and Sample Size ➤ One-way ANOVA**. Enter the number of treatments (a) in **Number of levels.** Click **Calculate power for each sample size** and enter the range of sample size to be examined in **Sample sizes** and true means of the treatments in **Means**. Enter also the population standard deviation in **Sigma**. Then click **Options**.



(2) In **Significance level** enter the probability of type I error (α).Then click **OK** twice.



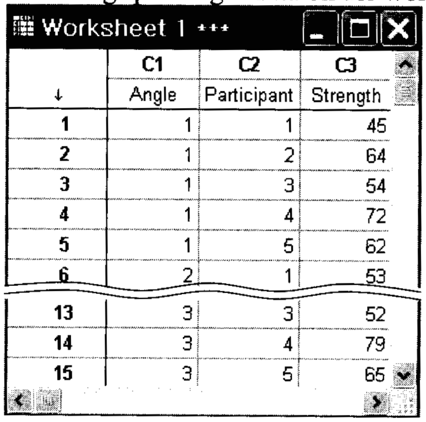
(3) Obtain the sample size required to satisfy the predefined test condition.



Example 13.6 **(Randomized Complete Block Design)**

(1) Choose **File ➤ New**, click **Minitab Project**, and click **OK**.

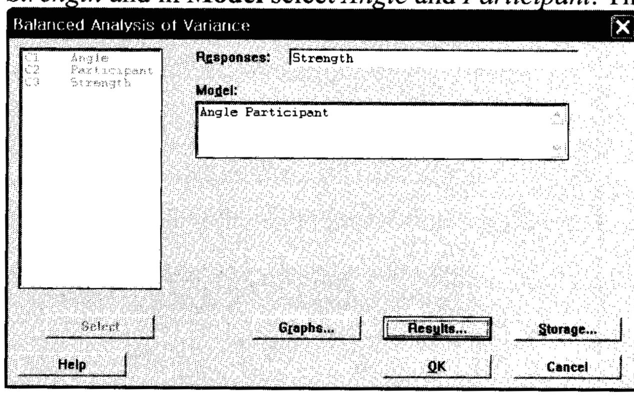
(2) Enter the grip strength data on the worksheet.



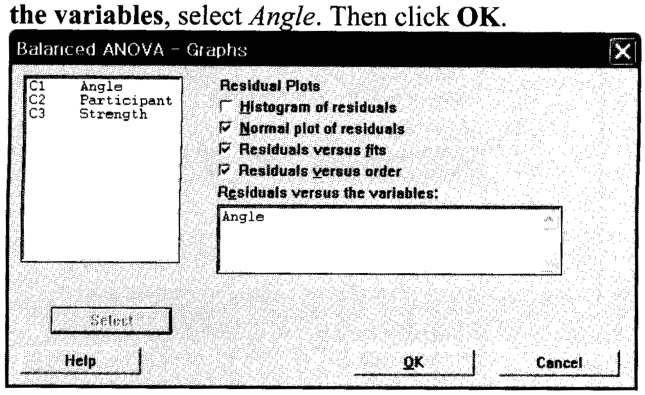
Worksheet 1 ***

↓	C1 Angle	C2 Participant	C3 Strength
1	1	1	45
2	1	2	64
3	1	3	54
4	1	4	72
5	1	5	62
6	2	1	53
13	3	3	52
14	3	4	79
15	3	5	65

(3) Choose **Stat ➤ ANOVA ➤ Balanced ANOVA**. In **Responses** select *Strength* and in **Model** select *Angle* and *Participant*. Then click **OK**.

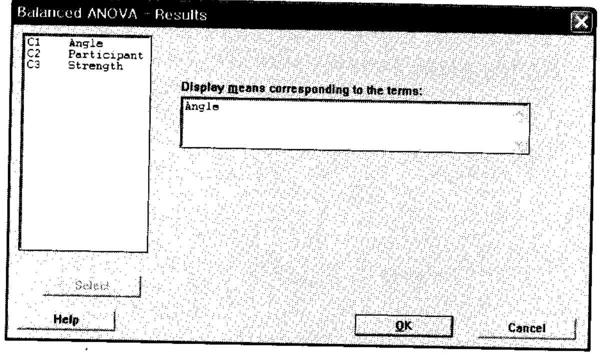


(4) Click **Graphs**. Under **Residual Plots**, check **Normal plot of residuals, Residuals versus fits,** and **Residuals versus order**. In **Residuals versus the variables**, select *Angle*. Then click **OK**.

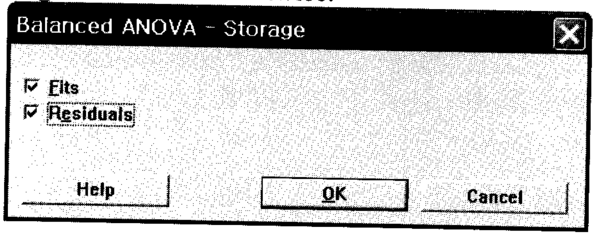


Example 13.6 (*cont.*)

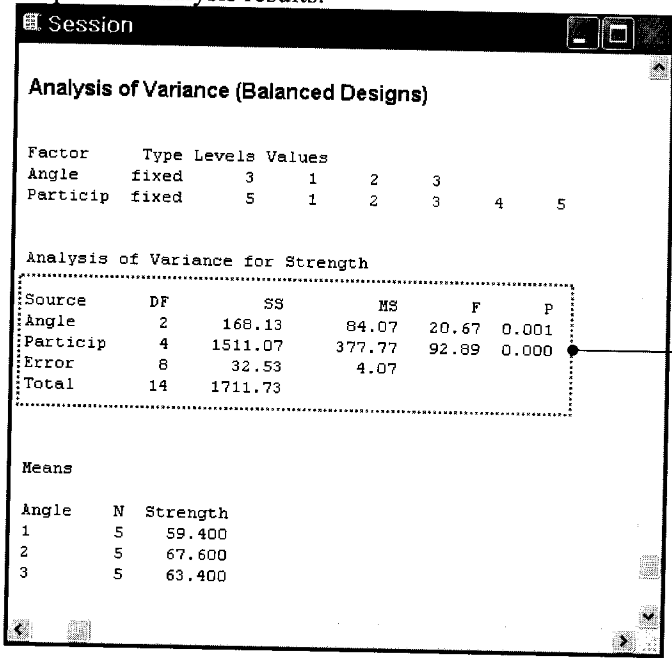
(5) Click **Results**. In **Display means corresponding to the terms**, select *Angle*. Then click **OK**.



(6) Click **Storage**. In **Display means corresponding to the terms**, select *Angle*. Then click **OK** twice.



(7) Interpret the analysis results.



Analysis of Variance (Balanced Designs)

Factor	Type	Levels	Values				
Angle	fixed	3	1	2	3		
Particip	fixed	5	1	2	3	4	5

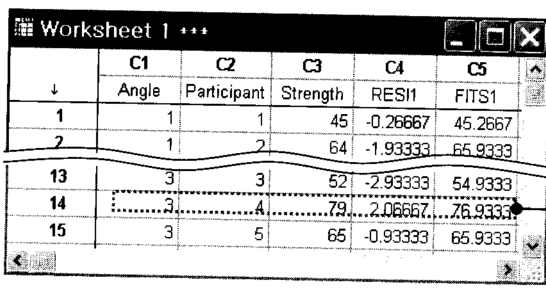
Analysis of Variance for Strength

Source	DF	SS	MS	F	P
Angle	2	168.13	84.07	20.67	0.001
Particip	4	1511.07	377.77	92.89	0.000
Error	8	32.53	4.07		
Total	14	1711.73			

13.6.1

Means

Angle	N	Strength
1	5	59.400
2	5	67.600
3	5	63.400



↓	C1 Angle	C2 Participant	C3 Strength	C4 RESI1	C5 FITS1
1	1	1	45	-0.26667	45.2667
2	1	2	64	-1.93333	65.9333
13	3	3	52	-2.93333	54.9333
14	3	4	79	2.06667	76.9333
15	3	5	65	-0.93333	65.9333

13.6.3

Answers to Exercises

Exercise 13.1

1. **(Test on the Significance of a Fixed-Effects Factor; Completely Randomized Design)**

The sums of squares for the analysis of variance are as follows:

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - \frac{1}{N} y_{..}^2 = 389{,}115 - \frac{2{,}785^2}{20} = 1{,}303.8$$

$$SS_{\text{Treatments}} = \frac{1}{n}\sum_{i=1}^{a} y_{i.}^2 - \frac{1}{N} y_{..}^2 = \frac{1{,}555{,}487}{4} - \frac{2{,}785^2}{20} = 1{,}060.5$$

$$SS_E = SS_T - SS_{\text{Treatments}} = 1{,}303.8 - 1{,}060.5 = 243.3$$

This analysis of variance is summarized in the following table:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	f_0
Coating type	1,060.5	4	265.1	16.3
Error	243.3	15	16.2	
Total	1,303.8	19		

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $\tau_1 = \tau_2 = \cdots = \tau_a = 0$

H_1: $\tau_i \neq 0$ for at least one i, $i = 1$ to 5

Step 2: Determine a **test statistic and its value.**

$$f_0 = \frac{SS_{\text{Treatments}}/(a-1)}{SS_E/(N-a)} = \frac{MS_{\text{Treatments}}}{MS_E} = \frac{265.1}{16.2} = 16.3$$

Step 3: Determine a **critical value(s) for α.**

$$f_{\alpha,a-1,N-a} = f_{0.05,5-1,20-5} = f_{0.05,4,15} = 3.06$$

Step 4: Make a **conclusion.**

Since $f_0 = 16.3 > f_{\alpha,a-1,N-a} = 3.06$, reject H_0 at $\alpha = 0.05$.

2. **(Confidence Interval on μ_i)**

$1 - \alpha = 0.95 \quad \Rightarrow \quad \alpha = 0.05$

95% two-sided CI on μ_1:

$$\bar{y}_{1.} - t_{\alpha/2,N-a}\sqrt{\frac{MS_E}{n}} \le \mu_1 \le \bar{y}_{1.} + t_{\alpha/2,N-a}\sqrt{\frac{MS_E}{n}}$$

$$\Rightarrow \; 145.0 - t_{0.05/2,20-5}\sqrt{\frac{16.2}{4}} \le \mu_1 \le 145.0 + t_{0.05/2,20-5}\sqrt{\frac{16.2}{4}}$$

$$\Rightarrow \; 145.0 - 2.131 \times 2.012 \le \mu_1 \le 145.0 + 2.131 \times 2.012$$

$$\Rightarrow \; 140.7 \le \mu_1 \le 149.3$$

Exercise 13.1 (*cont.*)

3. **(Confidence Interval on $\mu_i - \mu_j$)**

95% two-sided CI on $\mu_2 - \mu_3$:

$$\bar{y}_{2.} - \bar{y}_{3.} - t_{\alpha/2,N-a}\sqrt{\frac{2MS_E}{n}} \le \mu_2 - \mu_3 \le \bar{y}_{2.} - \bar{y}_{3.} + t_{\alpha/2,N-a}\sqrt{\frac{2MS_E}{n}}$$

$$\Rightarrow\ 145.3 - 131.5 - t_{0.05/2,20-5}\sqrt{\frac{2\times 16.2}{4}} \le \mu_2 - \mu_3 \le$$

$$145.3 - 131.5 + t_{0.05/2,20-5}\sqrt{\frac{2\times 16.2}{4}}$$

$$\Rightarrow\ 13.8 - 2.131\times 2.846 \le \mu_2 - \mu_3 \le 13.8 + 2.131\times 2.846$$

$$\Rightarrow\ 7.7 \le \mu_2 - \mu_3 \le 19.9$$

Exercise 13.2

(Least Significant Difference Method)

Step 1: Determine the **LSD** for α.

$$LSD = t_{\alpha/2,N-a}\sqrt{\frac{2MS_E}{n}} = t_{0.05/2,20-5}\sqrt{\frac{2\times 16.2}{4}} = 2.131\times 2.846 = 6.07$$

Step 2: Arrange the means of a treatments in **descending** order.

$\bar{y}_{2.} = \bar{y}_{5.} = 145.3$, $\bar{y}_{1.} = 145.0$, $\bar{y}_{3.} = 131.5$, and $\bar{y}_{4.} = 129.3$

Step 3: Compare the difference between the **largest** and **smallest** means with the LSD. Continue this comparison with the **next smallest** mean as long as the mean difference is greater than the LSD.

2, 5 vs. 4 = 145.3 – 129.3 = 16.0 > 6.07
2, 5 vs. 3 = 145.3 – 131.5 = 13.8 > 6.07
2, 5 vs. 1 = 145.3 – 145.0 = 0.3 < 6.07

Step 4: Continue Step 3 for the **next largest** mean until this iteration reaches the **second smallest** mean.

1 vs. 4 = 145.0 – 129.3 = 15.7 > 6.07
1 vs. 3 = 145.0 – 131.5 = 13.5 > 6.07
3 vs. 4 = 131.5 – 129.3 = 2.2 < 6.07

Step 5: Summarize the pairwise comparison results by arranging the a treatments in order of mean and then **underlining** treatments whose means are **not significantly different**.

$\underline{2\quad 5\quad 1}\quad \underline{3\quad 4}$

Exercise 13.3

(Calculation of Residual)

$$e_{24} = y_{24} - \hat{y}_{24} = y_{24} - \bar{y}_{2.} = 143 - 145.3 = -2.3 \quad \text{(overestimate)}$$

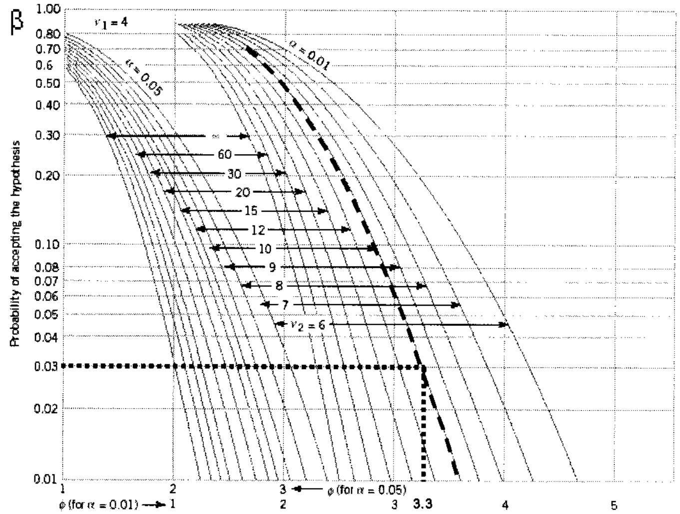
Exercise 13.4

(Sample Size Determination; Fixed-Effects Factor)

Use the operating characteristic curve of Figure 13-6(b) in MR for a fixed-effects factor with $\nu_1 = 4$ (= $a - 1 = 5 - 1$). From the OC curve, the power of the test according to n is found as follows:

Exercise 13.4 (*cont.*)

n	Φ	$\nu_2 = a(n-1)$	β	Power $(1-\beta)$
2	2.7	5	0.50	0.50
3	3.3	10	0.03	0.97



Thus, at least $n = 3$ replicates are needed to achieve the designated power of the test.

Exercise 13.5

(Estimation of Variance Components; Random-Effects Factor)

(1) $\hat{\sigma}^2 = MS_E = 16.2$

(2) $\hat{\sigma}_\tau^2 = \dfrac{MS_{\text{Treatments}} - MS_E}{n} = \dfrac{265.1 - 16.2}{4} = 62.2$

(3) $\hat{V}(Y_{ij}) = \hat{\sigma}_\tau^2 + \hat{\sigma}^2 = 62.2 + 16.2 = 78.4$

These variance estimates indicate that about 79% (= 62.2/78.4) of the total variability in tube conductivity is attributable to coating type.

Exercise 13.6

1. **(Test on the Significance of a Fixed-Effects Factor; Complete Block Design)**

The analysis of variance of the leaking current data is as follows:

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij}^2 - \frac{y_{..}^2}{ab} = 52.4 - \frac{27.7^2}{20} = 14.1$$

$$SS_{\text{Treatments}} = \frac{1}{b}\sum_{i=1}^{a} y_{i.}^2 - \frac{y_{..}^2}{ab} = \frac{219.4}{5} - \frac{27.7^2}{20} = 5.5$$

$$SS_{\text{Blocks}} = \frac{1}{a}\sum_{j=1}^{b} y_{.j}^2 - \frac{y_{..}^2}{ab} = \frac{167.8}{4} - \frac{27.7^2}{20} = 3.6$$

$$SS_E = SS_T - SS_{\text{Treatments}} - SS_{\text{Blocks}} = 14.1 - 5.5 - 3.6 = 5.0$$

Exercise 13.6 (*cont.*)

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	f_0
Channel Length	5.5	3	1.8	4.4
Channel Width	3.6	4	0.9	
Error	5.0	12	0.4	
Total	14.1	19		

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $\tau_1 = \tau_2 = \cdots = \tau_a = 0$

H_1: $\tau_i \neq 0$ for at least one i, $i = 1$ to 4

Step 2: Determine a **test statistic and its value.**

$$f_0 = \frac{SS_{\text{Treatments}}/(a-1)}{SS_E/(a-1)(b-1)} = \frac{MS_{\text{Treatments}}}{MS_E} = \frac{1.8}{0.4} = 4.4$$

Step 3: Determine a **critical value(s) for α.**

$$f_{\alpha,a-1,(a-1)(b-1)} = f_{0.05,4-1,(4-1)(5-1)} = f_{0.05,3,12} = 3.49$$

Step 4: Make a **conclusion.**

Since $f_0 = 4.4 > f_{\alpha,a-1,(a-1)(b-1)} = 3.49$, reject H_0 at $\alpha = 0.05$.

2. (**Least Significant Difference Method**)

Step 1: Determine the **LSD** for α.

$$LSD = t_{\alpha/2,(a-1)(b-1)}\sqrt{\frac{2MS_E}{b}} = t_{0.05/2,12}\sqrt{\frac{2\times 0.4}{5}} = 2.179\times 0.400 = 0.87$$

Step 2: Arrange the means of a treatments in **descending** order.

$\bar{y}_{3.} = 1.9$, $\bar{y}_{4.} = 1.9$, $\bar{y}_{2.} = 0.9$, and $\bar{y}_{1.} = 0.8$

Step 3: Compare the difference between the **largest** and **smallest** means with the LSD. Continue this comparison with the **next smallest** mean as long as the mean difference is greater than the LSD.

3, 4 vs. 1 = 1.9 – 0.8 = 1.1 > 0.87

3, 4 vs. 2 = 1.9 – 0.9 = 1.0 > 0.87

Step 4: Continue Step 3 for the **next largest** mean until this iteration reaches the **second smallest** mean.

2 vs. 1 = 0.9 – 0.8 = 0.1 < 0.87

Step 5: Summarize the pairwise comparison results by arranging the a treatments in order of mean and then **underlining** treatments whose means are **not significantly different.**

<u>3 4</u> <u>2 1</u>

3. (**Calculation of Residual**)

$$e_{21} = y_{21} - \hat{y}_{21} = y_{21} - (\bar{y}_{2.} + \bar{y}_{.1} - \bar{y}_{..}) = 0.8 - (0.9 + 0.9 - 1.4) = 0.4$$

14 Design of Experiment with Several Factors

OUTLINE

14-1 Introduction

This chapter presents statistical concepts and methods to design an experiment with multiple factors (≥ 2) and analyze the data obtained from the experiment. Most of the techniques introduced in Chapter 13 for a single-factor experiment are extended to this multiple-factor experiment.

14-3 Factorial Design

Learning Goals

- ❑ Explain the term *factorial design*.
- ❑ Distinguish between main and interaction effects.
- ❑ Explain why information of interactions between factors is more meaningful than that of the main effects when the interactions are significant.
- ❑ Explain the potential pitfall of a one-factor-at-a-time design when interactions exist.

Factorial Design

A factorial design is an experimental design where all combinations of factor levels under consideration are tested at each replicate. For example, if factors A and B have two and three levels, respectively, a factorial design of A and B runs six (= 2 × 3) combinations of the levels at each replicate.

Main vs. Interaction Effects

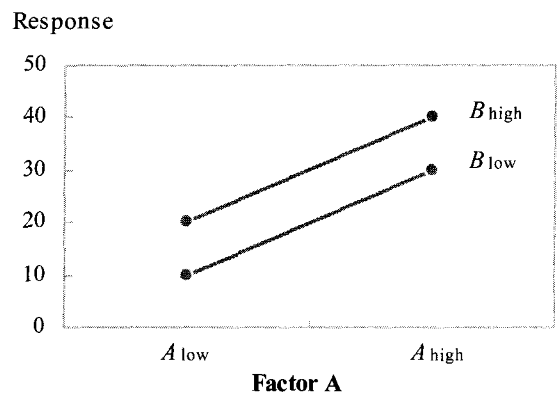
The effect of a factor indicates the change in response by a change in the level of the factor. Two types of factor effects are defined:

1. Main effect: The effect of a factor alone.
2. Interaction effect: The effect of an interaction between factors. An interaction exists between factors if the effect of one factor depends on the condition of the other factor(s).

Figure 14-1 shows that the main effect of A is independent of the level of B and vice versa, indicating the absence of an interaction between A and B. In other words, the main effects of A at the low and high levels of B are the same as 20 (i.e., 30 – 10 = 20 at the low level of B and 40 – 20 = 20 at the high level of B). Similarly, the main effects of B at the low and high levels of A are the same as 10. This insignificant AB interaction can be identified by calculating the difference of the diagonal averages:

$$AB = \frac{A_{\text{high}}B_{\text{high}} + A_{\text{low}}B_{\text{low}}}{2} - \frac{A_{\text{low}}B_{\text{high}} + A_{\text{high}}B_{\text{low}}}{2} = \frac{40+10}{2} - \frac{20+30}{2} = 0$$

Notice that the B_{high} and B_{low} lines in the figure are parallel to each other when the interaction of A and B is absent.

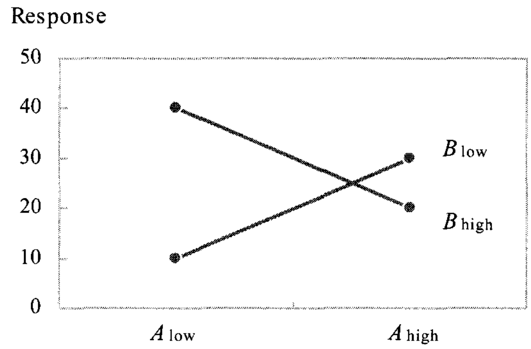


		Factor A			
		A_{low}	A_{high}	Totals	Averages
Factor B	B_{low}	10	30	40	20
	B_{high}	20	40	60	30
Totals		30	70	100	
Averages		15	35		25

Figure 14-1 Two-factor factorial experiment: no interaction.

On the other hand, Figure 14-2 illustrates that the main effect of A (or B) varies depending on the level of the other factor, indicating the presence of an interaction between A and B. The main effects of A at the low and high levels of B are 20 (= 30 – 10) and -20 (= 20 – 40) each, resulting in an average of zero as

Main vs. Interaction Effects (*cont.*)



		Factor A			
		A_{low}	A_{high}	Totals	Averages
Factor B	B_{low}	10	30	40	20
	B_{high}	40	20	60	30
Totals		50	50	100	
Averages		25	25		25

Figure 14-2 Two-factor factorial experiment: presence of an interaction.

the main effect of A. Similarly, the main effects of B at the low and high levels of A are 30 (= 40 – 10) and -10 (= 20 – 30) each, resulting in an average of 10 as the main effect of B. This significant AB interaction can be identified by a nonzero value of the diagonal average difference:

$$AB = \frac{A_{high}B_{high} + A_{low}B_{low}}{2} - \frac{A_{low}B_{high} + A_{high}B_{low}}{2} = \frac{20+10}{2} - \frac{40+30}{2} = -20$$

Notice that the B_{high} and B_{low} lines in the figure are not in parallel in the presence of an interaction between A and B.

As illustrated in Figure 14-2, a significant interaction can mask the significance of main effects. Accordingly, when interaction is present, information of the main effects is not meaningful any longer because the main effect of a factor depends on the condition of the other factor(s).

Necessity of Factorial Design

A factorial design is the only experimental design to check interactions between factors. When a significant interaction exists between factors, experimenting one factor at a time can produce inappropriate results. For example, in Figure 14-2, the maximum response is obtained at $A = A_{low}$ and $B = B_{high}$ by experimenting all the treatment combinations of A and B. But if the experiment had been designed first for A at $B = B_{low}$ and then for B at a time, the maximum response would have been concluded at $A = A_{high}$ and $B = B_{low}$ as follows:

(1) First the maximum response at $B = B_{low}$ would be found at $A = A_{high}$.
(2) Then by experimenting B at $A = A_{high}$, the maximum response would be achieved at $B = B_{low}$.

To avoid this kind of incorrect conclusion in the one-factor-at-a-time design, the factorial design should be used unless the absence of interactions is certain.

14-4 Two-Factor Factorial Experiments

Learning Goals

- ☐ Explain the model of a factorial design with two fixed-effects factors.
- ☐ Test the significance of the main and interaction effects in a two-fixed factor factorial design.
- ☐ Compare individual treatment means with each other by Fisher's LSD method.
- ☐ Calculate residuals in a two-factor factorial design.
- ☐ Explain a method to analyze a two-factor factorial design with a single replicate.

Context

1. **Factor**: Two fixed-effects factors (A with a treatments and B with b treatments)
2. **Repetition**: n replicates for each of ab treatment combinations; N (total number of observations) = nab (balanced design; see Table 14-1)
3. **Randomization**: Complete randomization

Table 14-1 Data Structure of a Two-Factor Factorial Design

		Factor A					
		1	2	…	a	Totals	Means
Factor B	1	y_{111} y_{112} ⋮ y_{11n}	y_{211} y_{212} ⋮ y_{21n}	…	y_{a11} y_{a12} ⋮ y_{a1n}	$y_{.1.}$	$\bar{y}_{.1.}$
	2	y_{121} y_{122} ⋮ y_{12n}	y_{221} y_{222} ⋮ y_{22n}	…	y_{a21} y_{a22} ⋮ y_{a2n}	$y_{.2.}$	$\bar{y}_{.2.}$
	⋮	⋮	⋮	…	⋮	⋮	⋮
	b	y_{1b1} y_{1b2} ⋮ y_{1bn}	y_{2b1} y_{2b2} ⋮ y_{2bn}	…	y_{ab1} y_{ab2} ⋮ y_{abn}	$y_{.b.}$	$\bar{y}_{.b.}$
Totals		$y_{1..}$	$y_{2..}$	…	$y_{a..}$	$y_{...}$	
Means		$\bar{y}_{1..}$	$\bar{y}_{2..}$	…	$\bar{y}_{a..}$		$\bar{y}_{...}$

Model

$$Y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ijk} \quad \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, b \\ k = 1, 2, \ldots, n \end{cases}$$

where: μ = overall mean

τ_i = effect of the i^{th} level of factor A, $\sum_{i=1}^{a} \tau_i = 0$

Model (*cont.*)

β_j = effect of the j^{th} level of factor B, $\sum_{j=1}^{b} \beta_j = 0$

$(\tau\beta)_{ij}$ = effect of interaction between A and B,

$$\sum_{i=1}^{a} (\tau\beta)_{ij} = 0 \text{ for } j = 1, 2, \ldots, b, \; \sum_{j=1}^{b} (\tau\beta)_{ij} = 0, \text{ for } i = 1, 2, \ldots, a$$

ε_{ijk} = random error ~ $i.i.d.\ N(0, \sigma^2)$

(Notes) 1. $Y_{ijk} \sim N(\mu_{ij}, \sigma^2)$

2. In the fixed-effects model, the treatment effects τ_i's, β_j's, and $(\tau\beta)_{ij}$'s are defined as deviations from the overall mean; thus, each sum of τ_i's, β_j's, and $(\tau\beta)_{ij}$'s is zero.

ANOVA

In the two-factor factorial design, the analysis of variance decomposes the total variability in Y (***SS_T***, **total sum of squares**; degrees of freedom = $N - 1$) into four components:

(1) Variability due to factor A (***SS_A***; degrees of freedom = $a - 1$)
(2) Variability due to factor B (***SS_B***; degrees of freedom = $b - 1$)
(3) Variability due to interaction AB (***SS_{AB}***; degrees of freedom = $(a - 1)(b - 1)$)
(4) Variability due to random error (***SS_E***; degrees of freedom = $ab(n - 1) = N - ab$)

In other words,

$$SS_T = SS_A + SS_B + SS_{AB} + SS_E$$

where:

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} (y_{ijk} - \bar{y}_{...})^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}^2 - \frac{y_{...}^2}{N}$$

$$SS_A = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} (\bar{y}_{i..} - \bar{y}_{...})^2 = \sum_{i=1}^{a} \frac{y_{i..}^2}{bn} - \frac{y_{...}^2}{N}$$

$$SS_B = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} (\bar{y}_{.j.} - \bar{y}_{...})^2 = \sum_{j=1}^{b} \frac{y_{.j.}^2}{an} - \frac{y_{...}^2}{N}$$

$$SS_{AB} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} (\bar{y}_{ij.} - \bar{y}_{...})^2 = \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{y_{ij.}^2}{n} - \frac{y_{...}^2}{N} - SS_A - SS_B$$

$$SS_E = SS_T - (SS_A + SS_B + SS_{AB})$$

By dividing each sum of squares by the corresponding degrees of freedom, the mean squares of A, B, AB, and error are determined as follows:

$$MS_A = \frac{SS_A}{a-1}$$

$$MS_B = \frac{SS_B}{b-1}$$

$$MS_{AB} = \frac{SS_{AB}}{(a-1)(b-1)}$$

ANOVA (*cont.*)

The expected values of these mean squares are

$$E(MS_A) = E\left(\frac{SS_A}{a-1}\right) = \sigma^2 + \frac{bn\sum_{i=1}^{a}\tau_i^2}{a-1}$$

$$E(MS_B) = E\left(\frac{SS_B}{b-1}\right) = \sigma^2 + \frac{an\sum_{j=1}^{b}\beta_j^2}{b-1}$$

$$E(MS_{AB}) = E\left(\frac{SS_{AB}}{(a-1)(b-1)}\right) = \sigma^2 + \frac{n\sum_{j=1}^{b}\sum_{i=1}^{a}(\tau\beta)_{ij}^2}{(a-1)(b-1)}$$

$$E(MS_E) = E\left(\frac{SS_E}{N-ab}\right) = \sigma^2$$ (Note) MS_E is an unbiased estimator of σ^2.

The ratios MS_A/MS_E, MS_B/MS_E, and MS_{AB}/MS_E have F distributions:

$$F = \frac{MS_A}{MS_E} \sim F(a-1, N-ab)$$

$$F = \frac{MS_B}{MS_E} \sim F(b-1, N-ab)$$

$$F = \frac{MS_{AB}}{MS_E} \sim F((a-1)(b-1), N-ab)$$

These F values become small as the effects of A, B, and AB are insignificant (i.e., the null hypotheses H_0: $\tau_i = 0$, $i = 1$ to a, H_0: $\beta_j = 0, j = 1$ to b, and H_0: $(\tau\beta)_{ij} = 0$ are true). The ANOVA results of the fixed factorial design are summarized in Table 14-2 to test the significance of the effects.

Table 14-2 ANOVA Table for a Two-Factor Factorial Design: Fixed-Effects Model

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
A	SS_A	$a-1$	MS_A	MS_A / MS_E
B	SS_B	$b-1$	MS_B	MS_B / MS_E
AB	SS_{AB}	$(a-1)(b-1)$	MS_{AB}	MS_{AB} / MS_E
Error	SS_E	$N-ab$	MS_E	
Total	SS_T	$N-1$		

Hypothesis Test (*F*-test)

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

$$\left.\begin{array}{ll} H_0: \tau_i = 0 & \text{for all } i\text{'s}, i = 1, 2, \ldots, a \\ H_1: \tau_i \neq 0 & \text{for at least one } i \end{array}\right\} \text{ for } A$$

$$\left.\begin{array}{ll} H_0: \beta_j = 0 & \text{for all } j\text{'s}, j = 1, 2, \ldots, b \\ H_1: \beta_j \neq 0 & \text{for at least one } j \end{array}\right\} \text{ for } B$$

$$\left.\begin{array}{ll} H_0: (\tau\beta)_{ij} = 0 & \text{for all combinations of } i\text{'s and } j\text{'s} \\ H_1: (\tau\beta)_{ij} \neq 0 & \text{for at least one combination of } i \text{ and } j \end{array}\right\} \text{ for } AB$$

Hypothesis Test (*cont.*)

Step 2: Determine a **test statistic and its value.**

$$F_0 = \frac{SS_A/(a-1)}{SS_E/(N-ab)} = \frac{MS_A}{MS_E} \sim F(a-1, N-ab) \quad \text{for } A$$

$$F_0 = \frac{SS_B/(b-1)}{SS_E/(N-ab)} = \frac{MS_B}{MS_E} \sim F(b-1, N-ab) \quad \text{for } B$$

$$F_0 = \frac{SS_{AB}/(a-1)(b-1)}{SS_E/(N-ab)} = \frac{MS_{AB}}{MS_E} \sim F((a-1)(b-1), N-ab) \quad \text{for } AB$$

Step 3: Determine a **critical value(s) for α.**

$$f_{\alpha,a-1,N-ab} \quad \text{for } A$$

$$f_{\alpha,b-1,N-ab} \quad \text{for } B$$

$$f_{\alpha,(a-1)(b-1),N-ab} \quad \text{for } AB$$

Step 4: Make a **conclusion**. Reject H_0 if

$$f_0 > f_{\alpha,a-1,N-ab} \quad \text{for } A$$

$$f_0 > f_{\alpha,b-1,N-ab} \quad \text{for } B$$

$$f_0 > f_{\alpha,(a-1)(b-1),N-ab} \quad \text{for } AB$$

It is recommended that the test for interaction be conducted first followed by the tests for main effects. If interaction is insignificant, the tests for main effects are meaningful; otherwise, the tests for main effects are meaningless because the effect of a factor is subject to the condition of the other factor(s).

Multiple Comparison

For significant fixed factors, the means of their individual treatments are compared by a multiple comparison method such as Fisher's least significant difference (LSD) method (see Section 13-2.3). When interaction *AB* is insignificant, the treatment means of each significant main effect are compared. When *AB* is significant, the treatment means of one factor (say *A*) at a particular level of the other factor (say *B*) are compared. For a fixed two-factor factorial design, the LSDs are determined as follows:

$$LSD = t_{\alpha/2,N-ab}\sqrt{\frac{2MS_E}{bn}} \quad \text{for } A$$

$$= t_{\alpha/2,N-ab}\sqrt{\frac{2MS_E}{an}} \quad \text{for } B$$

$$= t_{\alpha/2,N-ab}\sqrt{\frac{2MS_E}{n}} \quad \text{for } AB$$

Residual Analysis

Like the single-factor model (shown in Section 13-2.2), the two-factor factorial model has three assumptions on the error term: (1) normality, (2) constant variance, and (3) randomness. These assumptions can be checked by analyzing residuals ($e_{ijk} = y_{ijk} - \hat{y}_{ijk} = y_{ijk} - \bar{y}_{ij.}$) as explained in Section 13-2.5.

Factorial Design with a Single Replicate

Sometimes a two-factor factorial experiment has only one observation ($n = 1$) for each combination of the treatments. In this case, the error degrees of freedom become zero ($= N - ab = nab - ab$ for $n = 1$) and thus the error mean square is unavailable.

In this unreplicated design, the interaction mean square can be used as the error mean square, assuming the interaction effect is negligible. However, this no-interaction assumption can yield an improper conclusion if the interaction is significant; thus, data and residuals should be carefully examined to check the significance of the interaction.

Example 14.1 (MR 14-2)

An engineer suspects that the surface finish (Y) of metal parts is influenced by drying time (A) and type of paint (B). Three drying times ($a = 3$; 20, 25, and 30 min.) and two types of paint ($b = 2$) are specifically selected by the engineer and three parts ($n = 3$) are tested for each combination of the drying times and paint types, yielding the following:

		Drying Time (A; unit: min.)				
		20 (1)	25 (2)	30 (3)	Totals	Means
Paint Type (B)	1	74 64 50	73 61 44	78 85 92	621	69.0
	2	92 86 68	98 73 88	66 45 85	701	77.9
Totals		434	437	451	1,322	
Means		72.3	72.8	75.2		73.4

The following summary quantities are also provided for the surface finish data:

$$N = 18,\ \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}^2 = 101{,}598,\ \sum_{i=1}^{a} y_{i..}^2 = 582{,}726,\ \sum_{j=1}^{b} y_{.j.}^2 = 877{,}042,$$

$$\text{and } \sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij.}^2 = 298{,}066$$

1. **(Test on the Significance of Main and Interaction Effects; Two-Factor Factorial Design)** Test if the effects of A, B, and AB are significant each at $\alpha = 0.05$ by the analysis of variance.

The analysis of variance of the surface finish data is as follows:

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}^2 - \frac{y_{...}^2}{N} = 101{,}598 - \frac{1{,}322^2}{18} = 4{,}504.4$$

$$SS_A = \sum_{i=1}^{a} \frac{y_{i..}^2}{bn} - \frac{y_{...}^2}{N} = \frac{582{,}726}{6} - \frac{1{,}322^2}{18} = 27.4$$

$$SS_B = \sum_{j=1}^{b} \frac{y_{.j.}^2}{an} - \frac{y_{...}^2}{N} = \frac{877{,}042}{9} - \frac{1{,}322^2}{18} = 355.6$$

Example 14.1 (*cont.*)

$$SS_{AB} = \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{y_{ij.}^{2}}{n} - \frac{y_{...}^{2}}{N} - SS_A - SS_B$$
$$= \frac{298{,}066}{3} - \frac{1{,}322^2}{18} - 27.4 - 355.6 = 1{,}878.8$$
$$SS_E = SS_T - (SS_A + SS_B + SS_{AB}) = 4{,}504.4 - (27.4 + 355.6 + 1{,}878.8)$$
$$= 4{,}504.4 - (27.4 + 355.6 + 1{,}878.8) = 2{,}242.7$$

The ANOVA results are summarized in the following table:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	f_0
Drying Time (A)	27.4	2	13.7	0.07
Paint Type (B)	355.6	1	355.6	1.90
AB	1,878.8	2	939.4	5.03
Error	2,242.7	12	186.9	
Total	4,504.4	17		

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

$$\left.\begin{array}{ll} H_0\text{: } \tau_i = 0 & \text{for } i = 1 \text{ to } 3 \\ H_1\text{: } \tau_i \neq 0 & \text{for at least one } i \end{array}\right\} \text{ for } A$$

$$\left.\begin{array}{ll} H_0\text{: } \beta_j = 0 & \text{for } j = 1 \text{ and } 2 \\ H_1\text{: } \beta_j \neq 0 & \text{for at least one } j \end{array}\right\} \text{ for } B$$

$$\left.\begin{array}{ll} H_0\text{: } (\tau\beta)_{ij} = 0 & \text{for all combinations of } i\text{'s and } j\text{'s} \\ H_1\text{: } (\tau\beta)_{ij} \neq 0 & \text{for at least one combination of } i \text{ and } j \end{array}\right\} \text{ for } AB$$

Step 2: Determine a **test statistic and its value.**

$$f_0 = \frac{MS_A}{MS_E} = \frac{13.7}{186.9} = 0.07 \quad \text{for } A$$
$$f_0 = \frac{MS_B}{MS_E} = \frac{355.6}{186.9} = 1.90 \quad \text{for } B$$
$$f_0 = \frac{MS_{AB}}{MS_E} = \frac{939.4}{186.9} = 5.03 \quad \text{for } AB$$

Step 3: Determine a **critical value(s) for $\boldsymbol{\alpha}$.**

$$f_{\alpha,a-1,N-ab} = f_{0.05,3-1,18-6} = f_{0.05,2,12} = 3.89 \quad \text{for } A$$
$$f_{\alpha,b-1,N-ab} = f_{0.05,2-1,18-6} = f_{0.05,1,12} = 4.75 \quad \text{for } B$$
$$f_{\alpha,(a-1)(b-1),N-ab} = f_{0.05,(3-1)(2-1),18-6} = f_{0.05,2,12} = 3.89 \quad \text{for } AB$$

Step 4: Make a **conclusion.**

Since $f_0 = 5.03 > f_{\alpha,(a-1)(b-1),N-ab} = 3.89$, reject H_0 at $\alpha = 0.05$ for AB. Due to the significant interaction, tests on the main effects of A and B are meaningless.

Example 14.1 (*cont.*)

2. (**Multiple Comparison**) Compare the surface finish means of the three drying times for paint type 1. The corresponding surface means are $\bar{y}_{11.} = 62.7$, $\bar{y}_{21.} = 59.3$, and $\bar{y}_{31.} = 85.0$

Step 1: Determine the **LSD** for α.

$$LSD = t_{\alpha/2,N-ab}\sqrt{\frac{2MS_E}{n}} = t_{0.05/2,18-3\times2}\sqrt{\frac{2\times186.9}{3}} = 2.179\times11.16 = 24.32$$

Step 2: Arrange the means of a treatments in **descending** order.
$\bar{y}_{31.} = 85.0$, $\bar{y}_{11.} = 62.7$, and $\bar{y}_{21.} = 59.3$

Step 3: Compare the difference between the **largest** and **smallest** means with the LSD. Continue this comparison with the **next smallest** mean as long as the mean difference is greater than the LSD.
3 vs. 2 = 85.0 – 59.3 = 25.7 > 24.32
3 vs. 1 = 85.0 – 62.7 = 22.3 < 24.32

Step 4: Continue Step 3 for the **next largest** mean until this iteration reaches the **second smallest** mean.
1 vs. 2 = 62.7 – 59.3 = 3.40 < 24.32

Step 5: Summarize the pairwise comparison results by arranging the a treatments in order of mean and then **underlining** treatments whose means are **not significantly different.**

$$3 \quad \underline{1 \quad 2}$$

3. (**Calculation of Residual**) Calculate the residual of $y_{313} = 92$.

$e_{313} = y_{313} - \hat{y}_{313} = y_{313} - \bar{y}_{31.} = 92 - 85 = 7$ (underestimate)

Exercise 14.1 (MR 14-4)

An experiment is conducted to determine whether either firing temperature (A) or furnace position (B) affects the baked density (Y) of a carbon anode. Three firing temperatures ($a = 3$) and two furnace positions ($b = 2$) are specifically selected by the experimenter and three measurements ($n = 3$) are collected for each combination of the treatments. The measurements are as follows:

		Temperature (A; unit: °C)				
		800 (1)	825 (2)	850 (3)	Totals	Means
Position (B)	1	5.70	10.63	5.65		
		5.65	10.80	5.10	65.69	7.30
		5.83	10.43	5.90		
	2	5.28	9.88	5.26		
		5.47	10.26	5.38	62.10	6.90
		5.21	10.04	5.32		
Totals		33.14	62.04	32.61	127.79	
Means		5.52	10.34	5.44		7.10

The following summary quantities are also provided for the baked density data:

$$N = 18,\ \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}^{\,2} = 1{,}003.1,\ \sum_{i=1}^{a} y_{i..}^{\,2} = 6{,}010.6,\ \sum_{j=1}^{b} y_{.j.}^{\,2} = 8{,}171.6,$$

Exercise 14.1 (*cont.*)

and $\sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij.}^{\ 2} = 3{,}007.7$

1. Test if the effects of A, B, and AB are significant each at $\alpha = 0.05$ by the analysis of variance.
2. Test the differences between the baked density means of the three firing temperatures (A): $\bar{y}_{1..} = 5.52$, $\bar{y}_{2..} = 10.34$, and $\bar{y}_{3..} = 5.44$
3. Calculate the residual of $y_{213} = 10.43$.

14-5 General Factorial Experiments

Factorial Design with More Than Two Factors

An experiment can include more than two factors. The analysis of variance procedure for this general factorial experiment is just an extension of that of a two-factor factorial design but requires tedious computation; thus, a computer software program is mostly used for the analysis.

14-7 2^k Factorial Designs

14-7.1 2^2 Design

Learning Goals

- ❑ Identify the orthogonal contrasts of a 2^2 design.
- ❑ Estimate the main and interaction effects of a 2^2 factorial design.
- ❑ Test the significance of the main and interaction effects of a 2^2 factorial design.

2^k Design

A 2^k factorial design is a factorial design with k factors, each having two levels. The low and high levels of each factor are denoted by – (or '-1') and + (or '1'), respectively. A complete replicate of this design requires 2^k observations, which is the smallest number of observations for a complete factorial design with k factors.

Due to the minimum number of runs, the 2^k design is particularly useful for an experiment with many factors for screening purposes. Two treatments are chosen for each factor by assuming a linearity of the response over the range of the factor levels.

ANOVA for 2^2 Design

The totals of the treatment combinations in a 2^2 design (2^k design where $k = 2$ factors) are denoted by a series of lowercase letters: (1), a, b, and ab (see Figure 14-3). The presence (absence) of a letter indicates that the corresponding factor is at the high (low) level in the treatment combination. For example, the letter a indicates that factor A is at the high level and factor B is at the low level.

ANOVA for 2^2 Design (*cont.*)

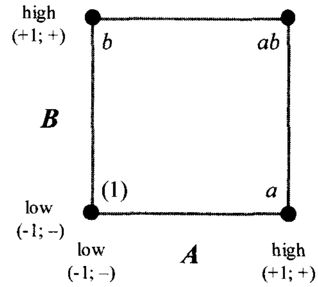


Figure 14-3 The 2^2 factorial design.

The main effects A and B and interaction AB with n replicates are estimated as follows:

$$A = \bar{y}_{A+} - \bar{y}_{A-} = \frac{a+ab}{2n} - \frac{b+(1)}{2n} = \frac{a+ab-b-(1)}{2n}$$

$$B = \bar{y}_{B+} - \bar{y}_{B-} = \frac{b+ab}{2n} - \frac{a+(1)}{2n} = \frac{b+ab-a-(1)}{2n}$$

$$AB = \frac{\bar{y}_{A+B+} + \bar{y}_{A-B-}}{2} - \frac{\bar{y}_{A+B-} + \bar{y}_{A-B+}}{2} = \frac{ab+(1)}{2n} - \frac{a+b}{2n} = \frac{ab+(1)-a-b}{2n}$$

The linear combinations in the numerators of the main and interaction equations are called **orthogonal contrasts** (denoted as C) as they have the following two properties:

(1) Each linear combination has coefficients $\{c_i\}$ whose sum is zero ($\sum_i c_i = 0$), and

(2) Any two contrasts have coefficients $\{c_i\}$ and $\{d_i\}$ whose sum of the products of the coefficients is zero ($\sum_i c_i d_i = 0$).

By using the orthogonal contrasts, the sums of squares for A, B, and AB are calculated:

$$SS_A = \frac{C_A^2}{n\sum_{i=1}^{4} c_i} = \frac{[a+ab-b-(1)]^2}{4n}$$

$$SS_B = \frac{C_B^2}{n\sum_{i=1}^{4} c_i} = \frac{[b+ab-a-(1)]^2}{4n}$$

$$SS_{AB} = \frac{C_{AB}^2}{n\sum_{i=1}^{4} c_i} = \frac{[ab+(1)-a-b]^2}{4n}$$

Note that the total and error sums of squares are calculated by

$$SS_T = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{n} y_{ijk}^2 - \frac{y_{...}^2}{N}$$

$$SS_E = SS_T - (SS_A + SS_B + SS_{AB})$$

ANOVA for 2^2 Design (*cont.*)

Since each contrast sum of squares has a single degree of freedom, the mean squares of A, B, and AB are equal to the corresponding sums of squares each. The mean squares of error is

$$MS_E = \frac{SS_E}{4(n-1)} = \frac{SS_E}{N-4}$$

The ratios MS_A/MS_E, MS_B/MS_E, and MS_{AB}/MS_E have F distributions:

$$F = \frac{MS_A}{MS_E} \sim F(1,4(n-1))$$

$$F = \frac{MS_B}{MS_E} \sim F(1,4(n-1))$$

$$F = \frac{MS_{AB}}{MS_E} \sim F(1,4(n-1))$$

These F values become small as the effects of A, B, and AB are insignificant (i.e., H_0: $\tau_i = 0$, H_0: $\beta_j = 0$, and H_0: $(\tau\beta)_{ij} = 0$ are true). The variation quantities from ANOVA are summarized in Table 14-3.

Table 14-3 ANOVA Table for a 2^2 Factorial Design

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
A	SS_A	1	MS_A	MS_A / MS_E
B	SS_B	1	MS_B	MS_B / MS_E
AB	SS_{AB}	1	MS_{AB}	MS_{AB} / MS_E
Error	SS_E	$N-4$	MS_E	
Total	SS_T	$N-1$		

Hypothesis Test (*F*-test)

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

$$\left.\begin{array}{ll} H_0: \ \tau_i = 0 & \text{for } i = 1 \text{ and } 2 \\ H_1: \ \tau_i \neq 0 & \text{for at least one } i \end{array}\right\} \text{ for } A$$

$$\left.\begin{array}{ll} H_0: \ \beta_j = 0 & \text{for } j = 1 \text{ and } 2 \\ H_1: \ \beta_j \neq 0 & \text{for at least one } j \end{array}\right\} \text{ for } B$$

$$\left.\begin{array}{ll} H_0: \ (\tau\beta)_{ij} = 0 & \text{for all combinations of } i\text{'s and } j\text{'s} \\ H_1: \ (\tau\beta)_{ij} \neq 0 & \text{for at least one combination of } i \text{ and } j \end{array}\right\} \text{ for } AB$$

Step 2: Determine a **test statistic and its value.**

$$F_0 = \frac{SS_A/1}{SS_E/4(n-1)} = \frac{MS_A}{MS_E} \sim F(1,4(n-1)) \quad \text{for } A$$

$$F_0 = \frac{SS_B/1}{SS_E/4(n-1)} = \frac{MS_B}{MS_E} \sim F(1,4(n-1)) \quad \text{for } B$$

$$F_0 = \frac{SS_{AB}/1}{SS_E/4(n-1)} = \frac{MS_{AB}}{MS_E} \sim F(1,4(n-1)) \quad \text{for } AB$$

Hypothesis Test (*cont.*)

Step 3: Determine a **critical value(s) for α.**

$f_{\alpha,1,4(n-1)}$ for A, B, and AB

Step 4: Make a **conclusion**. Reject H_0 if

$f_0 > f_{\alpha,1,4(n-1)}$ for A, B, and AB

Example 14.2

(**Calculation of Effects and Sums of Squares; 2^2 Design**) Consider the surface finish data in Example 14-1. Assume that only two drying times (20 and 30 min.; $a = 2$) and two types of paint ($b = 2$) were tested. The following summary quantities are provided for the surface finish data:

$n = 3$, $(1) = 188$, $a = 255$, $b = 246$, and $ab = 196$

Estimate the effects A, B, and AB and calculate the corresponding sums of squares.

The orthogonal contrasts of the 2^2 design for the surface finish data are

$$C_A = a + ab - b - (1) = 255 + 196 - 246 - 188 = 17$$
$$C_B = b + ab - a - (1) = 246 + 196 - 255 - 188 = -1$$
$$C_{AB} = ab + (1) - a - b = 196 + 188 - 255 - 246 = -117$$

Thus, the estimates of the main and interaction effects are

$$A = \frac{C_A}{n2^{k-1}} = \frac{17}{3 \times 2^{2-1}} = 2.8$$
$$B = \frac{C_B}{n2^{k-1}} = \frac{-1}{3 \times 2^{2-1}} = -0.2$$
$$AB = \frac{C_{AB}}{n2^{k-1}} = \frac{-117}{3 \times 2^{2-1}} = -19.5$$

Then, the corresponding sums of squares are

$$SS_A = \frac{C_A^2}{n2^k} = \frac{17^2}{3 \times 2^2} = 24.1$$
$$SS_B = \frac{C_B^2}{n2^k} = \frac{(-1)^2}{3 \times 2^2} = 0.1$$
$$SS_{AB} = \frac{C_{AB}^2}{n2^k} = \frac{(-117)^2}{3 \times 2^2} = 1{,}140.8$$

Exercise 14.2

Consider the carbon anode baked density data in Exercise 14-1. Assume that only two firing temperatures (800 and 850 °C; $a = 2$) and two furnace positions ($b = 2$) were tested. The following summary quantities are provided:

$n = 3$, $(1) = 17.18$, $a = 16.65$, $b = 15.96$, and $ab = 15.96$

Estimate the effects A, B, and AB and calculate the corresponding sums of squares.

14-7.2 2^k Design for $k \geq 3$ Factors

Learning Goals

- ❑ Identify the orthogonal contrasts of a 2^k design.
- ❑ Estimate main and interaction effects in a 2^k factorial design.

Orthogonal Contrasts of 2^k Design

As shown in the analysis of variance for a 2^2 design, orthogonal contrasts are useful to analyze a 2^k design. The effects of factors and their interactions and corresponding sums of squares can be calculated by using orthogonal contrasts:

$$\text{Effet} = \frac{\text{Contrast}}{n2^{k-1}}$$

$$SS = \frac{(\text{Contrast})^2}{n2^k}$$

To determine the orthogonal contrasts of a 2^k design in a systematic manner, a table of plus and minus signs is used (see Table 14-4 for a 2^3 design). This table includes columns of treatment combinations, identity (I), main effects, and interactions. In a main effect column, a plus (minus) sign indicates the high (low) level of the corresponding factor for the treatment condition (e.g., in the main effect column A, a plus is assigned to the treatment combination ac because the level of A is high. Once the signs of the main effects are established, the signs of each interaction column are determined by multiplying associated columns row by row (e.g., $AB = A \times B$). Except the identity column, each column includes an equal number of plus and minus signs.

By using the sign table, main and interaction effects and their sums of squares are calculated. For example, in Table 14-4, the main effect A and the corresponding sum of squares when $n = 2$ are determined as follows:

$$A = \frac{C_A}{n2^{k-1}} = \frac{a + ab + ac + abc - (1) - b - c - bc}{2 \times 2^{3-1}}$$

$$SS_A = \frac{C_A^2}{n2^k} = \frac{[a + ab + ac + abc - (1) - b - c - bc]^2}{2 \times 2^3}$$

Table 14-4 Signs for Orthogonal Contrasts in the 2^3 Design

Treatment	Effects							
Combinations	I	A	B	AB	C	AC	BC	ABC
(1)	+	−	−	+	−	+	+	−
a	+	+	−	−	−	−	+	+
b	+	−	+	−	−	+	−	+
ab	+	+	+	+	−	−	−	−
c	+	−	−	+	+	−	−	+
ac	+	+	−	−	+	+	−	−
bc	+	−	+	−	+	−	+	−
abc	+	+	+	+	+	+	+	+

Example 14.3

(Calculation of Effects and Sums of Squares; 2^k Design) Power grip strength is measured on a handle to examine the effects of three handle design factors: diameter (A), angle (B), and weight (C). Two treatments are chosen for each factor and a single replicate (n = 1) is run. The grip strength measurements (unit: kg) are as follows:

Treatment Combinations							
(1)	*a*	*b*	*ab*	*c*	*ac*	*bc*	*abc*
35	40	31	33	36	42	30	34

Estimate the main and interaction effects of this 2^3 design and calculate the corresponding sums of squares.

From Table 14-4, the orthogonal contrasts of the 2^3 design for the grip strength data are

$$C_A = a + ab + ac + abc - (1) - b - c - bc$$
$$= 40 + 33 + 42 + 34 - 35 - 31 - 36 - 30 = 17$$
$$C_B = b + ab + bc + abc - (1) - a - c - ac$$
$$= 31 + 33 + 30 + 34 - 35 - 40 - 36 - 42 = -25$$
$$C_{AB} = (1) + ab + c + abc - a - b - ac - bc$$
$$= 35 + 33 + 36 + 34 - 40 - 31 - 42 - 30 = -5$$
$$C_C = c + ac + bc + abc - (1) - a - b - ab$$
$$= 36 + 42 + 30 + 34 - 35 - 40 - 31 - 33 = 3$$
$$C_{AC} = (1) + b + ac + abc - a - ab - c - bc$$
$$= 35 + 31 + 42 + 34 - 40 - 33 - 36 - 30 = 3$$
$$C_{BC} = (1) + a + bc + abc - b - ab - c - ac$$
$$= 35 + 40 + 30 + 34 - 31 - 33 - 36 - 42 = -3$$
$$C_{ABC} = a + b + c + abc - (1) - ab - ac - bc$$
$$= 40 + 31 + 36 + 34 - 35 - 33 - 42 - 30 = 1$$

Thus, the estimates of the main and interaction effects are

$$A = \frac{C_A}{n2^{k-1}} = \frac{17}{1\times 2^{3-1}} = 4.25 \qquad B = \frac{C_B}{n2^{k-1}} = \frac{-25}{1\times 2^{3-1}} = -6.25$$

$$C = \frac{C_C}{n2^{k-1}} = \frac{3}{1\times 2^{3-1}} = 0.75 \qquad AB = \frac{C_{AB}}{n2^{k-1}} = \frac{-5}{1\times 2^{3-1}} = -1.25$$

$$AC = \frac{C_{AC}}{n2^{k-1}} = \frac{3}{1\times 2^{3-1}} = 0.75 \qquad BC = \frac{C_{BC}}{n2^{k-1}} = \frac{-3}{1\times 2^{3-1}} = -0.75$$

$$ABC = \frac{C_{ABC}}{n2^{k-1}} = \frac{1}{1\times 2^{3-1}} = 0.25$$

The corresponding sums of squares are

$$SS_A = \frac{C_A^2}{n2^k} = \frac{17^2}{1\times 2^3} = 36.13 \qquad SS_B = \frac{C_B^2}{n2^k} = \frac{(-25)^2}{1\times 2^3} = 78.13$$

$$SS_C = \frac{C_C^2}{n2^k} = \frac{3^2}{1\times 2^3} = 1.13 \qquad SS_{AB} = \frac{C_{AB}^2}{n2^k} = \frac{(-5)^2}{1\times 2^3} = 3.13$$

Example 14.3 (*cont.*)

$$SS_{AC} = \frac{C_{AC}^2}{n2^k} = \frac{3^2}{1 \times 2^3} = 1.13 \qquad SS_{BC} = \frac{C_{BC}^2}{n2^k} = \frac{(-3)^2}{1 \times 2^3} = 1.13$$

$$SS_{ABC} = \frac{C_{ABC}^2}{n2^k} = \frac{1^2}{1 \times 2^3} = 0.13$$

Exercise 14.3 (MR 14-13)

An engineer is interested in the effects of cutting speed (A), metal hardness (B), and cutting angle (C) on the life of a cutting tool. Two levels are chosen for each factor and two replicates ($n = 2$) of this 2^3 factorial design are run. The tool life data (unit: hours) are as follows:

Treatment Combination	Replicate I	II	Total	Average
(1)	221	311	532	266.0
a	325	435	760	380.0
b	354	348	702	351.0
ab	552	472	1024	512.0
c	440	453	893	446.5
ac	406	377	783	391.5
bc	605	500	1105	552.5
abc	392	419	811	405.5

Estimate the main and interaction effects of the factors and calculate the corresponding sums of squares.

14-7.3 Single Replicate of the 2^k Design

Learning Goals

- ❑ Explain a method to analyze a two-factor factorial design with a single replicate.
- ❑ Identify significant high-order interactions in a 2^k design by a normal probability plot.

2^k Design with a Single Replicate

Like the factorial design with a single replicate (presented in Section 14-4), an assumption is necessary for an unreplicated 2^k design to have the error mean square. Interactions higher than a second-order interaction are often assumed negligible and then used as the error mean square. Caution should be exercised to check if the interaction(s) used for the error term is significant.

Significant high-order interactions can be identified by constructing a normal probability plot of main and interaction effects. Negligible effects will be normally distributed with mean zero and variance σ^2 and thus lie along a straight line in the normal probability plot, whereas significant effects deviate from the straight line (see Figure 14-21 of MR).

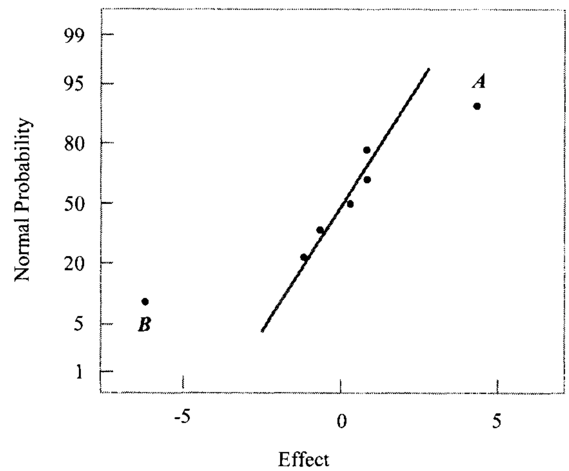
Example 14.4

(Normal Probability Plot of Effects) Consider the grip strength experiment with a single replicate in Example 14-3. The estimates of the main and interaction effects are as follows:

$A =$	4.25	$AC =$	0.75
$B =$	-6.25	$BC =$	-0.75
$C =$	0.75	$ABC =$	0.25
$AB =$	-1.25		

Construct a normal probability of these effects and check if the third-order interaction ABC is significantly important.

The normal probability plot of the main and interaction effects below indicates that the second- and third-order interactions (including ABC) are not important. Thus, ABC can be used as the error term to test the significance of the main and other second-order interaction effects.



Exercise 14.4 (MR 14-21)

An experiment has run a single replicate of a 2^4 design and calculated the following factor effects:

A = 80.25	AB = 53.25	ABC = -2.95
B = -65.50	AC = 11.00	ABD = -8.00
C = -9.25	AD = 9.75	ACD = 10.25
D = -20.50	BC = 18.36	BCD = -7.95
	BD = 15.10	$ABCD$ = -6.25
	CD = -1.25	

Construct a normal probability of these effects and check if the third- and fourth-order interactions are significantly important.

14-8 Blocking and Confounding in the 2^k Design

Learning Goals

- ☐ Construct 2^p blocks for a 2^k design.
- ☐ Estimate main and interaction effects of a 2^k design with 2^p blocks.

2^k Block Design and Confounding

Sometimes all treatment combinations cannot be run under homogeneous conditions for various reasons. For example, suppose that a 2^2 design (with 4 treatment combinations) requires a total of 16 hours (4 hours for each combination) to run a single replicate; thus, this experiment is conducted in two days (blocks) and randomization is restricted within each block. This kind of a factorial design is called **2^k design with 2^p blocks,** $p < k$.

Some effects are confounded with blocks in a 2^k block design. As an example, Figure 14-4 displays a 2^2 block design with interaction *AB* is confounded with the blocks. In this block design, the contrasts (*C*) to estimate the effects of *A*, *B*, and *AB* are

$$C_A = ab + a - b - (1) \qquad C_B = ab + b - a - (1) \qquad C_{AB} = ab + (1) - a - b$$

However, the *AB* contrast is identical to that of the block effect and thus *AB* is confounded with the blocks.

Block 1	Block 2
(1)	*a*
ab	*b*

Figure 14-4 A 2^2 block design.

Construction of Blocks

To construct a 2^k design with 2^p blocks, first, *p* effects should be selected for confounding. In general high-order interactions are chosen to confound with blocks; main and low-order interaction effects are less likely selected due to their importance. In addition to the *p* effects, $2^p - p - 1$ effects (interactions of the *p* effects; called **generalized interactions**) are confounded with blocks. For example, in a 2^3 design with 2^2 blocks, if *AB* and *BC* are selected for confounding, their generalized interaction *AC* ($= AB \times BC = AB^2C$ because $B^2 = I$) is also confounded. Care should be taken not to confound effects of interest.

Second, based on the selected *p* effects, **defining contrasts** (*L*) are formulated:

$$L = \alpha_1 x_1 + \alpha_{21} x_2 + \cdots + \alpha_k x_k$$

where: $$\alpha_i = \begin{cases} 0, & \text{if the } i^{\text{th}} \text{ factor is not in the counfounded effect} \\ 1, & \text{if the } i^{\text{th}} \text{ factor is in the counfounded effect} \end{cases}$$

$$x_i = \begin{cases} 0, & \text{if the } i^{\text{th}} \text{ factor is not in the treatment combination} \\ 1, & \text{if the } i^{\text{th}} \text{ factor is in the treatment combinaiton} \end{cases}$$

For example, in case that *AB* and *BC* are selected for confounding in a 2^3 design with 2^2 blocks, two defining contrasts are established from *AB* and *BC* as follows:

$$L_1 = x_1 + x_2 \quad \text{and} \quad L_2 = x_2 + x_3$$

where: $$x_i = \begin{cases} 0, & \text{if the } i^{\text{th}} \text{ factor is not in the treatment combination} \\ 1, & \text{if the } i^{\text{th}} \text{ factor is in the treatment combinaiton} \end{cases}$$

Construction of Blocks (*cont.*)

Lastly, treatment combinations having the same values of L (modulus 2) will be placed in the same block. For example, for the 2^3 block design with $L_1 = x_1 + x_2$ and $L_2 = x_2 + x_3$, the four blocks of the treatment combinations are determined according to their values of L_1 (mod 2) and L_2 (mod 2) as follows:

Treatment Combination	L_1 (mod 2) $[x_1 + x_2$ (mod 2)$]$	L_2 (mod 2) $[x_2 + x_3$ (mod 2)$]$	Block No.
(1)	(0 + 0) (mod 2) = 0	(0 + 0) (mod 2) = 0	1
a	(1 + 0) (mod 2) = 1	(0 + 0) (mod 2) = 0	2
b	(0 + 1) (mod 2) = 1	(1 + 0) (mod 2) = 1	3
ab	(1 + 1) (mod 2) = 0	(1 + 0) (mod 2) = 1	4
c	(0 + 0) (mod 2) = 0	(0 + 1) (mod 2) = 1	4
ac	(1 + 0) (mod 2) = 1	(0 + 1) (mod 2) = 1	3
bc	(0 + 1) (mod 2) = 1	(1 + 1) (mod 2) = 0	2
abc	(1 + 1) (mod 2) = 0	(1 + 1) (mod 2) = 0	1

ANOVA

The analysis of variance procedure by using orthogonal contrasts presented in Section 14-7.2 is applicable to the 2^k block design except treating the confounded effect(s). The confounded effect(s) is redefined as the block effect and the significance of the block effect can be tested by comparing with the error mean square. Recall that in case of a single replicate experiment, a high-order interaction(s) can be used as the error mean square.

Example 14.5

The effects of backrest angle (A) and location of lumbar support (B) on the disc pressure (unit: N) at the low back are under study. Two backrest angles (90° and 120°) and two lumbar support locations (1 and 5 cm) are selected in the experiment.

1. (**Construction of Blocks; 2^k Block Design**) Suppose that the four treatment combinations cannot be examined under the same condition. Establish two blocks with AB confounded for the 2^2 design.

The defining contrast is formulated from AB:

$$L = x_1 + x_2$$

where: $$x_i = \begin{cases} 0, & \text{if the } i^{\text{th}} \text{ factor is not in the treatment combination} \\ 1, & \text{if the } i^{\text{th}} \text{ factor is in the treatment combinaiton} \end{cases}$$

The four treatment combinations are assigned to two blocks according to the value of L (mod 2) as follows:

Treatment Combination	L (mod 2) $[x_1 + x_2$ (mod 2)$]$	Block No.
(1)	(0 + 0) (mod 2) = 0	1
a	(1 + 0) (mod 2) = 1	2
b	(0 + 1) (mod 2) = 1	2
ab	(1 + 1) (mod 2) = 0	1

2. (**Calculation of Effects and Sums of Squares**) The following disc pressure data are collected:

$$(1) = 620,\ a = 320,\ b = 380, \text{ and } ab = 155$$

Calculate the sums of squares for this 2^2 block design with AB confounded.

Example 14.5 (*cont.*)

The orthogonal contrasts of the 2^2 design for the disc pressure data are

$$C_A = a + ab - b - (1) = 320 + 155 - 380 - 620 = -525$$
$$C_B = b + ab - a - (1) = 380 + 155 - 320 - 620 = -405$$
$$C_{AB} = ab + (1) - a - b = 155 + 620 - 320 - 380 = 75$$

Thus, the estimates of the effects are

$$A = \frac{C_A}{n2^{k-1}} = \frac{-525}{2^{2-1}} = -525$$
$$B = \frac{C_B}{n2^{k-1}} = \frac{-405}{2^{2-1}} = -405$$
$$\text{Block effect} = AB = \frac{C_{AB}}{n2^{k-1}} = \frac{75}{2^{2-1}} = 75$$

The corresponding sums of squares are

$$SS_A = \frac{C_A^2}{n2^k} = \frac{(-525)^2}{1 \times 2^2} = 68{,}906.3$$
$$SS_B = \frac{C_B^2}{n2^k} = \frac{(-405)^2}{1 \times 2^2} = 41{,}006.3$$
$$SS_{\text{Block}} = SS_{AB} = \frac{C_{AB}^2}{n2^k} = \frac{75^2}{1 \times 2^2} = 1{,}406.3$$

Exercise 14.5 (MR 14-29)

Consider the tool life experiment in Exercise 14.3. Suppose that the eight treatment combinations cannot be examined under the same condition.

1. Establish four blocks with AB and AC confounded for the 2^3 design.
2. Calculate the sums of squares for this 2^3 design with four blocks. The orthogonal contrasts of the tool life length data have been calculated as follows:

$$C_A = 146,\ C_B = 674,\ C_C = 574,$$
$$C_{AB} = -90,\ C_{AC} = -954,\ C_{BC} = -194, \text{ and } C_{ABC} = -278$$

14-9 Fractional Replication of the 2^k Factorial Design

14-9.1 One Half Fraction of the 2^k Design

Learning Goals

- ❑ Describe the use of a fractional factorial design.
- ❑ Explain the terms *generator*, *defining relation*, *principal fraction*, *alternate fraction*, *aliasing*, and *dealiasing*.
- ❑ Construct a 2^{k-1} design.
- ❑ Estimate the main and interaction effects of a 2^{k-1} design.
- ❑ Identify the alias structure of a 2^{k-1} design.

2^{k-p} Fractional Factorial Design

A $1/2^p$ ($p < k$) fraction of a 2^k design (denoted as 2^{k-p}) can be run when high-order interactions are negligible. As the number of factors (= k) increases, the number of runs required increases geometrically but a large portion of the runs are used to estimate high-order interactions, which are often negligible. For example, a 2^5 design (k = five factors) which requires 32 runs uses, out of the 31 degrees of freedom, 5, 10, and 16 degrees of freedom to estimate the main effects, two-factor interactions, and three-factor and higher order interactions, respectively. When the three-factor and higher order interactions can be assumed negligible, a 2^{5-1} fractional factorial (which requires 16 runs) can be used to analyze the main effects and low-order interactions of the five factors.

Construction of 2^{k-1} Design

A 2^{k-1} design is constructed by selecting treatment combinations having the same sign (say, plus) on an effect column selected (called **generator**) in the table of signs for the 2^k design. For example, a 2^{3-1} design can be formed by selecting four treatment combinations (a, b, c, and abc) that have the plus sign on the ABC column in the sign table of the 2^3 design (Table 14-5). The fraction with the plus sign is called **principal fraction** and the other fraction is called **alternate fraction.**

Since the identity column has only the plus sign, a 2^{k-1} design has the following relations (called **defining relation**) between the identity and generator columns: the signs of the generator column are equal to those of the identity column for the principal fraction and the opposite becomes true for the alternate fraction. For example, a 2^{3-1} design with generator ABC has the defining relation $I = ABC$ for the principal fraction and $I = -ABC$ for the alternate fraction.

Table 14-5 Signs for the 2^3 Design

Treatment	Effects							
Combinations	I	A	B	AB	C	AC	BC	ABC
(1)	+	–	–	+	–	+	+	–
a	+	+	–	–	–	–	+	+
b	+	–	+	–	–	+	–	+
ab	+	+	+	+	–	–	–	–
c	+	–	–	+	+	–	–	+
ac	+	+	–	–	+	+	–	–
bc	+	–	+	–	+	–	+	–
abc	+	+	+	+	+	+	+	+

Another method to construct a 2^{k-1} design is available as follows:

Step 1: Prepare the sign table of a factorial design with $k - 1$ factors (called **basic design**). (e.g.) For a 2^{3-1} design in Table 14-6, the basic design starts with two factors (A and B).

Step 2: Add the k^{th} factor and identify the corresponding interaction by using the defining relation of the fractional design.

(e.g.) $I = ABC$ in Table 14-6

$\Rightarrow C{\cdot}I = ABC^2$

$\Rightarrow C = AB$ (for $C^2 = I$)

Step 3: Determine the signs of the interaction for the k^{th} factor column.

Step 4: Identify the treatment combinations of the 2^{k-1} design row by row.

Construction of 2^{k-1} Design (*cont.*)

Table 14-6 Construction of the 2^{3-1} Design with $I = ABC$

Run	Basic Design Treatment Combination	A	B	$C = AB$	Fractional Design Treatment Combination
1	(1)	−	−	+	c
2	a	+	−	−	a
3	b	−	+	−	b
4	ab	+	+	+	abc

Estimation of Effects

In a 2^{3-1} design with $I = ABC$ (principal fraction in Table 14-5), the linear combinations that are used to estimate the main effects and interactions have the following relationships:

$$l_A = \tfrac{1}{2}[a - b - c + abc] = l_{BC}$$
$$l_B = \tfrac{1}{2}[-a + b - c + abc] = l_{AC}$$
$$l_C = \tfrac{1}{2}[-a - b + c + abc] = l_{AB}$$

These relationships indicate that each liner combination estimates both a main effect and an interaction:

$$l_A = A + BC$$
$$l_B = B + AC$$
$$l_C = C + AB$$

In contrast, a 2^{3-1} design with $I = -ABC$ (alternate fraction in Table 14-5) produces the following estimates of the main effects and interactions:

$$l'_A = \tfrac{1}{2}[-(1) + ab + ac - bc] = A - BC$$
$$l'_B = \tfrac{1}{2}[-(1) + ab - ac + bc] = B - AC$$
$$l'_C = \tfrac{1}{2}[-(1) - ab + ac + bc] = C - AB$$

It is not possible to differentiate between A and BC, B and AC, C and AB. Thus, assuming the second-order interactions are negligible, the linear combinations l_A (or l'_A), l_B (or l'_B), and l_C (or l'_C) are used to estimate the main effects.

A normal probability plot can be used to identify significant interactions (see Section 14-7.3). Negligible interactions will lie on a straight line in the normal probability plot, whereas significant ones deviate far from the straight line.

Aliasing

In a fractional factorial design, some effects cannot be estimated separately and these effects are called **aliases**. For example, in a 2^{3-1} design with $I = ABC$, A and BC, B and AC, and C and AB are **aliased** with each other.

An alias of an effect is found by multiplying the defining relation to the effect. For example, a 2^{3-1} design with $I = ABC$ has the following alias structure:

$$A = A\cdot I = A\cdot ABC = A^2BC = BC \quad (\text{for } A^2 = I)$$
$$B = B\cdot I = B\cdot ABC = AB^2C = AC \quad (\text{for } B^2 = I)$$
$$C = C\cdot I = C\cdot ABC = ABC^2 = AB \quad (\text{for } C^2 = I)$$

Aliasing (*cont.*)

It is a common practice to form a fractional factorial design which aliases the main effects and low-order interactions with high-order interactions (which are mostly negligible).

Dealiasing

A sequence of fractional factorial designs can be performed to identify more detailed information of interactions. For example, a 2^{3-1} design with $I = ABC$ aliases the main effects with the two-factor interactions, assuming that the two-factor interactions are negligible. However, if this assumption is uncertain, the alternate fraction can be run after running the principal fraction.

By using the data from the principal and alternate fractions, the main effects and interactions can be estimated separately (called **dealiasing**) as follows:

$$\tfrac{1}{2}(l_A + l'_A) = \tfrac{1}{2}(A + BC + A - BC) = A \qquad \tfrac{1}{2}(l_A - l'_A) = BC$$
$$\tfrac{1}{2}(l_B + l'_B) = \tfrac{1}{2}(B + AC + B - AC) = B \qquad \tfrac{1}{2}(l_B - l'_B) = AC$$
$$\tfrac{1}{2}(l_C + l'_C) = \tfrac{1}{2}(C + AB + C - AB) = C \qquad \tfrac{1}{2}(l_C - l'_C) = AB$$

Projection of 2^{k-1} Design

When one or more factors are excluded from a 2^{k-1} design due to their lack of significance, the fractional design will project into a full factorial design to examine the active factors in detail. For example, a 2^{3-1} design will project into a 2^2 design as one factor is dropped.

Example 14.6

A study examines the effects of four factors on maximum acceptable weight (MAW; unit: kg) in lifting: object width (A), lifting level (B), lifting distance (C), and lifting frequency (D). A 2^{4-1} design with $I = ABCD$ is employed, yielding the following data:

Run	A	B	C	$D = ABC$	Treatment Combination	MAW
1	–	–	–	–	(1)	24
2	+	–	–	+	*ad*	32
3	–	+	–	+	*bd*	24
4	+	+	–	–	*ab*	16
5	–	–	+	+	*cd*	37
6	+	–	+	–	*ac*	15
7	–	+	+	–	*bc*	17
8	+	+	+	+	*abcd*	18

1. (**Alias Structure; 2^{k-1} Design**) Identify the alias structure of this 2^{4-1} design with $I = ABCD$.

The 2^{4-1} design with $I = ABCD$ has the following alias relationships:

$$A = A \cdot ABCD = A^2BCD = BCD$$
$$B = B \cdot ABCD = AB^2CD = ACD$$
$$C = C \cdot ABCD = ABC^2D = ABD$$
$$D = D \cdot ABCD = ABCD^2 = ABC$$
$$AB = AB \cdot ABCD = A^2B^2CD = CD$$
$$AC = AC \cdot ABCD = A^2BC^2D = BD$$
$$AD = AD \cdot ABCD = A^2BCD^2 = BC$$

Example 14.6 (*cont.*)

2. (**Estimation of Effects**) Estimate the factor effects. Which effects appear large?

The main effects and interactions are estimated as follows:

$$l_A = A + BCD = \tfrac{1}{4}(-24 + 32 - 24 + 16 - 37 + 15 - 17 + 18) = -5.25$$

$$l_B = B + ACD = \tfrac{1}{4}(-24 - 32 + 24 + 16 - 37 - 15 + 17 + 18) = -8.25$$

$$l_C = C + ABD = \tfrac{1}{4}(-24 - 32 - 24 - 16 + 37 + 15 + 17 + 18) = -2.25$$

$$l_D = D + ABC = \tfrac{1}{4}(-24 + 32 + 24 - 16 + 37 - 15 - 17 + 18) = 9.75$$

$$l_{AB} = AB + CD = \tfrac{1}{4}(24 - 32 - 24 + 16 + 37 - 15 - 17 + 18) = 1.75$$

$$l_{AC} = AC + BD = \tfrac{1}{4}(24 - 32 + 24 - 16 - 37 + 15 - 17 + 18) = -5.25$$

$$l_{AD} = AD + BC = \tfrac{1}{4}(24 + 32 - 24 - 16 - 37 - 15 + 17 + 18) = -0.25$$

(Note) The interactions are estimated by forming columns of AB, AC, and AD in the sign table.

The effects A, B, D, and AC are relatively large.

Exercise 14.6 (MR 14-32)

A 2^{4-1} design with $I = ABCD$ is used to study four factors in a chemical process. The factors are A = temperature, B = pressure, C = concentration, and D = stirring rate, and the response is filtration rate. The design and data are as follows:

Run	A	B	C	$D = ABC$	Treatment Combination	Filtration Rate
1	–	–	–	–	(1)	45
2	+	–	–	+	*ad*	100
3	–	+	–	+	*bd*	45
4	+	+	–	–	*ab*	65
5	–	–	+	+	*cd*	75
6	+	–	+	–	*ac*	60
7	–	+	+	–	*bc*	80
8	+	+	+	+	*abcd*	96

1. Estimate the factor effects. Which factor effects appear large?
2. Project this design into a full factorial including only important factors and provide a practical interpretation of the results.

14-9.2 Smaller Fractions: The 2^{k-p} Fractional Factorial

Learning Goals

- ☐ Explain the conditions of design resolutions III, IV, and V in a fractional factorial design.
- ☐ Construct a 2^{k-p} fractional factorial with a designated design resolution.
- ☐ Identify the alias structure of a fractional factorial design.

Design Resolution

The term *design resolution* indicates the pattern of aliases in a fractional factorial design. The alias patterns of resolution III, IV, and V designs are compared with each other in Table 14-7. A resolution III design (such as 2^{3-1} design with $I = ABC$) does not alias any main effects with other main effects but aliases main

Design Resolution (*cont.*)

effects with two-factor interactions and some two-factor interactions with each other. Next, a resolution IV design (such as 2^{4-1} design with $I = ABCD$) does not alias any main effects with other main effects or two-factor interactions but aliases two-factor interactions with other two-factor interactions. Lastly, a resolution V design (such as 2^{5-1} design with $I = ABCDE$) does not alias any main effects and two-factor interactions with other main effects and two-factor interactions. The resolution of a fractional factorial design is indicated with a subscript (e.g., 2_{III}^{3-1} denotes a 2^{3-1} design with resolution III).

As the design resolution becomes higher, more information about interactions is produced. For example, a resolution III design provides information about main effects only, while a resolution IV design provides information about two-factor interactions as well as main effects. Resolution III and IV designs are often used in experiments to screen factors.

Table 14-7 Alias Patterns of Resolution III, IV, and V Designs
(×: not aliased; O: aliased)

	Main Effects	Two-Factor Interactions
Main Effects	× (III) × (IV) × (V)	O (III) × (IV) × (V)
Two-Factor Interactions	O (III) × (IV) × (V)	O (III) O (IV) × (V)

Construction of 2^{k-p} Design

To achieve a designated resolution for a fractional design with k factors, a set of p generators should be selected properly. Table 14-29 of MR presents recommended generators for various 2^{k-p} designs ($k \le 15$), design resolutions (III, IV, and V), and runs ($n \le 128$). For example, the table recommends two design generators ($D = AB$ and $E = AC$, which are equivalent to the design relation $I = ABD = ACE$) for a 2_{III}^{5-2} design with $n = 8$ runs.

With the recommended generators of a fractional design, a **complete defining relation** is determined. Incorporating the generalized interactions of the generators (identified by multiplying the generators with each other; $2^p - p - 1$ generalized interactions are produced by p generators), the complete defining relation is formulated. For example, the two ($p = 2$) recommended generators, ABD and ACE, of the 2_{III}^{5-2} design produces one ($=2^p - p - 1 = 2^2 - 2 - 1$) generalized interaction:

$$ABD \cdot ACE = A^2BCDE = BCDE$$

Thus, the complete defining relation of the 2_{III}^{5-2} design becomes

$$I = ABD = ACE = BCDE$$

Construction of 2^{k-p} Design (*cont.*)

Like the method to construct a 2^{k-1} design, a 2^{k-p} design is constructed as follows (see Table 14-8 for an example):

Step 1: Prepare the sign table of a factorial design with $k-p$ factors (basic design).
Step 2: Add p columns for recommended design generators.
Step 3: Determine the signs of the p columns with the design generators.
Step 4: Identify the treatment combinations of the 2^{k-p} design row by row.

Table 14-8 Construction of the 2^{5-2}_{III} Design with $I = ABD = ACE = BCDE$

Run	A	B	C	$D = AB$	$E = AC$	Fractional Design Treatment Combination
1	–	–	–	+	+	*de*
2	+	–	–	–	–	*a*
3	–	+	–	–	+	*be*
4	+	+	–	+	–	*abd*
5	–	–	+	+	–	*cd*
6	+	–	+	–	+	*ace*
7	–	+	+	–	–	*bc*
8	+	+	+	+	+	*abcde*

Alias Structure of 2^{k-p} Design

The alias structure of a 2^{k-p} design is found by multiplying each effect to the complete defining relation. For example, the 2^{5-2}_{III} design with $I = ABD = ACE = BCDE$ has the following alias structure:

$A = BD = CE = ABCDE$
$B = AD = ABCE = CDE$
$C = ABCD = AE = BDE$
$D = AB = ACDE = BCE$
$E = ABDE = AC = BCD$
$BC = ACD = ABE = DE$
$CD = ABC = ADE = BE$

When a number of factors under consideration is large (say, $k \geq 5$), high-order (say, third- or higher-order) interactions are assumed to be negligible, which greatly simplifies the alias structure. For example, the alias structure above of the 2^{5-2}_{III} design with $I = ABD = ACE = BCDE$ can be simplified as follows by assuming that interactions higher than three factors are assumed negligible:

$A = BD = CE$
$B = AD = CDE$
$C = AE = BDE$
$D = AB = BCE$
$E = AC = BCD$
$BC = ACD = ABE = DE$
$CD = ABC = ADE = BE$

Example 14.7

1. (**Construction of a 2^{k-p} Design**) Construct a 2^{6-2}_{IV} fractional factorial design.

Table 14-29 of MR recommends two ($p = 2$) design generators, $E = ABC$ and $F = BCD$, for the 2^{6-2}_{IV} design, which yield the defining relation $I = ABCE = BCDF$. The two generators produce one generalized interaction:

$$ABCE \cdot BCDF = AB^2C^2DEF = ADEF$$

Example 14.7 (*cont.*)

Thus, the complete defining relation of the 2_{IV}^{6-2} design is

$I = ABCE = BCDF = ADEF$

By using the two design generators, the 2_{IV}^{6-2} design with $I = ABCE = BCDF = ADEF$ is established as follows:

Run	A	B	C	D	$E = ABC$	$F = BCD$	Fractional Design Treatment Combination
1	–	–	–	–	–	–	(1)
2	+	–	–	–	+	–	*ae*
3	–	+	–	–	+	+	*bef*
4	+	+	–	–	–	+	*abf*
5	–	–	+	–	+	+	*cef*
6	+	–	+	–	–	+	*acf*
7	–	+	+	–	–	–	*bc*
8	+	+	+	–	+	–	*abce*
9	–	–	–	+	–	+	*df*
10	+	–	–	+	+	+	*adef*
11	–	+	–	+	+	–	*bde*
12	+	+	–	+	–	–	*abd*
13	–	–	+	+	+	–	*cde*
14	+	–	+	+	–	–	*acd*
15	–	+	+	+	–	+	*bcdf*
16	+	+	+	+	+	+	*abcdef*

2. (**Alias Structure**) Identify the aliases of the fractional design, assuming that interactions higher than three factors are negligible.

Assuming that four-factor or higher-order interactions are negligible, the following aliases are found by multiplying each effect to the complete defining relation of the 2_{III}^{6-3} design:

$A = BCE = DEF$	$AB = CE$
$B = ACE = CDF$	$AC = BE$
$C = ABE = BDF$	$AD = EF$
$D = BCF = AEF$	$AE = BC = DF$
$E = ABC = ADF$	$AF = DE$
$F = BCD = ADE$	$BD = CF$
	$BF = CD$

Exercise 14.7 (MR 14-39)

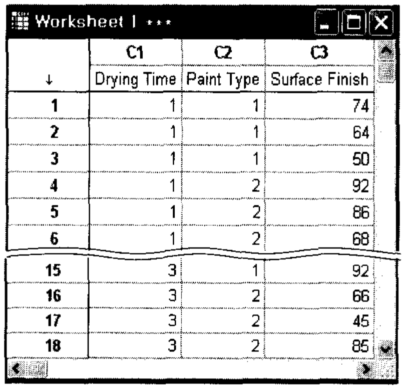
1. Construct a 2_{III}^{6-3} fractional factorial design.
2. Identify the aliases of the fractional design, assuming that only the main effects and two-factor interactions are of interest.

MINITAB Applications

Example 14.1 **(Two-Factor Factorial Design)**

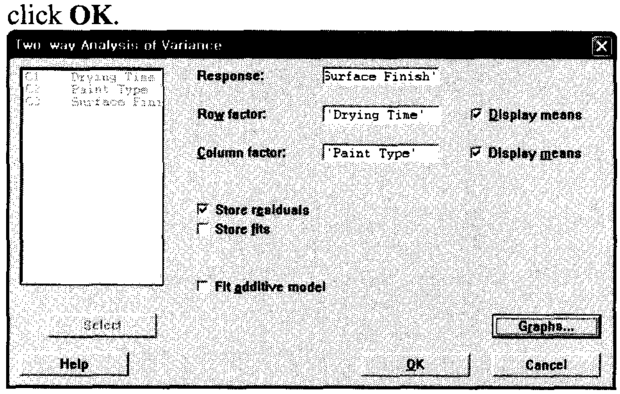
(1) Choose **File ➤ New**, click **Minitab Project**, and click **OK**.

(2) Enter the surface finish data on the worksheet.

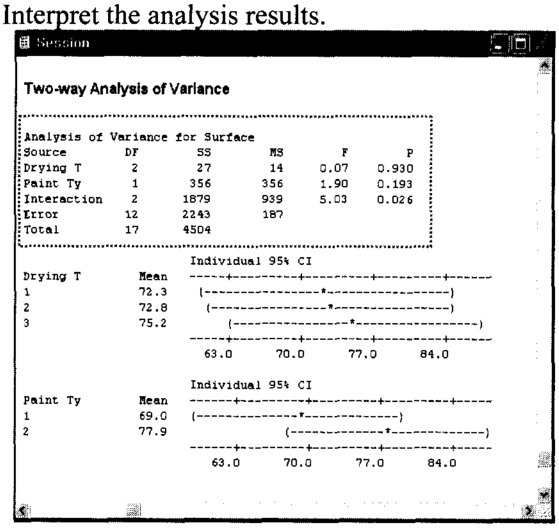


↓	C1 Drying Time	C2 Paint Type	C3 Surface Finish
1	1	1	74
2	1	1	64
3	1	1	50
4	1	2	92
5	1	2	86
6	1	2	68
15	3	1	92
16	3	2	66
17	3	2	45
18	3	2	85

(3) Choose **Stat ➤ ANOVA ➤ Two-way**. In **Response** select *Surface Finish* and in **Row factor** and **Column factor** select *Drying Time* and *Paint Type*, respectively. Check **Display means** and **Store residuals**. Then click **OK**.



(4) Interpret the analysis results.



Session

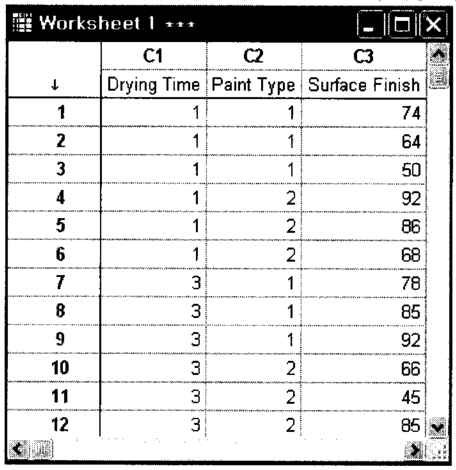
Two-way Analysis of Variance

```
Analysis of Variance for Surface
Source        DF       SS       MS        F        P
Drying T       2       27       14     0.07    0.930
Paint Ty       1      356      356     1.90    0.193
Interaction    2     1879      939     5.03    0.026
Error         12     2243      187
Total         17     4504

                         Individual 95% CI
Drying T       Mean   -----+---------+---------+---------+-----
1              72.3    (---------------*---------------)
2              72.8     (---------------*---------------)
3              75.2        (---------------*---------------)
                      -----+---------+---------+---------+-----
                          63.0      70.0      77.0      84.0

                         Individual 95% CI
Paint Ty       Mean   ------+---------+---------+---------+-----
1              69.0   (------------*------------)
2              77.9               (------------*------------)
                      ------+---------+---------+---------+-----
                           63.0      70.0      77.0      84.0
```

Example 14.2 **(2^2 Factorial Design)**

(1) Choose **File ➤ New**, click **Minitab Project**, and click **OK**.

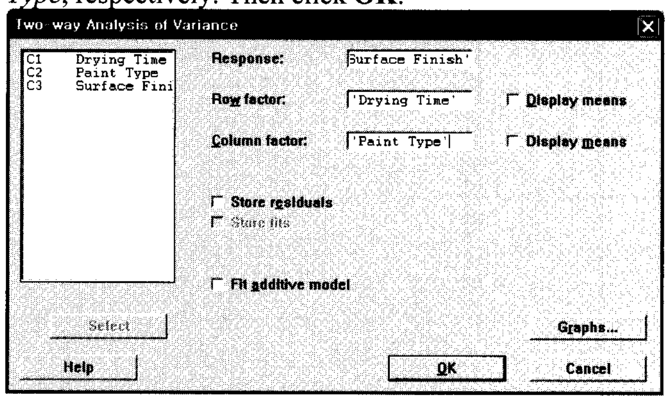
(2) Enter the surface finish data on the worksheet.



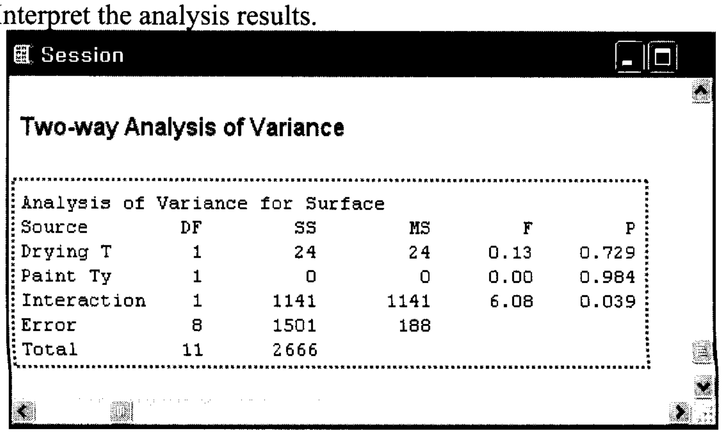
Worksheet 1 ***

↓	C1 Drying Time	C2 Paint Type	C3 Surface Finish
1	1	1	74
2	1	1	64
3	1	1	50
4	1	2	92
5	1	2	86
6	1	2	68
7	3	1	78
8	3	1	85
9	3	1	92
10	3	2	66
11	3	2	45
12	3	2	85

(3) Choose **Stat ➤ ANOVA ➤ Two-way**. In **Response** select *Surface Finish* and in **Row factor** and **Column factor** select *Drying Time* and *Paint Type*, respectively. Then click **OK**.



(4) Interpret the analysis results.



Session

Two-way Analysis of Variance

Analysis of Variance for Surface

Source	DF	SS	MS	F	P
Drying T	1	24	24	0.13	0.729
Paint Ty	1	0	0	0.00	0.984
Interaction	1	1141	1141	6.08	0.039
Error	8	1501	188		
Total	11	2666			

Example 14.4 **(Normal Probability Plot of Effects)**

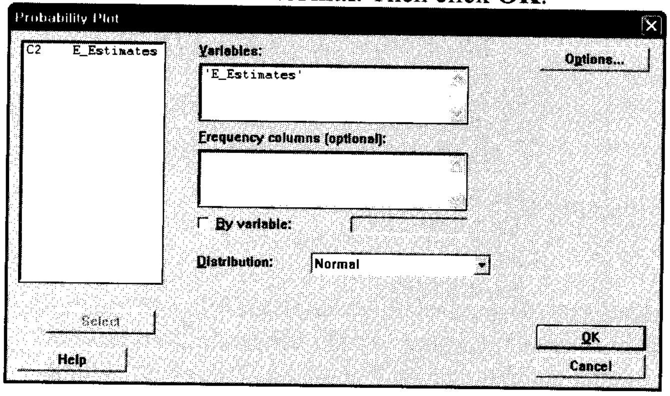
(1) Choose **File ➤ New**, click **Minitab Project**, and click **OK**.

(2) Enter the estimates of the effects on the worksheet.

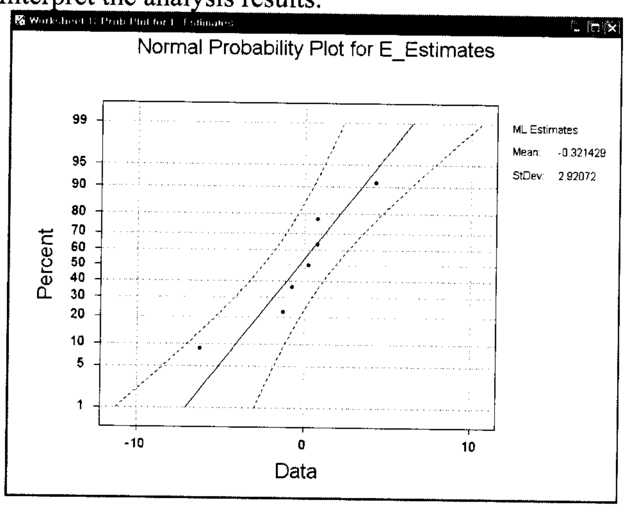
Worksheet 1 ***

↓	C1-T Effects	C2 E_Estimates
1	A	4.25
2	B	-6.25
3	AB	-1.25
4	C	0.75
5	AC	0.75
6	BC	-0.75
7	ABC	0.25

(3) Choose **Graph ➤ Probability Plot**. In **Variables** select *E_Estimates* and in **Distribution** select **Normal**. Then click **OK**.



(4) Interpret the analysis results.



Answers to Exercises

Exercise 14.1

1. **(Test on the Significance of Main and Interaction Effects; Two-Factor Factorial Design)**

The analysis of variance of the carbon anode density data is as follows:

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}^2 - \frac{y_{...}^2}{N} = 1{,}003.1 - \frac{127.79^2}{18} = 95.87$$

$$SS_A = \sum_{i=1}^{a} \frac{y_{i..}^2}{bn} - \frac{y_{...}^2}{N} = \frac{6{,}010.6}{6} - \frac{127.79^2}{18} = 94.53$$

$$SS_B = \sum_{j=1}^{b} \frac{y_{.j.}^2}{an} - \frac{y_{...}^2}{N} = \frac{8{,}171.6}{9} - \frac{127.79^2}{18} = 0.72$$

$$SS_{AB} = \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{y_{ij.}^2}{n} - \frac{y_{...}^2}{N} - SS_A - SS_B$$

$$= \frac{3{,}007.7}{3} - \frac{127.79^2}{18} - 94.53 - 0.72 = 0.08$$

$$SS_E = SS_T - (SS_A + SS_B + SS_{AB}) = 95.87 - (94.53 + 0.72 + 0.08) = 0.54$$

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	f_0
Firing Temperature (A)	94.53	2	47.27	1,056.1
Furnace Position (B)	0.72	1	0.72	16.0
AB	0.08	2	0.04	0.9
Error	0.54	12	0.04	
Total	95.87	17		

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

$$\left.\begin{array}{ll} H_0: & \tau_i = 0 \quad \text{for } i = 1 \text{ to } 3 \\ H_1: & \tau_i \neq 0 \quad \text{for at least one } i \end{array}\right\} \text{ for } A$$

$$\left.\begin{array}{ll} H_0: & \beta_j = 0 \quad \text{for } j = 1 \text{ and } 2 \\ H_1: & \beta_j \neq 0 \quad \text{for at least one } j \end{array}\right\} \text{ for } B$$

$$\left.\begin{array}{ll} H_0: & (\tau\beta)_{ij} = 0 \quad \text{for all combinations of } i\text{'s and } j\text{'s} \\ H_1: & (\tau\beta)_{ij} \neq 0 \quad \text{for at least one combination of } i \text{ and } j \end{array}\right\} \text{ for } AB$$

Step 2: Determine a **test statistic and its value.**

$$f_0 = \frac{MS_A}{MS_E} = \frac{47.27}{0.04} = 1{,}056.1 \quad \text{for } A$$

$$f_0 = \frac{MS_B}{MS_E} = \frac{0.72}{0.04} = 16.0 \quad \text{for } B$$

Example 14.1 (*cont.*)

$$f_0 = \frac{MS_{AB}}{MS_E} = \frac{0.04}{0.04} = 0.9 \quad \text{for } AB$$

Step 3: Determine a **critical value(s) for α.**

$$f_{\alpha,a-1,N-ab} = f_{0.05,3-1,18-6} = f_{0.05,2,12} = 3.89 \quad \text{for } A$$

$$f_{\alpha,b-1,N-ab} = f_{0.05,2-1,18-6} = f_{0.05,1,12} = 4.75 \quad \text{for } B$$

$$f_{\alpha,(a-1)(b-1),N-ab} = f_{0.05,(3-1)(2-1),18-6} = f_{0.05,2,12} = 3.89 \quad \text{for } AB$$

Step 4: Make a **conclusion.**

Since $f_0 = 0.9 \ngtr f_{\alpha,(a-1)(b-1),N-ab} = 3.89$, fail to reject H_0 at $\alpha = 0.05$ for AB, which indicates further tests on the main effects A and B are meaningful. Next, since $f_0 = 1{,}056.1 > f_{\alpha,a-1,N-ab} = 3.89$ for A and $f_0 = 16.0 > f_{\alpha,b-1,N-ab} = 4.75$ for B, we can conclude that both A (firing temperature) and B (furnace position) significantly affect the baked density of a carbon anode.

2. (**Multiple Comparison**)

Step 1: Determine the **LSD** for α.

$$LSD = t_{\alpha/2,N-ab}\sqrt{\frac{2MS_E}{bn}} = t_{0.05/2,18-3\times2}\sqrt{\frac{2\times0.04}{2\times3}} = 2.179\times0.116 = 0.25$$

Step 2: Arrange the means of a treatments in **descending** order.

$\bar{y}_{2..} = 10.34$, $\bar{y}_{1..} = 5.52$, and $\bar{y}_{3..} = 5.44$

Step 3: Compare the difference between the **largest** and **smallest** means with the LSD. Continue this comparison with the **next smallest** mean as long as the mean difference is greater than the LSD.

2 vs. 3 = 10.34 − 5.44 = 4.90 > 0.25
2 vs. 1 = 10.34 − 5.52 = 4.82 > 0.25

Step 4: Continue Step 3 for the **next largest** mean until this iteration reaches the **second smallest** mean.

1 vs. 3 = 5.52 − 5.44 = 0.08 < 0.25

Step 5: Summarize the pairwise comparison results by arranging the a treatments in order of mean and then **underlining** treatments whose means are **not significantly different.**

2 <u>1 3</u>

3. (**Calculation of Residual**)

$\bar{y}_{21.} = 10.62$

$e_{213} = y_{213} - \hat{y}_{213} = y_{213} - \bar{y}_{21.} = 10.43 - 10.62 = -0.19$ (overestimate)

Exercise 14.2 **(Calculation of Effects and Sums of Squares; 2^2 Design)**

The orthogonal contrasts of the 2^2 design for the baked density data are

$$C_A = a + ab - b - (1) = 16.65 + 15.96 - 15.96 - 17.18 = -0.53$$
$$C_B = b + ab - a - (1) = 15.96 + 15.96 - 16.65 - 17.18 = -1.91$$
$$C_{AB} = ab + (1) - a - b = 15.96 + 17.18 - 16.65 - 15.96 = 0.53$$

Thus, the estimates of the main and interaction effects are

$$A = \frac{C_A}{n2^{k-1}} = \frac{-0.53}{3\times 2^{2-1}} = -0.09 \qquad B = \frac{C_B}{n2^{k-1}} = \frac{-1.91}{3\times 2^{2-1}} = -0.32$$
$$AB = \frac{C_{AB}}{n2^{k-1}} = \frac{0.53}{3\times 2^{2-1}} = 0.09$$

The corresponding sums of squares are

$$SS_A = \frac{C_A^2}{n2^k} = \frac{(-0.53)^2}{3\times 2^2} = 0.02 \qquad SS_B = \frac{C_B^2}{n2^k} = \frac{(-1.91)^2}{3\times 2^2} = 0.30$$
$$SS_{AB} = \frac{C_{AB}^2}{n2^k} = \frac{0.53^2}{3\times 2^2} = 0.02$$

Exercise 14.3 **(Calculation of Effects and Sums of Squares; 2^k Design)**

The orthogonal contrasts of the 2^3 design for the tool life length data are

$$C_A = a + ab + ac + abc - (1) - b - c - bc$$
$$= 760 + 1{,}024 + 783 + 811 - 532 - 702 - 893 - 1{,}105 = 146$$
$$C_B = b + ab + bc + abc - (1) - a - c - ac$$
$$= 702 + 1{,}024 + 1{,}105 + 811 - 532 - 760 - 893 - 783 = 674$$
$$C_{AB} = (1) + ab + c + abc - a - b - ac - bc$$
$$= 532 + 1{,}024 + 893 + 811 - 760 - 702 - 783 - 1{,}105 = -90$$
$$C_C = c + ac + bc + abc - (1) - a - b - ab$$
$$= 893 + 783 + 1{,}105 + 811 - 532 - 760 - 702 - 1{,}024 = 574$$
$$C_{AC} = (1) + b + ac + abc - a - ab - c - bc$$
$$= 532 + 702 + 783 + 811 - 760 - 1{,}024 - 893 - 1{,}105 = -954$$
$$C_{BC} = (1) + a + bc + abc - b - ab - c - ac$$
$$= 532 + 760 + 1{,}105 + 811 - 702 - 1{,}024 - 893 - 783 = -194$$
$$C_{ABC} = a + b + c + abc - (1) - ab - ac - bc$$
$$= 760 + 702 + 893 + 811 - 532 - 1{,}024 - 783 - 1{,}105 = -278$$

Thus, the estimates of the main and interaction effects are

$$A = \frac{C_A}{n2^{k-1}} = \frac{146}{2\times 2^{3-1}} = 18.3 \qquad B = \frac{C_B}{n2^{k-1}} = \frac{674}{2\times 2^{3-1}} = 84.3$$
$$C = \frac{C_C}{n2^{k-1}} = \frac{574}{2\times 2^{3-1}} = 71.8 \qquad AB = \frac{C_{AB}}{n2^{k-1}} = \frac{-90}{2\times 2^{3-1}} = -11.3$$
$$AC = \frac{C_{AC}}{n2^{k-1}} = \frac{-954}{2\times 2^{3-1}} = -119.3 \qquad BC = \frac{C_{BC}}{n2^{k-1}} = \frac{-194}{2\times 2^{3-1}} = -24.3$$

Exercise 14.3 (*cont.*)

$$ABC = \frac{C_{ABC}}{n2^{k-1}} = \frac{-278}{2\times 2^{3-1}} = -34.8$$

The corresponding sums of squares are

$$SS_A = \frac{C_A^2}{n2^k} = \frac{146^2}{2\times 2^3} = 1{,}332.3 \qquad SS_B = \frac{C_B^2}{n2^k} = \frac{674^2}{2\times 2^3} = 28{,}392.3$$

$$SS_C = \frac{C_C^2}{n2^k} = \frac{574^2}{2\times 2^3} = 20{,}592.3 \qquad SS_{AB} = \frac{C_{AB}^2}{n2^k} = \frac{(-90)^2}{2\times 2^3} = 506.3$$

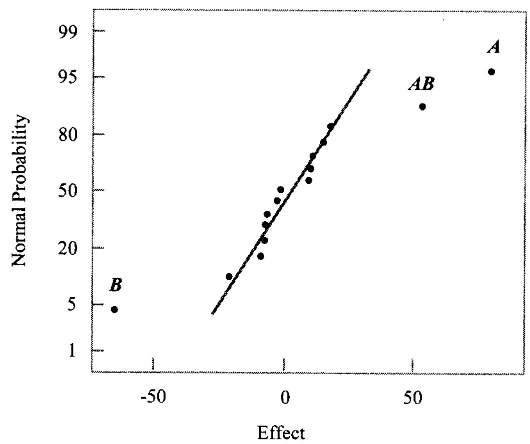
$$SS_{AC} = \frac{C_{AC}^2}{n2^k} = \frac{(-954)^2}{2\times 2^3} = 56{,}882.3 \qquad SS_{BC} = \frac{C_{BC}^2}{n2^k} = \frac{(-194)^2}{2\times 2^3} = 2{,}352.3$$

$$SS_{ABC} = \frac{C_{ABC}^2}{n2^k} = \frac{(-278)^2}{2\times 2^3} = 4{,}830.3$$

Exercise 14.4

(Normal Probability Plot of Effects)

The normal probability plot of the effects below indicates that all the third- and fourth-order interactions are not significant so that we can use these insignificant interactions to form the error mean square.



Exercise 14.5

1. **(Construction of Blocks; 2^k Block Design)**

The defining contrasts are formulated from AB and AC as follows:

$L_1 = x_1 + x_2$ and $L_2 = x_1 + x_3$

The eight treatment combinations are assigned to four blocks according to the values of L_1 (mod 2) and L_2 (mod 2) as follows:

Treatment Combination	L_1 (mod 2) [$x_1 + x_2$ (mod 2)]	L_2 (mod 2) [$x_1 + x_3$ (mod 2)]	Block No.
(1)	(0 + 0) (mod 2) = 0	(0 + 0) (mod 2) = 0	1
a	(1 + 0) (mod 2) = 1	(1 + 0) (mod 2) = 1	2
b	(0 + 1) (mod 2) = 1	(0 + 0) (mod 2) = 0	3
ab	(1 + 1) (mod 2) = 0	(1 + 0) (mod 2) = 1	4
c	(0 + 0) (mod 2) = 0	(0 + 1) (mod 2) = 1	4
ac	(1 + 0) (mod 2) = 1	(1 + 1) (mod 2) = 0	3
bc	(0 + 1) (mod 2) = 1	(0 + 1) (mod 2) = 1	2
abc	(1 + 1) (mod 2) = 0	(1 + 1) (mod 2) = 0	1

Exercise 14.5 (*cont.*)

2. (**Calculation of Effects and Sums of Squares**)

The generalized interaction of AB and AC is

$$AB \times AC = A^2BC = BC$$

Thus, the effects AB, AC, and BC are confounded with the blocks. The estimates of the effects are

$$A = \frac{C_A}{n2^{k-1}} = \frac{146}{2\times 2^{3-1}} = 18.3 \qquad B = \frac{C_B}{n2^{k-1}} = \frac{674}{2\times 2^{3-1}} = 84.3$$

$$C = \frac{C_C}{n2^{k-1}} = \frac{574}{2\times 2^{3-1}} = 71.8 \qquad ABC = \frac{C_{ABC}}{n2^{k-1}} = \frac{-278}{2\times 2^{3-1}} = -34.8$$

$$\text{Block effect} = AB + AC + BC = \frac{C_{AB} + C_{AC} + C_{BC}}{n2^{k-1}}$$

$$= \frac{-90 - 954 - 194}{2\times 2^{3-1}} = 154.8$$

The corresponding sums of squares are

$$SS_A = \frac{C_A^2}{n2^k} = \frac{146^2}{2\times 2^3} = 1{,}332.3 \qquad SS_B = \frac{C_B^2}{n2^k} = \frac{674^2}{2\times 2^3} = 28{,}392.3$$

$$SS_C = \frac{C_C^2}{n2^k} = \frac{574^2}{2\times 2^3} = 20{,}592.3 \qquad SS_{ABC} = \frac{C_{ABC}^2}{n2^k} = \frac{(-278)^2}{2\times 2^3} = 4{,}830.3$$

$$SS_{\text{Block}} = SS_{AB} + SS_{AC} + SS_{BC} = \frac{C_{AB}^2 + C_{AC}^2 + C_{BC}^2}{n2^k}$$

$$= \frac{(-90)^2 + (-954)^2 + (-194)^2}{2\times 2^3} = 59{,}740.8$$

Exercise 14.6

1. (**Estimation of Effects**)

The main effects and interactions are estimated as follows:

$$l_A = A + BCD = \tfrac{1}{4}(-45 + 100 - 45 + 65 - 75 + 60 - 80 + 96) = 19.0$$

$$l_B = B + ACD = \tfrac{1}{4}(-45 - 100 + 45 + 65 - 75 - 60 + 80 + 96) = 1.5$$

$$l_C = C + ABD = \tfrac{1}{4}(-45 - 100 - 45 - 65 + 75 + 60 + 80 + 96) = 14.0$$

$$l_D = D + ABC = \tfrac{1}{4}(-45 + 100 + 45 - 65 + 75 - 60 - 80 + 96) = 16.5$$

$$l_{AB} = AB + CD = \tfrac{1}{4}(45 - 100 - 45 + 65 + 75 - 60 - 80 + 96) = -1.0$$

$$l_{AC} = AC + BD = \tfrac{1}{4}(45 - 100 + 45 - 65 - 75 + 60 - 80 + 96) = -18.5$$

$$l_{AD} = AD + BC = \tfrac{1}{4}(45 + 100 - 45 - 65 - 75 - 60 + 80 + 96) = 19.0$$

Three factors A, C, and D show large effects on filtration rate, but factor B displays a relatively very small effect.

2. (**Projection into a Factorial Design**)

The 2^{4-1} design projects into a 2^3 design with a single replicate including factors A, C, and D. The main effects and interactions of this unreplicated 2^3 design can be easily determined by using the effect estimates of the 2^{4-1} design as follows:

Exercise 14.6 (*cont.*)

$A = 19.0$ $AC = -18.5$
$C = 14.0$ $AD = 19.0$
$D = 16.5$ $CD = -1.0$
$ACD = 1.5$

The effects A, C, D, AC, and AD have relatively large influence on filtration rate.

Exercise 14.7

1. **(Construction of a 2^{k-p} Design)**

Table 14-29 of MR recommends three ($p = 3$) design generators, $D = AB$, $E = AC$, and $F = BC$, for the 2^{6-3}_{III} design, which yield the defining relation $I = ABD = ACE = BCF$. These three generators produce four ($= 2^p - p - 1 = 2^3 - 3 - 1$) generalized interactions:

$ABD \cdot ACE = A^2BCDE = BCDE$
$ABD \cdot BCF = AB^2CDF = ACDF$
$ACE \cdot BCF = ABC^2EF = ABEF$
$ABD \cdot ACE\ BCF = A^2B^2C^2DEF = DEF$

Thus, the complete defining relation of the 2^{6-3}_{III} design is

$I = ABD = ACE = BCF = BCDE = ACDF = ABEF = DEF$

By using the three design generators, the 2^{6-3}_{III} design with $I = ABD = ACE = BCF = BCDE = ACDF = ABEF = DEF$ is established as follows:

Run	A	B	C	$D = AB$	$E = AC$	$F = BC$	Fractional Design Treatment Combination
1	–	–	–	+	+	+	*def*
2	+	–	–	–	–	+	*af*
3	–	+	–	–	+	–	*be*
4	+	+	–	+	–	–	*abd*
5	–	–	+	+	–	–	*cd*
6	+	–	+	–	+	–	*ace*
7	–	+	+	–	–	+	*bcf*
8	+	+	+	+	+	+	*abcdef*

2. **(Alias Structure)**

Assuming that three-factor or higher-order interactions are negligible, the following aliases are found by multiplying each effect to the complete defining relation of the 2^{6-3}_{III} design:

$A = BD = CE$ $B = AD = CF$
$C = AE = BF$ $D = AB = EF$
$E = AC = DF$ $F = BC = DE$
$AF = CD = BE$

15 Nonparametric Statistics

OUTLINE

15-1 Introduction

Learning Goals

- ❑ Explain nonparametric statistics in comparison with parametric statistics.

Nonparametric Statistics

Nonparametric methods do not use information of the distribution of the underlying population, whereas parametric methods (such as t and F tests) are based on the sampling distribution of a particular statistic for the parameter of interest. The nonparametric methods make no assumptions except the underlying distribution is continuous, using only signs and/or ranks—the original observations may need to be transformed. Therefore, the nonparametric statistics is considered a distribution-free method and does not include the procedure of parameter estimation (point and interval estimations).

Parametric and nonparametric statistics have their own advantages and disadvantages in terms of ease of use, time efficiency, applicability, and sensitivity (statistical power), as summarized in Table 15-1. Nonparametric tests are easier and quicker to perform and their applications are extended to the following contexts where parametric tests are inappropriate:

(1) **Underlying distribution unknown**: Information of the underlying distribution is unavailable.

(2) **Normal approximation inapplicable**: Normal approximation is not applicable when the underlying distribution is significantly deviated from a normal distribution and the sample size is small (say, $n < 30$) (see Section 7-5).

(3) **Categorical data**: Categorical data include nominal and ordinal data (see Section 9-7).

Nonparametric Statistics

However, nonparametric tests are less sensitive (statically powerful) in detecting a small departure from the hypothesized value because they do not use all the information of the sample. Therefore, the nonparametric methods usually require a larger sample size to achieve the same statistical power.

Table 15-1 Comparison of Parametric and Nonparametric Statistics

Criteria	Parametric	Nonparametric
Ease of use		○
Time efficiency		○
Applicability		○
Sensitivity	○	

(Note) ○: preferred

15-2 Sign Test

Learning Goals

- ❑ Read the sign test table.
- ❑ Test a hypothesis on the median ($\tilde{\mu}$) of a continuous distribution by the sign test.
- ❑ Calculate the *P*-value of a sign test statistic.
- ❑ Apply the normal approximation to the sign statistic when n is moderately large.
- ❑ Test a hypothesis on the median of differences ($\tilde{\mu}_D$) for paired samples by the sign test.

Sign Test

The sign test uses the plus and minus signs of the differences between the observations X_i and the hypothesized median $\tilde{\mu}_0$. The sign test assumes that the observations are drawn from a continuous distribution.

The sign test statistics R^-, R^+, and R are determined by the following procedure:

Step 1: Compute the differences $X_i - \mu_0$, $i = 1, 2, \ldots, n$.

Step 2: Count the numbers of the minus and plus signs of the differences

R^- = number of the minus signs

R^+ = number of the plus signs

$R = \min(R^-, R^+)$

Note that ties ($x_i = \tilde{\mu}_0$) are not included in the sign test because the probability that a value of X_i is exactly equal to $\tilde{\mu}_0$ is zero in a continuous distribution:

$$P(X_i = \tilde{\mu}_0) = 0$$

Sign Test Table

The sign test table (see Appendix Table VII in MR) provides critical values ($r^*_{\alpha, n}$) of a sign test statistic for various sample sizes (n) and levels of significance (α).

Sign Test Table (*cont.*)

(e.g.) Reading the sign test table

(1) $r^*_{0.05,10} = 1$

(2) $r^*_{0.025,30} = 9$

(3) $r^*_{0.005,25} = 5$

Inference Context for $\tilde{\mu}$

Parameter of interest: $\tilde{\mu}$ (median)

Test statistic of $\tilde{\mu}$: The appropriate sign statistic is selected depending on the type of H_1 as follows:

(1) $R^- =$ number of the minus signs of $X_i - \tilde{\mu}_0$, $i = 1, 2, \ldots, n$, for H_1: $\tilde{\mu} > \tilde{\mu}_0$

(2) $R^+ =$ number of the plus signs of $X_i - \tilde{\mu}_0$, $i = 1, 2, \ldots, n$, for H_1: $\tilde{\mu} < \tilde{\mu}_0$

(3) $R = \min(R^+, R^-)$ for H_1: $\tilde{\mu} \neq \tilde{\mu}_0$

***P*-value Calculation**

The sign statistics R^-, R^+, and R follow the binomial distribution $B(n, 0.5)$; therefore, the P-value (significance probability) of a sign statistic is

(1) $P = P(R^- \le r^-) = \sum_{x=0}^{r^-} B\left(n, \frac{1}{2}\right) = \sum_{x=0}^{r^-} \binom{n}{x}\left(\frac{1}{2}\right)^n$ for H_1: $\tilde{\mu} > \tilde{\mu}_0$

(2) $P = P(R^+ \le r^+) = \sum_{x=0}^{r^+} B\left(n, \frac{1}{2}\right) = \sum_{x=0}^{r^+} \binom{n}{x}\left(\frac{1}{2}\right)^n$ for H_1: $\tilde{\mu} < \tilde{\mu}_0$

(3) $P = 2 \times P(R \le r) = 2 \times \sum_{x=0}^{r} B\left(n, \frac{1}{2}\right) = 2 \times \sum_{x=0}^{r} \binom{n}{x}\left(\frac{1}{2}\right)^n$ for H_1: $\tilde{\mu} \neq \tilde{\mu}_0$

Recall that the null hypothesis H_0 is rejected if $P \le \alpha$ (see Section 9-1).

Sign Test Procedure for $\tilde{\mu}$

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $\tilde{\mu} = \tilde{\mu}_0$

H_1: $\tilde{\mu} \neq \tilde{\mu}_0$ for two-sided test

$\tilde{\mu} < \tilde{\mu}_0$ or $\tilde{\mu} > \tilde{\mu}_0$ for one-sided test

Step 2: Determine a **test statistic and its value.**

$R^- =$ number of the minus signs of $X_i - \tilde{\mu}_0$ for upper-sided test

$R^+ =$ number of the plus signs of $X_i - \tilde{\mu}_0$ for lower-sided test

$R = \min(R^-, R^+)$ for two-sided test

Step 3: Determine a **critical value(s) for $\boldsymbol{\alpha}$.**

$r^*_{\alpha/2,n}$ for two-sided test; $r^*_{\alpha,n}$ for one-sided test

Step 4: Make a **conclusion**. Reject H_0 if

$r^- \le r^*_{\alpha,n}$ for upper-sided test

$r^+ \le r^*_{\alpha,n}$ for lower-sided test

$r \le r^*_{\alpha/2,n}$ for two-sided test

Normal Approximation

Recall that a binomial distribution $B(n, p)$ approximates to a normal distribution with $\mu = np$ and $\sigma^2 = np(1 - p)$ if $np > 5$ and $n(1 - p) > 5$ (see Section 4-7). Since R^+ (R^- or R) follows $B(n, 0.5)$, if n is moderately large (say, $n > 10$), the sign statistic has approximately a normal distribution with mean and variance

$\mu = 0.5n$ and $\sigma^2 = 0.5^2 n$

Therefore, a hypothesis on the median by R^+ can be tested by using the z-test statistic

$$Z_0 = \frac{R^+ - 0.5n}{0.5\sqrt{n}}$$

Comparison to the t-Test

If the underlying distribution is normal or nonnormal but symmetric, the t-test is preferred to the sign test for better statistical power.

Example 15.1

Test scores of 12 students in a statistics class are as follows:

No.	1	2	3	4	5	6	7	8	9	10	11	12
Score (X)	85	75	96	74	87	71	73	80	75	90	94	86

1. (**Test on $\tilde{\mu}$; Sign Test**) Test if the median of the test scores is equal to 80 at $\alpha = 0.05$.

Step 1: State H_0 and H_1.

H_0: $\tilde{\mu} = 80$

H_1: $\tilde{\mu} \neq 80$

Step 2: Determine a **test statistic and its value.**

No.	1	2	3	4	5	6	7	8	9	10	11	12
x_i	85	75	96	74	87	71	73	80	75	90	94	86
$x_i - 80$	5	-5	16	-6	7	-9	-7	0	-5	10	14	6
Sign	+	-	+	-	+	-	-	*	-	+	+	+

* Excluded since $x_i = \tilde{\mu}_0$.

$r = \min(r^-, r^+) = \min(5, 6) = 5$

Step 3: Determine a **critical value(s) for α.**

$r^*_{\alpha/2,n} = r^*_{0.025,11} = 1$ (Note) number of ties = 1

Step 4: Make a **conclusion.**

Since $r = 5 \nleq r^*_{\alpha/2,n} = 1$, fail to reject H_0 at $\alpha = 0.05$.

2. (***P*-value Calculation**) Find the P-value of the sign statistic $r = 5$ for the two-sided test.

$$P = 2 \times P(R \leq 5) = 2 \times \sum_{x=0}^{5} \binom{11}{x} \frac{1}{2}^{11} \cong 1.0$$

Since $P = 1.0 \nleq \alpha = 0.05$, fail to reject H_0 at $\alpha = 0.05$.

Example 15.1 (*cont.*)

3. (**Normal Approximation**) Apply the normal approximation to the sign statistic R and draw a conclusion at $\alpha = 0.05$.

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $\tilde{\mu} = 80$

H_1: $\tilde{\mu} \neq 80$

Step 2: Determine a **test statistic and its value.**

$$z_0 = \frac{r - 0.5n}{0.5\sqrt{n}} = \frac{5 - 0.5 \times 11}{0.5 \times \sqrt{11}} = -0.30$$

Step 3: Determine a **critical value(s) for** $\boldsymbol{\alpha}$.

$$z_{\alpha/2} = z_{0.05/2} = z_{0.025} = 1.96$$

Step 4: Make a **conclusion.**

Since $|z_0| = 0.30 \ngtr z_{0.025} = 1.96$, fail to reject H_0 at $\alpha = 0.05$.

Exercise 15.1 (MR 15-3)

The impurity level (unit: ppm) is routinely measured in an intermediate chemical product. The following data are observed in an impurity test:

No.	1	2	3	4	5	6	7	8	9	10	11
Impurity (X)	2.4	2.5	1.7	1.6	1.9	2.6	1.3	1.9	2.0	2.5	2.6
No.	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**
Impurity (X)	2.3	2.0	1.8	1.3	1.7	2.0	1.9	2.3	1.9	2.4	1.6

1. Test if the median of the impurity level is <u>less</u> than 2.5 ppm at $\alpha = 0.05$.
2. Find the P-value of the sign statistic $r^+ = 2$ for the lower-sided test.
3. Apply the normal approximation to the sign statistic R^+ and draw a conclusion at $\alpha = 0.05$.

Inference Context for $\tilde{\mu}_D$

Parameter of interest: $\tilde{\mu}_D$ (median of the paired differences $D_j = X_{1j} - X_{2j}$)

Test statistic of $\tilde{\mu}_D$: The appropriate sign statistic is selected depending on the type of H_1 as follows:

(1) $R^- =$ number of the minus signs of $D_j, j = 1, 2, \ldots, n$ for H_1: $\tilde{\mu}_D > \delta_0$

(2) $R^+ =$ number of the plus signs of $D_j, j = 1, 2, \ldots, n$ for H_1: $\tilde{\mu}_D < \delta_0$

(3) $R = \min(R^-, R^+)$ for H_1: $\tilde{\mu}_D \neq \delta_0$

Sign Test Procedure for $\tilde{\mu}_D$

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $\tilde{\mu}_D = \delta_0$

H_1: $\tilde{\mu}_D \neq \delta_0$ for two-sided test

$\tilde{\mu}_D < \delta_0$ or $\tilde{\mu}_D > \delta_0$ for one-sided test

Sign Test Procedure for $\tilde{\mu}_D$ (*cont.*)

Step 2: Determine a **test statistic and its value.**

$R^- =$ number of the minus signs of D_j for upper-sided test

$R^+ =$ number of the plus signs of D_j for lower-sided test

$R = \min(R^-, R^+)$ for two-sided test

Step 3: Determine a **critical value(s) for α.**

$r^*_{\alpha/2,n}$ for two-sided test; $r^*_{\alpha,n}$ for one-sided test

Step 4: Make a **conclusion**. Reject H_0 if

$r^- \le r^*_{\alpha,n}$ for upper-sided test

$r^+ \le r^*_{\alpha,n}$ for lower-sided test

$r \le r^*_{\alpha/2,n}$ for two-sided test

Example 15.2

(Test on $\tilde{\mu}_D$; Sign Test) Scores of 10 students for two exams are below. Test H_0: $\tilde{\mu}_D = 0$ vs. H_1: $\tilde{\mu}_D \neq 0$ at $\alpha = 0.05$.

No.	1	2	3	4	5	6	7	8	9	10
Exam 1 (X_1)	85	75	96	74	87	71	73	80	75	90
Exam 2 (X_2)	90	73	98	80	92	72	78	85	75	92

Step 1: State H_0 and H_1.

H_0: $\tilde{\mu}_D = 0$

H_1: $\tilde{\mu}_D \neq 0$

Step 2: Determine a **test statistic and its value.**

No.	1	2	3	4	5	6	7	8	9	10
X_1	85	75	96	74	87	71	73	80	75	90
X_2	90	73	98	80	92	72	78	85	75	92
$D = X_1 - X_2$	-5	2	-2	-6	-5	-1	-5	-5	0	-2
Sign	-	+	-	-	-	-	-	-	*	-

* Excluded since $d_j = \delta_0$.

$$r = \min(r^-, r^+) = \min(8, 1) = 1$$

Step 3: Determine a **critical value(s) for α.**

$r^*_{\alpha/2,n} = r^*_{0.025,9} = 1$ (Note) number of ties = 1

Step 4: Make a **conclusion.**

Since $r = 1 \le r^*_{\alpha/2,n} = 1$, reject H_0 at $\alpha = 0.05$.

Exercise 15.2 (MR 15-12)

The diameter of a ball bearing was measured by 12 inspectors. The measurements, corresponding differences, and their signs are presented in the table on next page. Test if there is a significant difference between the medians of the populations of measurements by using two different calipers. Use $\alpha = 0.05$.

Exercise 15.2 (*cont.*)

Inspector	Measurement Caliper 1 (X_1)	Caliper 2 (X_2)	Difference ($D = X_1 - X_2$)	Sign (*: tie)
1	0.265	0.264	0.001	+
2	0.265	0.265	0.000	*
3	0.266	0.264	0.002	+
4	0.267	0.266	0.001	+
5	0.267	0.267	0.000	*
6	0.265	0.268	-0.003	-
7	0.267	0.264	0.003	+
8	0.267	0.265	0.002	+
9	0.265	0.265	0.000	*
10	0.268	0.267	0.001	+
11	0.268	0.268	0.000	*
12	0.265	0.269	-0.004	-

15-3 Wilcoxon Signed-Rank Test

Learning Goals

- ❑ Read the Wilcoxon signed-rank test table.
- ❑ Test a hypothesis on the mean (μ) of a continuous distribution by the Wilcoxon signed-rank test.
- ❑ Apply the normal approximation to the signed-rank statistic when n is moderately large.
- ❑ Test a hypothesis on the mean of differences (μ_D) for paired samples by the Wilcoxon signed-rank test.

Wilcoxon Signed-Rank Test

While the sign test uses only the plus and minus signs of the differences between the observations X_i and the hypothesized median $\tilde{\mu}_0$, the Wilcoxon signed-rank test uses both the <u>signs</u> of the differences between the observations X_i and the hypothesized mean μ_0 and the <u>ranks</u> of the absolute magnitudes of the differences. The signed-rank test assumes that the underlying distribution is <u>symmetric and continuous</u>, i.e., the mean equals the median.

The Wilcoxon signed-rank statistics W^-, W^+, and W are determined by the following procedure:

Step 1: Compute the differences $X_i - \mu_0$, $i = 1, 2, \ldots, n$.

Step 2: Rank the absolute differences $|X_i - \mu_0|$ in ascending order. For <u>ties</u>, use the average of the ranks that would be assigned if they differed.

Step 3: Assign the signs of the differences to the corresponding ranks.

Step 4: Calculate the sum of the signed ranks:

W^- = sum of the negative ranks

W^+ = sum of the positive ranks

$W = \min(W^-, W^+)$

Wilcoxon Signed-Rank Test Table

Like the sign test table, the Wilcoxon signed-rank test table (see Appendix Table VIII in MR) provides critical values ($w^*_{\alpha,n}$) of a signed-rank test statistic for various sample sizes (n) and levels of significance (α).

(e.g.) Reading the signed-rank test table

(1) $w^*_{0.05,10} = 10$

(2) $w^*_{0.025,20} = 52$

(3) $w^*_{0.005,15} = 15$

Inference Context for μ

Parameter of interest: μ

Test statistic of μ: The appropriate signed-rank statistic is selected depending on the type of H_1 as follows:

(1) W^- = sum of the <u>negative ranks</u> of $X_i - \mu_0$, $i = 1, 2, \ldots, n$ for H_1: $\mu > \mu_0$

(2) W^+ = sum of the <u>positive ranks</u> of $X_i - \mu_0$, $i = 1, 2, \ldots, n$ for H_1: $\mu < \mu_0$

(3) $W = \min(W^-, W^+)$ for H_1: $\mu \neq \mu_0$

Wilcoxon Signed-Rank Test Procedure for μ

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $\mu = \mu_0$

H_1: $\mu \neq \mu_0$ for two-sided test

$\mu < \mu_0$ or $\mu > \mu_0$ for one-sided test

Step 2: Determine a **test statistic and its value**.

W^- = sum of the negative ranks of $X_i - \mu_0$ for upper-sided test

W^+ = sum of the positive ranks of $X_i - \mu_0$ for lower-sided test

$W = \min(W^-, W^+)$ for two-sided test

Step 3: Determine a **critical value(s) for** $\boldsymbol{\alpha}$.

$w^*_{\alpha/2,n}$ for two-sided test; $w^*_{\alpha,n}$ for one-sided test

Step 4: Make a **conclusion**. Reject H_0 if

$w^- \le w^*_{\alpha,n}$ for upper-sided test

$w^+ \le w^*_{\alpha,n}$ for lower-sided test

$w \le w^*_{\alpha/2,n}$ for two-sided test

Normal Approximation

If the sample size is moderately large (say, $n > 20$), the signed-rank statistic W^+ (W^- or W) has approximately a normal distribution with mean and variance

$$\mu_{W^+} = \frac{n(n+1)}{4} \quad \text{and} \quad \sigma^2_{W^+} = \frac{n(n+1)(2n+1)}{24}$$

Therefore, a hypothesis on the single mean by W^+ can be tested by using the following z-statistic:

Normal Approximation (*cont.*)

$$Z_0 = \frac{W^+ - \dfrac{n(n+1)}{4}}{\sqrt{\dfrac{n(n+1)(2n+1)}{24}}}$$

Example 15.3

Test scores of 10 students in a statistics class are below.

No.	1	2	3	4	5	6	7	8	9	10
Score (X)	85	75	96	74	87	71	73	80	75	90

1. (**Test on μ; Wilcoxon Signed-Rank Test**) Test if the mean of the test scores is equal to 80 at $\alpha = 0.05$.

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $\mu = 80$

H_1: $\mu \neq 80$

Step 2: Determine a **test statistic and its value.**

No.	1	2	3	4	5	6	7	8	9	10
x_i	85	75	96	74	87	71	73	80	75	90
$x_i - 80$	5	-5	16	-6	7	-9	-7	0	-5	10
Sign	+	-	+	-	+	-	-	*	-	+
$\lvert x_i - 80 \rvert$	5	5	16	6	7	9	7	*	5	10
Rank	2	2	9	4	5.5	7	5.5	*	2	8
Signed rank	2	-2	9	-4	5.5	-7	-5.5	*	-2	8

* Excluded since $x_i = \mu_0$.

$$w = \min(w^-, w^+) = \min(20.5, 24.5) = 20.5$$

Step 3: Determine a **critical value(s) for α.**

$$w^*_{\alpha/2,n} = w^*_{0.025,9} = 5 \quad \text{(Note) number of ties} = 1$$

Step 4: Make a **conclusion.**

Since $w = 20.5 \not\leq w^*_{\alpha/2,n} = 5$, fail to reject H_0 at $\alpha = 0.05$.

2. (**Normal Approximation**) Apply the normal approximation to the signed-rank statistic W and draw a conclusion at $\alpha = 0.05$.

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $\mu = 80$

H_1: $\mu \neq 80$

Step 2: Determine a **test statistic and its value.**

$$z_0 = \frac{w - \dfrac{n(n+1)}{4}}{\sqrt{\dfrac{n(n+1)(2n+1)}{24}}} = \frac{20.5 - \dfrac{9 \times (9+1)}{4}}{\sqrt{\dfrac{9 \times (9+1) \times (2 \times 9 + 1)}{24}}} = -0.24$$

Example 15.3 (*cont.*)

Step 3: Determine a **critical value(s) for α.**

$$z_{\alpha/2} = z_{0.05/2} = z_{0.025} = 1.96$$

Step 4: Make a **conclusion.**

Since $|z_0| = 0.24 \ngtr z_{0.025} = 1.96$, fail to reject H_0 at $\alpha = 0.05$.

Exercise 15.3 (MR 15-23)

The impurity data in Exercise 14.1 is presented again as follows:

No.	1	2	3	4	5	6	7	8	9	10	11
Impurity (X)	2.4	2.5	1.7	1.6	1.9	2.6	1.3	1.9	2.0	2.5	2.6
No.	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**
Impurity (X)	2.3	2.0	1.8	1.3	1.7	2.0	1.9	2.3	1.9	2.4	1.6

1. Test if the mean of the impurity level is <u>less</u> than 2.5 ppm at $\alpha = 0.05$.
2. Apply the normal approximation to the signed-rank statistic W^+ and draw a conclusion at $\alpha = 0.05$.

Inference Context for μ_D

Parameter of interest: μ_D (mean of the paired differences $D_j = X_{1j} - X_{2j}$)

Test statistic of μ_D: The appropriate signed-rank statistic is selected depending on the type of H_1 as follows:

(1) W^- = sum of the negative ranks of $D_j, j = 1, 2, \ldots, n$ for H_1: $\tilde{\mu}_D > \delta_0$

(2) W^+ = sum of the positive ranks of $D_j, j = 1, 2, \ldots, n$ for H_1: $\tilde{\mu}_D < \delta_0$

(3) $W = \min(W^-, W^+)$ for H_1: $\tilde{\mu}_D \neq \delta_0$

Wilcoxon Signed-Rank Test Procedure for μ_D

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $\mu_D = \delta_0$

H_1: $\mu_D \neq \delta_0$ for two-sided test

$\mu_D < \delta_0$ or $\mu_D > \delta_0$ for one-sided test

Step 2: Determine a **test statistic and its value.**

W^- = sum of the negative ranks of D_j for upper-sided test

W^+ = sum of the positive ranks of D_j for lower-sided test

$W = \min(W^-, W^+)$ for two-sided test

Step 3: Determine a **critical value(s) for α.**

$w^*_{\alpha/2,n}$ for two-sided test; $w^*_{\alpha,n}$ for one-sided test

Step 4: Make a **conclusion.** Reject H_0 if

$w^- \le w^*_{\alpha,n}$ for upper-sided test

$w^+ \le w^*_{\alpha,n}$ for lower-sided test

$w \le w^*_{\alpha/2,n}$ for two-sided test

Example 15.4

(**Test on μ_D; Wilcoxon Signed-Rank Test**) The exam scores in Exercise 14.2 is presented again below. Test H_0: $\mu_D = 0$ vs. H_1: $\mu_D \neq 0$ at $\alpha = 0.05$.

No.	1	2	3	4	5	6	7	8	9	10
Exam 1 (X_1)	85	75	96	74	87	71	73	80	75	90
Exam 2 (X_2)	90	73	98	80	92	72	78	85	75	92

Step 1: State H_0 and H_1.

H_0: $\mu_D = 0$

H_1: $\mu_D \neq 0$

Step 2: Determine a **test statistic and its value**.

No.	1	2	3	4	5	6	7	8	9	10
X_1	85	75	96	74	87	71	73	80	75	90
X_2	90	73	98	80	92	72	78	85	75	92
$D = X_1 - X_2$	-5	2	-2	-6	-5	-1	-5	-5	0	-2
Sign	-	+	-	-	-	-	-	-	*	-
$\|D\|$	5	2	2	6	5	1	5	5	*	2
Rank	6.5	3	3	9	6.5	1	6.5	6.5	*	3
Signed rank	-6.5	3	-3	-9	-6.5	-1	-6.5	-6.5	*	-3

* Excluded since $d_j = \delta_0$.

$$w = \min(w^-, w^+) = \min(42, 3) = 3$$

Step 3: Determine a **critical value(s) for α**.

$$w^*_{\alpha/2,n} = w^*_{0.025,9} = 5 \quad \text{(Note) number of ties} = 1$$

Step 4: Make a **conclusion**.

Since $w = 3 \leq w^*_{\alpha/2,n} = 5$, reject H_0 at $\alpha = 0.05$.

Exercise 15.4 (MR 15-26)

Reconsider the ball bearing diameter data in Exercise 15.2. The measurements, corresponding differences, and their signed ranks are presented below. Test H_0: $\mu_D = 0$ vs. H_1: $\mu_D \neq 0$ at $\alpha = 0.05$.

Car	Measurement X_1	Measurement X_2	D ($X_1 - X_2$)	Sign	$\|D\|$	Rank	Signed Rank
1	0.265	0.264	0.001	+	0.001	2.0	2.0
2	0.265	0.265	0.000	*	*	*	*
3	0.266	0.264	0.002	+	0.002	4.5	4.5
4	0.267	0.266	0.001	+	0.001	2.0	2.0
5	0.267	0.267	0.000	*	*	*	*
6	0.265	0.268	-0.003	-	0.003	6.5	-6.5
7	0.267	0.264	0.003	+	0.003	6.5	6.5
8	0.267	0.265	0.002	+	0.002	4.5	4.5
9	0.265	0.265	0.000	*	*	*	*
10	0.268	0.267	0.001	+	0.001	2.0	2.0
11	0.268	0.268	0.000	*	*	*	*
12	0.265	0.269	-0.004	-	0.004	8.0	-8.0

* ties

15-4 Wilcoxon Rank-Sum Test

Learning Goals

- ☐ Read the Wilcoxon rank-sum test table.
- ☐ Test a hypothesis on the difference in mean of two continuous populations (μ_1 - μ_2) by the Wilcoxon rank-sum test.
- ☐ Apply the normal approximation to the rank-sum statistic when n_1 and n_2 are moderately large.

Wilcoxon Rank-Sum Test

The Wilcoxon rank-sum test (often called Mann-Whitney test) uses the ranks of the observations from two independent continuous populations X_1 and X_2 to test the difference in mean (μ_1 - μ_2).

Let $X_{11}, X_{12}, \dots, X_{1n_1}$ and $X_{21}, X_{22}, \dots, X_{2n_2}$ denote two independent random samples of sizes n_1 and n_2 ($n_1 \le n_2$) from the populations X_1 and X_2. Then, the Wilcoxon rank-sum statistics W_1 and W_2 are determined by the following procedure:

Step 1: Rank the $n_1 + n_2$ observations in ascending order. For ties, use the average of the ranks that would be assigned if they differed.

Step 2: Calculate the sum of the ranks for each sample:

W_1 = sum of the ranks in the smaller sample (size = n_1)

W_2 = sum of the ranks in the larger sample (size = n_2)

Note that W_1 and W_2 will be nearly equal for both the samples after adjusting for the difference in sample size.

Wilcoxon Rank-Sum Test Table

The Wilcoxon rank-sum test table (see Appendix Table IX in MR) provides critical values (w^*_{α,n_1,n_2}) of a rank-sum statistic for various sample sizes (n_1 and n_2 where $n_1 \le n_2$) and levels of significance (α = 0.05 and 0.01) for a two-sided test.

(e.g.) Reading the rank-sum test table for a two-sided test

(1) $w^*_{0.05/2,10,15} = w^*_{0.025,10,15} = 94$

(2) $w^*_{0.01/2,5,10} = w^*_{0.005,5,10} = 19$

Inference Context for μ_1 - μ_2

Parameter of interest: $\mu_1 - \mu_2$

Test statistic of $\mu_1 - \mu_2$: The appropriate rank-sum statistic is selected depending on the type of H_1 as follows:

(1) W_1 = sum of the ranks of $X_{1j}, j = 1, 2, \dots, n_1$ for H_1: $\mu_1 - \mu_2 < \delta_0$

(2) W_2 = sum of the ranks of $X_{2j}, j = 1, 2, \dots, n_2$ for H_1: $\mu_1 - \mu_2 > \delta_0$

(3) W_1 or W_2 for H_1: $\mu_1 - \mu_2 = \delta_0$

Wilcoxon Rank-Sum Test Procedure for μ_1 - μ_2

Step 1: State H_0 and H_1.

H_0: $\mu_1 - \mu_2 = \delta_0$

H_1: $\mu_1 - \mu_2 \neq \delta_0$ for two-sided test

$\mu_1 - \mu_2 > \delta_0$ or $\mu_1 - \mu_2 < \delta_0$ for one-sided test

Wilcoxon Rank-Sum Test Procedure (*cont.*)

Step 2: Determine a **test statistic and its value.**

W_1 = sum of the ranks of X_{1j} for lower-sided test

W_2 = sum of the ranks of X_{2j} for upper-sided test

W_1 or W_2 for two-sided test

Step 3: Determine a **critical value(s) for α.**

$w^*_{\alpha/2,n_1,n_2}$ for two-sided test; w^*_{α,n_1,n_2} for one-sided test

Step 4: Make a **conclusion**. Reject H_0 if

$w_1 \le w^*_{\alpha,n_1,n_2}$ for lower-sided test

$w_2 \le w^*_{\alpha,n_1,n_2}$ for upper-sided test

$w_1 \le w^*_{\alpha/2,n_1,n_2}$ or $w_2 \le w^*_{\alpha/2,n_1,n_2}$ for two-sided test

Normal Approximation

When n_1 and n_2 are moderately large (say, n_1 and $n_2 > 8$), the rank-sum statistic W_1 (or W_2) has approximately a normal distribution with mean and variance

$$\mu_{W_1} = \frac{n_1(n_1+n_2+1)}{2} \quad \text{and} \quad \sigma^2_{W_1} = \frac{n_1 n_2(n_1+n_2+1)}{12}$$

Therefore, a hypothesis on the difference in mean by W_1 can be tested by using the z-statistic

$$Z_0 = \frac{W_1 - \mu_{W_1}}{\sigma_{W_1}} = \frac{W_1 - \dfrac{n_1(n_1+n_2+1)}{2}}{\sqrt{\dfrac{n_1 n_2(n_1+n_2+1)}{12}}}$$

Example 15.5

The life lengths (unit: hour) of INFINITY (X_1) and FOREVER (X_2) light bulbs are under study. Suppose that no information is available about the underlying distributions of X_1 and X_2. A random sample selected from each brand results in the following:

No.	1	2	3	4	5	6	7	8	9	10
INFINITY (X_1)	760	810	775	680	730	725	740	700	720	-
FOREVER (X_2)	795	780	820	770	810	765	800	790	750	730

1. (**Test on $\mu_1 - \mu_2$; Wilcoxon Rank-Sum Test**) Test if the mean life lengths of the two brands are the same at $\alpha = 0.05$.

Step 1: State H_0 and H_1.

H_0: $\mu_1 - \mu_2 = 0$

H_1: $\mu_1 - \mu_2 \neq 0$

Step 2: Determine a **test statistic and its value.**

No.	1	2	3	4	5	6	7	8	9	10
X_1	760	810	775	680	730	725	740	700	720	-
Rank	9	17.5	12	1	5.5	4	7	2	3	-
X_2	795	780	820	770	810	765	800	790	750	730
Rank	15	13	19	11	17.5	10	16	14	8	5.5

$w_1 = 61$ and $w_2 = 129$

Example 15.5 (*cont.*)

Step 3: Determine a **critical value(s) for α.**

$$w^*_{\alpha/2,n_1,n_2} = w^*_{0.05/2,9,10} = 65$$

Step 4: Make a **conclusion.**

Since $w_1 = 61 \le w^*_{\alpha/2,9,10} = 65$, reject H_0 at $\alpha = 0.05$.

2. (**Normal Approximation**) Apply the normal approximation to the rank-sum statistic W_1 and draw a conclusion at $\alpha = 0.05$.

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $\mu_1 - \mu_2 = 0$

H_1: $\mu_1 - \mu_2 \neq 0$

Step 2: Determine a **test statistic and its value.**

$$z_0 = \frac{w_1 - \dfrac{n_1(n_1+n_2+1)}{2}}{\sqrt{\dfrac{n_1 n_2 (n_1+n_2+1)}{12}}} = \frac{61 - \dfrac{9\times(9+10+1)}{2}}{\sqrt{\dfrac{9\times 10\times(9+10+1)}{12}}} = -2.37$$

Step 3: Determine a **critical value(s) for α.**

$$z_{\alpha/2} = z_{0.025} = 1.96$$

Step 4: Make a **conclusion.**

Since $|z_0| = 2.37 > z_{0.025} = 1.96$, reject H_0 at $\alpha = 0.05$.

Exercise 15.5 (MR 15-30)

The manufacturer of a hot tub is interested in testing two different types of heating elements. The heating element that produces a larger heat gain (unit: °F) after 15 minutes will be preferable. Ten observations of the heat gain for each heating element are obtained as follows:

No.	1	2	3	4	5	6	7	8	9	10
Element 1 (X_1)	25	27	29	31	30	26	24	32	33	38
Element 2 (X_2)	31	33	32	35	34	29	38	35	37	30

1. Test if one heating element is superior to the other at $\alpha = 0.05$.
2. Apply the normal approximation to the rank-sum statistic W_1 and draw a conclusion at $\alpha = 0.05$.

15-5 Nonparametric Methods in the Analysis of Variance

Learning Goals

- ❑ Perform the Kruskal-Wallis test in the single-factor analysis of variance (ANOVA).
- ❑ Explain use of ranks in ANOVA.

Kruskal-Wallis Test

The Kruskal-Wallis test is a nonparametric alternative to the parametric F-test of the single-factor analysis of variance (ANOVA). While the parametric ANOVA model, $Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$, $i = 1, 2, \ldots, a$ and $j = 1, 2, \ldots, n_i$, assumes the error

Kruskal-Wallis Test (*cont.*)

terms ε_{ij} are normally distributed with mean zero and variance σ^2, the nonparametric ANOVA method assumes that the ε_{ij} have the same continuous distribution for all the factor levels $i = 1, 2, \ldots, a$.

The Kruskal-Wallis statistic H is determined by the following procedure:

Step 1: Rank all $N\ (=\sum_{i=1}^{a} n_i\ )$ observations in ascending order. For ties, use the average of the ranks that would be assigned if they differed.

Step 2: Calculate the sum ($R_{i.}$) of the ranks R_{ij} in the i^{th} treatment, $i = 1, 2, \ldots, a$.

Step 3: Calculate the Kruskal-Wallis statistic

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{a} \frac{R_{i.}^{\ 2}}{n_i} - 3(N+1) \quad \text{for observations without ties}$$

$$= \frac{1}{S^2}\left[\sum_{i=1}^{a} \frac{R_{i.}^{\ 2}}{n_i} - \frac{N(N+1)^2}{4}\right] \quad \text{for observations with ties}$$

$$\text{where } S^2 = \frac{1}{N-1}\left[\sum_{i=1}^{a}\sum_{j=1}^{n_i} R_{ij}^{\ 2} - \frac{N(N+1)^2}{4}\right]$$

χ^2 Approximation

The Kruskal-Wallis static approximates to a χ^2-distribution with $a - 1$ degrees of freedom when the sample sizes n_i are moderately large ($n_i \geq 6$ for $a = 3$; $n_i \geq 5$ for $a > 3$).

Inference Context

Parameters of interest: $\tau_1, \tau_2, \ldots, \tau_a$

Test statistic of $\tau_1, \tau_2, \ldots, \tau_a$

(1) **Case 1: Observations with ties**

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{a} \frac{R_{i.}^{\ 2}}{n_i} - 3(N+1)$$

(2) **Case 2: Observations without ties**

$$H = \frac{1}{S^2}\left[\sum_{i=1}^{a} \frac{R_{i.}^{\ 2}}{n_i} - \frac{N(N+1)^2}{4}\right]$$

Kruskal-Wallis Test Procedure

Step 1: State H_0 and H_1.

H_0: $\tau_1 = \tau_2 = \cdots = \tau_a = 0 \quad (\mu_1 = \mu_2 = \cdots = \mu_a)$

H_1: $\tau_i \neq 0$ for at least one i, $i = 1, 2, \ldots, a$

Step 2: Determine a **test statistic and its value.**

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{a} \frac{R_{i.}^{\ 2}}{n_i} - 3(N+1) \quad \text{for observations without ties}$$

$$= \frac{1}{S^2}\left[\sum_{i=1}^{a} \frac{R_{i.}^{\ 2}}{n_i} - \frac{N(N+1)^2}{4}\right] \quad \text{for observations with ties}$$

$$\text{where } S^2 = \frac{1}{N-1}\left[\sum_{i=1}^{a}\sum_{j=1}^{n_i} R_{ij}^{\ 2} - \frac{N(N+1)^2}{4}\right]$$

Kruskal-Wallis Test Procedure (*cont.*)

Step 3: Determine a **critical value(s) for α.**

$$\chi^2_{\alpha,a-1}$$

Step 4: Make a **conclusion**. Reject H_0 if

$$h \geq \chi^2_{\alpha,a-1}$$

***F*-test on Ranks**

If the ordinary F-test were to be applied to the ranks instead of the original data in the single-factor ANOVA, the F-test statistic would be

$$F_0 = \frac{H/(a-1)}{(N-1-H)/(N-a)}$$

Note that F_0 is closely related to the Kruskal-Wallis statistic H, which indicates that the Kruskal-Wallis test is approximately equivalent to the parametric ANOVA to the ranks. Accordingly, the F-test on the ranks can be applied when a nonparametric alternative to the analysis of variance is unavailable.

It is recommended that the F-test be performed on both the original data and the ranks when the normality assumption and/or the effect of an outlier(s) are concerned. If both the test results are similar, use of the original data is satisfactory; otherwise, use of the ranks is preferred.

Example 15.6

(**Nonparametric ANOVA; Kruskal-Wallis Test**) Maximum power grip forces (unit: lb.) of a 45-year-old male at four elbow angles (0°, 45°, 90°, and 135°) are obtained below. Test if elbow angle has a significant effect on maximum power grip strength at $\alpha = 0.05$.

Elbow angle	Maximum grip force (lb.)				
0° (Y_1)	65	68	62	66	69
45° (Y_2)	70	72	69	75	68
90° (Y_3)	67	63	68	64	67
135° (Y_4)	58	60	59	63	60

Step 1: State H_0 and H_1.

H_0: $\tau_1 = \tau_2 = \cdots = \tau_a = 0$

H_1: $\tau_i \neq 0$ for at least one i, i = 1, 2, 3, 4

Step 2: Determine a **test statistic and its value.**

Elbow angle	Maximum grip force (lb.)					$R_{i.}$
0° (y_{1j})	65	68	62	66	69	
Rank (r_{1j})	9	14	5	10	16.5	54.5
45° (y_{2j})	70	72	69	75	68	
Rank (r_{2j})	18	19	16.5	20	14	87.5
90° (y_{3j})	67	63	68	64	67	
Rank (r_{3j})	11.5	6.5	14	8	11.5	51.5
135° (y_{4j})	58	60	59	63	60	
Rank (r_{4j})	1	3.5	2	6.5	3.5	16.5

Example 15.6 (*cont.*)

$$s^2 = \frac{1}{N-1}\left[\sum_{i=1}^{a}\sum_{j=1}^{n_i} r_{ij}^{\;2} - \frac{N(N+1)^2}{4}\right]$$

$$= \frac{1}{20-1}\left[2{,}866 - \frac{20\times(20+1)^2}{4}\right] = 34.79$$

$$h = \frac{1}{s^2}\left[\sum_{i=1}^{a}\frac{r_{i.}^{\;2}}{n_i} - \frac{N(N+1)^2}{4}\right]$$

$$= \frac{1}{34.79}\left[\left(\frac{54.5^2}{5}+\frac{87.5^2}{5}+\frac{51.5^2}{5}+\frac{16.5^2}{5}\right) - \frac{20\times(20+1)^2}{4}\right]$$

$$= 14.52$$

Step 3: Determine a **critical value(s) for α.**

$$\chi^2_{\alpha,a-1} = \chi^2_{0.05,4-1} = \chi^2_{0.05,3} = 7.81$$

Step 4: Make a **conclusion.**

Since $h = 14.52 > \chi^2_{\alpha,a-1} = 7.81$, reject H_0 at $\alpha = 0.05$.

Exercise 15.6 (MR 15-37)

The tensile strengths (unit: lb./in.^{2}) of cements prepared by four different mixing techniques (A, B, C, and D) are measured below. Test if the difference in mixing technique affects the tensile strength at $\alpha = 0.05$.

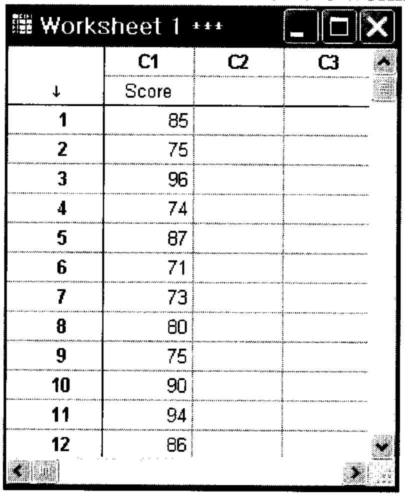
Mixing technique	Tensile strength (lb./in.^{2})				
A (Y_1)	3,129	3,000	2,865	2,890	2,971
B (Y_2)	3,200	3,000	2,975	3,150	3,081
C (Y_3)	2,800	2,900	2,985	3,050	2,934
D (Y_4)	2,600	2,700	2,600	2,765	2,666

MINITAB Applications

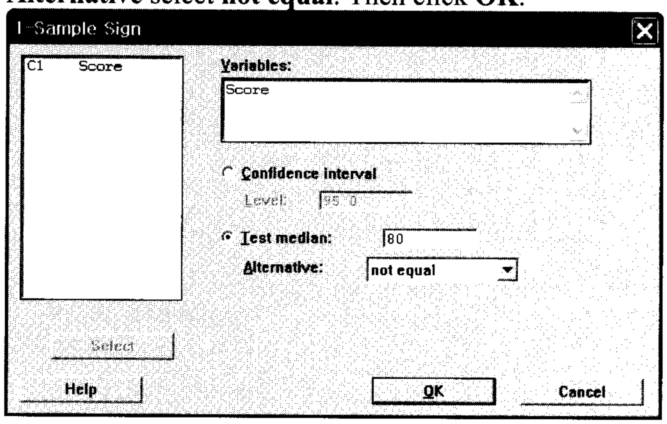
Example 15.1 **(Inference on $\tilde{\mu}$; Sign Test)**

(1) Choose **File ➤ New**, click **Minitab Project**, and click **OK**.

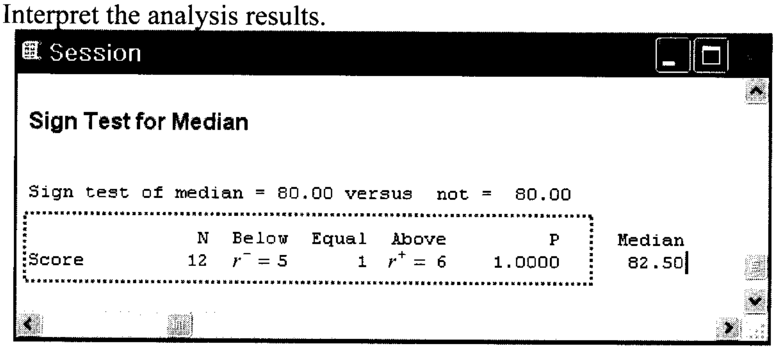
(2) Enter the test score data on the worksheet.



(3) Choose **Stat ➤ Nonparametrics ➤ 1-Sample Sign**. In **Variables** select *Score*. Check **Test median**, enter '80' (hypothetical median), and in **Alternative** select **not equal**. Then click **OK**.



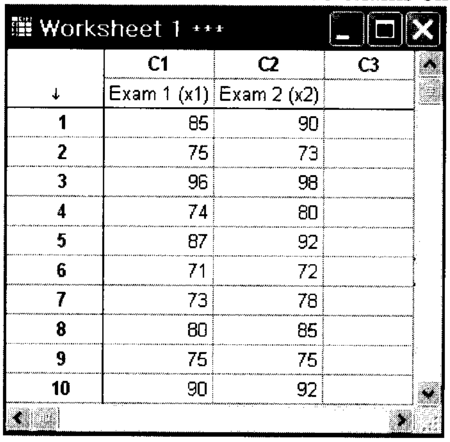
(4) Interpret the analysis results.



Example 15.2 **(Inference on $\tilde{\mu}_D$; Sign Test)**

(1) Choose **File ➤ New**, click **Minitab Project**, and click **OK**.

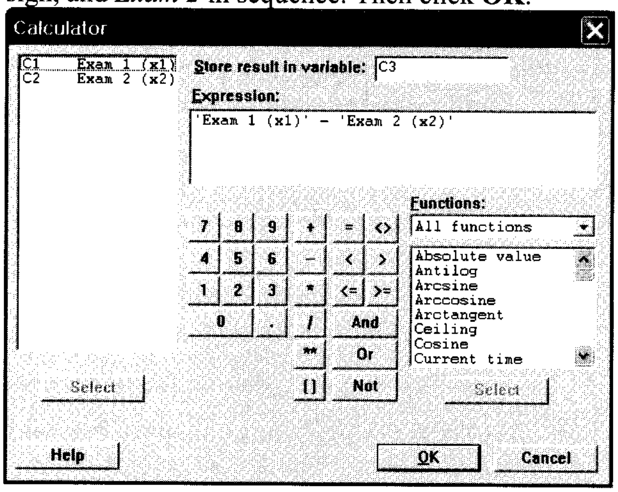
(2) Enter the score data of the two exams on the worksheet.



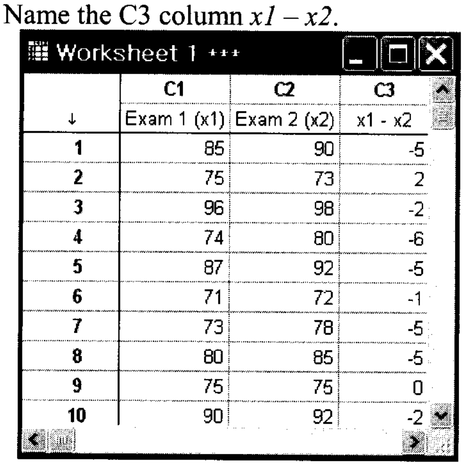
Worksheet 1 ***

	C1	C2	C3
↓	Exam 1 (x1)	Exam 2 (x2)	
1	85	90	
2	75	73	
3	96	98	
4	74	80	
5	87	92	
6	71	72	
7	73	78	
8	80	85	
9	75	75	
10	90	92	

(3) Choose **Calc ➤ Calculator**. In **Store result in variable**, enter *C3* as a column to store calculation results. In **Expression**, click *Exam 1*, minus sign, and *Exam 2* in sequence. Then click **OK**.



(4) Name the C3 column $x1 - x2$.

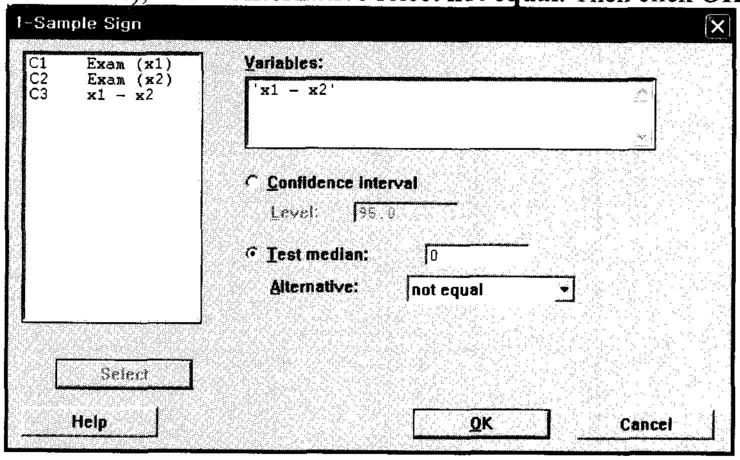


Worksheet 1 ***

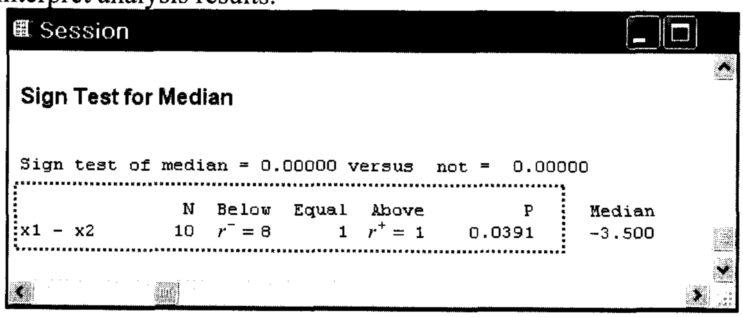
	C1	C2	C3
↓	Exam 1 (x1)	Exam 2 (x2)	x1 - x2
1	85	90	-5
2	75	73	2
3	96	98	-2
4	74	80	-6
5	87	92	-5
6	71	72	-1
7	73	78	-5
8	80	85	-5
9	75	75	0
10	90	92	-2

Example 15.2 (*cont.*)

(5) Choose **Stat ➤ Nonparametrics ➤ 1-Sample Sign**. In **Variables** select *x1 – x2*. Check **Test median**, enter '0' (hypothetical median of the score differences), and in **Alternative** select **not equal**. Then click **OK**.



(6) Interpret analysis results.



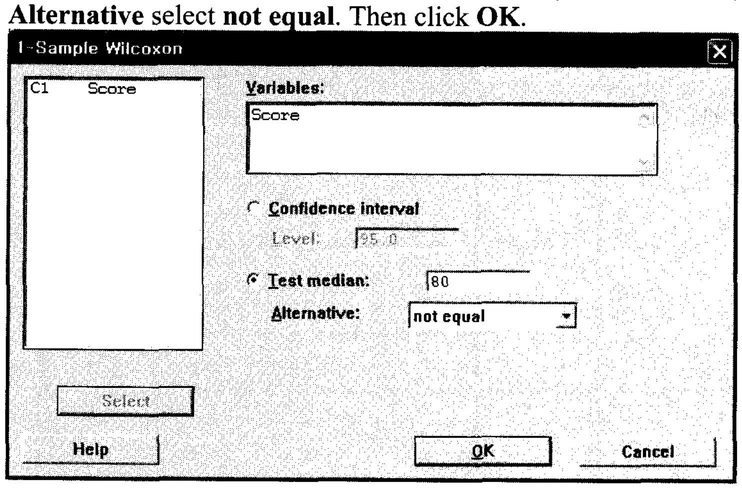
Example 15.3

(Inference on μ; Wilcoxon Signed-Rank Test)

(1) Choose **File ➤ New**, click **Minitab Project**, and click **OK**.

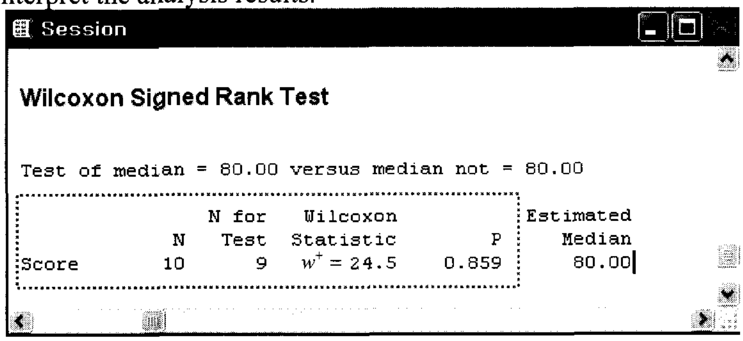
(2) Enter the test score data on the worksheet.

(3) Choose **Stat ➤ Nonparametrics ➤ 1-Sample Wilcoxon**. In **Variables** select *Score*. Check **Test median**, enter '80' (hypothetical mean), and in **Alternative** select **not equal**. Then click **OK**.



Example 15.3 (*cont.*)

(4) Interpret the analysis results.



Example 15.4 **(Inference on μ_D; Wilcoxon Signed-Rank Test)**

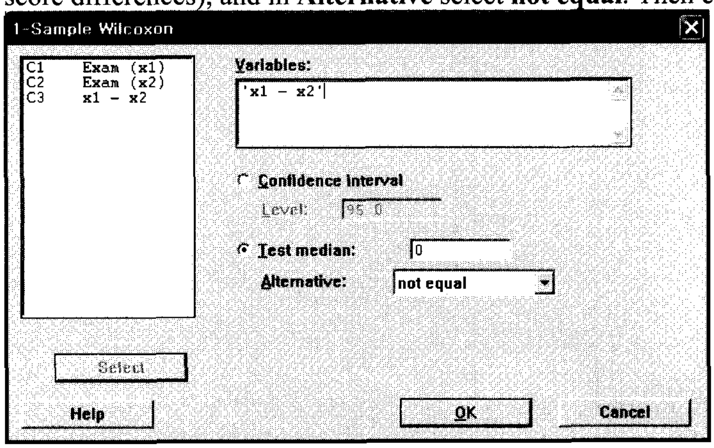
(1) Choose **File ➤ New**, click **Minitab Project**, and click **OK**.

(2) Enter the score data of the two exams on the worksheet.

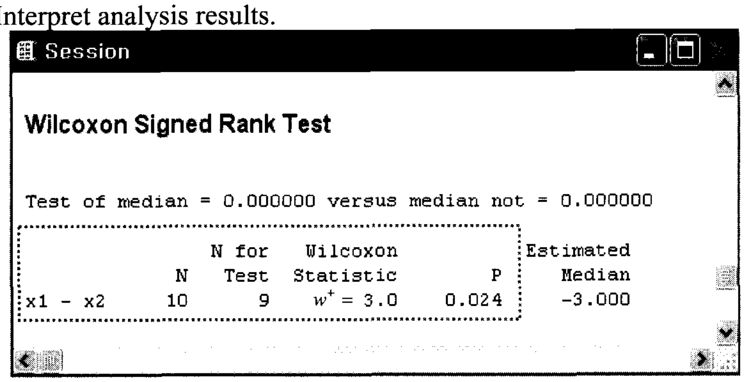
(3) Choose **Calc ➤ Calculator**. In **Store result in variable**, enter *C3* as a column to store calculation results. In **Expression**, click *Exam 1*, minus sign, and *Exam 2* in sequence. Then click **OK**.

(4) Name the C3 column $x1 - x2$.

(5) Choose **Stat ➤ Nonparametrics ➤ 1-Sample Wilcoxon**. In **Variables** select $x1 - x2$. Check **Test median**, enter '0' (hypothetical mean of the score differences), and in **Alternative** select **not equal**. Then click **OK**.



(6) Interpret analysis results.



Example 15.5 **(Inference on μ_1 - μ_2; Wilcoxon Rank-Sum Test)**

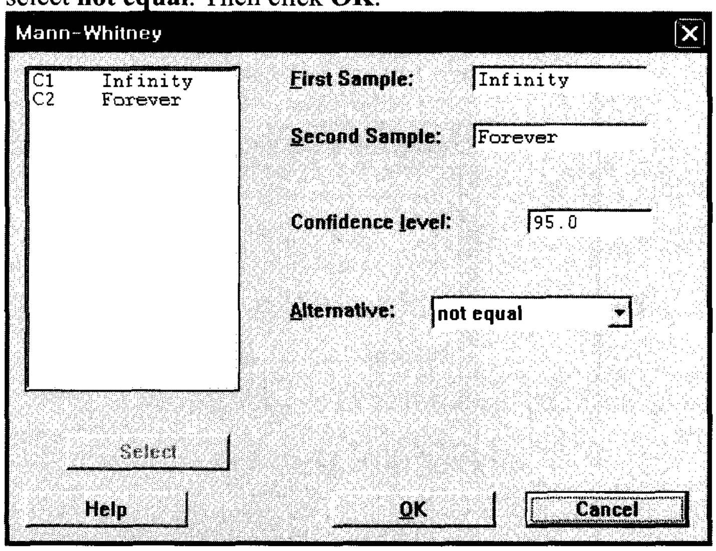
(1) Choose **File ➤ New**, click **Minitab Project**, and click **OK**.

(2) Enter the life length data of the two light bulb brands on the worksheet.

Worksheet 1 ***

↓	C1 Infinity	C2 Forever	C3
1	760	795	
2	810	780	
3	775	820	
4	680	770	
5	730	810	
6	725	765	
7	740	800	
8	700	790	
9	720	750	
10		730	

(3) Choose **Stat ➤ Nonparametrics ➤ Mann-Whitney**. In **First Sample** select *Infinity*, in **Second Sample** select *Forever*, and in **Alternative** select **not equal**. Then click **OK**.



(4) Interpret the analysis results.

Session

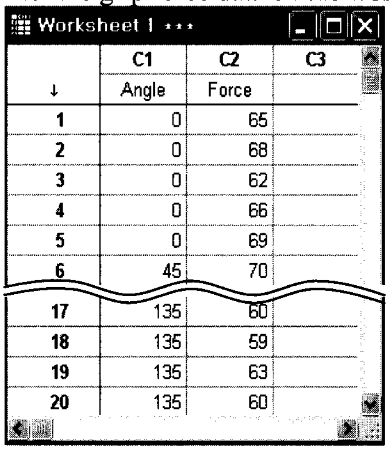
Mann-Whitney Confidence Interval and Test

```
Infinity   N =   9     Median =      730.00
Forever    N =  10     Median =      785.00
Point estimate for ETA1-ETA2 is      -47.50
95.5 Percent CI for ETA1-ETA2 is (-79.99,-10.01)
W = 61.0 = w1
Test of ETA1 = ETA2  vs  ETA1 not = ETA2 is significant at 0.0200
The test is significant at 0.0199 (adjusted for ties)
```

Example 15.6 **(Nonparametric ANO VA; Kruskal-Wallis Test)**

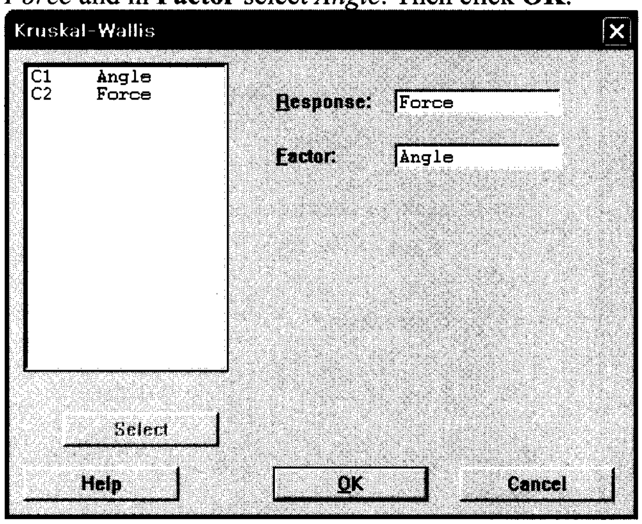
(1) Choose **File ➤ New**, click **Minitab Project**, and click **OK**.

(2) Enter the grip force data on the worksheet.

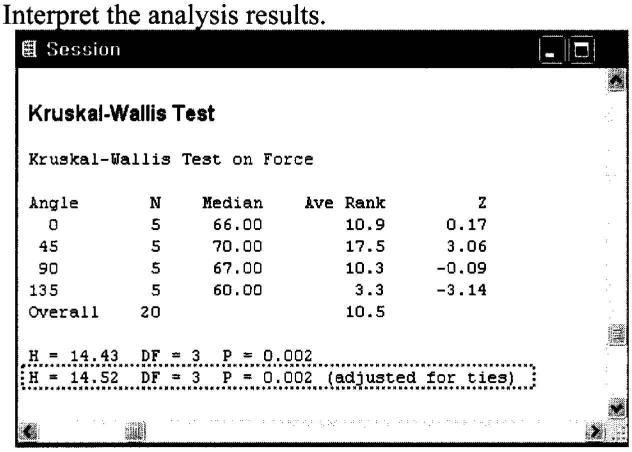

Worksheet 1 ***

	C1	C2	C3
↓	Angle	Force	
1	0	65	
2	0	68	
3	0	62	
4	0	66	
5	0	69	
6	45	70	
17	135	60	
18	135	59	
19	135	63	
20	135	60	

(3) Choose **Stat ➤ Nonparametrics ➤ Kruskal-Wallis**. In **Response** select *Force* and in **Factor** select *Angle*. Then click **OK**.



(4) Interpret the analysis results.


Session

Kruskal-Wallis Test

Kruskal-Wallis Test on Force

```
Angle      N   Median   Ave Rank       Z
  0        5    66.00       10.9    0.17
 45        5    70.00       17.5    3.06
 90        5    67.00       10.3   -0.09
135        5    60.00        3.3   -3.14
Overall   20                10.5

H = 14.43  DF = 3  P = 0.002
H = 14.52  DF = 3  P = 0.002 (adjusted for ties)
```

Answers to Exercises

Exercise 15.1

1. **(Test on $\tilde{\mu}$; Sign Test)**

Step 1: State H_0 and H_1.

H_0: $\tilde{\mu} = 2.5$

H_1: $\tilde{\mu} < 2.5$

Step 2: Determine a **test statistic and its value.**

No.	1	2	3	4	5	6	7	8	9	10	11
x_i	2.4	2.5	1.7	1.6	1.9	2.6	1.3	1.9	2.0	2.5	2.6
$x_i - 2.5$	-0.1	0.0	-0.8	-0.9	-0.6	0.1	-1.2	-0.6	-0.5	0.0	0.1
Sign	-	*	-	-	-	+	-	-	-	*	+
No.	12	13	14	15	16	17	18	19	20	21	22
x_i	2.3	2.0	1.8	1.3	1.7	2.0	1.9	2.3	1.9	2.4	1.6
$x_i - 2.5$	-0.2	-0.5	-0.7	-1.2	-0.8	-0.5	-0.6	-0.2	-0.6	-0.1	-0.9
Sign	-	-	-	-	-	-	-	-	-	-	-

* Excluded since $x_i = \tilde{\mu}_0$.

$r^+ = 2$

Step 3: Determine a **critical value(s) for α.**

$r^*_{\alpha,n} = r^*_{0.05,20} = 5$ (Note) number of ties = 2

Step 4: Make a **conclusion.**

Since $r^+ = 2 \le r^*_{\alpha,n} = 5$, reject H_0 at $\alpha = 0.05$.

2. **(*P*-value Calculation)**

$$P = P(R^+ \le 2) = \sum_{x=0}^{2} \binom{20}{x} \frac{1}{2}^{20} = 0.0002$$

Since $P = 0.0002 < \alpha = 0.05$, reject H_0 at $\alpha = 0.05$.

3. **(Normal Approximation)**

Step 1: State H_0 and H_1.

H_0: $\tilde{\mu} = 2.5$

H_1: $\tilde{\mu} < 2.5$

Step 2: Determine a **test statistic and its value.**

$$z_0 = \frac{r^+ - 0.5n}{0.5\sqrt{n}} = \frac{2 - 0.5 \times 20}{0.5 \times \sqrt{20}} = -3.58$$

Step 3: Determine a **critical value(s) for α.**

$z_\alpha = z_{0.05} = z_{0.05} = 1.65$

Step 4: Make a **conclusion.**

Since $z_0 = -3.58 < -z_{0.05} = -1.65$, reject H_0 at $\alpha = 0.05$.

Exercise 15.2

(Test on $\widetilde{\mu}_D$; Sign Test)

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $\widetilde{\mu}_D = 0$

H_1: $\widetilde{\mu}_D \neq 0$

Step 2: Determine a **test statistic and its value.**

$r = \min(r^-, r^+) = \min(2, 6) = 2$

Step 3: Determine a **critical value(s) for α.**

$r^*_{\alpha/2,n} = r^*_{0.025,8} = 0$ (Note) number of ties = 4

Step 4: Make a **conclusion.**

Since $r = 2 \not\leq r^*_{\alpha/2,n} = 0$, fail to reject H_0 at $\alpha = 0.05$.

Exercise 15.3

1. **(Test on μ; Wilcoxon Signed-Rank Test)**

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $\mu = 2.5$

H_1: $\mu < 2.5$

Step 2: Determine a **test statistic and its value.**

No	x_i	$x_i - 2.5$	Sign	$\lvert x_i - 2.5\rvert$	Rank	Signed rank
1	2.4	-0.1	-	0.1	2.5	-2.5
2	2.5	0.0	*	*	*	*
3	1.7	-0.8	-	0.8	15.5	-15.5
4	1.6	-0.9	-	0.9	17.5	-17.5
5	1.9	-0.6	-	0.6	11.5	-11.5
6	2.6	0.1	+	0.1	2.5	2.5
7	1.3	-1.2	-	1.2	19.5	-19.5
8	1.9	-0.6	-	0.6	11.5	-11.5
9	2.0	-0.5	-	0.5	8.0	-8.0
10	2.5	0.0	*	*	*	*
11	2.6	0.1	+	0.1	2.5	2.5
12	2.3	-0.2	-	0.2	5.5	-5.5
13	2.0	-0.5	-	0.5	8.0	-8.0
14	1.8	-0.7	-	0.7	14.0	-14.0
15	1.3	-1.2	-	1.2	19.5	-19.5
16	1.7	-0.8	-	0.8	15.5	-15.5
17	2.0	-0.5	-	0.5	8.0	-8.0
18	1.9	-0.6	-	0.6	11.5	-11.5
19	2.3	-0.2	-	0.2	5.5	-5.5
20	1.9	-0.6	-	0.6	11.5	-11.5
21	2.4	-0.1	-	0.1	2.5	-2.5
22	1.6	-0.9	-	0.9	17.5	-17.5

* Excluded since $x_i = \mu_0$.

$w^+ = 5$

Exercise 15.3 (*cont.*)

Step 3: Determine a **critical value(s) for α.**

$w^*_{\alpha,n} = w^*_{0.05,20} = 60$ (Note) number of ties = 2

Step 4: Make a **conclusion.**

Since $w^+ = 5 \le w^*_{\alpha,n} = 60$, reject H_0 at $\alpha = 0.05$.

2. (**Normal Approximation**)

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $\mu = 2.5$

H_1: $\mu < 2.5$

Step 2: Determine a **test statistic and its value.**

$$z_0 = \frac{w^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} = \frac{5 - \frac{20\times(20+1)}{4}}{\sqrt{\frac{20\times(20+1)\times(2\times 20+1)}{24}}} = -3.73$$

Step 3: Determine a **critical value(s) for α.**

$z_\alpha = z_{0.05} = 1.65$

Step 4: Make a **conclusion.**

Since $z_0 = -3.73 < -z_{0.05} = -1.65$, reject H_0 at $\alpha = 0.05$.

Exercise 15.4

(**Test on μ_D; Wilcoxon Signed-Rank Test**)

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $\mu_D = 0$

H_1: $\mu_D \ne 0$

Step 2: Determine a **test statistic and its value.**

$w = \min(w^-, w^+) = \min(14.5, 21.5) = 14.5$

Step 3: Determine a **critical value(s) for α.**

$w^*_{\alpha/2,n} = w^*_{0.025,8} = 3$ (Note) number of ties = 4

Step 4: Make a **conclusion.**

Since $w = 14.5 \not\le w^*_{\alpha/2,n} = 3$, fail to reject H_0 at $\alpha = 0.05$.

Exercise 15.5

1. (**Test on μ_1 - μ_2; Wilcoxon Rank-Sum Test**)

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $\mu_1 - \mu_2 = 0$

H_1: $\mu_1 - \mu_2 \ne 0$

Exercise 15.5 (*cont.*)

Step 2: Determine a **test statistic and its value.**

No.	1	2	3	4	5	6	7	8	9	10
X_1	25	27	29	31	30	26	24	32	33	38
Rank	2	4	5.5	9.5	7.5	3	1	11.5	13.5	19.5
X_2	31	33	32	35	34	29	38	35	37	30
Rank	9.5	13.5	11.5	16.5	15	5.5	19.5	16.5	18	7.5

$w_1 = 77$ and $w_2 = 133$

Step 3: Determine a **critical value(s) for α.**

$w^*_{\alpha/2,n_1,n_2} = w^*_{0.05/2,10,10} = 78$ for two-sided test

Step 4: Make a **conclusion.**

Since $w_1 = 77 \le w^*_{\alpha/2,10,10} = 78$, reject H_0 at $\alpha = 0.05$.

2. **(Normal Approximation)**

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $\mu_1 - \mu_2 = 0$

H_1: $\mu_1 - \mu_2 \neq 0$

Step 2: Determine a **test statistic and its value.**

$$z_0 = \frac{w_1 - \dfrac{n_1(n_1+n_2+1)}{2}}{\sqrt{\dfrac{n_1 n_2 (n_1+n_2+1)}{12}}} = \frac{77 - \dfrac{10\times(10+10+1)}{2}}{\sqrt{\dfrac{10\times 10\times(10+10+1)}{12}}} = -2.12$$

Step 3: Determine a **critical value(s) for α.**

$z_{\alpha/2} = z_{0.025} = 1.96$

Step 4: Make a **conclusion.**

Since $|z_0| = 2.12 > z_{0.025} = 1.96$, reject H_0 at $\alpha = 0.05$.

Exercise 15.6

(Nonparametric ANOVA; Kruskal-Wallis Test)

Step 1: State $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

H_0: $\tau_1 = \tau_2 = \cdots = \tau_a = 0$

H_1: $\tau_i \neq 0$ for at least one i, $i = 1, 2, 3, 4$

Step 2: Determine a **test statistic and its value.**

Mixing technique	Tensile strength (lb./in.^{2})					$R_{i.}$
A (y_{1j})	3,129	3,000	2,865	2,890	2,971	
Rank (r_{1j})	18	14.5	7	8	11	58.5
B (y_{2j})	3,200	3,000	2,975	3,150	3,081	
Rank (r_{2j})	20	14.5	12	19	17	82.5
C (y_{3j})	2,800	2,900	2,985	3,050	2,934	
Rank (r_{3j})	6	9	13	16	10	54.0
D (y_{4j})	2,600	2,700	2,600	2,765	2,666	
Rank (r_{4j})	1.5	4	1.5	5	3	15.0

Exercise 15.6 (*cont.*)

$$s^2 = \frac{1}{N-1}\left[\sum_{i=1}^{a}\sum_{j=1}^{n_i} r_{ij}^{\ 2} - \frac{N(N+1)^2}{4}\right]$$

$$= \frac{1}{20-1}\left[2{,}869 - \frac{20\times(20+1)^2}{4}\right] = 34.95$$

$$h = \frac{1}{s^2}\left[\sum_{i=1}^{a}\frac{r_{i.}^{\ 2}}{n_i} - \frac{N(N+1)^2}{4}\right]$$

$$= \frac{1}{34.95}\left[\left(\frac{58.5^2}{5}+\frac{82.5^2}{5}+\frac{54.0^2}{5}+\frac{15.0^2}{5}\right) - \frac{20\times(20+1)^2}{4}\right]$$

$$= 13.42$$

Step 3: Determine a **critical value(s) for α.**

$$\chi^2_{\alpha,a-1} = \chi^2_{0.05,4-1} = \chi^2_{0.05,3} = 7.81$$

Step 4: Make a **conclusion.**

Since $h = 13.42 > \chi^2_{\alpha,a-1} = 7.81$, reject H_0 at $\alpha = 0.05$.