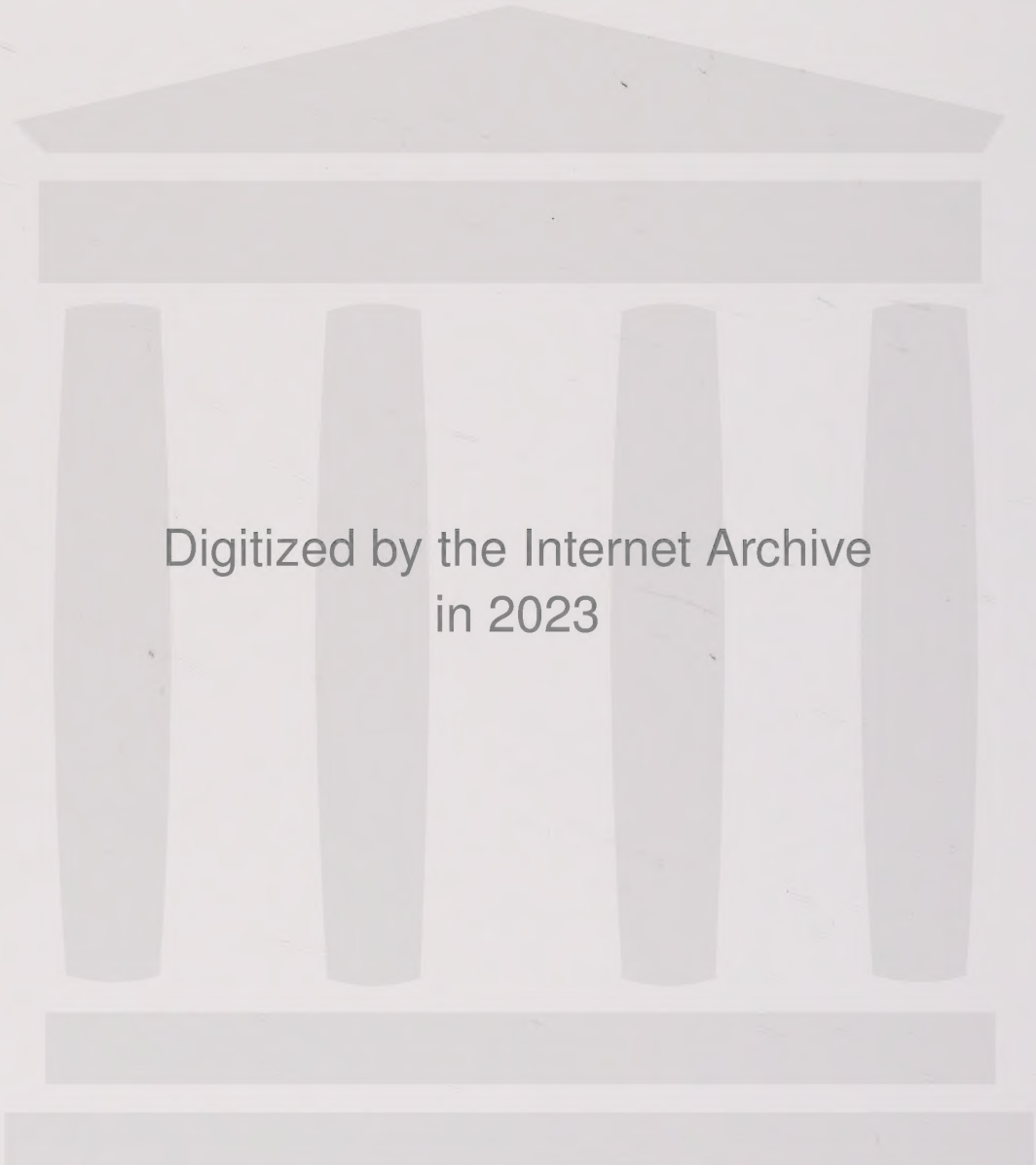
INTRODUCTION TO BUSINESS ANALYSIS



FRED HULME CAROLYN HULME TURNER


Digitized by the Internet Archive
in 2023

https://archive.org/details/introductiontobu0000fred

INTRODUCTION TO BUSINESS ANALYSIS



FRED HULME ● CAROLYN HULME TURNER

Kendall Hunt
publishing company

Cover image © Shutterstock.com

Kendall Hunt
publishing company

www.kendallhunt.com
Send all inquiries to:
4050 Westmark Drive
Dubuque, IA 52004-1840

Copyright © 2020 by Kendall Hunt Publishing Company

ISBN 978-1-7924-2293-5

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the copyright owner.

Published in the United States of America

Table of Contents

CHAPTER 1 Summarizing Data with Tables and Charts 1

CHAPTER 2 Describing Data—Numerical Summaries 37

CHAPTER 3 Risk Analysis—Probability 73

CHAPTER 4 Sampling and Sampling Distributions 117

CHAPTER 5 Statistical Estimation—Confidence Intervals 141

CHAPTER 6 Hypothesis Testing 169

CHAPTER 7 Analysis of Count Data—Chi-Square Applications 197

CHAPTER 8 Analysis of Variance ANOVA 215

Preface

Fred Hulme wrote this textbook to introduce the fundamental concepts of quantitative business analysis to students. It is written with the knowledge that most business school students have not had a full course in statistics and especially not one focusing on business applications. The problems lend themselves to the analytical side more than the mathematical side with the idea that the course is a tool for business students and not an end unto itself. An emphasis is placed on using computer software to solve realistic problems in terms of complexity and size.

Chapters 1 and 2 involve summarizing data through tables, charts, and numerical calculations. These summaries provide information used in decision making. Techniques for condensing large amounts of data into a meaningful summary are presented. Both Excel and MegaStat are used within the text, but the same concepts can easily be understood using JMP, SAS, and R. Proficiency in Microsoft Excel is often requested by organizations that recruit university business school interns and graduates.

Business decisions often involve choosing from several alternatives. In order to decide among several alternatives, one must be able to estimate the consequences of each alternative. The consequences often depend upon factors that are subject to chance. The uncertainty cannot be eliminated, but risk analysis methods, which allow you to estimate the risk in your decisions, exist. These tools are discussed in Chapter 3.

The concepts of gathering samples, analyzing the data, and making inferences about the populations are covered in Chapters 4 through 6. The topics of sampling procedures, probability, interval estimation, and hypothesis testing are all covered.

The last two chapters cover the analysis of count data using Chi Square techniques and comparing more than two means with Analysis ofVariance designs. Understanding both methodologies boost the analytical skills of business students.

PowerPoint slides, Data Files, and a Study Guide and Formula Sheets by Shaun Eide is included at the end of the textbook.

CHAPTER

1

Summarizing Data with Tables and Charts



© Kendall Hunt Publishing Company

Introduction

Business Data Analysis (Quantitative Business Analysis [QBA] at Baylor) involves the collection and summarizing of data such that the summaries can be used in a decision-making process. The summaries may describe the data, make comparisons of groups, and make predictions from the data. There are two broad categories of business statistics procedures: (1) **descriptive statistics** and (2) **inferential statistics**. Descriptive statistics involves creating tables, charts, and numeric summaries such as averages to condense what is usually a large amount of information. Inferential statistics involves using data from a **sample** to draw conclusions about the larger **population** or **process** from which the sample was taken.

An example of descriptive statistics is shown in Table 1.1. The table summarizes the responses of a randomly selected sample of 240 customers as to the brand of athletic shoes they purchased and their level of satisfaction with the shoes. The frequencies are as follows:

		Nike	New Balance	Adidas	Reebok
	Low	25	7	7	2
Satisfaction					
Level	Moderate	41	14	8	6
	High	76	27	21	6

Table 1.1 A Two-Factor Frequency Table
Source: Fred Hulme

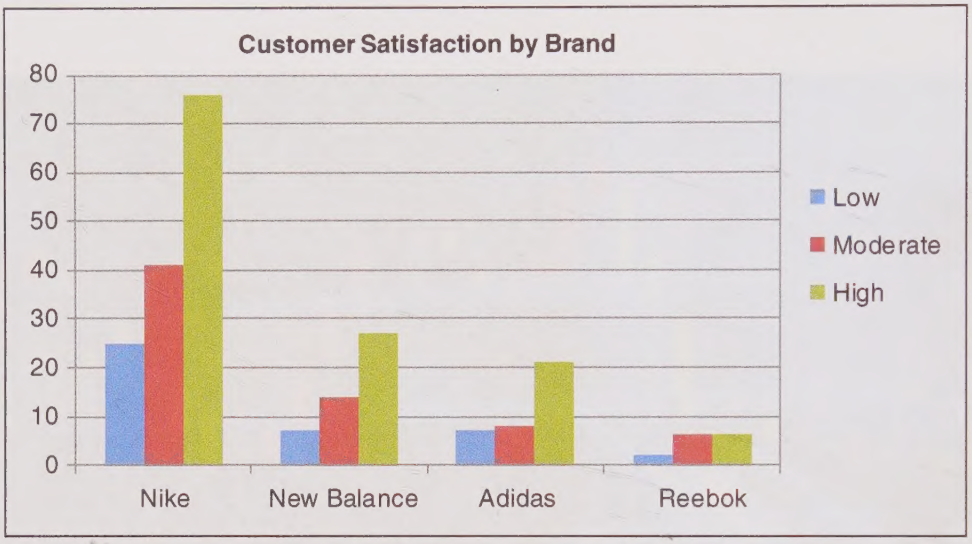


Figure 1.1 Column Chart for Comparisons
Source: Fred Hulme

Figure 1.1 displays the relationship between customer satisfaction and shoe brand, which is also an example of descriptive statistics.

Inferential statistical procedures use the results of descriptive statistics to generalize beyond a **sample** to a larger **population** or **process** from which the sample was drawn. Data often arises from the results of a survey of individuals. For example, the management of a fast food chain might be interested in determining the total number of dollars that Baylor students spend each year eating in Waco fast food restaurants. The fast food chain would probably also like to know the fast food preferences of Baylor students. Both information would be helpful in estimating how successful a fast food chain might be if located near the Baylor campus. The management of the fast food chain probably does not have the time or money to question every single Baylor student. However, time and money to question 300 students might be available. Techniques exist for randomly choosing a representative sample of 300 students from a population of 17,200 Baylor students. An estimate or guess for the total dollars spent by all 17,200 Baylor students per year can be made from the amount spent by the sample of 300 students. You would not expect the estimate to be perfect. There would be some difference between the estimate based on the sample and the unknown total amount spent by all 17,200 students. The difference between the two amounts is referred to as **sampling error**. The word, "error," does not refer to the fact that a mistake has been made. It simply means that you have based your findings on incomplete yet representative data. Procedures exist for estimating the maximum amount of sampling error. You will be able to state with a large degree of confidence or surety how much difference there is between your estimate (a **statistic**) and the unknown true answer (a **parameter**).

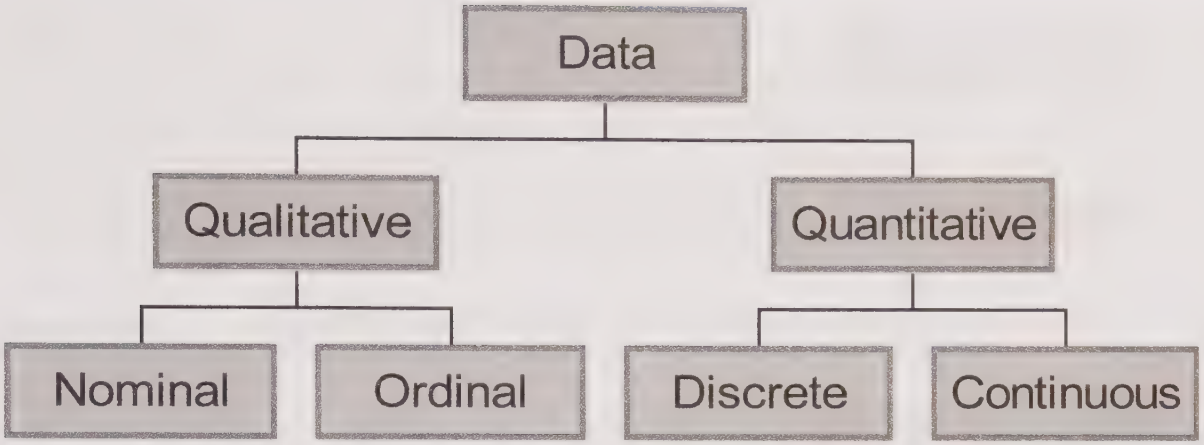


Figure 1.2 Data Classification
Source: Fred Hulme

Types of Data

Data can be categorized as either **qualitative** or **quantitative** as in Figure 1.2. A qualitative variable simply records an attribute or category. A qualitative variable can have values containing any characters—letters, numbers, special characters. A person's gender would be an example of a qualitative variable. A code of "F" for female and "M" for male could represent the groups. A person's marital status (single or married) is an example of a qualitative variable. Most qualitative variables simply represent categories; however, some qualitative variables can be ordered on some scale. The classification of public high schools in Texas (1A, 2A, 3A, 4A, and 5A) is an example of a qualitative variable that includes a ranking or ordering. The possible classifications are categories, but they have an inherent ordering.

Measurement Scale	Example
Nominal	Type of car
	1 = Ford
	2 = General Motors
	3 = Chrysler
	4 = Foreign
Ordinal	Rating of a restaurant
	1 = Below average
	2 = Average
	3 = Above average
Interval	Temperature in Fahrenheit
Ratio	Length, weight, sales in $

Table 1.2 Scales of Measurement
Source: Fred Hulme



© Kendall Hunt Publishing Company

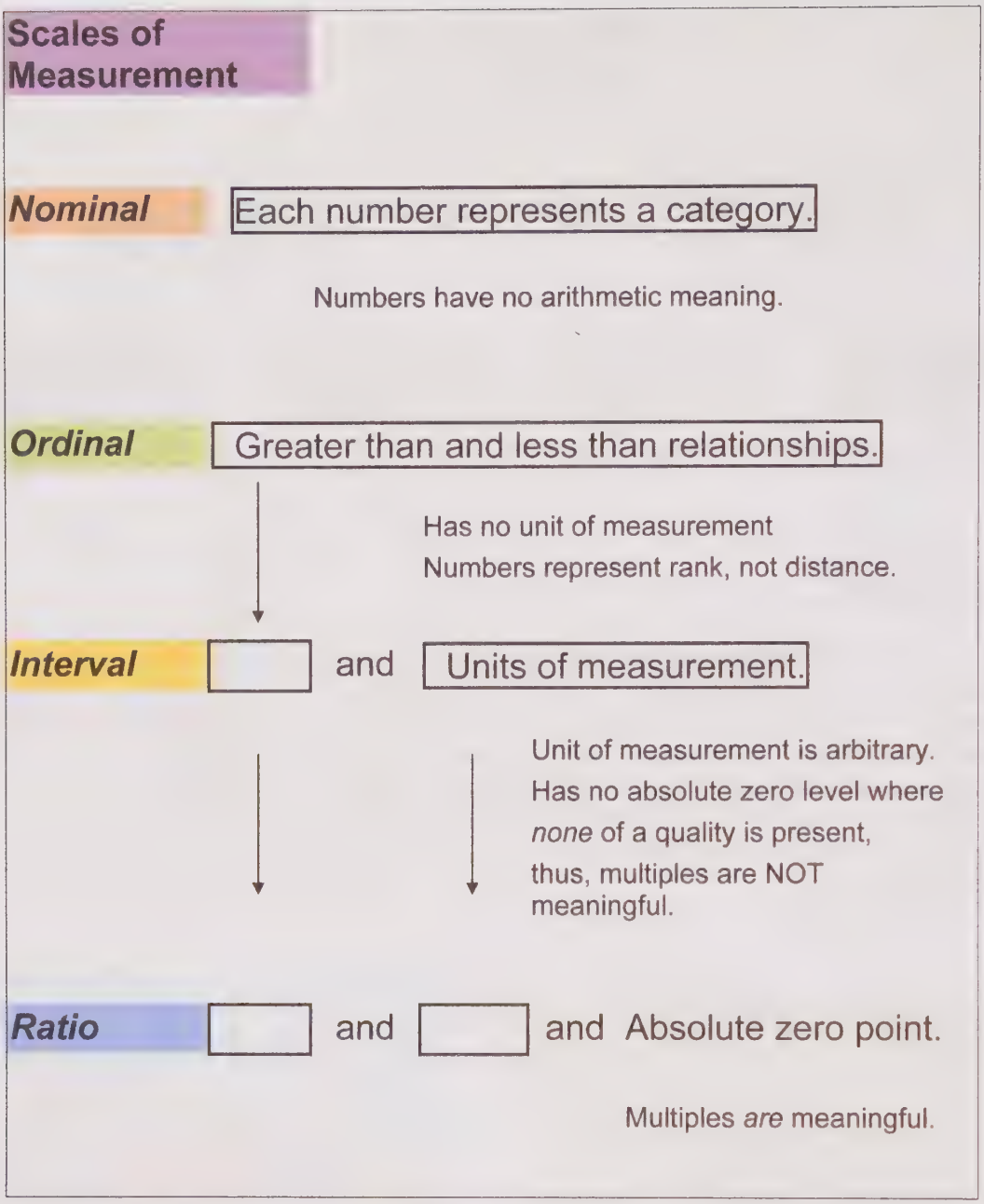


Figure 1.3 Scales of Measurement Diagram
Source: Fred Hulme

A quantitative variable results from a count or a measurement. Quantitative variables have values that contain numeric digits only. A person's age and salary would be examples of quantitative variables. A quantitative variable can be further divided into **discrete** and **continuous**. A discrete variable usually results from a count being taken, which typically yields an integer value 0, 1, 2, 3, ... but not fractions. More precisely the values of a discrete variable can assume only a limited number of values. A continuous variable results from a measurement of a quantity. A person's weight and college grade point average are examples of a continuous variable. A continuous variable conceptually can assume an infinite number of possible values. Sometimes the measuring device limits the possible attainable values for a continuous variable, but the variable is still considered a continuous variable.

Another common way of categorizing numeric data is through the **measurement scale** of the numbers in Table 1.2 and Figure 1.3.

Understanding the type of data presented in a problem is that not all summarizing methods are appropriate for all types of data. A chart for instance that is appropriate for summarizing a nominal, qualitative variable may not be useful in summarizing a continuous, quantitative variable. In summarizing data effectively, one begins by asking the question—"What type of data needs to be summarized?" The appropriate procedure can then be chosen for the summary.

Example of a Microsoft Excel Data Spreadsheet

e-Tronics, a web-based business selling video and sound equipment, personal computers, and other electronic products, advertises by mailing brochures to its customers listing its products and specifying its web address. For a sample of seventy-five customers, there is data on the following variables:



© Kendall Hunt Publishing Company

Gender:	Coded as "M" for males, "F" for females
Age:	Age of customer
Income:	Combined annual household income
Children:	Number of children living with customer
Last Year:	Total amount of purchases made from e-Tronics last year
Current Purchases:	Total amount of current purchases made from e-Tronics

An example row of data is:

Gender	Age	Income	Children	Last Year	Current Purchases
M	51	$97,300	1	$0.00	$2,850.50

Table 1.3 e-Tronics Sample Record

Source: Fred Hulme

This data recorded in Table 1.3 indicates a 51-year-old male customer with a household income of $97,300. This customer has one child living at home and spent $2,850.50 in Current Purchases but made no purchases with e-Tronics last year. The data for each customer would occupy one row in the spreadsheet. Each field or data element such as the household income would occupy one column in the spreadsheet. You would enter this data into a spreadsheet with the first row containing headings that can be used in defining the data. The first 15 rows and last 15 rows of data are displayed from the spreadsheet in Table 1.4.

Summaries of the data could be useful in answering questions such as: (1) Does the customer's gender seem to matter with regard to the amount spent with e-Tronics?; (2) Should e-Tronics target more of its advertising on younger or older customers?; (3) Does e-Tronics seem to be doing a good job of retaining customers from the previous year?

e-Tronics Data

	A	B	C	D	E	F
1	**Gender**	**Age**	**Income ($)**	**Children**	**Last Year ($)**	**Current Purchases ($)**
2	M	51	97,300	1	0.00	2,850.50
3	M	42	59,100	2	390.06	452.04
4	M	38	62,700	2	413.82	549.09
5	F	55	39,500	0	869.00	1,207.51
6	F	55	80,000	0	2,640.00	2,381.59
7	M	52	83,600	2	551.76	419.43
8	F	45	32,500	0	715.00	1,059.43
9	M	26	51,200	0	1,126.40	1,552.12
10	M	37	21,900	0	481.80	330.77
11	M	65	59,400	0	1,306.80	2,039.06
12	M	31	101,700	1	0.00	1,015.89
13	F	35	80,600	0	2,659.80	3,407.56
14	M	42	63,300	2	0.00	493.51
15	M	54	55,500	0	1,221.00	1,029.58
61	F	53	10,600	2	0.00	403.02
62	M	43	84,100	1	1,850.20	2,121.45
63	F	55	12,600	0	415.80	289.33
64	F	71	93,400	2	0.00	2,401.67
65	M	50	49,200	3	108.24	113.14
66	M	34	19,500	0	0.00	685.74
67	F	26	56,000	3	369.60	286.35
68	F	23	47,500	0	1,567.50	1,016.96
69	F	71	15,900	0	0.00	486.74
70	M	32	74,400	0	2,455.20	2,271.24
71	M	32	73,300	0	0.00	2,033.36
72	M	21	64,000	3	140.80	221.41
73	F	75	63,300	2	417.78	529.53
74	M	69	63,000	0	1,386.00	2,291.52
75	F	40	110,000	0	3,630.00	4,000.00

Table 1.4 Sample Data
Source: Fred Hulme

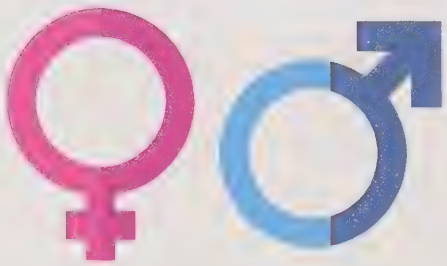
Summarizing Data through Frequency Tables

A **frequency table** or **distribution** is one of the simplest methods for summarizing both quantitative and qualitative data. A frequency table for qualitative data reduces the original data to a list of the possible categories for the data and the *frequencies or counts* of the data belonging to the respective categories. The raw data for the gender of the customers in the e-Tronics data is:

M	M	F	M	M	F	F	M
M	F	M	F	M	F	M	M
M	M	M	F	M	F	F	F
F	M	M	F	F	F	F	M
F	F	F	M	F	F	M	F
M	M	F	F	F	M	M	
F	F	M	M	M	F	F	
M	M	F	M	M	M	F	
M	M	F	M	F	M	F	
M	M	F	F	F	F	M	

Table 1.5 Gender Data
Source: Fred Hulme

From the raw data in Table 1.5, can you quickly determine the number of female and male employees? The answer is probably not, and as the amount of data gets larger, this type of summarization gets more difficult. A frequency table would show that there are only two categories for gender, Female and Male, and give the number of customers in each category. Sometimes relative and cumulative frequencies are provided in frequency tables like Table 1.6.



© Kendall Hunt Publishing Company

Category	Frequency	Relative Frequency (%)	Cumulative Frequency	Cumulative (%)
Female	37	49.33	37	49.33%
Male	38	50.67	75	100.00%

Table 1.6 Frequency Table for Gender
Source: Fred Hulme

A frequency table for a quantitative variable is constructed much the same way if there are not too many possible values for the variable. The frequency table for the number of children living with the customer for the e-Tronics data is shown in Table 1.7.



© Kendall Hunt Publishing Company

Number of Children	Frequency	Relative Frequency (%)	Cumulative Frequency	Cumulative (%)
0	40	53.33	40	53.33
1	14	18.67	54	72.00
2	14	18.67	68	90.67
3	7	9.33	75	100.00

Table 1.7 Descriptive Frequency Table
Source: Fred Hulme

If a quantitative variable assumes many possible values, then a frequency table that enumerates every possible value would not provide much summarization. In this case, a series of equally spaced intervals are constructed and the frequencies within the intervals are counted rather than each individual value.

The intervals would be constructed such that they would contain the smallest value through the largest value. Usually the number of intervals is from as few as five and as many as fifteen. Too many intervals would not provide enough summarization; too few intervals will provide too much summarization.

A frequency distribution using six intervals for the customer age is in Table 1.8



© Kendall Hunt Publishing Company

Age	Frequency	Cumulative Frequency	Relative Frequency	Cumulative (%)
<25	7	7	9.3	9.3
≥25 and <35	19	26	25.3	34.7
≥35 and <45	17	43	22.7	57.3
≥45 and <55	13	56	17.3	74.7
≥55 and <65	9	65	12.0	86.7
≥65 and <75	9	74	12.0	98.7
≥75 and <85	1	75	1.3	100.0

Table 1.8 Frequency Table for Age
Source: Fred Hulme

You should be able to glance at a frequency distribution like Table 1.8 and get a feel for where the values of a variable fall. From the previous frequency distribution, you can quickly tell that around 75% of the customers were between 25 and 55 years old.

By summarizing data with an interval-based frequency distribution you lose the identity of the original data. Using the previous frequency distribution without having the original data, we would not know the actual nine numbers that fell in the interval with a lower limit of 55. They could be as small as 55 or as large as 64.

Summarizing Data through Frequency Tables with MegaStat

MegaStat is an add-in program to Microsoft Excel that extends Excel's statistical capabilities. The various MegaStat options are listed below.

Qualitative Frequency Table Example

To create a frequency distribution for *Gender*, choose the MegaStat command, **Frequency Distributions/ Qualitative** in Figure 1.4. The following dialog box will be presented:

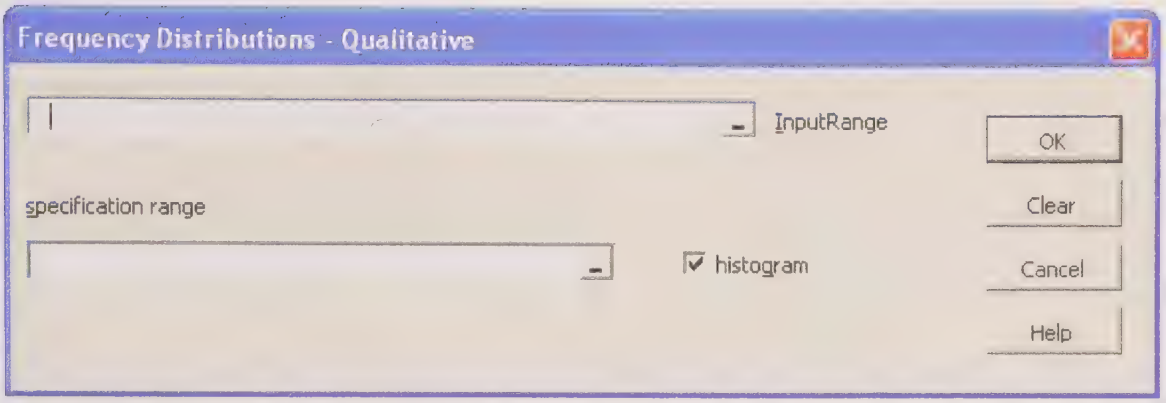


Figure 1.4 MegaStat Dialog Box
Source: Fred Hulme

Since the values for *Gender* are in cells A1 through A76, you can either drag over those cells to define the Input Range or type the cell range: A1:A76 as shown in Figure 1.5. The specification range is requesting a consecutive range of cells that contain all the possible outcomes for the data. In the case of *Gender*, that would be the choices—"F" and "M". Cells A6 and A7 in our data satisfy this condition. If you cannot find consecutive cells that contain all the possible outcomes, you could have entered the outcomes in some blank cells prior to issuing the **Frequency Distributions** command. A new worksheet, **Output**, will be created to store the results of the command.

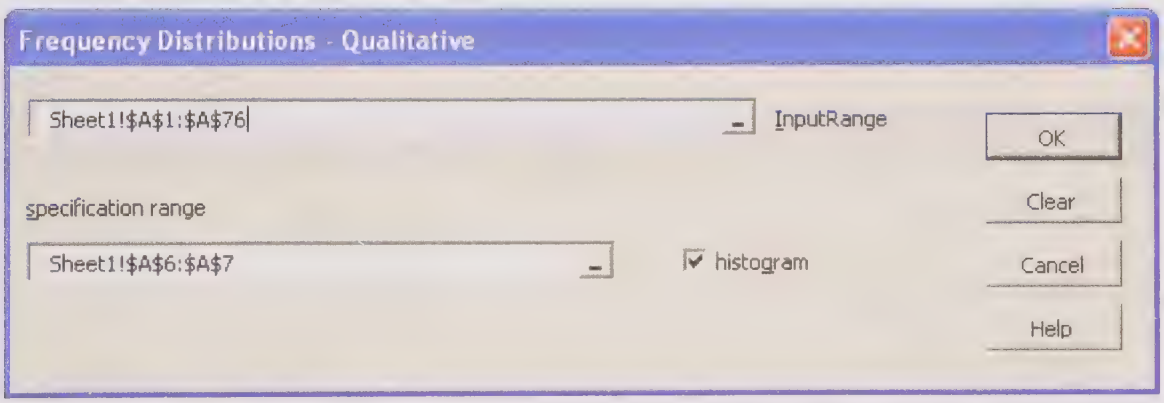


Figure 1.5 Typing in Ranges
Source: Fred Hulme

MegaStat Results

Frequency Distribution—Qualitative		
Gender	**Frequency**	**%**
F	37	49.3
M	38	50.7
	75	100.0

Table 1.9 Gender Frequency
Source: Fred Hulme



© Kendall Hunt Publishing Company

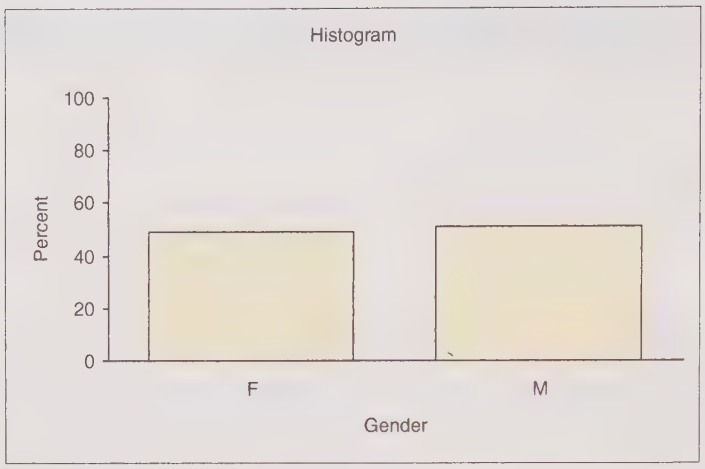


Figure 1.6 Gender Chart
Source: Fred Hulme

The frequency distribution and accompanying column chart in Table 1.9 and Figure 1.6 show that there is approximately an equal distribution of females and males in the e-Tronics data.

Quantitative Frequency Table Example

The appropriate MegaStat command for creating a frequency distribution for *Current Purchases* would be—**Frequency Distributions/Quantitative**.

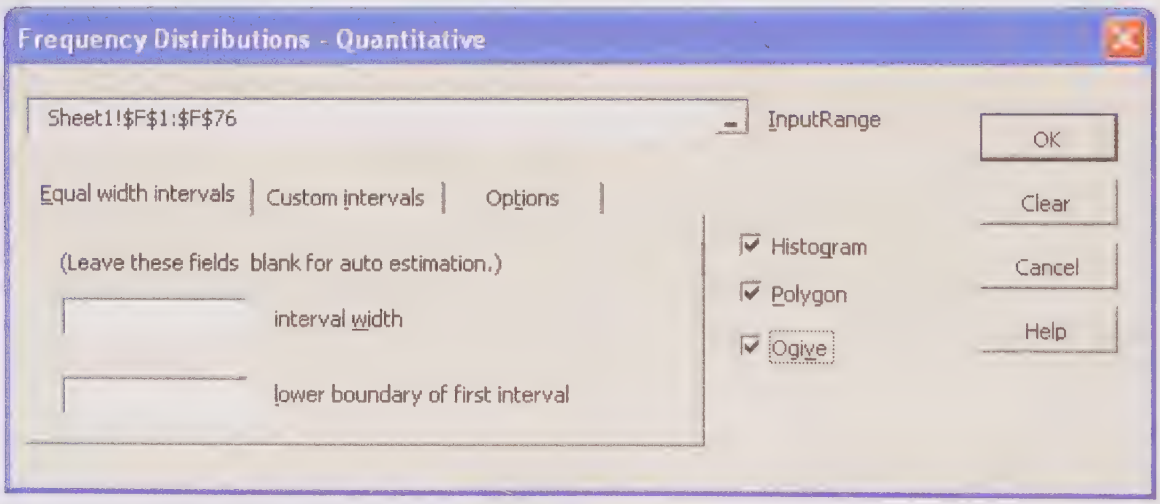


Figure 1.7 MegaStat Dialog Box
Source: Fred Hulme

You can control the interval width and the lower limit of the first interval by entering their values. See Figure 1.7. If you leave those entries blank, MegaStat will decide their values based on the data. A column chart of the relative frequencies known as a **histogram** is the default graphical summary of the frequency table data. Optional graphical summaries are a **polygon** and an **ogive**.

MegaStat Results

Frequency Distribution—Quantitative								
	Current Purchases						Cumulative	
Lower		Upper	Midpoint	Width	Frequency	%	Frequency	%
0	<	500	250	500	24	32.0	24	32.0
500	<	1,000	750	500	12	16.0	36	48.0
1,000	<	1,500	1,250	500	12	16.0	48	64.0
1,500	<	2,000	1,750	500	9	12.0	57	76.0
2,000	<	2,500	2,250	500	11	14.7	68	90.7
2,500	<	3,000	2,750	500	3	4.0	71	94.7
3,000	<	3,500	3,250	500	2	2.7	73	97.3
3,500	<	4,000	3,750	500	1	1.3	74	98.7
4,000	<	4,500	4,250	500	1	1.3	75	100.0
					75	100.0		

Table 1.10 Current Purchases
Source: Fred Hulme

The following questions can be answered from Table 1.10.

(a) How many customers have a *Current Purchase* amount less than $500? What percentage?
Answers: 24 customers, 32.0%

(b) How many of the customers have a *Current Purchase* amount greater than or equal to $1,500 and less than $2,000? What percentage?
Answers: 9 customers, 12.0%

(c) How many of the customers have a *Current Purchase* amount greater than or equal to $3,500? What percentage?
Answers: 2 customers, 2.6%

(d) How many of the customers have a *Current Purchase* amount greater than or equal to $1,000 and less than $3,000? What percentage?
Answers: 35 employees, 46.7%

Graphical Summaries of Frequency Table Data

A **histogram** provides a visual alternative to the table display of the frequency table. You can quickly glance at the histogram and get a feel for where most of the values of a distribution occur. You also get some indication of the central location and whether the values of the distribution are approximately symmetric around the central location.

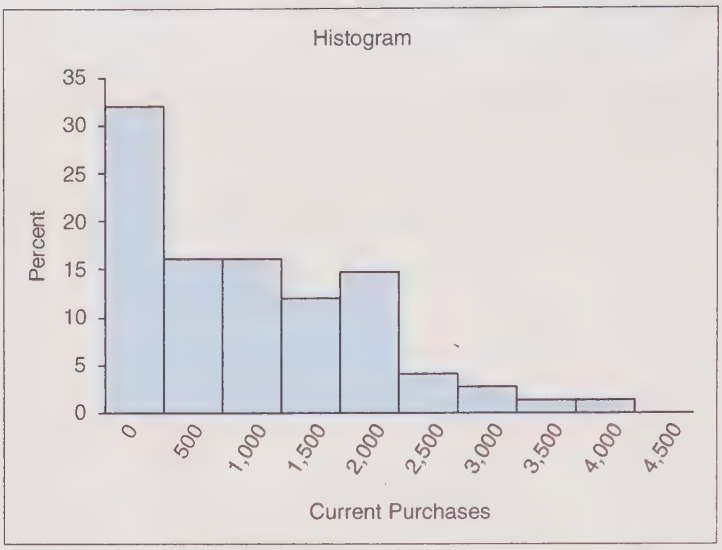


Figure 1.8 Current Purchases Histogram
Source: Fred Hulme

A distribution that is nonsymmetric is sometimes referred to as a **skewed** distribution. A distribution with fewer values on the low (left) end is called **skewed left** or **skewed negatively**; a distribution with fewer values on the high (right) end is called **skewed right** or **skewed positively**. Whether a distribution is approximately symmetric or skewed affects the choices made in properly summarizing the data with other measures that will be discussed later.

The distribution of the 75 values of *Current Purchases* would be considered skewed right based on the shape of the histogram as in Figure 1.8. This implies that there are more customers on the low end of the range for *Current Purchases* than on the high end of the range. The histogram reflects that most of the customers spent less than $2,500 Current Purchases.

An alternative graphical summary to the histogram is a **frequency polygon** in Figure 1.9. A frequency polygon uses a line chart to represent the relative frequencies. An advantage of a frequency polygon is that you can overlay two or more frequency polygons on the same chart to compare multiple distributions.

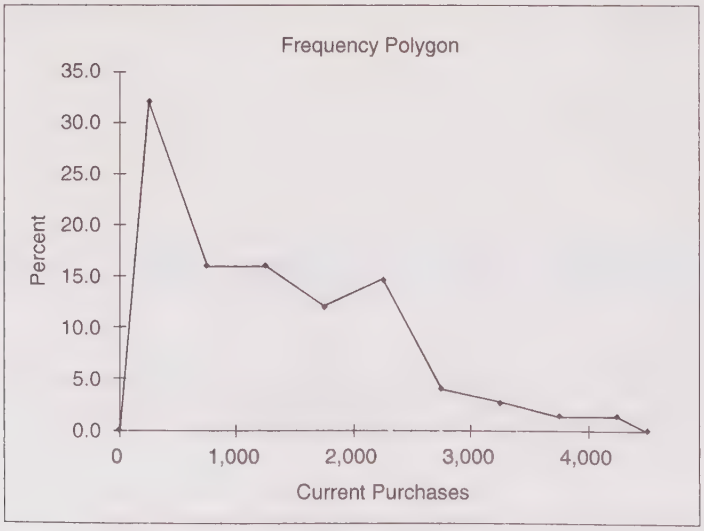


Figure 1.9 Current Purchases Frequency Polygon
Source: Fred Hulme

Another chart that sometimes accompanies a frequency distribution is a line chart of the "Cumulative %'s" plotted against the upper limit of each interval. This chart is called an **ogive** (rhymes with "alive") found in Figure 1.10. From this chart you can visually approximate the answer to questions such as—"What percent of the customers spent less than a particular amount?"

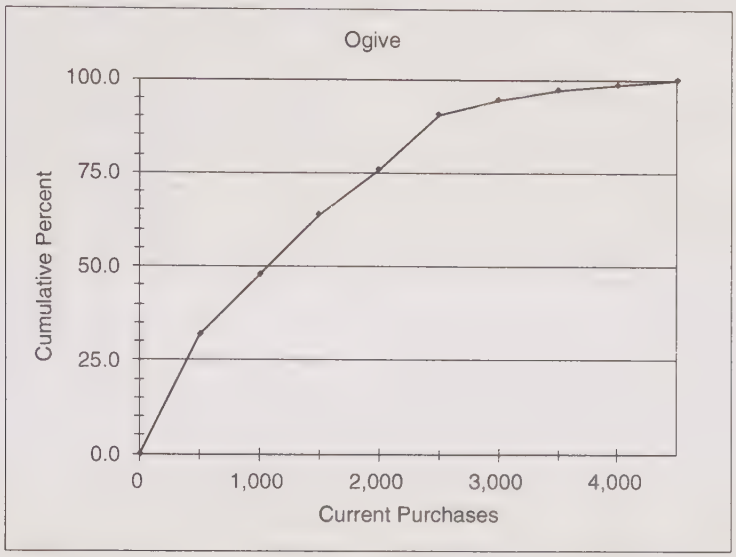


Figure 1.10 Current Purchases OGIVE
Source: Fred Hulme

Using the "ogive" chart, approximately 75% of the customers would have a Current Purchase amount less than $2,000.

Comparing Groups through Frequency Polygons Using MS Excel

Frequency polygons are particularly useful for comparing distributions. For example, if you want to compare the percentages of the frequencies of the *Current Purchases* by females versus males, you would begin by sorting the data into the two groups—female and male. Using MegaStat, you could then create a separate frequency distribution of the *Current Purchases* for both females and males found in Table 1.11.

The results of those frequency distributions using intervals of 0–500, 500–1,000, 1,000–1,500, 1,500–2,000, 2,000–2,500, 2,500–3,000, 3,000–3,500, 3,500–4,000, 4,000–4,500 are:



© Kendall Hunt Publishing Company

	Male			Female
Amount Spent	%		Amount Spent	%
0–500	23.7		0–500	40.5
500–1,000	10.5		500–1,000	21.6
1,000–1,500	18.4		1,000–1,500	13.5
1,500–2,000	15.8		1,500–2,000	8.1
2,000–2,500	23.7		2,000–2,500	5.4
2,500–3,000	5.3		2,500–3,000	2.7
3,000–3,500	0.0		3,000–3,500	5.4
3,500–4,000	2.6		3,500–4,000	0.0
4,000–4,500	0.0		4,000–4,500	2.7

Table 1.11 Comparing Values
Source: Fred Hulme

You would then need to copy into consecutive columns the midpoints for each interval and the corresponding percentages for female and male. A spreadsheet with this information is given below.

	A	B	C
1		Male	Female
2	$0	0.0	0.0
3	$250	23.7	40.5
4	$750	10.5	21.6
5	$1,250	18.4	13.5
6	$1,750	15.8	8.1
7	$2,250	23.7	5.4
8	$2,750	5.3	2.7
9	$3,250	0.0	5.4
10	$3,750	2.6	0.0
11	$4,250	0.0	2.7

Table 1.12 Data Comparing Purchases
Source: Fred Hulme

Steps in Creating a Comparison of Two Frequency Polygons

1. Select the cells containing the midpoints for each interval and the corresponding frequency percentages for the groups to be compared in Table 1.12. Include the labels found in the top row in order to have appropriate legends for the chart. You would highlight or select cells, A1 through C11, for the above example.
2. Select the **Insert** tab and choose the "**Line**" chart option.
3. With the chart selected, the tab, **Chart Tools/Layout**, will allow you to add a **Chart Title** and **Axis Titles**.

Frequency polygons compare the distributions of the *Current Purchases* by the female customers versus the male customers are displayed below in Figure 1.11.

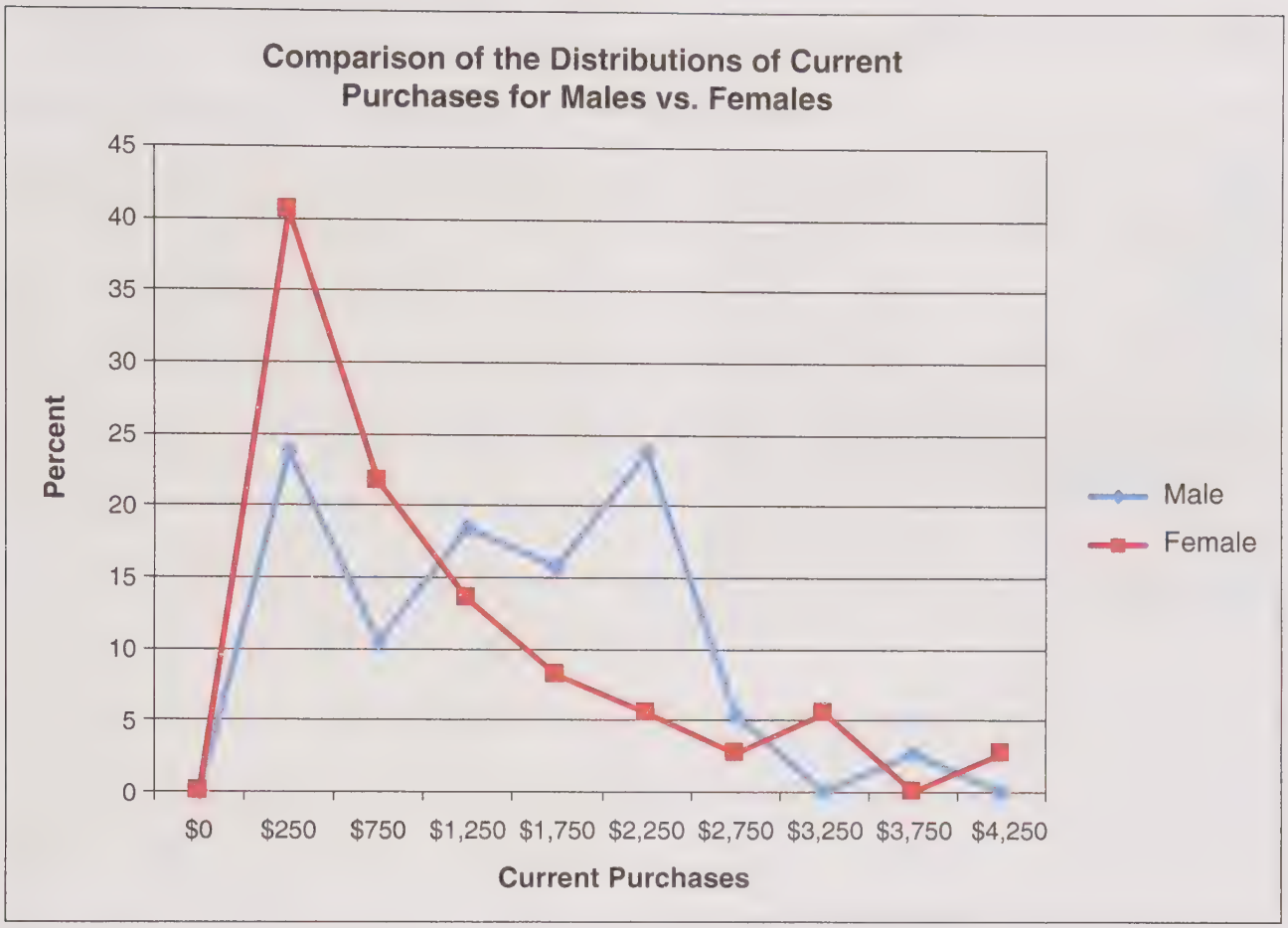


Figure 1.11 Comparison of Purchases by Gender
Source: Fred Hulme

Box Plots

An alternative graphical summary for a histogram is a **box plot**. A box plot is based on the following "five number" summary.

(a) the minimum value
(b) the first quartile, Q_1 (25% of the numbers are less than Q_1)
(c) the median, Q_2 (50% of the numbers are less than Q_2)
(d) the third quartile, Q_3 (25% of the numbers are greater than Q_3)
(e) the maximum value.

From these values, the interquartile range, IQR, can be calculated: IQR = $(Q_3 - Q_1)$.

A box plot has four components: (1) a rectangular **box** that indicates the range of the middle 50% of the data values, (2) a **line** inside the box indicating the median value of the data, (3) horizontal lines (**whiskers**) drawn to the right and left of the box indicating degree of symmetry, and (4) an optional pair of vertical lines (**fences**) drawn on either side of the whiskers indicating extreme values or outliers.

Box Plot Calculations

The right side of the box is drawn at the third quartile, Q_3. The left side of the box is drawn at the first quartile, Q_1.

$$\text{Inner fences} = [Q_1 - 1.5 \times (\text{IQR}),\ Q_3 + 1.5 \times (\text{IQR})]$$

$$\text{Outer fences} = [Q_1 - 3 \times (\text{IQR}),\ Q_3 + 3 \times (\text{IQR})]$$

The right-hand whisker is drawn to the most extreme data value that is less than or equal to the inner, upper fence. The left-hand whisker is drawn to the most extreme data value that is greater than or equal to the inner, lower fence. Values that fall beyond the inner fences but within the outer fences are classified as "**low or high outliers**"; values that fall outside the outer fences are classified as "**low or high extremes**." A big part of analyzing data involves looking for results that cannot be explained by chance alone. Values occurring between the inner and outer fences are considered moderately unusual. These values are labeled "**outliers**" by MegaStat. Values outside the outer fences should rarely occur. These values are labeled "**extreme**" values by MegaStat.

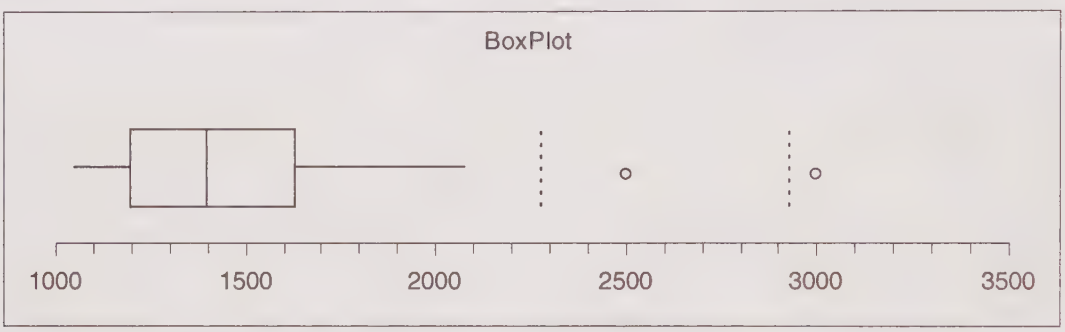


Figure 1.12 Current Purchases Box Plot

Source: Fred Hulme

The "Boxplot" in Figure 1.12 indicates that the range of the middle 50% of the data values is between 1,200 and 1,625 approximately. This range is indicated by the width of the box—the IQR. The median (the middle value) is indicated to be approximately 1,400. The upper whisker is longer than the lower whisker indicating the distribution of data values is skewed right (or positively skewed.) There is a "high outlier" indicated at a data value of 2,500 and a "high extreme" indicated at a data value of 3,000.

e-Tronics Box Plot

From the "e-Tronics" customer data,

	Current Purchases
Minimum	$42.50
First quartile, *Q*1	$463.47
Median, *Q*2	$1,016.96
Third quartile, *Q*3	$1,943.48
Maximum	$4,000.00
Outer upper fence	6,383.51
Inner upper fence	4,163.50
Upper Whisker	4,000.00
Maximum	4,000.00
Third quartile	1,943.48
Median	1,016.96
First quartile	463.47
Minimum	42.50
Lower whisker	42.50
Inner lower fence	−1,756.55
Outer lower fence	−3,976.56
Interquartile Range	1,480.01

Table 1.13 Descriptive Statistics
Source: Fred Hulme



© Rawpixel.com/Shutterstock.com

Using the five-number summary in Table 1.13, you can calculate the interquartile range and the limits for the box plot.

$$\text{Interquartile range} = Q_3 - Q_1$$
$$\text{Interquartile range} = \$1{,}943.48 - \$463.47$$
$$\text{Interquartile range} = \$1{,}480.01$$

From the previous definitions for the inner and outer fences,

$$\text{Inner fences} = [\$463.47 - 1.5 \times (\$1{,}480.01), \$1{,}943.48 + 1.5 \times (\$1{,}480.01)]$$
$$\text{Inner fences} = [-\$1{,}756.55, \$4{,}163.50]$$

$$\text{Outer fences} = [\$463.47 - 3 \times (\$1{,}480.01), \$1{,}943.48 + 3 \times (\$1{,}480.01)]$$
$$\text{Outer fences} = [-\$3{,}976.56, \$6{,}383.51]$$

Box Plot of Current Purchases from MegaStat

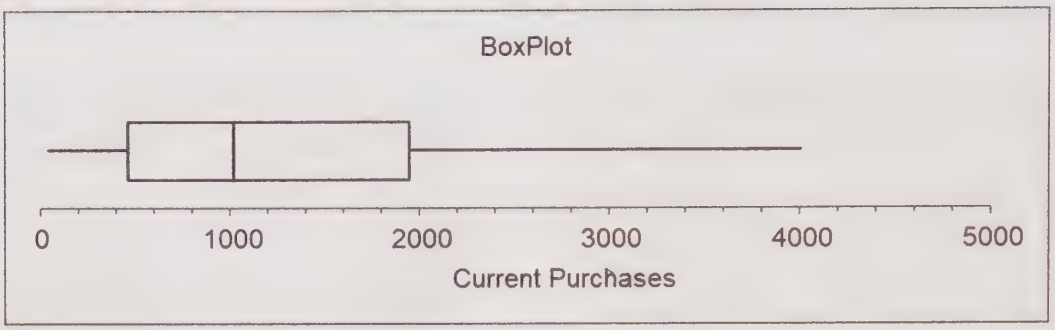


Figure 1.13 Box Plot of Current Purchases
Source: Fred Hulme

The box plot in Figure 1.13 indicates the range of the middle 50% of the *Current Purchases* is from approximately $450 to $1,950. The distribution of the purchase amounts is skewed right based on the relative lengths of the whiskers. None of the 75 purchase amounts would be considered unusual. The whiskers would extend all the way to the minimum and maximum values of the data. MegaStat suppresses the display of the "fences" when there are no unusual values indicated.

Summarizing Data through Cross Tabulation (Two-Factor Frequency Table) with MegaStat

You may want to see how two qualitative or discrete quantitative variables interact with one another. A two-dimensional table listing the frequencies for each combination of the two variables can give an insight into possible interactions. These tables are referred to as "two-**factor frequency tables**" because the frequencies are broken down according to two factors. The possible outcomes of one factor constitute the rows of the table; the possible outcomes of the other factor define the columns of the table. A two-factor frequency table is also sometimes referred to as a **cross tabulation** of the data.

From the e-Tronics customer data, you might be interested if there is a relationship between gender and the number of children living with the customer. To answer this question, you would need the counts or frequencies of the customers for the eight combinations of gender by the number of children.

e-Tronics Cross Tabulation

The MegaStat command for creating a cross tabulation in Figure 1.14 for *Gender* and *Children* would be: **Chi-Square/Crosstab—Cross tabulation**



© Kendall Hunt Publishing Company

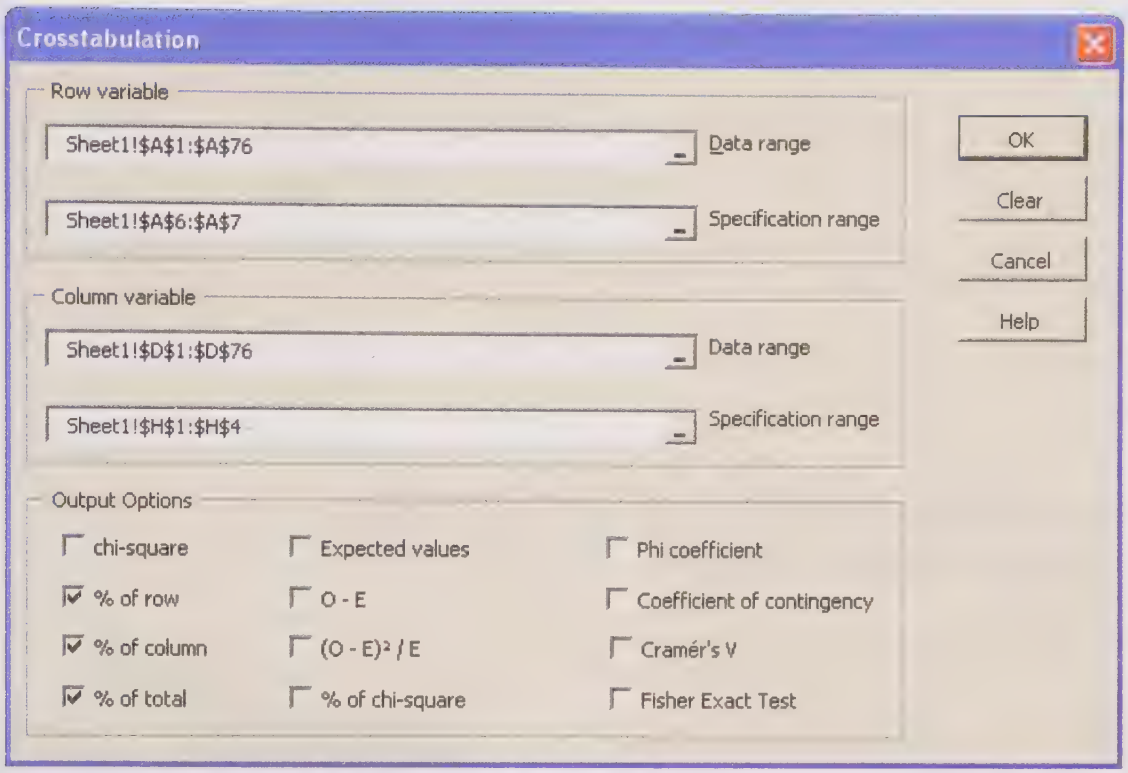


Figure 1.14 Chi Square Dialog Box
Source: Fred Hulme

MegaStat Results

Cross tabulation							
			Children				
			0	1	2	3	Total
	F	Observed	18	7	7	5	37
		% of row	48.6	18.9	18.9	13.5	100.0
		% of column	45.0	50.0	50.0	71.4	49.3
Gender		% of total	24.0	9.3	9.3	6.7	49.3
	M	Observed	22	7	7	2	38
		% of row	57.9	18.4	18.4	5.3	100.0
		% of column	55.0	50.0	50.0	28.6	50.7
		% of total	29.3	9.3	9.3	2.7	50.7
	Total	Observed	40	14	14	7	75
		% of row	53.3	18.7	18.7	9.3	100.0
		% of column	100.0	100.0	100.0	100.0	100.0
		% of total	53.3	18.7	18.7	9.3	100.0

Table 1.14 Chi Square/ Crosstabs Results
Source: Fred Hulme

The following questions can be answered from the "cross tabulation" in Table 1.14 (two-factor frequency table) for *Gender* by *Children*.

(a) What fraction of all customers are female with no children living at home? What percentage?
Answers: $\frac{18}{75}$ customers, 24.0%

(b) Of the female customers, what fraction have no children living at home? What percentage?
Answers: $\frac{18}{37}$ customers, 48.6%

(c) Of the customers with no children living at home, what fraction are female? What percentage?
Answers: $\frac{18}{40}$ customers, 45.0%

Summarizing Quantitative Data through Two-dimensional Plots

With business data, you are often interested in seeing if there is a relationship between two numeric variables. As the values of one variable increase, do the values of the other variable tend to increase, decrease, or not change? A two-dimensional (scatter) plot can help give you a visual answer to the above question. You should plot the variable to be predicted (*Y*) on the vertical axis and the variable you are predicting from (*X*) on the horizontal axis. If there appears to be a relationship between the two variables, then a mathematical model or equation can be estimated from the data that attempts to explain the relationship or pattern of change.



© Kendall Hunt Publishing Company

Scatterplot for e-Tronics

Using data about the customers of e-Tronics, a mail order electronics company, the initial data summary necessary to answer if the customer's household income data may be used as an accurate predictor of the amount spent by the Current Purchases would be a two-dimensional plot or chart of the two variables. This type of plot is referred to as a "scatter" chart or "*XY*" chart in Microsoft Excel or "scatterplot" in MegaStat.

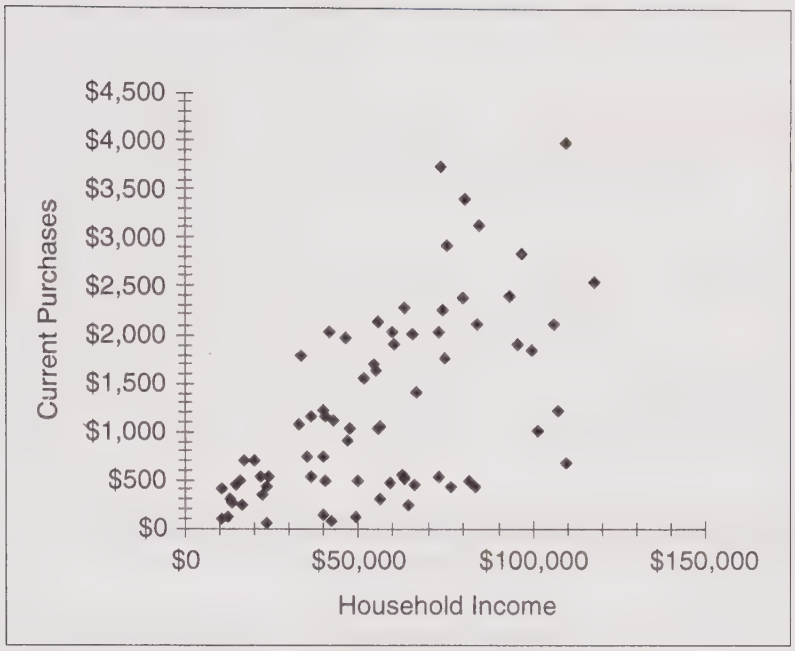


Figure 1.15 Scatterplot of Current Purchases vs Household Income

Source: Fred Hulme

From the "scatterplot" in Figure 1.15, the amount of current purchases and the household income appears to display a positive, linear (straight line) relationship between the two variables. A linear relationship does not require the points on the scatterplot to fall exactly on an upward or downward sloping line. The scatter points just have to be oriented around a sloping line.

Exercises

Short Answer/Discussion

1. Contrast descriptive vs. inferential statistical procedures.

2. Tables, charts, and numerical summaries are examples of statistical procedures.

3. Inferential statistical procedures involve providing information that can be used to draw conclusions from sample data.

4. The mean (average) gas mileage for all 2012 Honda Civics would be referred to as a _____

5. The mean gas mileage of a sample of only 30 Honda Civics is referred to as _____

6. The difference between the mean (average) gas mileage between all 2012 Honda Civics and the mean gas mileage of a sample of only 30 Honda Civics is referred to as _____

7. Explain the difference between qualitative and quantitative data.

8. For a qualitative variable, explain the difference between a nominal versus an ordinal variable. Give an example of each.

9. Explain the difference between discrete and continuous quantitative data.

10. For the following variables, classify them as either qualitative or quantitative. If the variable is qualitative, then qualify the variable as nominal or ordinal. If the variable is quantitative, then qualify the variable as discrete or continuous.
    a. Whether a car passes or fails an inspection
    b. Length of a bolt
    c. Number of defective fuses in a box of 100
    d. How you rate a restaurant: (1) below average, (2) average, or (3) above average
    e. Brand of automobile

11. Why is it important to understand the type of data that you are summarizing?

12. What is the purpose of a frequency distribution or table?

13. How do frequency distributions for quantitative variables that have many potential values differ from a frequency distribution for a qualitative variable?

14. What is the problem with using too many intervals in a frequency distribution for a continuous quantitative variable?

15. Explain what is meant by the term, "relative frequency," with respect to a frequency distribution. Why are "relative frequencies" sometimes more useful than raw frequencies?

16. What type of questions about numeric data can be answered with a frequency distribution of the data?

17. What type of data summary can be used to see if two qualitative variables interact with each other?

18. Why is a two-way table sometimes referred to as a cross tabulation?

19. What name is often used for a column chart or graph derived from the frequencies contained in a frequency distribution?

20. What is meant by the term, "skewness?"

21. A variable that has a distribution that is skewed left will have more frequencies on which end of the distribution?

22. Would you expect the estimated value of the homes in a large city to form a skewed distribution? If so, in what direction?

23. Why might you choose to use a frequency polygon rather than a histogram to display graphically the information conveyed in a frequency distribution?

24. What type of questions can be answered from viewing an ogive associated with a frequency distribution? For a symmetric variable with a mound-shaped or bell-shaped distribution, what would you expect the ogive to look like?

25. What percentage of the data is contained within the "box" of a box plot?

26. The vertical line within the "box" of a box plot indicates the _____ of the data.

27. The relative length of the "whiskers" of a box plot indicates the _____ of the data.

28. For a distribution of values that is skewed to the right, you would expect the right whisker to be _____ than the left whisker.

29. For a distribution of values that is approximately symmetric, you would expect the whiskers to be approximately _____ in length.

30. How is an "outlier" data value indicated in a box plot? How is an "extreme" data value indicated in a box plot?

Problems

1. The following data in Table 1.15 pertains to 30 graduates from U.S. colleges who have obtained their first job. The data represents the college major of the graduates where the data is coded by:

   1 = Education, 2 = Business, 3 = Social Science, 4 = Engineering/Science/Math

2	4	4	3	2
3	3	2	3	2
1	2	1	4	1
1	2	1	2	1
2	1	2	2	3
2	1	2	1	4

Table 1.15 Data for Major in College
Source: Fred Hulme



© Kendall Hunt Publishing Company

   a. What type of data is this?
   b. Produce a frequency distribution for the data.

2. The data in Table 1.16 represents the starting salaries in thousands of dollars for the 30 graduates from problem (1).

20.9	22.5	21.5	25.0	23.5
27.2	22.0	21.6	22.7	25.5
20.2	23.4	26.0	20.6	21.3
22.8	23.8	24.4	24.6	23.2
21.8	22.2	23.9	28.6	28.5
25.3	21.0	22.4	26.6	20.4

Table 1.16 Data for Starting Salaries
Source: Fred Hulme

   a. What type of data is this?
   b. Produce a frequency distribution for the data using five equally spaced intervals.

3. The data in Table 1.17 represents the starting salary of 200 graduates from problem (2), which is summarized below in a frequency distribution.

Salary	Frequency	Relative Frequency (%)	Cumulative Frequency	Cumulative %
≥ 20 and < 22	63	31.50	63	31.50%
≥ 22 and < 24	66	33.00	129	64.50%
≥ 24 and < 26	41	20.50	170	85.00%
≥ 26 and < 28	21	10.50	191	95.50%
≥ 28 and < 30	7	3.50	198	99.00%
≥ 30 and < 32	2	1.00	200	100.00%

Table 1.17 Starting Salaries
Source: Fred Hulme

a. How many graduates had a starting salary less than $22,000?
b. What percentage of the graduates had a starting salary less than $22,000?
c. How many graduates had a starting salary greater than or equal to $26,000 and less than $28,000?
d. What percentage of the graduates had a starting salary greater than or equal to $26,000 and less than $28,000?
e. How many graduates had a starting salary less than $28,000?
f. What percentage of the graduates had a starting salary less than $28,000?
g. How many graduates had a starting salary greater than or equal to $28,000?
h. What percentage of the graduates had a starting salary greater than or equal to $28,000?

4. The data in Table 1.18 is from problem (1) with an additional factor of the region (north, south, east, or west) where the graduate's college was located. This variable is coded as follows: 1 = North, 2 = South, 3 = East, 4 = West.

Graduate	Major	Region	Graduate	Major	Region	Graduate	Major	Region
1	2	1	11	1	3	21	4	2
2	3	4	12	1	2	22	2	3
3	1	3	13	4	2	23	2	2
4	1	2	14	2	1	24	1	3
5	2	3	15	1	1	25	2	2
6	2	4	16	1	3	26	2	1
7	4	1	17	2	2	27	1	1
8	3	2	18	2	2	28	1	2
9	2	2	19	3	3	29	3	3
10	2	2	20	3	1	30	4	4

Table 1.18 Frequency Table for Graduates by Major and Region
Source: Fred Hulme

Create a two-factor frequency table for this data. For the frequency of the cell that defines a business major from a college in the south, calculate:
a. the percentage of the total graduates,
b. the percentage of all the business majors
c. the percentage of all the graduates from a college in the south.

5. Table 1.19 gives the "five number summary" for the grade point average data from Problem 3.

Minimum	2.0
First quartile, Q_1	2.6
Median, Q_2	2.8
Third quartile, Q_3	3.0
Maximum	3.5

Table 1.19 GPA Statistics
Source: Fred Hulme

a. What grade point average is such that 1/4 of the values are smaller and 3/4 are larger?
b. Calculate the interquartile range.
c. Calculate the inner fences for the box plot of the data.
d. Calculate the outer fences for the box plot of the data.
e. Draw the box plot for the grade point averages.
f. Would either the minimum or maximum value be considered moderately unusual? Extremely unusual?

Self-Check

Questions 1–8 refer to the following study.
50 Texas car owners were asked several questions about their cars.


© Kendall Hunt Publishing Company

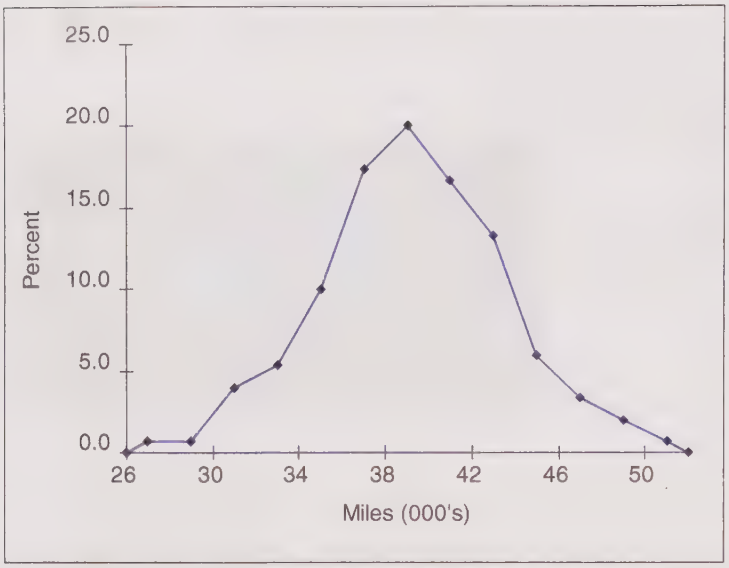
1. If the 50 motorists were classified as owning (O) vs. leasing (L) their car, what type of data would the response be?
   a. Qualitative, nominal
   b. Qualitative, ordinal
   c. Quantitative, discrete
   d. Quantitative, continuous
   e. None of the above

2. The number of miles between oil changes reported by each car owner would be what type of data?
   a. Qualitative, nominal
   b. Qualitative, ordinal
   c. Quantitative, discrete
   d. Quantitative, continuous
   e. None of the above

3. What type of data would the number of times each month that the car owners report that they check the air pressure in their tires?
   a. Qualitative, nominal
   b. Qualitative, ordinal
   c. Quantitative, discrete
   d. Quantitative, continuous
   e. None of the above

4. If the motorists rate the service received from their car dealership using the following choices (Above average, Average, Below average), what type of data would the responses be?
   a. Qualitative, nominal
   b. Qualitative, ordinal
   c. Quantitative, discrete
   d. Quantitative, continuous
   e. None of the above

5. The chart in Figure 1.16 shows the distribution of the relative frequencies of the number of miles driven before the tires on the cars need replacing in thousands of miles. What is the name of this type of summary?
   a. Histogram
   b. Frequency polygon
   c. Ogive
   d. *XY* scatterplot
   e. None of the above





© Kendall Hunt Publishing Company

Figure 1.16 Miles Driven
Source: Fred Hulme

6. Using the chart above, the shape of the distribution of the number miles in thousands driven before the tires on the cars need replacing is best described as
   a. skewed left
   b. skewed right
   c. approximately symmetric
   d. uniform
   e. none of the above

7. The chart in Figure 1.17 displays the cumulative %'s of the frequencies of the number miles driven before the tires on the cars need replacing. What is the name of this type of chart?
   a. Histogram
   b. Frequency polygon
   c. Ogive
   d. *XY* scatterplot
   e. None of the above

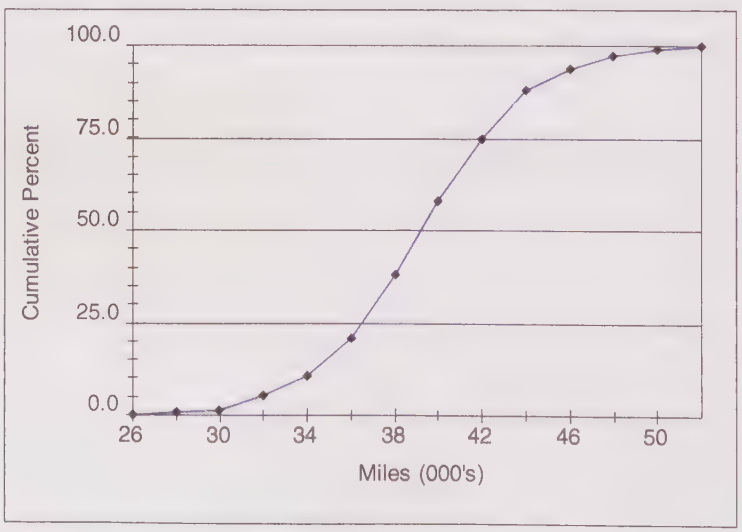


Figure 1.17 Cumulative Miles Driven
Source: Fred Hulme

8. Using the chart above, approximate the third quartile of the distribution of the number miles in thousands driven before the tires on the cars need replacing.
   a. 34 **b.** 38 **c.** 42 **d.** 46


© Kendall Hunt Publishing Company

An airline is interested in determining whether customer satisfaction is related to the on-board service the passenger receives. The data is in Table 1.20.

			Service	
		Snacks Only	Meals and Snacks	Snacks, Meals, and Entertainment
Passenger	High	2	15	17
Satisfaction	Moderate	16	11	10
	Low	21	15	5

Table 1.20 On-board Service Survey Data
Source: Fred Hulme

9. What fraction of all the passengers surveyed gave a flight a "High" satisfaction level and "Snacks Only" were served?
   a. 2/39 **b.** 2/34 **c.** 2/112 **d.** None of the above

10. If a flight had a service level of "Snacks Only," what fraction of the passengers gave a "High" satisfaction level?
    a. 2/39 **b.** 2/34 **c.** 2/112 **d.** None of the above

11. If the passengers gave a "High" satisfaction response, what fraction were on "Snacks Only" flights?
    a. 2/39 **b.** 2/34 **c.** 2/112 **d.** None of the above

Questions 12-16 pertain to the "Box Plot" in Figure 1.18 for the age of 250 customers of a business.

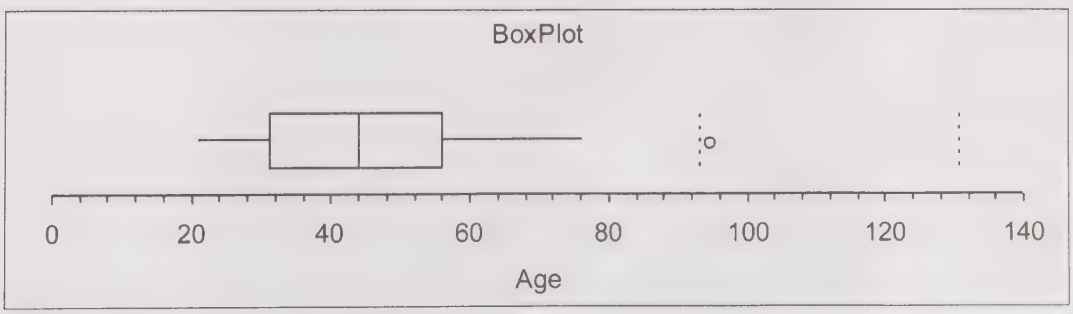


Figure 1.18 Box Plot of Customers' Ages
Source: Fred Hulme

12. Approximate median age of the customers?
a. 20 **b.** 32 **c.** 44 **d.** 56

13. Shape of the distribution of the ages?
a. Approximately symmetric **b.** Skewed left **c.** Skewed right **d.** Uniform

14. An "outlier" or unusual value around 95 years is indicated.
a. True **b.** False

15. The width of the "box" is equal to the
a. First quartile **b.** Median **c.** Third quartile **d.** IQR

16. 25% of the customers have an age that exceeds _____ years.
a. 20 **b.** 32 **c.** 44 **d.** 56

17. The appropriate data summary that could be used to determine if there is a relationship between the number of hours that students study for an exam and the numeric grade on the exam would be a(n)
a. Scatterplot
b. Histogram
c. Frequency polygon
d. Two-factor frequency table
e. Ogive

18. The appropriate data summary that could be used to determine if a student's gender is related to the student's major in college would be a
a. Scatterplot
b. Histogram
c. Frequency polygon
d. Two-factor frequency table
e. Ogive

Questions 19-25 pertain to Table 1.21 which summarizes data taken from measuring the amount of liquid detergent filled in bottles. The target amount was 50 oz. Using the frequency distribution, answer the following questions.



© Kendall Hunt Publishing Company

Ounces	Frequency	Relative Frequency (%)	Cumulative Frequency	Cumulative (%)
49.88 < 49.90	1	1.00	1	1.00
49.90 < 49.92	3	3.00	4	4.00
49.92 < 49.94	6	6.00	10	10.00
49.94 < 49.96	11	11.00	21	21.00
49.96 < 49.98	14	14.00	35	35.00
49.98 < 50.00	23	23.00	58	58.00
50.00 < 50.02	21	21.00	79	79.00
50.02 < 50.04	11	11.00	90	90.00
50.04 < 50.06	4	4.00	94	94.00
50.06 < 50.08	3	3.00	97	97.00
50.08 < 50.10	3	3.00	100	100.00

Table 1.21 Frequency Table of Ounces per Bottle
Source: Fred Hulme

19. How many bottles were included in the study?

20. Discuss the shape of the distribution of the amounts. The choices would be: approximately symmetric, skewed right, or skewed left.

21. Which interval would contain that amount such ½ of the bottles contained less than this amount and ½ of the bottles contained more than this amount?
[_____, _____]

22. What % of the bottles contained 49.94 oz or more but less than 50.04 oz?

23. What % of the bottles had < 49.90 oz or ≥ 50.08 oz? _____

24. Give the lower and upper limit of the interval that contained the most number of bottles?
[_____, _____]

25. If the manufacturer of this product wants to advertise that based on sample tests that 90% of the bottles will have greater than or equal to XX.XX oz. of liquid, what would the value of XX.XX be?
_____

Chapter 1—End of Chapter Answers

Short Answer/Discussion

1. Descriptive statistics' procedures involve creating tables, charts, and numeric summaries such as averages to condense what is usually a large amount of information. Inferential statistics' procedures involve using data from a **sample** to draw conclusions about the larger **population** or **process** from which the sample was taken.

2. True

3. True

4. Parameter

5. Statistic

6. Sampling error

7. A qualitative variable simply records an attribute or category; a quantitative variable results from a count or a measurement. Quantitative variables have values that contain numeric digits only.

8. A nominal variable essentially is a categorical variable. An example is type of housing: apartment, dorm room, duplex, single family dwelling. An ordinal variable expresses relationships that are greater than or less than. An example is the quality of food at Penland cafeteria: poor, average, good, great.

9. A discrete variable usually results from a count being taken. A continuous variable results from a measurement of a quantity. Discrete data has integer values only. Continuous data can be divided an infinite number of times.

10. **a.** qualitative, nominal
    b. quantitative, continuous
    c. quantitative, discrete
    d. qualitative, ordinal
    e. qualitative, nominal

11. So you know the appropriate tools you can use to most effectively summarize the data.

12. To get a quick summary of the data.

13. A frequency distribution for qualitative data will have the same number of categories as the qualitative variable. A frequency distribution for a quantitative variable may be arranged with an infinite number of categories or intervals.

14. It does not summarize the data enough.

15. Relative frequencies show the values as a percentage. They allow one to make summary statements of comparison very easily.

16. You can easily determine how many observations or what percentage of observations fall in an interval or category.

17. Cross tabulation or two-way table.

18. It is possible for either variable to be on the row or column, thus the name cross tabulation.

19. Histogram

20. Skewness is one term that may be used to tell the statistician something about the shape of the distribution. A distribution may be positively (skewed to the right) or negatively (skewed to the left) or not skewed at all.

21. Right

22. yes, right

23. So that you can overlay two or more frequency polygons on the same chart to compare multiple distributions.

24. What percent is less than or greater than a particular value? Left half of a bell-shaped distribution.

25. 50%

26. median

27. distribution shape

28. longer

29. equal

30. It appears outside the inner fence. Extreme data value appears outside the outer fence.

Answers to Problems

1(a) Qualitative/nominal

1(b)

Frequency Distribution—Qualitative			
Data	Frequency	%	
1	9	30.0	
2	12	40.0	
3	5	16.7	
4	4	13.3	
	30	100.0	

Chapter 1 Homework problem 1
Table 1.15 data

Table 1.22 Answer to Problem 1b
Source: Fred Hulme

2(a) quantitative/continuous

2(b)

Frequency Distribution—Quantitative						Chapter 1 Homework problem 2		
							Table 1.16 data	
	Data						Cumulative	
Lower		Upper	Midpoint	Width	Frequency	%	Frequency	%
20	<	22	21	2	9	30.0	9	30.0
22	<	24	23	2	11	36.7	20	66.7
24	<	26	25	2	5	16.7	25	83.3
26	<	28	27	2	3	10.0	28	93.3
28	<	30	29	2	2	6.7	30	100.0
					30	100.0		

Table 1.23 Answer to Problem 2b
Source: Fred Hulme

3(a) 63

3(b) 31.5%

3(c) 21

3(d) 10.5%

3(e) 191

3(f) 95.5%

3(g) 9

3(h) 4.5%

4(a) 20%

4(b) 50%

4(c) 50%

5(a) 2.6

5(b) .4

5(c) (2, 3.6)

5(d) (1.4, 4.2)

5(e) See notes from class

5(f) Yes, no.

Self-Check Answers

1. a	2. d	3. c	4. b	5. b	6. c	7. c	8. c	9. c	10. a
11. b	12. c	13. c	14. a	15. d	16. b	17. a	18. d	19. 100	

20. approximately symmetric

21. 49.98, 50.00

22. 80%

23. 4%

24. 49.98, 50.00

25. 49.94

CHAPTER

2

Describing Data—Numerical Summaries



© s4svisuals/Shutterstock.com

In addition to tables and charts, data may also be summarized through numerical calculations. Replacing the original data with a few key measures explains a set of data for an observer. For quantitative variables, measures exist that describe the **central location** and **variation** or **spread** of the data. Measures of central location are used to answer the question—"What is a typical value of the data or what value does most of the data rally around?" Commonly used measures of central location are the arithmetic **mean** (average), the **median**, and the **mode**. Measures of central location are often used to compare groups of data. Measures of variation answer how spread out the data is. Groups of data may be equivalent based on a measure of central location but drastically different based on a measure of variation. For example, the mean or average rate of return of two investments, A and B, over the last ten years may both be 8.5% (see chart below), but the rate of return of investment A is more spread out over the ten years. Looking at Figure 2.1 investment A would be considered riskier or less stable than investment B.


© Travis Wolfe/Shutterstock.com

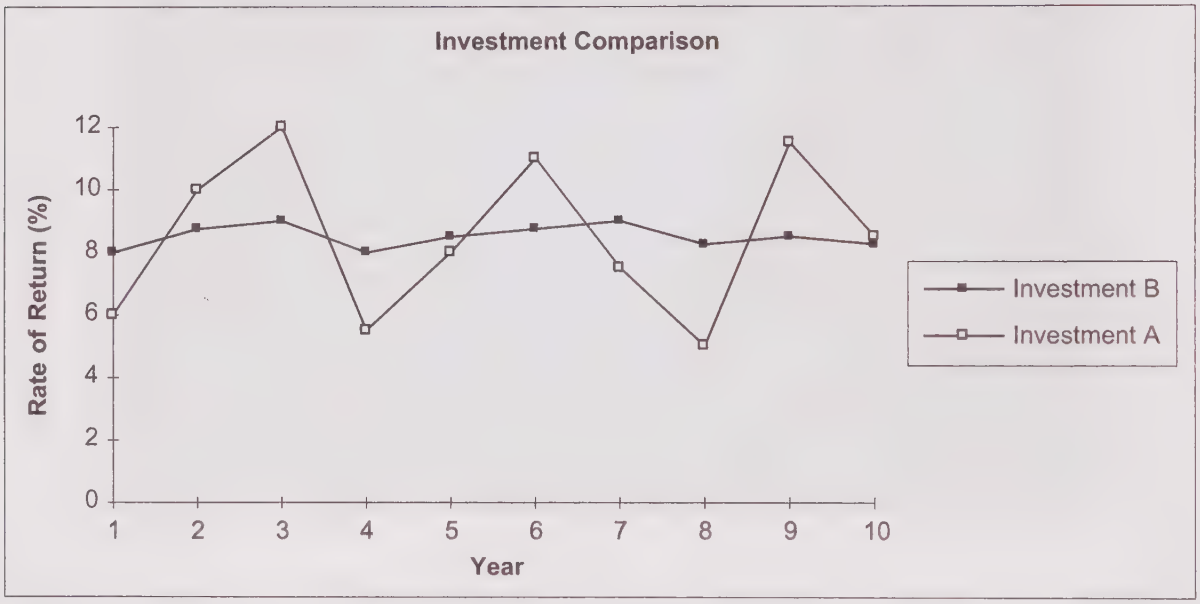


Figure 2.1 Graphical Comparison of Investments
Source: Fred Hulme

Common measures of variation or dispersion are the **range**, the **standard deviation**, and the **interquartile range**.

Measures of Central Location

The most common measure of central location is the **arithmetic mean** or **average**. To calculate a mean or average of a set of numbers, you simply sum or add up the numbers and divide the sum by the amount of numbers.

$$\text{Mean or average} = \text{Sum} \div \text{Count}$$

A more formal expression for calculating the mean of a variable named, x, that has n values is:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{(x_1 + x_2 + x_3 + \ldots + x_n)}{n}$$

Calculating Mean Waiting Time



© Sorbis/Shutterstock.com

Let x represent the number of minutes that a customer waits in line to be served at a cafeteria.

Data: 5 10 8 12 20 6 5 10 10

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = (5 + 10 + 8 + 12 + 20 + 6 + 5 + 10 + 10)/9$$

$$\bar{x} = 86/9$$

$$\bar{x} = 9.56 \text{ minutes}$$

A common naming convention for means or averages is to express the mean of a variable, x, as an x with a bar over it, $(\bar{x})$. This symbol is read as x bar, which represents the mean of a variable named, x. The units of a mean are the same units as the original data.

You will often hear the mean of some variable being quoted in mass media such as newspaper articles or television news programs. The reason for the popularity of the mean in representing a larger set of numbers is:

(a) most people have an intuitive understanding of a mean or average;
(b) the mean is unique and always exists for a quantitative variable;
(c) the mean uses information from all the values of a variable;
(d) means can be used in further statistical analysis for comparisons and predictions; then the reliability of the comparisons and predictions can be assessed using the means;
(e) a mean provides a stable measure of central location in that a mean is not as sensitive to chance variation as some other measures.

The calculation of a simple mean or average assumes that each data value should receive an equal amount of importance. Some situations require a varying degree of importance for some of the values. This situation requires what is known as a **weighted mean or average**.

Calculating a Weighted Mean



© Artem Kovalenco/Shutterstock.com

If a class has a homework component and an exam component where the homework is to count 30% and the exams count 70%, the weighted mean grade for a student with an 80 homework average and a 90 exam average would be:

Weighted Mean Grade = $(30 * 80 + 70 * 90)/(30 + 70) = 87$.

The simple mean would be $(80 + 90)/2 = 85$.

The formula for a weighted mean is:

$$\textbf{Weighted Mean} = \bar{x} = \frac{\sum_{i=1}^{k} w_i x_i}{\sum_{i=1}^{k} w_i}$$

A mean is not a good measure of central location if the data contains some atypical values or outliers. The mean is a distorted measure in the presence of outliers. The mean will be drawn toward the outliers and misrepresent the overall data. An example of this would be the mean salary of employees in a small business. The owner/manager might make $75,000 a year, and the other employees might make between $18,000 and $25,000. The mean salary would be inflated due to the one high salary and not be representative of all salaries. In data that contains outliers, a better measure of central location would be the **median**. The median, $\tilde{x}$, is calculated by ordering the data from the lowest to the highest and choosing the middle value. This calculation is straightforward if you have an odd number of values because you will have a unique middle number. If you have an even number of numbers, find the midpoint of the two middle values to be the median.

Calculating a Median

Using the cafeteria waiting time data

Data:	5	10	8	12	20	6	5	10	10
Ordered Data:	5	5	6	8	10	10	10	12	20

The median, $\tilde{x}$, would equal 10. You can calculate the position of the median by computing, $(n + 1)/2$, where n is the number of numbers. For the above data, the position of the median is given by $(9 + 1)/2 = 5$th number in order.

If we change the 20 in the data to say 50, the median will still be equal to 10. The mean would go up, but the median is not sensitive to extreme values.

The third measure of central location is known as the **mode**. The mode is simply that number that occurs the most number of times. The mode may not be a unique value in that there could be a tie for the number that occurs with the highest frequency. For some data the mode may not exist if no two numbers repeat. This could be quite likely for continuous data.

Calculating a Mode



© Sorbis/Shutterstock.com

Using the cafeteria waiting time data

Data:	5	10	8	12	20	6	5	10	10

The mode would be equal to 10 in that the value of "10" occurs with highest frequency—3 times.

Measures of Variation or Spread

The simplest measure of variation or the spread of data is the **range**. The range is defined as the largest number minus the smallest number. The more spread out the values of a variable, the larger the range.

Finding the Range

Using the cafeteria waiting time data

Data: 5 10 8 12 20 6 5 10 10

The range equals 15 ⟶ (20 − 5). One problem with the range is that the range does not lend itself to be used widely with further statistical procedures. The ease of computation is the greatest strength of the range. Quality control charts that monitor if a process is in control use the range as a measure of dispersion. The units of the range are the same as the units of the original data.

Another measure of variation is a calculated value known as the **variance**. The variance measures the dispersion of a set of numbers around the mean of the numbers. The variance is a nonnegative number that becomes larger for sets of numbers that are more widely dispersed. The formula for the variance is:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

The symbol, s^2, is often used to represent the variance of a sample set of numbers.7

Calculating the Standard Deviation

Using the cafeteria waiting time data in Table 2.1, the standard deviation may be found.

Data: 5 10 8 12 20 6 5 10 10

X	$X - \bar{X}$	$(X - \bar{X})^2$
5	(5 – 9.56) = –4.56	20.7936
10	(10 – 9.56) = 0.44	0.1936
8	(8 – 9.56) = –1.56	2.4336
12	(12 – 9.56) = 2.44	5.9536
20	(20 – 9.56) = 10.44	108.9936
6	(6 – 9.56) = –3.56	12.6736
5	(5 – 9.56) = –4.56	20.7936
10	(10 – 9.56) = 0.44	0.1936
10	(10 – 9.56) = 0.44	0.1936
	$\sum_{i=n}^{n}(x_i - \bar{x})^2 =$	172.2224

Table 2.1 Calculating a Standard Deviation
Source: Fred Hulme

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

$$s^2 = 172.2224/(9-1)$$

$$s^2 = 21.53 \text{ min}^2$$

A computationally more efficient formula for the variance is:

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}$$

Using this formula with the cafeteria waiting time data in Table 2.2.

X	X^2
5	25
10	100
8	64
12	144
20	400
6	36
5	25
10	100
10	100
$\sum_{i=1}^{n} x_i = 86$	$\sum_{i=1}^{n} x_i^2 = 994$

Table 2.2 Summarizing Data
Source: Fred Hulme

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}$$

$$s^2 = \frac{994 - \frac{(86)^2}{9}}{9-1}$$

$$s^2 = \frac{994 - 821.778}{8}$$

$$s^2 = \frac{172.222}{8}$$

$$s^2 = 21.53 \text{ min}^2$$

There are a couple of problems with using the variance as a measure of dispersion. The units of the variance are the squared units of the original data. Also, it is hard to have any intuition about the relative size of the variance. Another measure known as the **standard deviation** overcomes these problems. The standard deviation is defined as the square root of the variance. The standard deviation is often denoted by the letter, s, where,

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

or

$$s = \sqrt{s}$$

For the cafeteria waiting time data,

$$s = \sqrt{21.53} \text{ minutes}$$

$$s = 4.64 \text{ minutes.}$$

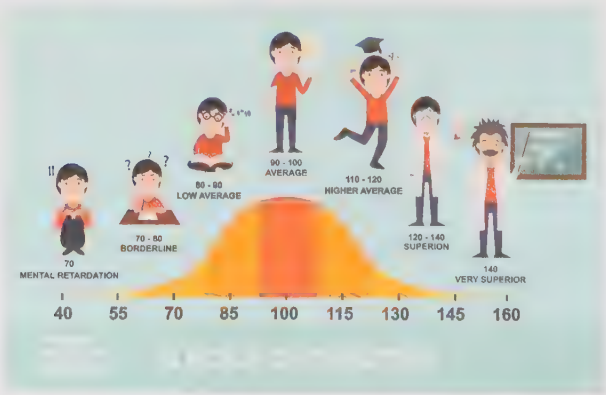
Chebyshev's Theorem



© miami beach forever/Shutterstock.com

A theorem or rule from mathematics provides an intuitive feel for the relative size of the standard deviation measure. The rule is named after a nineteenth century Russian mathematician, **P. L. Chebyshev**. The rule states that if you calculate for any set of numbers, the mean ±3 standard deviations, that the resulting limits will contain at least 8/9 (approximately 90%) of the possible values. This rule applies in all cases. The distribution of the numbers can be either symmetric or skewed. The rule is usually conservative and often produces limits that contain 100% of the values. Therefore, by knowing the mean and standard deviation of a variable, you can calculate a range of reasonable values that the variable will assume most of the time.

Using Chebyshev's Theorem



© jehsomwang/Shutterstock.com

Assuming the mean IQ for a large group of people is 100 with a standard deviation of 10, what upper and lower limit will contain most of the IQ's?

Limits: Mean − (3 × Standard Deviation), Mean + (3 × Standard Deviation)

Limits: 100 − 3 (10), 100 + 3(10)

Limits: 100 − 30, 100 + 30

Limits: 70, 130

You would expect at least 8/9 (approximately 90%) of all the people to have an IQ between 70 and 130. Therefore, it would be considered rare if a person had an IQ of 150 or more from this population. The standard deviation when used in combination with the mean of a variable can be used to establish a range of believable values. If you observe many values of the variable outside the believable range, you begin to wonder if something has changed. The exact statement of Chebyshev's Theorem is:

The proportion of any numerical distribution that lies within k standard deviations of the mean is at least $1 - \frac{1}{k^2}$, where k is any positive number larger than 1.

For an intuitive feel for the meaning of the measure, standard deviation, when presented with the standard deviation of a numeric variable, calculate the value of the standard deviation times six. The resulting number will be an approximation for the range (maximum − minimum) for the variable. You will probably be able to attach more intuition as to what the range of a variable conveys as opposed to the standard deviation. For the above example, you would approximate the range of the IQ's by six times ten or sixty points.

Empirical Rule

If a numeric variable has a **mound-shaped** or **bell-shaped** frequency distribution, the range defined by the mean plus or minus one standard deviation will contain approximately 68% of the possible values of the variable. Within two standard deviations of the mean there should be approximately 95% of the values (Figure 2.2), and within 3 standard deviations there should be approximately 100% of the values (Figure 2.3).

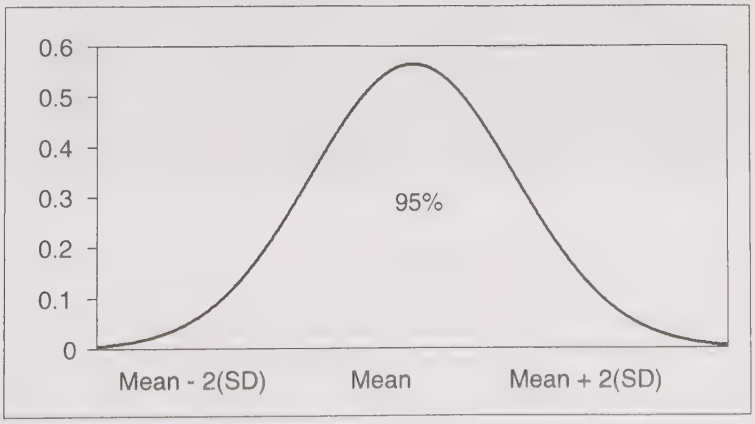


Figure 2.2 Empirical Rule 2s

Source: Fred Hulme

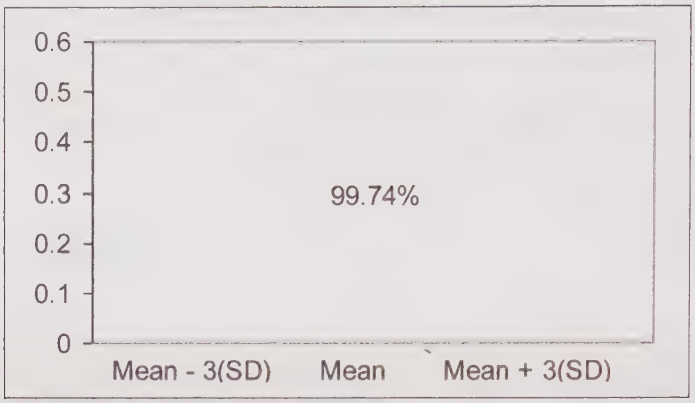


Figure 2.3 Empirical Rule 3s
Source: Fred Hulme

This rule is known as the **empirical rule**. Therefore, if you can assume that the distribution is bell-shaped, you can increase the estimate of the proportion of numbers that will fall between the limits defined by multiples of the standard deviation around the mean.

Applications to Quality Control



© Dusit/Shutterstock.com

A major application of the measures of central location and variation is in the field of quality assurance or quality control. A major tool used in monitoring the quality of the output of some system or process is a **control chart**. A control chart is a chart or graph of data that can be used to determine if a process should be considered "in control". Outliers and trends in the data can be detected visually from a control chart. Upper and lower limits are established in developing a control chart. As long as the process yields results falling between the acceptable limits, you conclude the process is "in control." Once the process begins producing results outside the acceptable

limits, you may want to stop the process and determine what is causing the unusual results. With any process, you will expect some variation in the results, but the amount of variation should be within some reasonable range.

Several types of control charts are used in quality control; however, the most frequently used is a control chart for the mean of a process. The limits for the control chart are calculated differently depending upon what is known about the process.

Control chart for a mean (process standard deviation, σ, is known)

Lower control limit (LCL) and upper control limit (UCL) are given by:

$$LCL = \mu - \frac{3\sigma}{\sqrt{n}}$$

$$UCL = \mu + \frac{3\sigma}{\sqrt{n}}$$

The above limits should contain approximately 99.74% of the sample means taken from the process based on a sample size of n observations or items.

Finding the *LCL* and *UCL*



© Avatar_023/Shutterstock.com

A machine is used to fill liquid in a bottle. The process is considered in control if it appears the mean amount filled in each bottle is 16 oz. The process standard deviation is known to be 0.06 oz. The process is monitored each hour by taking a sample of 36 bottles and measuring the amount of fill in each bottle. The mean of the 36 sampled bottles is then calculated. If the sample mean falls within the prescribed limits, the process is considered in control.

The lower and upper control limits are given by:

$$LCL = 16 - \frac{3 * 0.06}{\sqrt{36}} \qquad LCL = 16 + \frac{3 * 0.06}{\sqrt{36}}$$

$$LCL = 16 - .03 \qquad UCL = 16 + .03$$

$$LCL = 15.97oz. \qquad UCL = 16.03 \text{ oz}$$

Table 2.3 represents the sample means of 12 samples (each sample contains measurements made on 36 randomly selected bottles, i.e., the first sample mean of 15.98 is not the amount of liquid in a single bottle, but the mean or average of 36 bottles.)

Sample	Sample Mean
1	15.98
2	16.02
3	16.01
4	15.99
5	15.98
6	16.01
7	15.98
8	16.00
9	16.02
10	16.03
11	16.04
12	16.05

Table 2.3 Control Chart Data
Source: Fred Hulme

Figure 2.4 depicts the control chart for the mean with the LCL, UCL, and the 12-sample means plotted.

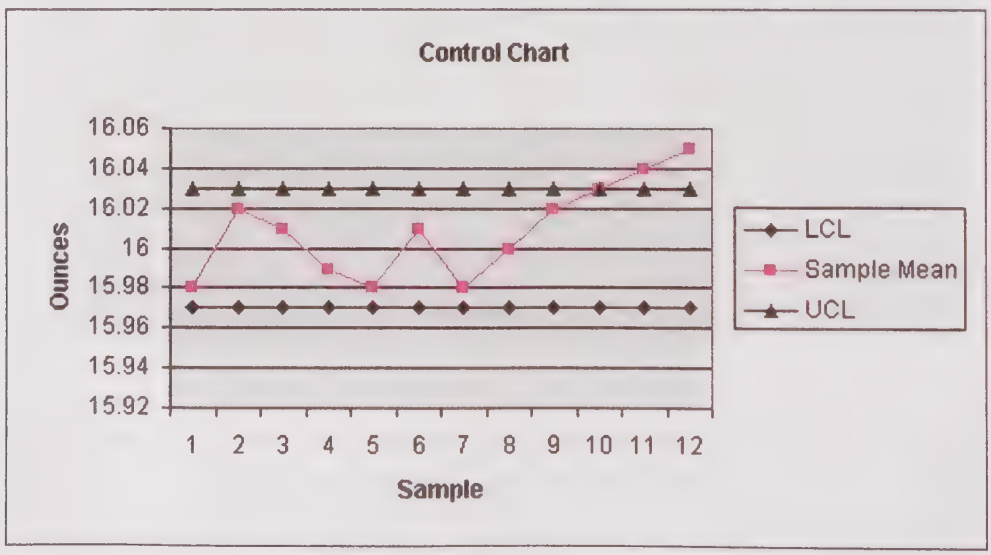


Figure 2.4 "Out of Control Process" Control Chart
Source: Fred Hulme

Interpretation of the control chart

The means of samples 1–9 are within the prescribed limits implying that the process was in control while those samples were collected. Sample 10 is right on the UCL, and samples 11 and 12 are clearly above the UCL. If the process were still in control while samples 10–12 were being collected, there was only a 0.0026 chance (less than 3 out of a 1,000) of getting any one of the three-sample means. The increasing trend displayed by samples 9–12 may indicate that the machine is falling more and more out of control.

Control chart for a mean (process standard deviation, σ, is unknown)

In most new processes, the process standard deviation will be unknown. In these cases, you must estimate the process standard deviation. Repeated random samples taken while the process is considered under control and calculating either the sample standard deviations or sample ranges will suffice. The sample range is used more often than the sample standard deviation because of the lack of statistical training of the personnel making the calculations.

Lower control limit (LCL) and upper control limit (UCL) are given by:

$$LCL = \mu - A_2\bar{R} \qquad UCL = \mu + A_2\bar{R}$$

where μ is either known or assumed or estimated by repeated sampling while the process is considered under control, and $\bar{R}$ is the average of the ranges from the repeated samples. For a sample size of 10, the factor A_2 is 0.308. Tables exist for the factors of A_2 using different sample sizes.

MPG Control Chart



© Liz Van Steenburgh/Shutterstock.com

Turner Trucking recently bought a large fleet of new trucks. The trucks are all identical in make and model. A control chart is desired to help monitor the efficiency of the trucks based on gas mileage. Although the trucks are identical models, the gas mileage will vary from truck to truck. In order to establish the limits for a control chart for the mean gas mileage of the trucks, five groups of ten trucks each were observed with respect to gas mileage. The results are given below in Table 2.4.

Truck	Group 1	Group 2	Group 3	Group 4	Group 5
1	17.59	15.54	16.15	16.93	16.06
2	17.14	15.05	15.67	16.85	16.48
3	16.03	15.42	15.65	16.80	16.05
4	16.64	15.92	16.25	15.97	17.09
5	16.00	15.40	15.90	16.51	16.53
6	16.45	15.41	16.19	16.33	15.93
7	16.30	16.03	15.94	13.53	14.63
8	15.82	15.57	16.59	16.36	15.35
9	15.06	15.97	15.28	16.43	15.45
10	16.32	14.84	15.99	15.46	16.37
Mean	16.34	15.52	15.96	16.12	15.99
Maximum	17.59	16.03	16.59	16.93	17.09
Minimum	15.06	14.84	15.28	13.53	14.63
Range	2.53	1.19	1.31	3.40	2.46
Group Mean, $\bar{\bar{X}}$		15.98			
Average Range, $\bar{R}$		2.18			

Table 2.4 Truck Statistics
Source: Fred Hulme

Lower control limit (LCL) and upper control limit (UCL) are given by:

$$LCL = \bar{\bar{X}} - A_2\bar{R} \qquad UCL = \bar{\bar{X}} + A_2\bar{R}$$

$$LCL = 15.98 - 0.308 * 2.18 \qquad UCL = 15.98 + 0.308 * 2.18$$

$$LCL = 15.98 - 0.67 \qquad UCL = 15.98 + 0.67$$

$$LCL = 15.31 \text{ mpg} \qquad UCL = 16.65 \text{ mpg}$$

Other Numerical Summaries

Percentiles

Another common numerical summary is a measure known as a percentile. Students' college entrance exam (SAT or ACT) percentile scores were reported in addition to the raw scores. A percentile score corresponds to the percentage of scores that are less than a particular raw score. For instance, the percentile score for a composite ACT score of 27 might be 85%. This means that 85% of the students in the comparison group made less than 27 on the composite score. Commonly reported percentiles are the 5th, 10th, 25th (first quartile or Q_1), 50th (the median or Q_2), 75th (third quartile or Q_3), 90th, and 95th.



© Brian A Jackson/Shutterstock.com

Exam Score Percentiles

In a class of nine students, assume the following test scores.

Score:	92	76	52	87	79	83	86	74	68
Ordered Scores:	52	68	74	76	79	83	86	87	92

In Table 2.5 since 4 of the scores are less than 79 and 4 of the scores are greater than 79, the percentile for a score of 79 would be 4/8 * 100% or 50%. Since 7 of the scores are less than 87 and only 1 score is greater than 87, the percentile for a score of 87 would be 7/8 * 100% or 87.5%.

	Sorted		
Score	Score	Rank	Percentile (%)
92	92	1	100.0
76	87	2	87.5
52	86	3	75.0
87	83	4	62.5
79	79	5	50.0
83	76	6	37.5
86	74	7	25.0
74	68	8	12.5
68	52	9	0.0

Table 2.5 Percentile Chart

Source: Fred Hulme

Interquartile Range

As mentioned earlier for variables with skewed distributions, the more appropriate measure of central location is the median or 50th percentile as opposed to the arithmetic mean. The arithmetic mean will be biased toward the atypical values (outliers) of the distribution. In this case, it does not seem appropriate to measure the variation or spread of a distribution relative to the mean.

A value that is often used to measure the spread of skewed data is the **interquartile range, IQR,** which is the difference between the 75th and 25th percentile.

$$\text{Interquartile Range} = 75\text{th percentile} - 25\text{th percentile}$$

$$\text{or Interquartile Range} = Q_3 - Q_1$$

The interquartile range measures the range of the middle 50% of the values in a distribution. The median plus or minus a constant multiple of the interquartile range is often used to represent a range of reasonable values. Values occurring outside these boundaries are further scrutinized in an attempt to explain the unusual variation associated with them.

IQR for Exam Scores

Using the test score data from the previous example, the 75th percentile or Q_3 equals 86. The 25th percentile or Q_1 equals 74. Therefore, the interquartile range is equal to 12 (86 − 74). This means that the middle 50% of the test scores has a range of 12 points. The interquartile range should not be confused with the median, the 50th percentile. The median score, 79, is that score such that 50% of the scores are less than it and 50% of the scores are greater than it.



© Monkey Business Images/Shutterstock.com

Calculating Numerical Summaries with Microsoft Excel

Microsoft Excel has over 70 stand-alone statistical functions available for describing data. A list of the more commonly used functions and their required inputs is given below and in Table 2.6.

Function Name	Purpose of Function	
=COUNT(range of cells)	Counts the number of numbers in a range of numbers	
=MIN(range of cells)	Calculates the smallest number from a range of numbers	
=MAX(range of cells)	Calculates the largest number from a range of numbers	
=AVERAGE(range of cells)	Calculates the average or mean of a range of numbers	
=MEDIAN(range of cells)	Calculates the middle number of a range of numbers	
=MODE(range of cells)	Calculates the most common number from a range of numbers	
=STDEV(range of cells)	Calculates the standard deviation a range of numbers	
=QUARTILE(range of cells, *k*)	Calculates the *k*th quartile of a range of numbers where $k = 0, 1, 2, 3,$ or 4	
=PERCENTILE(range of cells, k)	Calculates the *k*th percentile of a range of numbers, where $0 \le k \le 1$	

Table 2.7 uses the Excel functions to calculate various statistics on Current Purchases for e-Tronics.

=COUNT(Purch2011)	75
=MIN(Purch2011)	42.50
=MAX(Purch2011)	4,000.00
=AVERAGE(Purch2011)	1,217.66
=MEDIAN(Purch2011)	1,016.96
=MODE(Purch2011)	#N/A
=VAR(Purch2011)	916,037.78
=STDEV(Purch2011)	957.10
=QUARTILE(Purch2011, 1)	463.47
=QUARTILE(Purch2011, 3)	1,943.48
=PERCENTILE(Purch2011, 0.25)	463.47
=PERCENTILE(Purch2011, 0.75)	1,943.48

Table 2.6 Excel Commands
Source: Fred Hulme

Using the above functions, a table of numeric summaries for the "e-Tronics" customer data is given below.

	Age	Income	Children	Last Year	Current Purchases
Count	75	75	75	75	75
Minimum	21	10,600	0	0	42.50
Maximum	75	118,300	3	3,630	4,000
Mean	43.04	55,552	0.84	719.39	1,217.66
Median	40	55,500	0	473	1,016.96
Mode	55	63,300	0	0	#N/A
Standard deviation	15.69	28,168.14	1.04	804.30	957.10
First quartile	28	36,000	0	0	463.47
Third quartile	54.5	74,650	2	1,212.20	1,943.48

Table 2.7 Excel Output for e-Tronics
Source: Fred Hulme

Salary Distributions



© Cartoon Resource/Shutterstock.com

In a salary study, fifty employees were randomly sampled, and their annual salary recorded. The employees included primarily workers paid an hourly wage, but some salaried managers were also included. The annual salaries for the 50 employees are in Table 2.8.

$48,157	$27,516	$53,321	$15,155	$33,971
43,739	21,103	31,631	58,052	35,165
22,433	22,828	36,476	42,042	51,242
29,297	59,456	24,819	20,007	34,415
20,556	26,483	28,809	33,218	30,486
39,274	30,080	27,948	16,596	41,062
72,292	31,072	44,947	18,127	37,087
26,100	37,874	24,218	52,002	47,680
32,210	39,088	46,021	23,578	18,652
64,787	42,171	25,659	19,306	21,830

Table 2.8 Salary Data
Source: Fred Hulme

From Figure 2.5, you can see that the distribution of salaries is skewed right. The higher salaries of the managers are reflected in the upper tail of the distribution. These few higher salaries would distort the mean or average salary. Therefore, the mean would not be a good measure of the typical salary.

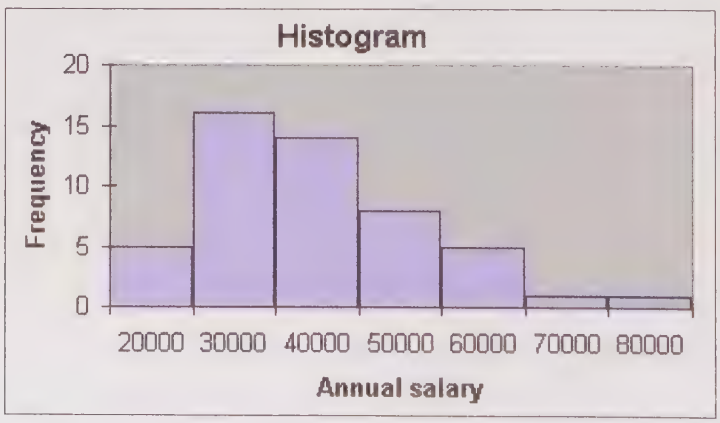


Figure 2.5 Histogram of Salaries
Source: Fred Hulme

Using the Excel functions with the "Salary" data results in Table 2.9.

	Salary
Count	50
Minimum	15,155
Maximum	72,292
Mean	34,601
Median	31,921
Mode	#N/A
Variance	175,363,075
Standard deviation	13,242
First quartile	24,368
Third quartile	42,139

Table 2.9 Excel Output for Salaries
Source: Fred Hulme

You can also use the **MegaStat** command, **Descriptive Statistics** in Figure 2.6, to calculate various numerical summaries.

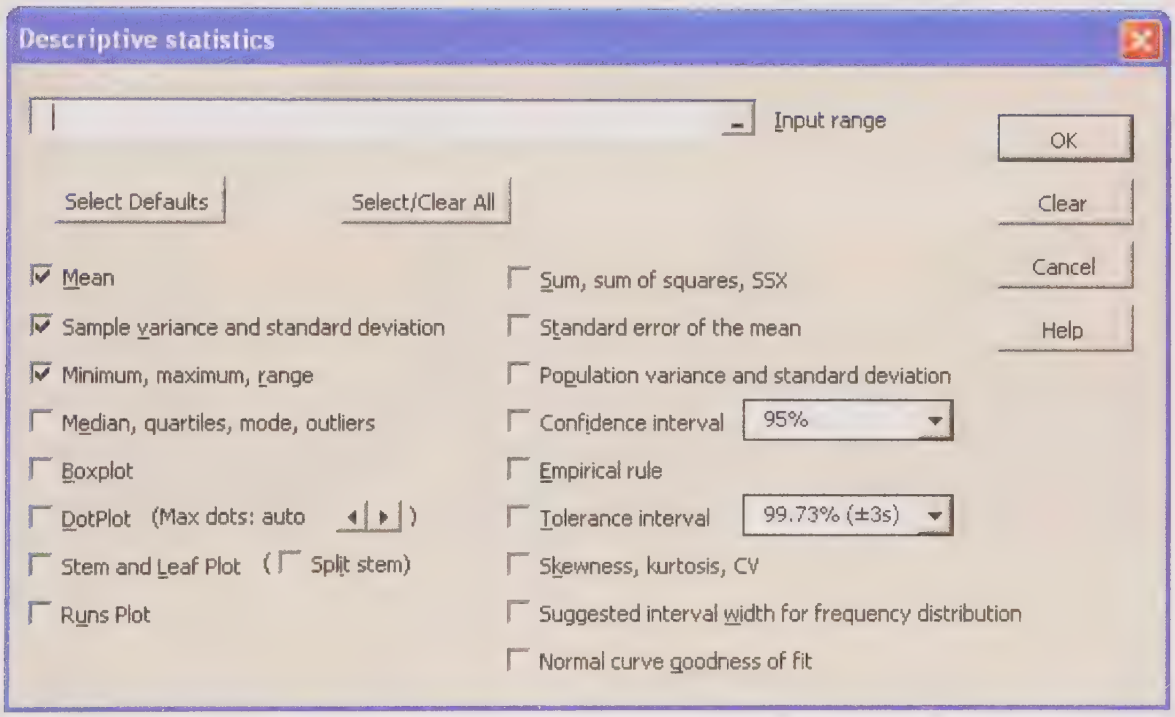


Figure 2.6 MegaStat Descriptive Statistics Dialog Box
Source: Fred Hulme

The options for calculating the mean, the variance and standard deviation, minimum, maximum, and the range are chosen automatically. Other options such as median, quartiles, mode, outliers, Boxplot, and Empirical rule are also often chosen to provide numerical summaries.

The MegaStat "Descriptive Statistics" dialog box below in Figure 2.7 has "input range" defined as F1:F76 since the values for Current Purchases from the "e-Tronics" data were located in cells F1 through F76.

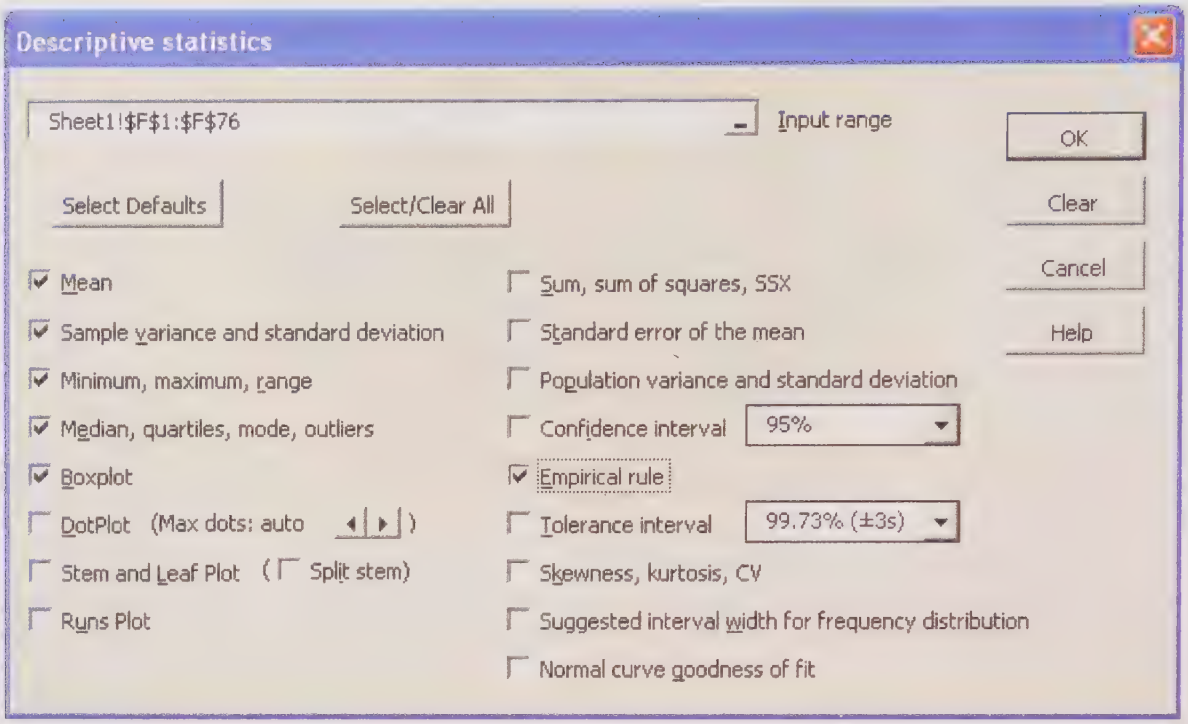


Figure 2.7 Selection of Statistics
Source: Fred Hulme

The results of the **MegaStat** command, **Descriptive Statistics,** for Current Purchases are displayed in Table 2.10.

Descriptive Statistics				
	Current Purchases			**Current Purchases**
Count	75		Mean – 3s	-1,653.6395
Mean	1,217.6564		Mean + 3s	4,088.9523
Sample variance	916,037.7791		Percent in interval (99.73%)	100.0%
Sample standard deviation	957.0986			
Minimum	42.5		1st quartile	463.4700
Maximum	4000		Median	1,016.9600
Range	3,957.5		3rd quartile	1,943.4800
			Interquartile range	1,480.0100
Empirical rule			Mode	#N/A
Mean – 1s	260.5578			
Mean + 1s	2,174.7550		Low extremes	0
Percent in interval (68.26%)	73.3%		Low outliers	0
Mean – 2s	-696.5409		High outliers	0
Mean + 2s	3,131.8537		High extremes	0
Percent in interval (95.44%)	94.7%			

Table 2.10 MegaStat Output for e-Tronics
Source: Fred Hulme

Comparing Two or More Groups Based on Means

One application of statistical methods involves comparing two or more groups or treatments of data. For example, did one teaching method outperform another method? Of three gasoline additives, which one improved gas mileage the most? Which advertisement stimulated sales the most? To begin to answer these questions you should summarize the data for the different groups or treatments in a comparative manner to get a feel for any differences if they exist.

Comparing Purchases by Gender



© fizkes/Shutterstock.com

Does the gender of the customer make a difference with respect to the mean or average Current Purchases? You can ask Excel to calculate the mean or average amount for the males versus the females with the command, **Insert/PivotTable/PivotTable**. You would define the "Row" fields of the pivot table by the variable, ***Gender***, and you would define the "Value" fields of the pivot table by the amount of Current Purchases. You would also need to change the type of summary desired from "**Sum of Current Purchases**" to "**Average of Current Purchases**." You would use the "**Value Field Settings**" option to change the summarizing method from a sum to an average. See Figure 2.8.



Figure 2.8 Building a Pivot Table
Source: Fred Hulme

The result of the **PivotTable** command would be as follows in Table 2.11.

Average of Current Purchases	
Gender	**Total**
F	1,030.16
M	1,400.22
Grand Total	1,217.66

Table 2.11 Pivot Table
Source: Fred Hulme

You would probably conclude that there is a significant difference between the average (mean) Current Purchases by the males, $1,400.22 versus the females, $1,030.16. The gender seems to matter with respect to how much the customer spends on average based on Table 2.11.

Exercises

Short Answer/Discussion

1. What are the common measures of central location for quantitative data?

2. What are the common measures of variation or dispersion for quantitative data?

3. When describing data through numerical summaries, why is it dangerous to only take central location into account?

4. Why is the arithmetic mean or average so widely used as a measure of central location of quantitative data?

5. Give an example where a simple mean is not a sufficient measure of central location.

6. Why is the median a more appropriate measure of central location than the mean when the data has "outliers"?

7. With regard to the mean and standard deviation of a quantitative variable, within what range would you expect most of the values of the variable to fall?

8. Explain what information a "control chart" conveys about a process that can be measured with a numeric outcome?

9. What numeric summaries are used in constructing a control chart for the mean of a process when the process standard deviation is unknown?

10. What information does a percentile score convey? For example, if a person's percentile score on an exam is eighty, what does this mean?

11. What is another name for the 75th percentile of a set of numbers?

12. What is the definition of the interquartile range? What does it represent? Is it a measure of central location or dispersion?

13. Which Microsoft Excel function can be used to compare the means of a variable for two or more groups?

Problems

1. For the data below, calculate the following summary statistics: the mean, median, mode, range, variance, and standard deviation.

   1 3 3 2 6

2. Suppose the mean length of employment of a salesclerk in a major department store chain is 1.6 years with a standard deviation of six months. Mary has been working as a salesclerk for 30 months. Is her length of employment unusual? **(Help: Calculate if Mary's length of employment is more than 3 standard deviations away from the mean length of employment.)**

3. The number of days for a business to receive an order (sometimes referred to as lead time) can be summarized by the following frequency distribution based on the last 50 orders in Table 2.12.

Lead time (days)	3	4	5
Frequency of orders	5	30	15

Table 2.12 Lead Time Demand
Source: Fred Hulme

   Calculate the "weighted mean" lead time. (Help: Use the "frequency of orders" as the weights in the weighted mean formula.)

4. A rental car company wants to use control charts in monitoring the number of miles driven on its "unlimited" mileage contracts. The company samples ten rentals a day for five days. The mean and range for each of the five days are listed in Table 2.13.

Day	Mean	Range
1	512.4	133.7
2	511.2	115.0
3	482.3	134.6
4	500.9	225.0
5	496.3	97.3

Table 2.13 Rental Car Data
Source: Fred Hulme

   Calculate the upper and lower control limit for the mean based on the sample data.

5. A study made by the manager of a mail-order business recorded the length of time (in minutes) to take telephone orders.

11.9	11.0	8.7	10.5	9.1	9.2	8.8	11.9	4.9	11.0
10.8	6.0	10.8	10.8	6.5	3.5	4.2	9.7	10.9	6.2

Table 2.14 Telephone Order Data
Source: Fred Hulme

Using the following relationship, approximate the standard deviation of the length of time (in minutes) that was needed to take the orders.

$$\text{Standard Deviation} \approx \frac{\text{Range}}{6}$$

6. Table 2.15 contains numerical summaries on the college grade point average (GPA) of 200 recent graduates.

Mean	2.799
Median	2.8
Mode	2.8
Standard deviation	0.2506
Sample variance	0.0628
Range	1.5
Minimum	2

Table 2.15 College GPAs
Source: Fred Hulme

a. What is the average grade point average?

b. What grade point average is such that 1/2 of the values are smaller and 1/2 are larger?

c. What value for the grade point average occurred the most number of times?

d. Give three measures of dispersion or variation in the grade point average. Give the name as well as the value of the measures.

e. What does the fact that the mean, median, and mode are all approximately the same value (2.8) indicate about the data?

7. Another sample of 200 graduates was taken with the gender and the college grade point average of the graduate recorded in Table 2.16. Summaries of the count, mean, and standard deviation of the two groups defined by gender are below.

Gender	Count	Mean GPA	SD GPA
Female	80	2.97	0.265
Male	120	2.60	0.232

Table 2.16 Pivot Table of GPAs by Gender
Source: Fred Hulme

a. How many of the 200 graduates were female? Male?

b. Would you consider the two groups different based on mean grade point average?

c. Would you consider the two groups different based on the dispersion or variation of the grade point average?

8. The mean score of an exam is 75 with a standard deviation of 5. Answer (a) and (b) using **Chebyshev's Rule**.

a. At least 8/9 of the scores will fall between what two scores?

b. At least what fraction of the scores will fall between (65, 85)?

9. An oil change business uses an average or mean of 300 quarts of oil per day with a standard deviation of 20 quarts. Answer the following questions assuming the daily usage has a mound-shaped (bell-shaped) distributed. (Help: Apply the "**Empirical Rule**.")

a. Estimate the fraction of the days that the demand will be between 260 and 340 quarts.

b. Estimate the fraction of the days that the demand will be between 240 and 360 quarts.

Questions 10–15 pertain to the data in Table 2.17and 2.18. A study was made of homes that sold within the last six months. The following data was collected for each home.

(1) Selling price in dollars—(PRICE)

(2) Size of home in square feet—(SQFT)

(3) Age of home in years—(AGE)

(4) Number of bedrooms—(BEDRMS)

(5) Location (A, B, C, D, E, F, G)—(LOCATION)

(6) Whether the home has an attached garage or not (Yes or No)—(GARAGE)

Descriptive Statistics for "PRICE"

Price	
Mean	$90,030
Median	85,721
Mode	#N/A
Standard deviation	27,176
Range	119,710
Minimum	44,301
Maximum	164,011
Sum	11,253,808
Count	125
Q_1—First quartile	68,531
Q_3—Third quartile	108,496
IQR—Interquartile range	39,965

Table 2.17 Home Data
Source: Fred Hulme

Mean and Standard Deviation of "PRICE" by "GARAGE"

Garage	Data	Total
N	Count of PRICE	55
	Average of PRICE	$ 81,935
	SD of PRICE	$ 21,057
Y	Count of PRICE	70
	Average of PRICE	$ 96,391
	SD of PRICE	$ 29,782
Total Count of PRICE		125
Total Average of PRICE		$ 90,030
Total SD of PRICE		$ 27,176

Table 2.18 Garage Presence by Price Statistics
Source: Fred Hulme

10. What price divides the distribution in half (i.e., 1/2 of the orders are processed in less time than this number and 1/2 are greater)?
    a. \$90,030
    b. \$68,531
    c. \$39,965
    d. \$85,721
    e. \$108,496

11. 75.00% of the homes had a price less than
    a. \$90,030
    b. \$68,531
    c. \$39,965
    d. \$85,721
    e. \$108,496

12. The "middle 50%" of the price of the homes has what range?
    a. \$90,030
    b. \$68,531
    c. \$39,965
    d. \$85,721
    e. \$108,496

13. Assume that 125 prices have been ordered from the lowest to the highest. The median for these prices would be which number?
    a. The 62nd price in order
    b. The 63rd price in order
    c. The average of the 62nd and 63rd numbers in order
    d. The difference between Q_3 and Q_1

14. The standard deviation of the price of the homes measures the amount of spread or variation around the _____ of the prices.
    a. median b. mode c. mean or average d. range

15. Does having an attached garage or not seem to make much difference in the mean value of the homes? Why or why not?

Self-Check

Questions 1–10 pertain to the following study.
A study was made by the manager of a mail-order business. The data represents the length of time (in minutes) that was needed to take the orders of customers who called in an order by telephone. The first 20 data values are given below (not all of the data.) Use Tables 2.19 and 2.20 and Figure 2.9 to answer questions 1–10.

11.9	11.0	8.7	10.5	9.1	9.2	8.8	11.9	4.9	11.0
10.8	6.0	10.8	10.8	6.5	3.5	4.2	9.7	10.9	6.2

Table 2.19 Mail-order data
Source: Fred Hulme

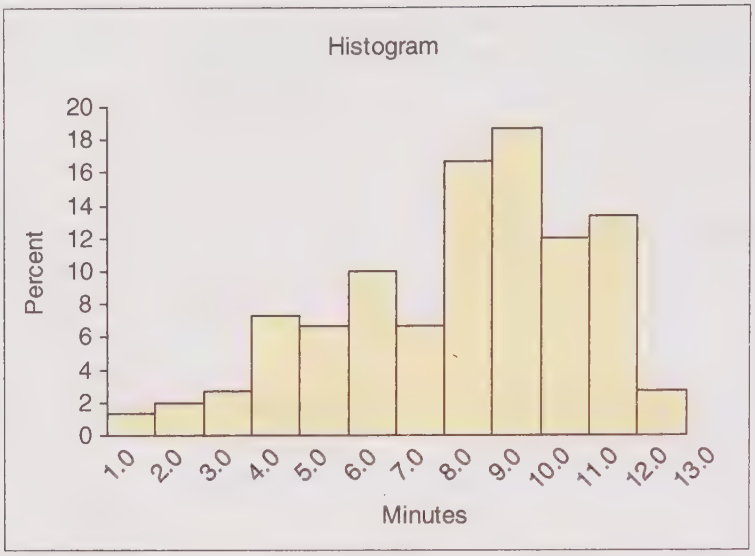


Figure 2.9 Mail-Order Minutes
Source: Fred Hulme

Descriptive Statistics					
	Minutes			**Empirical Rule**	
Count	150			Mean – 1s	5.727
Mean	8.273			Mean + 1s	10.818
Sample variance	6.479			Percent in interval (68.26%)	63.3%
Sample standard deviation	2.545			Mean – 2s	3.182
Minimum	1.6			Mean + 2s	13.363
Maximum	12.9			Percent in interval (95.44%)	96.0%
Range	11.3			Mean – 3s	0.636
1st quartile	6.500			Mean + 3s	15.909
Median	8.650			Percent in interval (99.73%)	100.0%
3rd quartile	10.300				
Interquartile range	3.800				
Mode	10.800				

Table 2.20 Descriptive Statistics for Mail-Order Minutes
Source: Fred Hulme

1. How many orders are involved in the data summaries?

2. Discuss the shape of the distribution of the length of time to process an order.

3. What length of time to process an order divides the distribution in half (i.e., 1/2 of the orders are processed in less time than this number and 1/2 are greater)?

4. 25.00 % of the orders were processed in more than this number of minutes.

5. The "middle 50%" of the values for the length of time to process an order has what range in minutes?

6. Assume that the numbers have been ordered from the lowest to the highest. Explain how the median for these numbers would be determined?

7. Assume an "outlier" or atypical order time is defined as an order time that is **more than 2 standard deviations** above or below the mean order time. What percentage of order times in the data would be classified as outliers?

8. The standard deviation of the length of time to process an order measures the amount of spread or variation around the _____ of the numbers.

9. Is it guaranteed that what fraction of the minutes will fall within 3 standard deviations of the mean number of minutes? (This question makes no assumption about the underlying distribution of the number of minutes needed to take the orders of customers.) 2 standard deviations?

10. What does the mode = 10.8 minutes represent?

11. Find the median for the following set of numbers.

7 11 6 13 13

12. Use the formula and table below to calculate the standard deviation for the following set of numbers.

4, 9, 17

$$\text{Standard Deviation} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

where where $\bar{x}$ = the mean or average of the x's

X	$X - \bar{X}$	$(X - \bar{X})^2$

13. What measure would best describe the "central location" or representative value for the annual salaries of the faculty and administrators of a large public-school district? Choose between mean or median.

14. Using the rule that states that for any set of numbers, the range defined by:

Mean $\pm$ 3 (Standard Deviation) will contain at least 8/9's (approximately 90%) of the numbers, what is the approximate relationship between the range of a set of numbers and the standard deviation of the numbers? (Fill in the blank below.)

Range $\approx$ _____ $\times$ (Standard Deviation)

15. If a bottling company fills bottles with a mean of 12 oz of liquid with a standard deviation of 0.1 oz, at least 8/9's of the bottles will have how many oz in them? Give a range for this answer—(lower limit, upper limit).

Chapter 2—End of Chapter Answers

Short Answer/Discussion

1. Mean, Median, mode
2. Range, variance, standard deviation
3. Because central location does NOT give you information regarding the spread of the data.
4. People have an intuitive understanding of the mean, it is unique, it is simple to calculate, and it may be used for further statistical analysis of a variable.
5. A simple mean assumes that each data value has the same amount of importance. There are many situations when this is not the case.
6. The mean is heavily influenced by outliers or extreme values. The median is NOT.
7. Plus or minus 3 standard deviations.
8. Let's one know whether the process is in control or not.
9. The mean and the range
10. The person scored higher than 80% of the others taking the exam.
11. Q_3
12. $Q_3 - Q_1$, the middle 50% of the observations, dispersion
13. Pivot Table

Problems

1. Mean = 3 Mode = 3 Range = 5

   Median = 3 Variance = 3.5 Standard deviation = 1.87
2. $30 < 37.2$, this is NOT unusual
3. 4.2
4. 457.155; 544.085
5. 1.4

6a. 2.799

6b. 2.8

6c. 2.8

6d. Range = 1.5, $s = 0.2506$, $s^2 = 0.0628$

6e. symmetrical, approximately normal

7a. 80, 120 males

7b. Yes, mean differences is .37 standard deviations

7c. No, standard deviations are close to each other

8a. 60, 90

8b. 75%

9a. 95%

9b. 99%

10. D

11. E

12. C

13. B

14. C

15. Yes, the average sale price with a garage is much larger than the average sale price without a garage.

Self-Check

1. 150
2. Skewed to the left
3. 8.650
4. 10.3
5. 3.8
6. Average the 75th and 76th observation once the data has been put in order from the smallest to the largest observation

7. 4%

8. Mean

9. 89%, 75%

10. Observation that occurred the most often

11. 11

12. Mean = 10, SD = 6.56

13. Median (the distribution is negatively skewed)

14. 6

15. 11.7, 12.3

CHAPTER

3

Risk Analysis—Probability



© Nomad_Soul/Shutterstock.com

Business decisions often involve choosing from several alternatives. In order to make a decision among several alternatives, one must be able to estimate the consequences of each alternative. The consequences often depend upon some factors that are subject to chance or uncertainty. The uncertainty cannot be eliminated, but risk analysis methods exist, which allow you to estimate the risk in your decisions.

Decisions under Uncertainty

EXAMPLE 1

The PC Solution (PCS), a Texas-based retail computer store, is faced with fierce competition in terms of both price and service. PCS is particularly concerned about the number of stockouts occurring on a popular type computer. Stockouts are very costly because when PCS cannot supply this particular computer, the customer promptly buys it from a competing store and PCS loses the sale. There is usually no opportunity to backorder the computer because the customer does not want to wait. PCS may also lose future business from that customer

as well. PCS measures the effects of stockouts in terms of service level, or the percentage of total demand that can be satisfied from inventory.

PCS sells an average of six computers a day. However, sales can vary on any given day. From the previous 100 sales' days, the frequency distribution for the number of units sold is in Table 3.1.

Units	0	1	2	3	4	5	6	7	8	9	10
Days	1	2	4	6	9	14	18	22	16	6	2

Table 3.1 PCS Demand
Source: Fred Hulme

The manufacturer of this computer is located in Washington state. Although it takes an average of four days for PCS to receive an order from the manufacturer, the time it takes to receive an order is also subject to uncertainty. The number of days to receive an order (sometimes referred to as **lead time**) can be summarized by the following frequency distribution based on the last 50 orders in Table 3.2.

Lead time (days)	3	4	5
Frequency of orders	10	30	10

Table 3.2 PCS Lead Time Demand
Source: Fred Hulme

PCS is faced with two decisions to make: (1) how many computers to request with each order—the **order quantity**, and (2) when to place the order—the **reorder point**. The reorder point is that stock level that would trigger a new order. PCS does not worry about stocking out of the computer immediately after receiving an order. However during the time from when an order is placed until the order is received (lead time), there is some level of anxiety. One way to guard against stockouts and improve service level is to increase the reorder point for the item so that more inventory is on hand to meet the demand during the lead time. PCS wants to determine the order quantity and reorder point that results in an average service level of at least 95%. The cost of ordering and holding the inventory should also be taken into consideration. The solution to PCS's problem will be addressed in latter examples.

If the daily demand for the computer and the lead time were both known with 100% surety, PCS would know exactly how many computers to order and when. If this were the case, this would be considered a **decision under certainty**. For instance, if PCS always sold six computers per day and it always took four days to receive a new shipment, then PCS would place an order when the stock level reached 24 units (six computers/day times four days). Twenty-four units would cover the sales while waiting for the next order to arrive.

In most business decisions not all the factors are known ahead of time with 100% surety. Based on the past data concerning sales demand and lead time, these factors are not constant. When crucial factors in a decision are not known, you have a situation known as a **decision under uncertainty**. You can sometimes quantify the uncertainty with a concept known as **probability**. A probability is a numeric measure of uncertainty. A probability in its simplest form is a ratio.

$$\text{Probability that an event will occur} = \frac{\text{Number of outcomes in which the event occurs}}{\text{Total number of outcomes}}$$


© Pavel L Photo and Video/Shutterstock.com

Using the lead time data from Example 1,

$$\Pr(\text{Lead time} = 4 \text{ days}) = \frac{30 \text{ orders}}{50 \text{ orders}}$$

$$\Pr(\text{Lead time} = 4 \text{ days}) = 0.60.$$

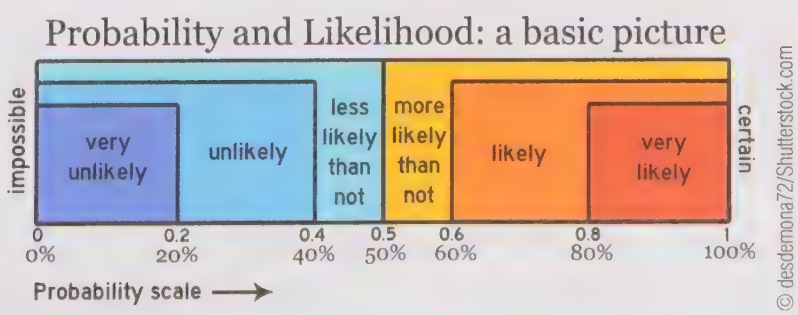
From these results you would say that there is a 60% likelihood or probability that the lead time for a single order will be four days. This measure known as a probability quantifies the uncertainty in an event such as the lead time for an order.

$$\Pr(\text{Daily sales} = 6 \text{ computers}) = \frac{18 \text{ days}}{100 \text{ days}}$$

Using the sales demand data from Example 1,

$$\Pr(\text{Daily sales} = 6 \text{ computers}) = 0.18.$$

Sometimes probabilities are expressed in terms of **odds**. If there are **a** outcomes that favor an event happening and **b** outcomes that do not favor the event happening, you say the odds of the event occurring are **a to b**. You would say the odds of the event **not** happening are **b to a**. Therefore the odds of the lead time equaling 4 days is 30 to 20 or in its reduced form, 3 to 2.



© desdemona72/Shutterstock.com

If the odds of event A occurring are a to b, then

$$\Pr(\text{A occurring}) = \frac{a}{a+b}$$

If the odds of event A occurring are a to b, then

$$\Pr(\text{A not occurring}) = \frac{b}{a+b}$$

Rules of Probability

1. The probability of any event must be a positive real number between 0 and 1.

2. A probability of 1 implies that the event is certain to occur; a probability of 0 means that the event is certain not to occur. The more certain you are that an event is to occur, the closer its probability will be to 1. The more certain you are that an event will not occur, the closer its probability of occurrence will be to 0.

3. If two events are mutually exclusive (both events cannot occur at the same time), the probability that one or the other event will occur equals the sum of the individual probabilities. This fact is sometimes expressed as

$$\Pr(A \text{ or } B) = \Pr(A) + \Pr(B)$$

or

$$\Pr(A \cup B) = \Pr(A) + \Pr(B).$$

A more general rule than the above is

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$$

Where $\Pr(A \cap B)$ is the probability that both A and B will occur jointly. If A and B are mutually exclusive, then $\Pr(A \cap B) = 0$. The above rule is sometimes referred to as the "addition rule."

4. The sum of the probabilities of an event A and its complement A' is equal to 1 or

$$\Pr(A) + \Pr(A') = 1.$$

Another way of expressing this concept is $\Pr(A) = 1 - \Pr(A')$. The complement of an event is that all the possible outcomes not included in that event.

5. If two events, A and B, are independent, then

$$\Pr(A \text{ and } B) = \Pr(A) \times \Pr(B)$$

or

$$\Pr(A \ \ B) = \Pr(A) \times \Pr(B).$$

For two events to be independent, it means that the occurrence of one event does not affect the chances of the occurrence of the other event. The above rule is sometimes referred to as the "multiplication rule."

6. The conditional probability of A occurring given that B has occurred is given by

$$\Pr(A \mid B) = \Pr(A \cap B) \div \Pr(B)$$

where $\Pr(A \mid B)$ is read as "the probability of A given B."
Therefore

$$\Pr(A \cap B) = \Pr(B) \times \Pr(A \mid B).$$

EXAMPLE 2

Assume you are inspecting the wrapping of candy bars coming off a production line. The quality of the wrapping of an individual bar either passes inspection or fails the inspection. From previous inspections when the wrapping process is considered "in control" it can be shown that usually 5 out of 100 bars have a defective wrapping.



© wanpatsorn/Shutterstock.com

a. If a single bar is selected at random, what is the probability that the wrapping is defective (D)?

$$\Pr(D) = \frac{5}{100} = 0.05$$

b. If a single bar is selected at random, what is the probability that the wrapping is not defective (D')?

$$\Pr(D') = 1 - \Pr(D)$$

$$\Pr(D') = 1 - 0.05 = 0.95$$

c. If two bars are selected at random, what is the probability that they **both** have defective wrapping?

Let D_1 be the event that the first bar is defective; let D_2 be the event that the second bar is defective. If you consider that the outcome on the first bar has no impact on the result of the second bar (i.e., D_1 and D_2 are independent events), then

$$\Pr(D_1 \text{ and } D_2) = \Pr(D_1) * \Pr(D_2)$$

Therefore, $\Pr(D_1 \text{ and } D_2) = 0.05 * 0.05 = 0.0025$

d. If two bars are selected at random, what is the probability that one or the other or both have defective wrapping?

$$\Pr(D_1 \text{ or } D_2) = \Pr(D_1) + \Pr(D_2) - \Pr(D_1 \text{ and } D_2)$$

$$\Pr(D_1 \text{ or } D_2) = 0.05 + 0.05 - 0.0025 = 0.0975$$

EXAMPLE 3



© bliderpool/Shutterstock.com

The following table shows the 2,500 wage employees of the ABC Corporation, classified by gender and by opinion on a proposal to emphasize fringe benefits rather than wage increase in impending contract discussions in Table 3.3.

		Opinion		
	Approves	**Neutral**	**Opposed**	**Total**
Gender				
Male	900	150	450	1,500
Female	300	100	600	1,000
Total	1,200	250	1,050	2,500

Table 3.3 Survey Data by Gender
Source: Fred Hulme

Let A represent the event that an employee approves the proposal; let N represent neutral; and let O represent opposed. Let M represent the event that an employee is a male, and let F represent female.

a. If an employee is picked at random, what is the probability that the employee approves the proposal?

$$\Pr(A) = \frac{1,200}{2,500} = 0.48$$

b. If an employee is picked at random, what is the probability that the employee is a female?

$$\Pr(F) = \frac{1{,}000}{2{,}500} = 0.40$$

c. If an employee is picked at random, what is the probability that the employee is a female **and** favors the proposal?

$$\Pr(A \text{ and } F) = \frac{300}{2{,}500} = 0.12$$

Notice that $\Pr(A \text{ and } F) \neq \Pr(A) * \Pr(F) =$, because these two events are **not** independent.

d. **Given** that an employee picked at random is female, what is the probability that the employee approves the proposal?

$$\Pr(A|F) = \Pr(A \text{ and } F) \div \Pr(F)$$

$$\Pr(A|F) = \frac{0.12}{0.40} = 0.30$$

This probability is called a "conditional" probability. Conditional probabilities are used in the field known as decision analysis where probabilities are revised based on additional information. A company may be planning an expansion for which its management has some feel for the odds or likelihood of success. Before carrying out the expansion, the company might conduct a survey of its customers as to their anticipated demand in the future. The company could then revise its estimates of the odds of success for the expansion based on the results of the customer survey. This revision of probabilities uses a formula that is known as **Bayes' Rule**.

EXAMPLE 4

The amount charged for life insurance depends on the age of the insured. The older the insured person, the more expensive the coverage becomes. As a person ages, the chance that he/she will have a claim against the policy will increase. The following formula can be used to calculate the probability if a person age x lives n more years:

$${}_np_x = \frac{l_{x+n}}{l_x} = 0.30, \text{ where } l_x = \text{the number living at age } x \text{ from the mortality table below in Table 3.4.}$$



© Jirsak/Shutterstock.com

Mortality Table				
	Number			Number
Age	Living		Age	Living
0	100,000		25	97,567
.	.		.	.
5	98,788		30	96,980
.	.		31	96,851
10	98,667		.	.
.	.			
20	98,122		65	79,040
21	98,018			

Table 3.4 Mortality Table
Source: Fred Hulme

Using the mortality (survival) Table 3.4, find the following probabilities:

a. that a person age 30 will live for 1 year

$$ {}_1p_{30} = \frac{l_{30+1}}{l_{30}} = \frac{96,851}{96,980} = 0.9987 $$

b. that a person age 30 will not live for 1 more year: $1 - 0.9987 = 0.0013$

c. that a person age 30 will live to age 65

$$ {}_{35}p_{30} = \frac{l_{30+35}}{l_{30}} = \frac{79,040}{96,980} = 0.8150 $$

d. that a person age 30 will not live to age 65: $1 - 0.8150 = 0.1950$

Probability Distributions

A **probability distribution** gives you the likelihood or probability of potential values for a random variable (a variable whose values are subject to chance). Probability distributions can either be classified as discrete or continuous. The random variable for a discrete distribution can take on only a finite number of values; the random variable for a continuous distribution can assume an infinite number of values. Theoretical probability distributions can be used to "model" a random variable. Histograms of different distributions exist that allow a lay person to identify a likely distribution. The central tendency, dispersion, and shape of the distribution is taken into account when choosing an appropriate distribution. A decision under uncertainty can be made if probabilities can be estimated or a probability distribution selected for the uncertain factors.

EXAMPLE 5





© Viktoria Kurpas/Shutterstock.com

Assume you are responsible for measuring the rate of defective items from a manufacturing process that typically produces 10% defective items. If you draw three items from the process and label each item as *P* (Pass) or *F* (Fail), the eight possible combinations and their respective probabilities are in Table 3.5.

If you define a random variable, *X*, to be the number of defective items out of a sample of three items, you would get the following probability distribution in Table 3.6.

Combination	Probability
PPP	.90 * .90 * .90 = 0.729
PPF	.90 * .90 * .10 = 0.081
PFP	.90 * .10 * .90 = 0.081
PFF	.90 * .10 * .10 = 0.009
FPP	.10 * .90 * .90 = 0.081
FPF	.10 * .90 * .10 = 0.009
FFP	.10 * .10 * .90 = 0.009
FFF	.10 * .10 * .10 = 0.001

Table 3.5 Pass/Fail Probability
Source: Fred Hulme

X	Pr(X)
0	0.729
1	0.243
2	0.027
3	0.001

Table 3.6 Defective Probability Chart
Source: Fred Hulme



© tool2530/Shutterstock.com

One use of the above table could be that if you began getting 2 or more defective items out of the three sampled items (which should occur only 0.027 + 0.001 = 0.028 of the time), you might begin to doubt if the manufacturing process is still producing at a 10% defective rate. A probability distribution can be used to predict the expected pattern of variation of a random variable. You cannot always eliminate variation in a manufacturing process that leads to defective items, but you can predict the amount of allowable variation that you should expect given a certain process defective rate.

EXAMPLE 6 QUALITY CONTROL: ACCEPTANCE SAMPLING

The monitoring of the quality of manufactured items usually involves some inspection procedure. Two reasons could lead to a decision of inspecting less than 100% of the manufactured items:

a. the total number of items produced is too large to test every single item, or

b. the inspection process is intrusive to the point of rendering the item useless after the inspection.

This inspection process is sometimes referred to as **acceptance sampling**. A sample of size, **n**, is taken from a "lot" of finished product. Each sampled item is tested and either passes or fails the inspection. If more than a certain number of sampled items, **c**, fail the inspection, then the quality of the entire lot is rejected. Steps are then taken to discover the problem that led to what appears to be a higher than acceptable defective rate in the entire lot. A sampling plan is determined by the sample size, n, and the acceptance number, c. For example if $n = 50$ and $c = 1$, then you would randomly inspect 50 items from the lot. If you find one or fewer items failing the inspection, you would declare the quality of the entire lot that you have sampled as being acceptable. If you are the producer of the items, you would ship the entire lot to your customers. The outgoing true defective rate of the lot would hopefully be small enough to meet the requirements of your customers.

The effectiveness of the sampling plan is based on the sampling plan accepting lots with a low defective rate with a high probability, and it failing to accept lots with a high defective rate with a high probability. In order to determine this effectiveness, you must be able to calculate the probability distribution for the number of defective items out of n items where there is a probability of Π that any one of the items is defective. This type of random variable is referred to as a **binomial variable**. You can use the rules of probability to show that

$$\Pr(X = k) = C(n, k)\Pi^k(1 - \Pi)^{n-k}$$

where

(a) $C(n, k) = \dfrac{n!}{k!(n - k)!}$

(b) Π is the probability that any one item is defective (the true defective rate for the process)
(c) n is the number of items sampled
(d) k is the number of defective items out of n.

Using the sampling plan ($n = 50, c = 1$) and a true defective rate of 5%,

$\Pr(X = 0) = C(50, 0)(0.05)^0(1 - 0.05)^{50-0}$
$\Pr(X = 0) = \dfrac{50!}{0!(50 - 0)!}(0.05)^0(0.95)^{50}$
$\Pr(X = 0) = (0.95)^{50} = 0.076945$
$\Pr(X = 1) = C(50, 1)(0.05)^1(1 - 0.05)^{50-1}$
$\Pr(X = 1) = \dfrac{50!}{1!(50 - 1)!}(0.05)^1(0.95)^{49}$
$\Pr(X = 1) = 50(0.05)(0.95)^{49} = 0.202487$
$\Pr(X \geq 2) = 1 - [\Pr(X = 0) + \Pr(X = 1)]$
$\Pr(X \geq 2) = 1 - [0.076945 + 0.202487]$
$\Pr(X \geq 2) = 1 - 0.279432$
$\Pr(X \geq 2) = 0.720568$

Figure 3.1 Calculating Probability of x Defectives Out of n

Source: Fred Hulme

If you define a random variable, X, to be the number of defective items out of a sample of 50 items taken from a lot with a defective rate of 5%, you would get the following probability distribution. See Figure 3.1 and Table 3.7

X	Pr(X)
0	0.076945
1	0.202487
≥2	0.720568

Table 3.7 Probabilities
Source: Fred Hulme

From the probability distribution, a sampling plan of ($n = 50, c = 1$) with lots that are 5% defective,

$$\Pr(\text{Accepting the lot}) = \Pr(X = 0 \text{ or } X = 1)$$
$$\Pr(\text{Accepting the lot}) = \Pr(X = 0) + \Pr(X = 1)$$
$$\Pr(\text{Accepting the lot}) = 0.076945 + 0.202487$$
$$\Pr(\text{Accepting the lot}) = 0.279432$$
$$\Pr(\text{Not Accepting the lot}) = \Pr(X \geq 2) = 0.720568$$

A series of probability distributions could be calculated for various values of the true lot defective rate. From these probability distributions, the probability of accepting the lot could be calculated. Table 3.8 displays those probabilities.

π = defective rate (%)	Pr(Accepting the lot)
0	1
1	0.910565
2	0.735771
3	0.555280
4	0.400481
5	0.279432
6	0.190003
7	0.126493
8	0.082712
9	0.053238
10	0.033786
11	0.021165
12	0.013099
13	0.008015

Table 3.8 Acceptance Table
Source: Fred Hulme

A line chart of this information is known as an **operating characteristic** (OC) curve shown in Figure 3.2. This chart allows you to check the effectiveness of a sampling plan at various levels of incoming quality. If you change either n, the sample size, or c, the acceptance number, the OC curve will change.

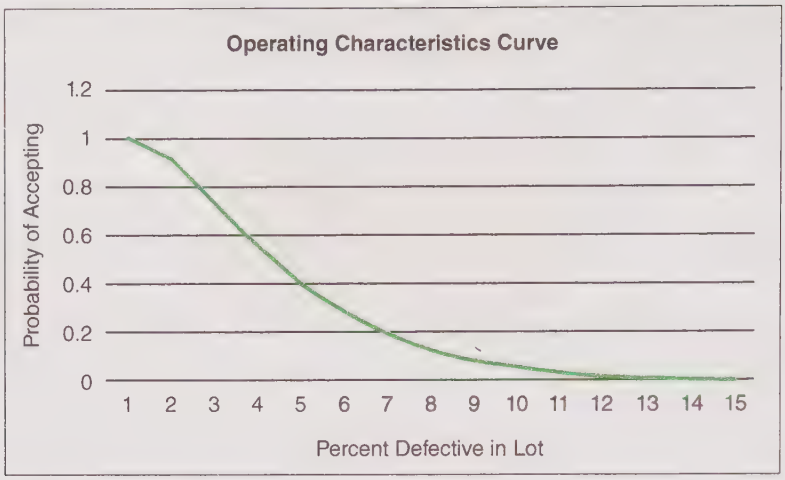


Figure 3.2 OC Curve
Source: Fred Hulme

Mathematical Expectation

One method of incorporating the probabilistic nature of a decision is to use a concept known as the **expected value** of a random variable. The expected value of a random variable is defined as the "long run average" of that variable. In mathematical definitions the expected value of a discrete random variable is:

$$E(X) = [X^{\star}\Pr(X)]$$

The expected value, $E(X)$, is simply the weighted average of the possible values of a random variable with the probability of occurrence as the weighting factor.

EXAMPLE 7

If you buy one of 1,000 raffle tickets for a color television set worth $650, the probability of winning the raffle would be 1/1,000 or 0.001. The probability of not winning therefore would be 1 − .001 = 0.999. The winnings could be represented by a random variable, X, with the following probability distribution in Table 3.9.



© Dzm1try/Shutterstock.com

X	Pr(X)
$0	0.999
$650	0.001

Table 3.9 Raffle Probability
Source: Fred Hulme

By applying the formula above, the "expected" winnings would be:

$$E(\text{Winnings}) = 650^{\star}(0.001) + 0^{\star}(0.999)$$

$$E(\text{Winnings}) = \$0.65$$

The "fair" price to charge for each ticket to cover the cost of the prize would be 65 cents per ticket. The actual ticket price would probably be larger to cover administrative costs and profit for the promoters. Another way of thinking of what the mathematical expectation represents for this example is that 999 of the tickets will not pay anything at all, one of the tickets will pay $650, so that altogether the 1,000 tickets pay $650, or on the average 65 cents per ticket. This is what is called the mathematical expectation for each ticket.

EXAMPLE 8

An oil company who has an option on a particular piece of land must decide whether to drill on the land before the expiration of the option or give up its rights. If the company drills, it believes its costs to be $200,000. If it finds oil, it expects to receive $1 million in revenue. There will be no revenue if it does not find oil. The company believes that the probability of finding oil if it drills on this piece of land is 0.25 and that the probability of not finding oil if it drills is 0.75.



© huyangshu/Shutterstock.com

a. Construct a **decision tree** to represent the decision.

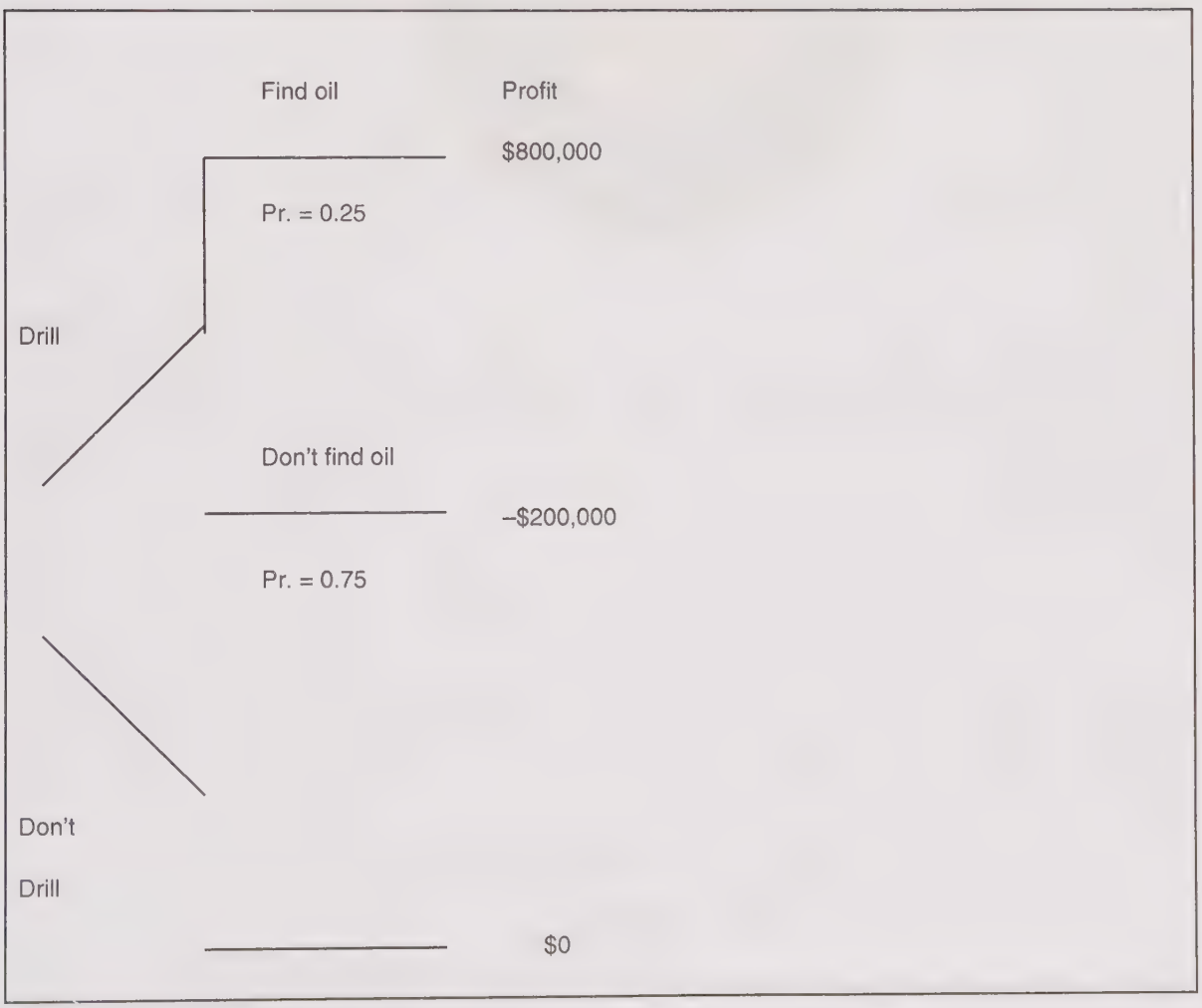


Figure 3.3 Decision Tree to Drill or Not Drill

Source: Fred Hulme

b. Based on maximizing the Expected Monetary Value (EMV), what decision should be made? Use Figure 3.3

$$\text{EMV(Drilling)} = \$800{,}000^{\star}(0.25) + -\$200{,}000^{\star}(0.75) = \$50{,}000$$

$$\text{EMV(Not drilling)} = \$0$$

Based on maximizing the Expected Monetary Value (EMV), the decision would be to drill.

EXAMPLE 9



© DeiMosz/Shutterstock.com

Using the information from Example 5, you can calculate the expected or average outgoing quality of the lots subjected to the acceptance sampling. Based on the results of the sampling, a lot is either passed on through or rejected. If it is rejected, assume that an exhaustive, 100% inspection is done and the outgoing lots have zero defectives. Therefore the outgoing quality of the lots (fraction defective) is either the incoming level or zero. Using the sampling plan

($n = 50, c = 1$) and a true defective rate for the lot of 5 %,

$$\Pr(\text{Accepting the lot}) = 0.279432$$

$$\Pr(\text{Not accepting the lot}) = 0.720568.$$

Therefore 27.94% of the time the lot is accepted and the outgoing quality is 5% defectives; 72.06% of the time the lot is not accepted and a 100% inspection is done, which removes all of the defectives from the lot giving a 0% defective rate. If this inspection process were to be carried out over and over, the expected or average outgoing quality (AOQ) would be:

$$E(\text{Outgoing defective rate}) = 5\% \star (0.279432) + 0\% \star (0.720568)$$

$$E(\text{Outgoing defective rate}) = 1.397\%.$$

Therefore acceptance sampling would reduce the defective rate of 5% to an average of 1.397%.

The following table contains the average outgoing quality (AOQ) for lots with defective rates from 0% to 14%.

Π = defective rate (%)	Pr(Accepting the lot)	AOQ (%)
0	1	0.000
1	0.910565	0.911
2	0.735771	1.472
3	0.55528	1.666
4	0.400481	1.602
5	0.279432	1.397
6	0.190003	1.140
7	0.126493	0.885
8	0.082712	0.662
9	0.053238	0.479
10	0.033786	0.338
11	0.021165	0.233
12	0.013099	0.157
13	0.008015	0.104
14	0.004851	0.068

Table 3.10 Average Outgoing Quality Level
Source: Fred Hulme

From Table 3.10, you can estimate the maximum AOQ regardless of the defective rate of the incoming lots. The average outgoing quality limit (AOQL) for the sampling plan ($n = 50, c = 1$) is approximately 1.7 %. No matter what the level of incoming quality of the lots entering the inspection process is, the average outgoing defective rate will be no more than approximately 1.7% defective. The peak in Figure 3.4 demonstrates this concept.

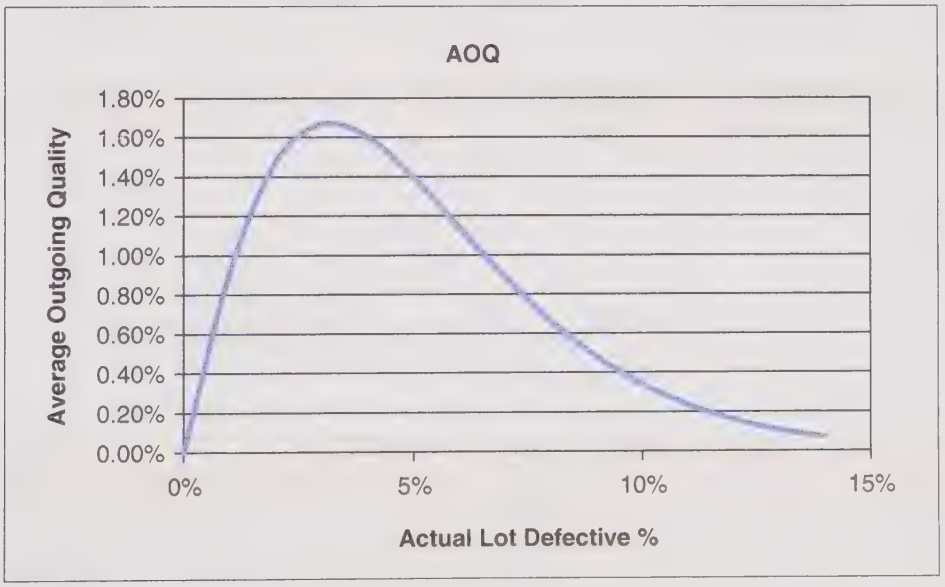


Figure 3.4 AOQL Chart
Source: Fred Hulme

Decision Analysis

Structuring a decision problem

Step 1 Identify the relevant alternatives or strategies (actions).

Step 2 Determine the conditions that are outside the control of the decision-maker that will determine the consequences of a given strategy. These conditions, referred to as states of nature or events, must be mutually exclusive and collectively exhaustive (that is, they cannot overlap and they must include all possible conditions.)

Step 3 Develop a payoff table or loss table that shows the payoff or loss resulting from each possible combination of strategies and events.

Step 4 If possible, estimate the probability that each event will occur.

Step 5 Evaluate each alternative according to some decision criteria and select the best alternative.

Four Basic Decision Criteria Under Uncertainly (This is a situation where it is impossible to assess the probability of future events.)

Maximax (the optimistic approach)

This procedure is to identify the best payoff in the entire payoff table and choose the alternative that corresponds to that payoff.

Maximin (the conservative approach)

This procedure is to determine the worst possible outcome for each alternative and choose the alternative that has the "best" of these worst possible outcomes.

Minimax Regret

This procedure requires building a "regret matrix." A "regret" is the difference between a given payoff and the best payoff for a given event. Then determine the worst possible regret for each alternative and choose the alternative that has the "best" of the worst regrets.

Laplace

The Laplace procedure assumes that all possible events are equally likely to occur. The procedure is to calculate the average (mean) payoff for each alternative and choose the alternative with the highest payoff.

Decision-Making Under Risk

Expected Value Criteria



© mangostock/Shutterstock.com

This procedure assumes that you can estimate the likelihood or probability of the states of nature (events). The procedure is to calculate the "expected values" of each alternative by summing the "weighted payoffs" for each alternative and select the alternative with the highest expected value. The "weighted payoffs" are calculated by multiplying the payoff by the probability for the corresponding state of nature.

EXAMPLE 10

A company must decide how many dozen sandwiches to order for the concession stand at a soccer match. Each dozen costs $8.00 and sells for $18.00. Any sandwiches leftover at the end of the game are of no value. Based on experience over the last 100 games with similar weather forecast, the following demand distribution was constructed:

Demand (in dozens)	1	2	3	4
No. of games	40	30	20	10

Step 1: Complete the profit matrix or "payoff" table as in Table 3.11.

		# Sold at Game (doz.)		
# Stocked (doz.)	1	2	3	4
1	$10.00	$10.00	$10.00	$10.00
2	$ 2.00	$20.00	$20.00	$20.00
3	−$ 6.00	$12.00	$30.00	$30.00
4	−$14.00	$ 4.00	$22.00	$40.00

Table 3.11 Payoff Table

Source: Fred Hulme

Step 2: Find Maximum and Minimum Possible Values as in Table 3.12.

		# Sold at Game (doz.)				
# Stocked (doz.)	1	2	3	4	Max	Min
1	$10.00	$10.00	$10.00	$10.00	$10.00	**$10.00**
2	$ 2.00	$20.00	$20.00	$20.00	$20.00	$ 2.00
3	–$ 6.00	$12.00	$30.00	$30.00	$30.00	–$ 6.00
4	–$14.00	$ 4.00	$22.00	$40.00	**$40.00**	–$14.00

Table 3.12 Max and Min Profits
Source: Fred Hulme

MAXIMAX Criteria: Choose the alternative that corresponds to the largest profit: Stock 4 dozen.
MAXIMIN Criteria: Choose the alternative that maximizes the smallest profit: – Stock 1 dozen.
Study the "Regret" Matrix shown in Table 3.13.

		# Sold at Game (doz.)			
# Stocked (doz.)	1	2	3	4	Maximum Regret
1	$0.00	$10.00	$20.00	$30.00	$30.00
2	$8.00	$0.00	$10.00	$20.00	$20.00
3	$16.00	$8.00	$0.00	$10.00	**$16.00**
4	$24.00	$16.00	$8.00	$0.00	$24.00

Table 3.13 Regret Matrix
Source: Fred Hulme

MINIMAX Regret Criteria: Choose the alternative that minimizes the maximum regret: Stock 3 dozen.

		# Sold at Game (doz.)			
# Stocked (doz.)	1	2	3	4	Average
1	$10.00	$10.00	$10.00	$10.00	$10.00
2	$ 2.00	$20.00	$20.00	$20.00	$15.50
3	–$ 6.00	$12.00	$30.00	$30.00	$16.50
4	–$14.00	$ 4.00	$22.00	$40.00	$13.00

Table 3.14 Laplace Criteria
Source: Fred Hulme

Laplace Criteria: Choose the alternative that maximizes the simple average across the events: Stock 3 dozen shown in Table 3.14.

		# Sold at Game (doz.)			Expected
# Stocked (doz.)	1	2	3	4	Value
1	$10.00	$10.00	$10.00	$10.00	$10.00
2	$ 2.00	$20.00	$20.00	$20.00	$12.80
3	–$ 6.00	$12.00	$30.00	$30.00	$10.20
4	–$14.00	$ 4.00	$22.00	$40.00	$ 4.00

Table 3.15 Expected Value
Source: Fred Hulme

Expected Value Criteria: Choose the alternative that maximizes the expected profit across the events: Stock 2 dozen. See Table 3.15.

Expected Value with Perfect Information: EVwPI

$$EVwPI = (10 \times 0.40) + (20 \times 0.30) + (30 \times 0.20) + (40 \times 0.10) = \$20.00$$

Expected Value of Perfect Information = $EVwPI - EV = \$20.00 - \$12.80 = \$7.20$

Normal Distribution

An often used statistical distribution for business modeling is the normal distribution or bell-shaped distribution. A graph of a normally distributed variable is given below. The graph is sometimes referred to as a normal curve shown in Figure 3.5.

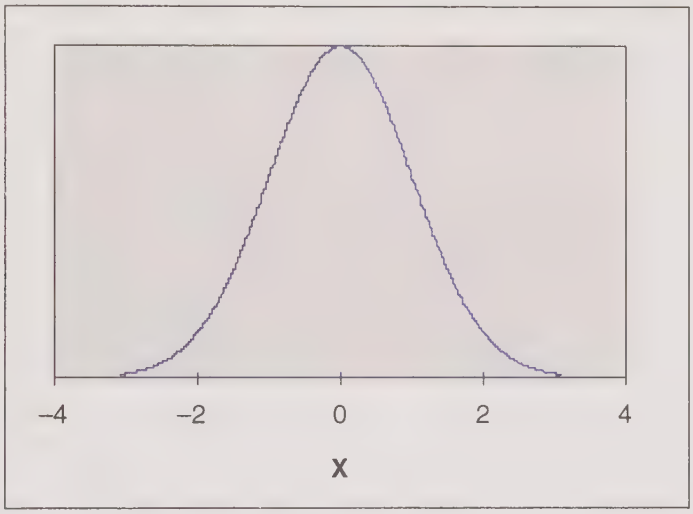


Figure 3.5 Normal Distribution
Source: Fred Hulme

A normally distributed variable is characterized by two factors (1) the mean and (2) the standard deviation. The values of the variable are symmetrically distributed about the mean. The standard deviation controls how spread out the values of the variable are. If a normally distributed variable has a mean of zero and a standard deviation of one, it is referred to as a standard normal random variable. The graph above is for a standard normal distribution.

Different random phenomena such as people's heights and IQ's have approximate normal distributions. The next histogram represents the distribution of IQ scores for 400 individuals. Its shape is approximately bell-shaped with an estimated mean of 100 and a standard deviation of 10 shown in Figure 3.6.

A random phenomenon such as IQ's that display a bell-shaped histogram can be modeled with a theoretical normal distribution. The relative frequencies that are represented by the rectangles of the histogram can be approximated by the areas under a continuous normal curve from one point to another. The theoretical normal curve that could be used to model the above histogram is given in Figure 3.7. You would estimate the mean and standard deviation from the 400 IQ scores.

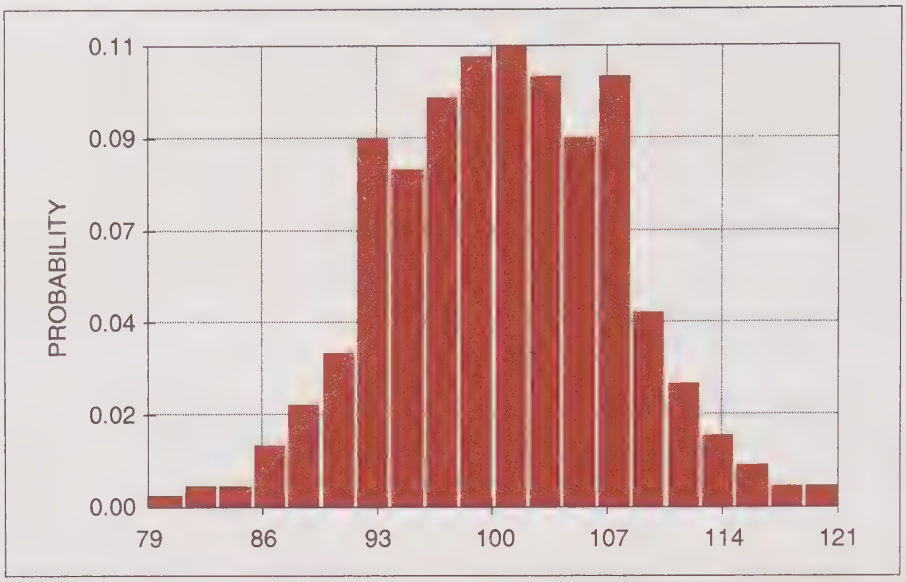


Figure 3.6 IQ Distribution
Source: Fred Hulme

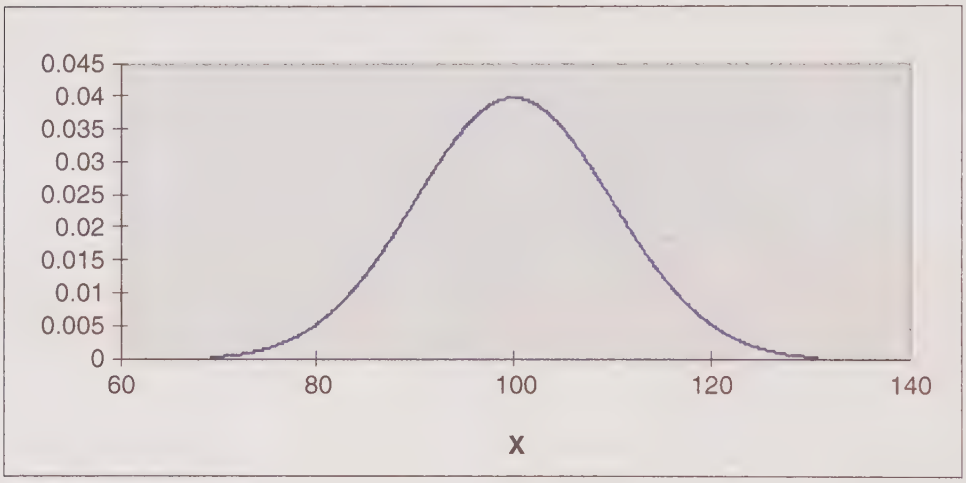


Figure 3.7 Normal Curve
Source: Fred Hulme

An estimate of the proportion of individuals that have an IQ score between 100 and 110 would be the area under the normal curve shown above, bounded between $x = 100$ and $x = 110$. You can use a table to find the areas under a normal curve, or you can use the Excel calculator, =NORMDIST.

The precise mathematical function for a normally distributed variable is

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

This formula will give the height of a particular normal curve at a value of x. It is the area under the normal curve between two values of x that is of main interest in that the area represents a probability.

Properties of a normal distribution (curve):

a. the total area under the curve is 1 or 100%

b. the curve is symmetric around the mean

c. the area under the curve to the right of the mean is 0.5 or 50%

d. the area under the curve to the left of the mean is 0.5 or 50%

The areas under the standard normal distribution are given Table 3.16. The numbers in the body of the table represent the area from the mean to a point to the right of the mean. These areas represent the probability that a random variable with a standard normal distribution will take on a certain range of values.

Standard Normal Distribution

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

Table 3.16 *Z* Values

Source: Fred Hulme

EXAMPLE 11

Find the probability that a random variable, Z, with a standard normal distribution will take on a value between 0 and 2.33. To find this answer, find the intersection of the value of 2.3 under the left-hand column headed by Z and the value of 0.03 across the top row of Table 3.16. The number at this intersection is 0.4901. In symbols, this probability is represented as

$$\Pr(0 \le Z \le 2.33) = 0.4901.$$

EXAMPLE 12

Assume that the random variable, Z, has a normal distribution with a mean of zero and a standard deviation of one.

a. $\Pr(0 \le Z \le 2.00) = 0.4772$

b. $\Pr(Z \ge 2.00) = 0.5000 - 0.4772 = 0.0228$

c. $\Pr(1.00 \le Z \le 2.00) = 0.4772 - 0.3413 = 0.1359$

d. $\Pr(-2.00 \le Z \le 2.00) = 0.4772 + 0.4772 = 0.9544$

e. $\Pr(-1.00 \le Z \le 2.00) = 0.3413 + 0.4772 = 0.8185$

f. $\Pr(Z \le -2.00 \text{ or } Z \ge 2.00) = 0.0228 + 0.0228 = 0.0456$

Most normally distributed variables do not have both the mean equal to zero and the standard deviation equal to one. However by using the following "standardizing" formula given below, you can convert a nonstandard normally distributed variable into a standard normally distributed variable.

$$Z = \frac{(X - \text{Mean})}{\text{Standard Deviation}}$$

EXAMPLE 13



© TRMK/Shutterstock.com

In a study involving the amount of time necessary for assembly line workers to complete a product, it was determined that this time was a normal random variable with a mean of 15 minutes and a standard deviation of 2 minutes. Sometimes this is written symbolically as:

$$\text{Assembly Time} \sim N(\mu = 15, \sigma = 2).$$

This is read as: "Assembly time is distributed as a normally distributed variable with a mean (μ) of 15 minutes and a standard deviation (σ) of 2 minutes."

What fraction of the items would have an assembly time between 15 and 19?

SOLUTION

The above question symbolically stated is: $\Pr(15 \le \text{Assembly Time} \le 19)$?

You would "standardize" both the 15 and 19 using the above formula.

$$Z_1 = \frac{15 - 15}{2} = 0 \qquad Z_2 = \frac{19 - 15}{2} = 2.00$$

Therefore, $\Pr(15 \le \text{Assembly Time} \le 19) = \Pr(0 \le Z \le 2.00)$.

From the standard normal table, $\Pr(0 \le Z \le 2.00) = 0.4772$. Therefore you would expect about 47.72% of the items to have an assembly time between 15 and 19 minutes.

EXCEL SOLUTION

The Excel function, **=NORMDIST(*X*, Mean, Standard Deviation, True)**, calculates the probability that the normally distributed variable is $\le X$.

$$\Pr(15 \le \text{Assembly Time} \le 19) = \text{NORMDIST}(19, 15, 2, \text{True}) - 0.5000 = 0.4772$$

EXAMPLE 14



© Kiev.Victor/Shutterstock.com

The distribution of the number of days the guests stay at a resort can be approximated with a normal distribution with a mean of 8.5 days and a standard deviation of 2.2 days. Out of 1,000 guests at the resort, how many can be expected to stay between 5 and 10 days?

SOLUTION

$$\Pr(5 \le \textit{Days} \le 10) = \Pr\left(\frac{5 - 8.5}{2.2} \le Z \le \frac{10 - 8.5}{2.2}\right) = \Pr(-1.59 \le Z \le .68)$$

$$\Pr(-1.59 \le Z \le .68) = 0.2517 + 0.4441 = 0.6958$$

Therefore you would expect approximately 696 guests out of 1,000 to stay between 5 and 10 days.

EXCEL SOLUTION

$$\Pr(5 \le \text{Length of Stay} \le 10) = \text{NORMDIST}(10, 8.5, 2.2, \text{True}) - \text{NORMDIST}(5, 8.5, 2.2, \text{True})$$

$$\Pr(5 \le \text{Length of Stay} \le 10) = 0.7523 - 0.0558 = 0.6965$$

EXAMPLE 15

Using the information from Example 11, what would the assembly time be such that only 5% of the items had an assembly time greater than that time?

SOLUTION

You would need to find the standard normal score, Z_0, such that $\Pr(Z \ge Z_0) = 0$. From the abbreviated standard normal table below, $Z_0 = 1.64$. You would then set the standardizing formula equal to 1.64 and solve for the score. Use Table 3.17.

$$\text{Score} = 15 + (1.645 \times 2) = 18.29 \text{ minutes}$$

Standard Normal Distribution

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545

Table 3.17 Partial *Z* Table
Source: Fred Hulme

EXCEL SOLUTION

The Excel function, =**NORMINV(Probability, Mean, Standard Deviation)** will find the value of the normally distributed variable that places the input for "Probability" to the left of the value.

$$\text{Assembly Time} = \text{NORMINV}(0.95, 15, 2) = 18.29 \text{ minutes}$$

Exercise

Short Answer/Discussion

1. A probability is a number that can range between _____ and _____ .

2. The probability of an event occurring can often be thought of as the _____ of the number of ways in which the event can occur to the total number of possible outcomes.

3. If the probability of event A is $\frac{1}{3}$, then the probability of the event, not A (the complement of A) would equal _____ .

4. What would be the complement of the event that a manufactured item is defective?

5. Five manufactured items are inspected and labeled as either passing or failing the inspection. If the random variable, X, is defined as the number of items failing the inspection, what are the possible values of X?

6. Using the random variable, X, defined in Question (5), the sum of the probabilities of all the possible outcomes of X should equal _____ .

7. What does it mean for two events to be independent?

8. If two events, A and B, are independent, then the probability that both A and B occur equals the probability of A _____ the probability of B.

9. What does it mean for two events, A and B, to be mutually exclusive?

10. Given two events, A and B, what is the formula for the probability that A or B occurs?

11. What are the characteristics of a binomial random variable?

12. Would a binomial random variable be considered a discrete or continuous random variable?

13. What is the purpose of the statistical quality control procedure known as acceptance sampling?

14. What role does the acceptance number play in an acceptance sampling plan?

15. Does the Excel calculator, =BINOMDIST, calculate the probability of accepting or rejecting an entire shipment based on only a sample of items?

16. What is the name of the chart that can be used to summarize the effectiveness of a particular acceptance sampling plan?

17. What is another name for the "long run" average value of a random variable?

18. The "expected value" of a discrete random variable is an example of a "weighted average". In this case, what provides the "weights" in the calculation?

19. What two factors describe a normally distributed variable?

20. A standard normal random variable has a mean equal to _____ and standard deviation equal to _____ .

21. A normal distribution is symmetric with respect to its _____ .

22. What does "standardizing" a value of a normally distributed variable mean?

23. You would expect to find practically 100% of all standard (Z) scores for a normally distributed variable to fall between what two values?

24. Which Excel calculator can be used to find probabilities for a range of a normally distributed variable?

25. Draw and diagram the normal curve that the Excel calculator, =NORMDIST(400, 500, 50, TRUE) defines.

Problems

1. A consumer research organization studied the level of customer service provided by gasoline stations carrying advertised brands of gasoline and those carrying only unadvertised brands. In its report. The following table reflects the number of stations in each category in Table 3.18.


© FUN FUN PHOTO/Shutterstock.com

	Low Level of Service	Moderate-to-High Level of Service
Advertised brands	18	28
Unadvertised brands	22	12

Table 3.18 Gas Station Brand/Service Table
Source: Fred Hulme

Let L and L' represent the events of low level of service and moderate-to-high level of service, respectively. Let A and A' represent the events of advertised brands and unadvertised brands, respectively. If a motorist drives into one of these 80 stations at random, find the following probabilities using the data from Table 3.18.

(a) $\Pr(A)$ (b) $\Pr(A')$
(c) $\Pr(A \text{ and } L)$ (d) $\Pr(A \mid L)$

2. A marketing research firm is conducting a preliminary analysis of the potential market for a new monthly magazine for the Internet. The firm has compiled Table 3.19 containing demographic information from a survey of 10,000 people.

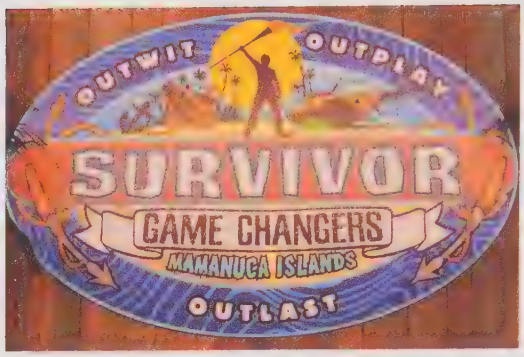

© Kaspars Grinvalds/Shutterstock.com

Income	Age < 40	Age ≥ 40	Total
< $50,000	5,203	1,125	6,328
≥ $50,000	2,652	1,020	3,672
Total	7,855	2,145	10,000

Table 3.19 Age/Income Table
Source: Fred Hulme

a. What is the probability that a randomly chosen person from this population will have a yearly gross income of less than $50,000?
b. What is the probability that a randomly chosen person from this population will be 40 years or older?
c. What is the probability that a randomly chosen person from this population will have a yearly gross income of less than $50,000 and be 40 years or older?
d. What is the probability that a randomly chosen person from this population will have a yearly gross income of less than $50,000, given that they are 40 years or older?

3. Using the mortality (survival) table below, find the following probabilities using Table 3.20.



© Kathy Hutchins/Shutterstock.com

Mortality Table				
	Number			Number
Age	Living		Age	Living
0	100,000		25	97,567
.	.		.	.
5	98,788		30	96,980
.	.		31	96,851
10	98,667		.	.
.	.			
20	98,122		65	79,040
21	98,018			

Table 3.20 Mortality Rates
Source: Fred Hulme

a. that a person age 20 will live for 1 year
b. that a person age 20 will not live for 1 more year
c. that a person age 25 will live to age 65
d. that a person age 25 will not live to age 65
e. that a husband and wife of ages 25 will both live to age 65

4. The demand for white latex paint at a hardware store has always been 0, 1, 2, 3, 4, or 5 gallons per day. Over the past 300 working days, the owner notes the following frequencies of demand:


© Norman Pgson/Shutterstock.com

Quantity Demanded (gallons)	0	1	2	3	4	5
Number of Days	20	60	75	90	30	25

Convert the frequency distribution into a probability distribution by calculating the probability of each possible outcome.

5. Decide whether the following can be probability functions. Explain your answers.

   a. $f(x) = \frac{1}{2}$ for $x = 1, 2,$ or 3

   b. $f(x) = \frac{x}{10}$ for $x = 1, 2, 3,$ or 4

   c. $f(x) = \frac{x - 4}{5}$ for $x = 1, 3, 5, 7,$ or 9

6. You are in charge of monitoring the quality of a candy bar manufacturing process. Every hour you sample 2 candy bars and decide if they pass inspection. Let X = the number of defective bars out of the 2 sampled bars. If the manufacturing process normally has a 5% defective rate, find the probability distribution of X.

7. A software business advertises by sending direct-mail advertisements to prospective buyers. From experience it is known that out of 10 businesses that receive an advertisement, one will buy a copy of the software. Twenty businesses are sent the advertisement.



© Everyonephoto Studio/Shutterstock.com

   a. What is the expected number of buyers? (Hint: The expected number equals the number of advertisements sent times the probability that any one business receiving the advertisement will purchase a copy of the software.)
   b. What is the probability that no copies of the software will be sold?
   c. What is the probability that at least one copy of the software will be sold?
   d. What is the probability that exactly two copies of the software will be sold?

8. An automobile manufacturer buys 1,000 radios per order from a supplier. When each shipment of 1,000 radios arrives, the automobile manufacturer takes a random sample of 10 radios from the shipment. If more than one radio in the sample is defective, the automobile manufacturer rejects the shipment and sends all of the radios back to the supplier.



© Radu Bercan/Shutterstock.com

   a. If 0.5% of all the radios in the shipment are defective (i. e., the chance that any one radio is defective is 0.5%), find the probability that none of the radios in the sample of ten are defective.
   b. If 0.5% of all the radios in the shipment are defective, find the probability that exactly one of the ten radios sampled will be defective.
   c. If 0.5% of all the radios in the shipment are defective, find the probability that the entire shipment will be accepted?
   d. If 0.5% of all the radios in the shipment are defective, find the probability that the entire shipment will be rejected?

9. The following table gives the probabilities that a person browsing in a book store will buy 0, 1, 2, 3, or 4 books as shown in Table 3.21.

Books	0	1	2	3	4
Probability	0.40	0.30	0.15	0.10	0.05

Table 3.21 Book Buys
Source: Fred Hulme

How many books can a person browsing in the book store be expected to buy?

10. An importer has arranged for a shipment of seasonal merchandise that promises a profit of $120,000 with a probability of 0.60 or a loss of $45,000 (due to late delivery of the shipment) with a probability of 0.40. What is the importer's mathematical expectation?

11. A company is considering drilling for oil at a certain location. Given the preliminary geological samples, the company believes its choices have the given probabilities in Table 3.22.

Drilling Outcome	Profit	Probability
Oil	$400,000	0.20
Natural gas	$300,000	0.10
Neither	−$50,000	0.70

Table 3.22 Drilling Decision
Source: Fred Hulme

What is the expected profit for drilling at this location?

12. A firm is trying to decide whether to raise the price of its product. It estimates that if the product is price inelastic (i.e., a price increase and resultant drop in sales will result in an increase in total revenue), it will increase profits by $800,000. If the product is price elastic (i.e., a price increase and resultant drop in sales will result in a decrease in total revenue), it will decrease profits by $500,000. Studies indicate that the probability that the price is elastic is 0.60.


© VectorMine/Shutterstock.com

a. Draw a decision tree depicting this problem.

b. Find the decision that maximizes the "expected" change in profits.

Problems 13 to 15 pertain to the following information.
A party store buys birthday cakes from a bakery. The party store pays the bakery $3.00 per cake and then attempts to sell the cakes for $12.00 per cake. The cakes must be ordered one day in advance, any cakes left unsold at the end of the day are sold to the "Thrift Bakery" for a salvage price of $1.00 per cake. Use Table 3.23 to answer questions 13 to 15.

		# of cakes sold	
# of cakes stocked	4	5	6
4		???	
5			
6	???		???

Table 3.23 Cake Demand
Source: Fred Hulme


© Melnikov Sergey/Shutterstock.com

13. Calculate the profit if the party store purchases or "stocks" 4 cakes and the demand turned out to be 5 cakes.
A. $34 **B.** $36 **C.** $45 **D.** $48

14. Calculate the profit if the party store purchases or "stocks" 6 cakes but sells only 4 of the cakes.
A. $32 **B.** $36 **C.** $48 **D.** $54

15. Calculate the profit if the party store purchases or "stocks" 6 cakes and sells all 6 of them.
A. $32 **B.** $36 **C.** $48 **D.** $54

16. A vendor plans on selling barbecue beef by the pound at a large convention. For each pound of beef that she stocks and sells, her profit is 65 cents. If she stocks more beef than she can sell, she can still recover from a complete loss by selling any unsold meat to an animal shelter for a net loss of 15 cents per pound. She estimates demand in pounds to be 10% of the total attendance for the convention. The total attendance for the convention is estimated in Table 3.24.


© Jag_cz/Shutterstock.com

Attendance	Probability (%)
60,000	20
80,000	60
100,000	20

Table 3.24 BBQ Demand
Source: Fred Hulme

a. If the total attendance turns out to be 60,000 and the vendor stocked 8,000 pounds of beef, calculate her profit.
b. If the total attendance turns out to be 80,000 and the vendor stocked 8,000 pounds of beef, calculate her profit.
c. If the total attendance turns out to be 100,000 and the vendor stocked 8,000 pounds of beef, calculate her profit.
d. Using the results from (a) to (c), calculate the "expected" profit if the vendor stocks 8,000 pounds of beef.

Use Table 3.25 to answer questions 17 to 23:

The national sales manager for a T-shirt business supplies all salespersons with the payoff (profit) table shown below, giving the potential profit generated when a retailer purchases from 1 to 4 dozen T-shirts. The probability of demand for 1 to 4 dozen T-shirts is also shown.

PAYOFF TABLE				
	Future Demands			
	1	2	3	4
Probabilities of demands	.2	.4	.3	.1
Purchase Quantities				
1	120	120	120	120
2	0	240	240	240
3	-120	120	360	360
4	-240	0	240	480

Table 3.25 T-Shirt Payoff Table
Source: Fred Hulme


© ecco/Shutterstock.com

17. Which "Purchase Quantity" amount would be associated with the "**MAXIMAX**" decision criteria?
A. 1 dozen **B.** 2 dozen **C.** 3 dozen **D.** 4 dozen

18. Which "Purchase Quantity" amount would be associated with the "**MAXIMIN**" decision criteria?
A. 1 dozen **B.** 2 dozen **C.** 3 dozen **D.** 4 dozen

19. Which "Purchase Quantity" amount would be associated with the "**Laplace**" decision criteria?
A. 1 dozen **B.** 4 dozen **C.** be indifferent between 2 or 3 dozen

20. What is the "expected" payoff (profit) if a retailer purchases 4 dozen T-shirts?
A. 72 **B.** 120 **C.** 168 **D.** 192

21. How many dozen T-shirts should be purchased to yield the highest expected payoff (profit)?
A. 0 **B.** 1 **C.** 2 **D.** 3 **E.** 4

22. How large is the "expected value with perfect information"?
A. 192 **B.** 276 **C.** 84 **D.** 120 **E.** 480

23. How large is the "expected value of perfect information"?
A. 192 **B.** 276 **C.** 84 **D.** 120 **E.** 480

24. Assume that the random variable, Z, has a "standard" normal distribution with a mean of zero and a standard deviation of one. Use Figure 3.8 to aid you.

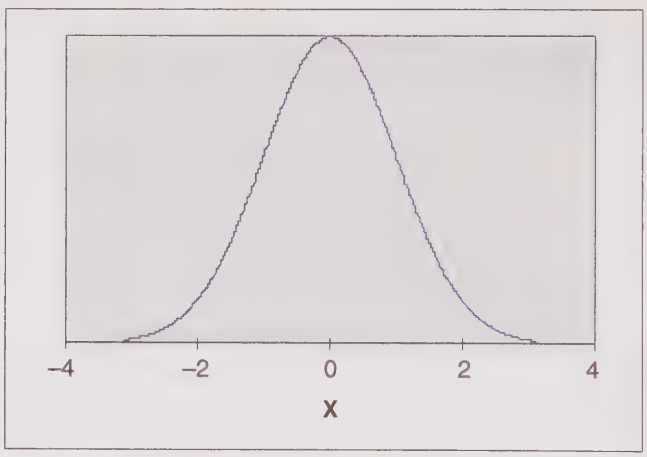


Figure 3.8 Normal Curve
Source: Fred Hulme

a. $\Pr(0 \le Z \le 1.00)$
$\Pr(Z \ge 1.00)$
b. $\Pr(2.00 \le Z \le 3.00)$
c. $\Pr(-1.00 \le Z \le 1.00)$
d. $\Pr(-1.00 \le Z \le 3.00)$
e. $\Pr(Z \le -1.00 \text{ or } Z \ge 1.00$

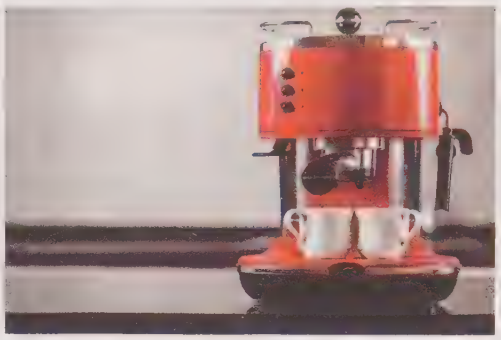
25. A coffee machine is set to fill cups with an average of 6 ounces. Assume that the amount is normally distributed with a standard deviation of 0.5 ounces.

a. Pr(6.00 ≤ Amount ≤ 7.00)

b. Pr(Amount ≥ 7.00)

c. Pr(5.00 ≤ Amount ≤ 7.00)

d. Pr(Amount ≤ 5.00)



© Melica/Shutterstock.com

26. From question (25), find the amount such that only 5% of the cups would have that amount or more.

Questions 27 to 34 pertain to the following description.
An oil change business uses an average or mean of 300 quarts of oil per day with a standard deviation of 20 quarts. Answer the following questions assuming the daily usage is normally (bell-shaped) distributed.



© Lesterman/Shutterstock.com

27. The probability is _____ that the daily usage is between 300 and 350 quarts.

28. The probability is _____ that the daily usage is greater than 350 quarts.

29. The probability is _____ that the daily usage is less than 350 quarts.

30. The probability is _____ that the daily usage is between 250 and 350 quarts.

31. The probability is _____ that the daily usage is between 330 and 350 quarts.

32. If the business begins a day with 350 quarts, what % of days should they have enough oil to meet the customer demand?

33. How many quarts should the business have at the beginning of the day such that they would expect to run out only 1% of the days?

34. Give the upper and lower limits of the "middle 50%" of the daily oil usage.

Self-Check

A manufacturer has 3 shifts—day, afternoon, and night. Fifty items were tested and classified in Table 3. 26.



© Phanu D Pongvanit/Shutterstock.com

	Day	Afternoon	Night		Total
Defective	3	2	1		6
Acceptable	17	13	14		44
Total	20	15	15		50

Table 3.26 Shift Quality
Source: Fred Hulme

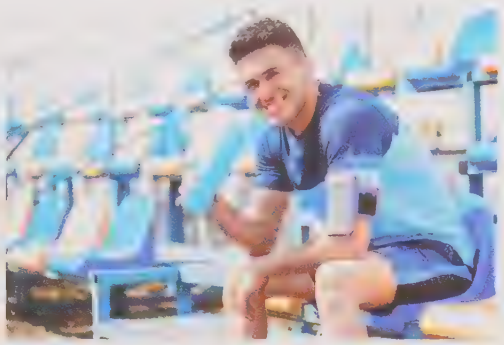
1. If one item is chosen at random, find the probability that the item was manufactured on the day shift.

2. If one item is chosen at random, find the probability that the item was defective.

3. If one item is chosen at random, find the probability that the item was manufactured on the day shift and was defective.

4. If an item is produced on the day shift, what is the probability that it is defective?

Cases of a new fitness drink sell for $20 each. The cost per case is $8 and any unsold cases may be returned to the supplier, who will refund the cost for each case returned minus $2 per case for handling and storage.



Profit Matrix			Actual Demand (cases)	
		5	6	7
	5			
# of cases ordered	6			
(Purchase Quantity)	7			

Table 3.27 Fitness Drink Demand
Source: Fred Hulme

5. Fill in the profits in Table 3.27.

6. Assume the probability distribution of demand is:

Demand (# of cases)	5	6	7
Probability	0.20	0.30	0.50

Table 3.28 Fitness Drink Probability
Source: Fred Hulme

a. Using Tables 3.27 and 3.28, which "Purchase Quantity" amount would be associated with the "**MAXIMAX**" decision criteria?

b. Which "Purchase Quantity" amount would be associated with the "**MAXIMIN**" decision criteria?

c. Which "Purchase Quantity" amount would be associated with the "**Laplace**" decision criteria?

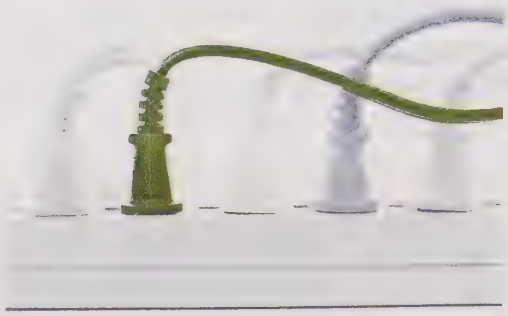
d. What is the expected profit for a purchase quantity of seven cases?

e. How many cases should be purchased to yield the highest expected profit?

f. How large is the "expected value with perfect information"?

e. How large is the "expected value of perfect information"?

An electrical products distributor buys 1,000 power strips from a supplier. When each shipment of 1,000 power strips arrives, the distributor takes a random sample of 15 power strips from the shipment. If more than one power strip in the sample is defective, the distributor rejects the shipment and sends all of the power strips back to the supplier.



7. If 1% of all the power strips in the shipment are defective (i.e., the chance that any one power strip is defective is 1%), find the probability that none of the 15 power strips in the sample are defective.

8. If 1% of all the power strips in the shipment are defective, find the probability that exactly one of the 15 power strips sampled will be defective.

9. If 1% of all the power strips in the shipment are defective, find the probability that the entire shipment will be accepted?

An airline has a plane that has 250 seats. Due to "no-shows", the airline has a policy of selling more tickets than the plane has seats. This creates a chance that the flight will be "overbooked" meaning that a person with a valid reservation will be "bumped" from the flight. Assume the chance of a no-show for a ticket holder is 6% and the airline sells 260 tickets.



10. What is the chance that the flight will have exactly 250 passengers?

11. What is the chance that the flight will have enough seats? **(Hint: Use the Excel calculator, =BINOMDIST.)**

12. What is the chance that the flight will be "overbooked?"

In a study involving the amount of time necessary for assembly line workers to complete a product, it was determined that this time was a normal random variable with a mean of 15 minutes and a standard deviation of 2 minutes.



13. The probability is _____ that a product is assembled in between 15 and 17 minutes.

14. The probability is _____ that a product is assembled in more than 17 minutes.

15. The probability is _____ that a product is assembled in more than 12 minutes.

16. The probability is _____ that a product is assembled in less than 12 minutes.

17. The probability is _____ that a product is assembled between 17 and 19 minutes.

18. The probability is _____ that a product is assembled between 13 and 19 minutes.

19. 20% of the products require more than _____ minutes for assembly.

20. The "middle 80%" of products are assembled between what two times?

Chapter 3 – End of Chapter Answers

Short Answer/Discussion

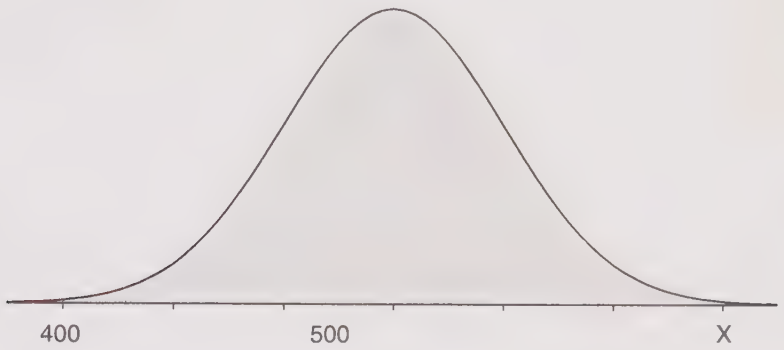
1. 0 and 1
2. ratio
3. $\frac{2}{3}$
4. That the item is nondefective
5. X = 0, 1, 2, 3, 4, 5
6. One
7. Independent events mean that the occurrence of one event does not affect the chance of the occurrence of the other event.
8. times
9. Pr(A and B) = 0; the events can't both occur at the same time
10. Pr(A and B) = Pr(A) × Pr($B|A$)
11. Two possible outcomes on each trial; n trials are observed; the probability of a "success" stays constant from trial to trial
12. discrete
13. To decide if a large shipment (lot) should be accepted (retained) or sent back to the manufacturer.
14. The acceptance number is the maximum number of defectives allowed in the sample and the entire shipment still be accepted.
15. accepting
16. operating characteristic curve
17. expected value
18. The probabilities provide the "weights".
19. mean and standard deviation
20. zero; one
21. mean

22. Standardizing formula: $Z = \dfrac{(X - \text{Mean})}{\text{Standard Deviation}}$; the "$Z$ score" gives the number of units of standard deviation that a value is above or below the mean

23. $(-3, +3)$

24. =NORMDIST

25.



Problems

(1a)	0.575
(1b)	0.425
(1c)	0.225
(1d)	0.450
(2a)	0.633
(2b)	0.215
(2c)	0.113
(2d)	0.524

3a. $\frac{98,018}{98,122} = 0.9989$

3b. $1 - 0.9989 = 0.0011$

3c. $\frac{79,040}{97,567} = 0.8101$

3d. $1 - 0.8101 = 0.1999$

3e. $(0.1999)(0.1999) = 0.0400$

4.

Quantity Demanded	Probability
0	0.067
1	0.200
2	0.250
3	0.300
4	0.100
5	0.083

5a. No—the probabilities sum to a number greater than one

5b. Yes—all the properties of a probability function are satisfied

5c. No—some of the probabilities will be negative

6.

X = # of defective bars	Probability
0	0.9025
1	0.0950
2	0.0025
	1.0000

Let X = the number of businesses out of 20 that buy the software after receiving the advertisement.

7a. $E(X) = 2$

7b. $\Pr(X = 0) = 0.1216$

7c. $\Pr(X \geq 1) = 0.8784$

7d. $\Pr(X = 2) = 0.2852$

Let X = the number of defective radios in the sample of 10

8a. $\Pr(X = 0) = (1 - 0.005)^{10} = 0.9511$

8b. $\Pr(X = 1) = 10(1 - 0.005)^{9}(0.005)^{1} = 0.0478$

8c. $\Pr(X \leq 1) = 0.9511 + 0.0478 = 0.9989$

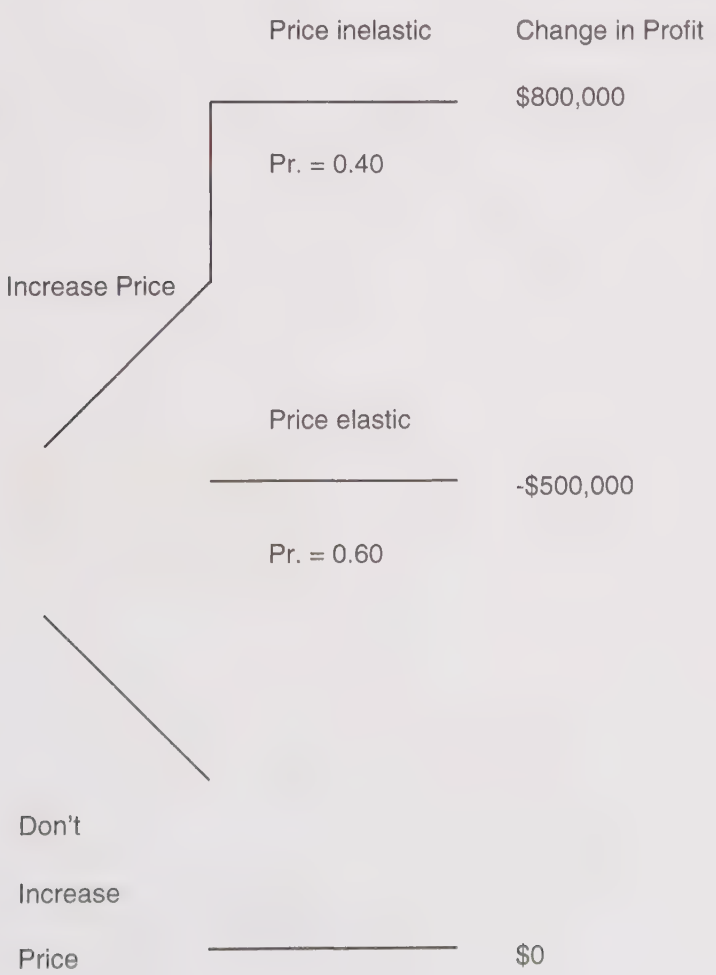
8d. $\Pr(X > 1) = 1 - \text{Prob}(X \leq 1) = 1 - 0.9989 = 0.0011$

9. $E(\#\text{ of books sold}) = 1.1$ books

10. $E(\text{Profit}) = \$120{,}000(0.60) + (-\$45{,}000)(0.40) = \$54{,}000$

11. $E(\text{Profit}) = \$400{,}000(0.20) + \$300{,}000(0.10) + (-\$50{,}000)(0.70) = \$75{,}000$

12a. Construct a **decision tree** to represent the decision.



12b. Expected(Change in profit given a price increase) = \$800,000(0.40) + (−\$500,000)(0.60)

Expected(Change in profit given a price increase) = \$20,000

The decision that maximizes the "expected change in the profits" would be to increase the price of the product.

13. B

14. A

15. \$54

16a. 6,000(0.65) − 2,000(0.15) = \$3,600

16b. 8,000(0.65) = \$5,200

16c. 8,000(0.65) = \$5,200

16.d. Expected Profit = \$3,600(0.20) + \$5,200(0.60) + \$5,200(0.20) = \$4,880

17. D

18. A

19. C

20. A

21. C

22. B

23. C

24a. 0.3413

24b. 0.1587 =0.5 − 0.3413

24c. 0.0215 =0.4987 − 0.4772

24d. 0.6826 =0.3413 + 0.3413

24e. 0.8400 =0.4987 + 0.3413

24f. 0.3174 =0.1587 + 0.1587

25a. 0.4772

25b. 0.0228 =0.5 − 0.4772

25c. 0.9544 =0.4772 + 0.4772

25d. 0.0228 =0.5 − 0.4772

26. 6.82 =6 + (1.64*0.5)

27. 0.4938

28. 0.0062

29. 0.9938

30. 0.9876

31. 0.0606

32. 0.9938

33. 346.6 =300 + (2.33*20)

34. 313.4 =300 + (0.67*20)

286.6 = 300 − (0.67*20)

Self-Check

1. 20/50 = 0.40

2. 6/50 = 0.12

3. 3/50 = 0.06

4. 3/20 = 0.15

5.

	5	6	7
5	$60	$60	$60
6	$58	$72	$72
7	$56	$70	$84

6a. Purchase Quantity = 7 cases

6b. Purchase Quantity = 5 cases

6c. Purchase Quantity = 7 cases

6d. $74.20

6e. Purchase Quantity = 7 cases

6f. $75.60

6e. $1.40

Let X = the number of defective power strips in the sample of 15

7. $\Pr(X = 0) = (1 - 0.01)^{15} = 0.8601$

8. $\Pr(X = 1) = 15(1 - 0.01)^{14}(0.01)^{1} = 0.1303$

9. $\Pr(X \leq 1) = 0.8601 + 0.1303 = 0.9904$

Let X = the number of ticketholders who show up for the flight

10. $\Pr(X = 250)$ =BINOMDIST(250, 260, 94%, False) $= 0.0378$

11. $\Pr(X \leq 250)$ =BINOMDIST(250, 260, 94%, True) $= 0.9524$

12. $\Pr(X > 250) = 1 - \Pr(X \leq 250) = 1 - 0.9524 = 0.0476$

Let X = number of minutes required to assemble a product

13. $\Pr(15 < X < 17) = 0.3413$

14. $\Pr(X > 17) = 0.500 - 0.3413 = 0.1587$

15. $\Pr(X > 12) = 0.5000 + .4332 = 0.9332$

16. $\Pr(X < 12) = 0.5000 - .4332 = 0.0668$

17. $\Pr(17 < X < 19) = 0.4772 - 0.3413 = 0.1359$

18. $\Pr(13 < X < 19) = 0.3413 + 0.4772 = 0.8185$

19. $X = 15 + 0.84(2) = 16.68$ minutes

20. $X = 15 + 1.28(2) = 17.56$ minutes

$X = 15 - 1.28(2) = 12.44$ minutes

CHAPTER

4

Sampling and Sampling Distributions



© Bakhtiar Zein/Shutterstock.com

Introduction

This chapter introduces how you can use data from a sample to draw conclusions about the larger population from which the sample was taken. Data often arises from the results of a survey of individuals. For example, the management of a fast food chain might be interested in determining the total number of dollars that Baylor students spend each year eating in Waco fast food restaurants. The fast food chain would also like to know the fast food preferences of the Baylor students. Both the information would be helpful in estimating how successful a particular fast food chain might be if located near the Baylor campus. The management of the fast food chain probably does not have the time or money to question every single Baylor student. Time and money to question 300 students might be available. Techniques exist for randomly choosing a representative sample of 300 students from a population of 17,000 Baylor students. An estimate or guess for the total dollars spent by all 17,000 Baylor students per year can be made from the amount spent by the sample of 300 students. You would not expect the estimate to be perfect. There would be some difference between the estimate based on the sample and the unknown total amount spent by all 17,000 students. The difference between the two amounts is called sampling error. The word, "error," does not refer to the fact that a mistake has been made. It simply means that you have based your findings on incomplete yet representative data. Procedures exist for estimating the maximum amount

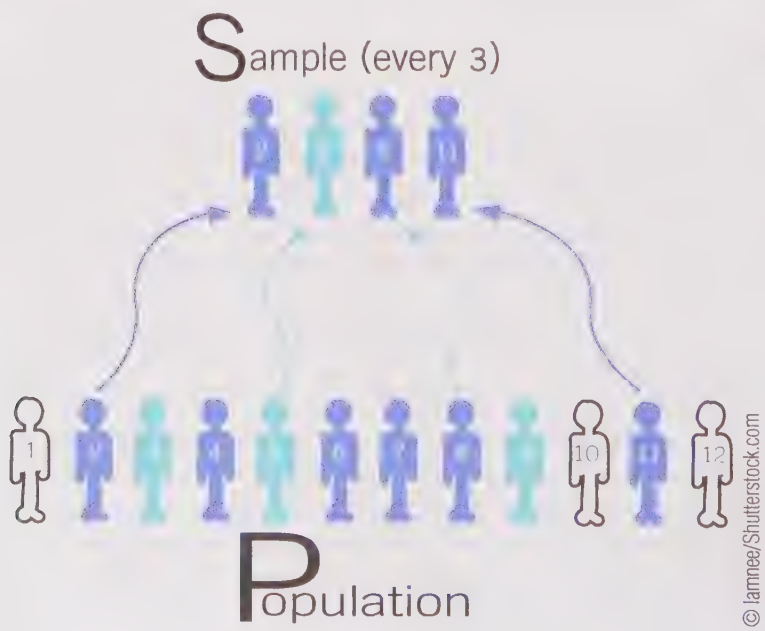
of sampling error. You will be able to state with a large degree of confidence how much difference there is between your estimate (a statistic) and the true answer (a parameter).

Terms

Random variable — A characteristic of interest such as the amount spent per year by a Baylor student eating at fast food restaurants in Waco.
Experimental unit — The entity on which a random variable is measured such as each randomly selected Baylor student.
Population — An entire collection of individuals, objects, or measurements whose characteristics are of some interest. The entire Baylor student body would be considered a population.
Parameter — A characteristic of a population. The total or average amount spent per year by all Baylor students eating at fast food restaurants in Waco is an example of a parameter. A parameter is typically unknown but constant.
Sample — A portion or subset of a population.
Statistic — A characteristic of a sample. The total or average amount spent per year by 300 randomly selected Baylor students eating at fast food restaurants in Waco would be considered a statistic. A statistic is variable depending onv the particular sample selected. The value of the statistic will be known once the sample is drawn. You use a statistic to provide an estimate or guess of a parameter. It is the value of the parameter that you are really interested in, but you usually have to accept an estimate of the parameter based on the statistic.
Sampling error — The difference between a parameter and a statistic. An estimate of the maximum amount of sampling error is based on the distribution of possible values that the statistic can assume from theoretical repeated samplings. The distributions of common statistics such as sample means and sample proportions are known. These distributions are referred to as sampling distributions.

Sampling Methods

There are four major sampling methods—(1) simple random, (2) systematic, (3) stratified, and (4) cluster.



Simple Random Sampling

A simple random sample is selected by a procedure such that each sample has the same chance of occurring as any other sample. Another characteristic of a simple random sample is that each element in the population has an equal chance of being selected. One way of achieving a simple random sample if you have a list of the units in the population is to assign a unique sequence of random digits to each unit. You can then either use a table of published random digits or computer-generated random digits to actually select a specific sample. Many statistical procedures assume that a simple random sample (SRS) of data is available for analysis.

Systematic Sampling

Assume you want to select a sample of 200 names from a master list of 5,000 names. One strategy would be to randomly select an integer between 1 and 25. Let's say that randomly selected integer is the number 10. The first sampled item would correspond to the 10th name from the list. Then you would include every 25th name after the 10th name—the 35th name, the 60th name, etc. Taking every 25th name was found by dividing the 5,000 by the 200. This process will not have exactly the same characteristics that a simple random sample has. If the master list does not have some recurring period, then a particular systematic sample will mimic closely the qualities of a simple random sample.

Stratified Sampling



© 4 PM production/Shutterstock.com

If the population that is being sampled can be divided into two or more groups in which the variable of interest is more homogeneous within the groups than in general. You then take either a simple random sample or maybe a systematic sample independently within each group. This procedure can have the effect of reducing the overall sampling error that could be achieved through simple random sampling across the entire population. For example, if you want to estimate the mean number of hours that all Baylor undergraduate students work in part-time jobs, it would probably be helpful to "stratify" (group) the students based on their classification—freshmen, sophomore, junior, and senior. You would then sample separately within each classification. The different classifications would most likely be homogeneous within each classification with respect to the number of hours worked in part-time jobs.

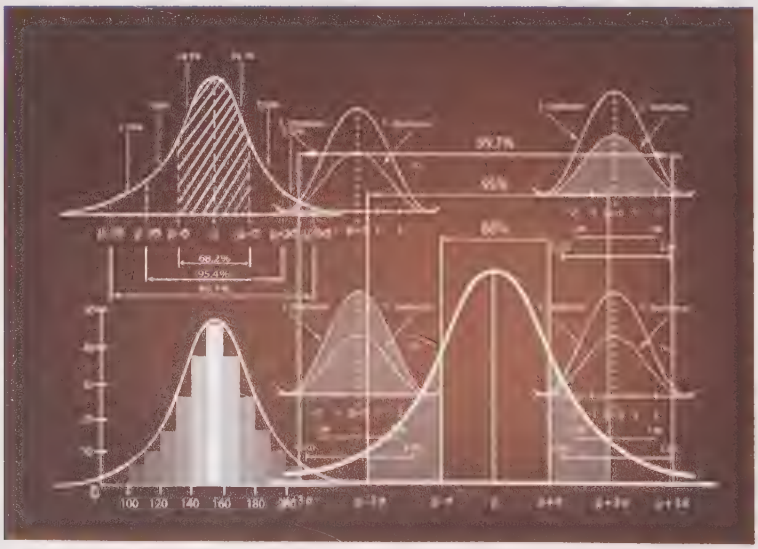
Cluster Sampling



© netsign33/Shutterstock.com

How do you sample if a list of the entire population does not exist? The answer is a sampling method known as cluster sampling. Cluster sampling involves at least two stages of sampling. For example, if you want to sample the households of a large city, you could first compile a list of all the residential streets of the city and take a simple random sample or systematic sample from this list. Then from each of just the sampled streets, you would compile an exhaustive list of the households on those streets. You would then take a simple random sample or systematic sample from the households of the sampled streets. This procedure may lead to a slightly higher sampling error than strict simple random sampling because you are using incomplete data at two stages rather than just one. However, the cost of sampling has a chance of being considerably lower with cluster sampling.

Sampling Distribution of the Sample Mean, $\bar{X}$



© lamnee/Shutterstock.com

Although there are some applications where you can assume that a random variable has an approximate normal distribution, the more widely used application of the normal distribution is in the theory of sampling distributions. A **sampling distribution** is the theoretical distribution of all the possible values that a statistic can assume with repeated sampling of a population. The two most common sampling distributions are those for the sample mean and sample proportion. A theorem from mathematics known as the **Central Limit Theorem** states that for a sufficiently large enough sample size ($n \geq 30$), the sampling distribution of the sample mean is approximately normally distributed or bell-shaped. The mean of the sampling distribution is equal to the mean of the population from which you are sampling. The standard deviation of the sampling distribution of the sample mean (**standard error**) is equal to the population standard deviation divided by the square root of the sample size. A similar theory exists for the sampling distribution of a sample proportion.

Central Limit Theorem—If n, the sample size, is sufficiently large, the sampling distribution of the sample mean can be approximated closely with a normal distribution. No requirement is placed on the shape of the population from which you are sampling. It is not necessary that the population from which you are sampling be normally distributed.

EXAMPLE 1

Assume the number of flashes before the battery of a digital camera needs to be recharged has a process distribution with a mean, μ, of 500 flashes and a standard deviation, σ, of 30 flashes. If repeated random samples of 36 batteries are tested and the sample mean number of flashes calculated, what distribution would you expect for the possible sample means?

SOLUTION



© Aliaksei Marozau/Shutterstock.com

Using the Central Limit Theorem, you would assume that the values of the sample means would be approximately normally distributed with a mean of 500 flashes and a standard deviation (sometimes referred to as the **standard error of** the **sample mean**) equal to (30 ÷ the square root of 36) or 5.

$$\textbf{Standard Error of the Sample Mean} = \frac{\sigma}{\sqrt{n}} = \frac{30}{\sqrt{36}} = 5 \text{ flashes}$$

A corresponding "margin of error" of the sampling estimation can be given by:

$$\textbf{Approximate 95\% Margin of Error} = 2\frac{\sigma}{\sqrt{n}} = 2\frac{30}{\sqrt{36}} = 10 \text{ flashes}$$

The maximum difference between any of the possible sample means and the population mean is equal to the "margin of error" with an approximate 95% surety. If you only have a sample of data from a population or process, you can state that your particular sample mean does not differ from the unknown population mean by more than the "margin of error" with an approximate 95% surety. There would be only an approximate 5% chance that any sample mean differs from the unknown population mean by more than the margin of error.

Inferential Calculations for the Mean of a Quantitative Variable

Point Estimates of μ and σ: $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$ $\quad s = \sqrt{\dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$

Standard Error of the Mean: $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}} \approx \dfrac{s}{\sqrt{n}}$

Bound (Margin) on Sampling Error: $(z \times \sigma_{\bar{x}})$

Sample Size Determination: $n = \left(\dfrac{z \times \sigma}{E}\right)^2$

EXAMPLE 2

A rental agent who is anxious to rent a location for a restaurant in a shopping center assures a restaurant owner that during an average business day at least 3,500 persons pass through the mall. Being a cautious businessperson, the restaurant owner conducts his own study and obtains a sample mean (average) of 3,350 persons with a sample standard deviation of 436 persons for 32 randomly selected days. Is the sample mean result (3,350) too far away from the claimed process or population mean (3,500) to cause the restaurant owner to doubt the rental agent's claim?



© somnai damrongpanich/Shutterstock.com

In determining if the restaurant owner should doubt the rental agent, you would compare the sample average based on a random sample of 32 days against the "claimed" population average. The population average would be the average of all days.

It is the population average or mean that the restaurant owner would be most concerned with and not the sample average value. You might hastily jump to the conclusion that since the sample average of 3,350 is less than the claimed population average, then the claimed average is too high. This line of reasoning ignores the sampling variation of the sample average. If you took another sample of 32 randomly selected days, you would probably get a value different from 3,350 due to sampling variation.

The question to be answered is—"Is it likely that the population average is actually 3,500 persons per day and that a sample average of only 3,350 or less occurs?

From the Central Limit Theorem, you can develop an estimate for how much different any one sample mean should differ from the unknown population mean with a high probability—say approximately 95%.

Calculations

1. Point estimate of μ is $\bar{x} = 3{,}350$ shoppers
2. Standard error of the mean is $\sigma_{\bar{x}} \approx \frac{s}{\sqrt{n}} = \frac{436}{\sqrt{32}} = 77.07$

   (You rarely know the population standard deviation, so you estimate its value with the sample standard deviation.)
3. Approximate 95% bound on sampling error is $2\sigma_{\bar{x}} = 2 \times 77.07 = 154.14$

   Margin of error $\approx$ 154 shoppers

The multiplier of the "2" in the formula comes from the fact that approximately 95% of the values of a normally distributed variable lies within 2 standard deviations of the mean. You can use other multipliers to alter the level of surety using Table 4.1.

Confidence Level (%)	Z Factor from the Standard Normal Table
80	1.28
90	1.64
95	1.96
Approx. 95	2.00
98	2.33
99	2.58

Table 4.1 Common *Z* values for Confidence Levels

Source: Fred Hulme

Since the observed sample mean of 3,350 shoppers is within the "margin of error" if the true population mean is 3,500 shoppers, then the sample mean would not be considered extremely unusual. This fact would lead you to conclude that the sample data is **not** in enough conflict to reject the claim that the true, unknown population mean equals 3,500 shoppers per day. If the sample mean had not fallen within the expected margin of error, you would have concluded that the claim was wrong.

You can use the Excel statistical function,

=CONFIDENCE(alpha, standard_dev, size)

to calculate the "margin of error." In the CONFIDENCE function,

alpha = 100% minus the level of confidence desired (i.e., use alpha =5% for a 95% level of confidence)

standard_dev = either the population standard deviation, σ, or an estimate of the population standard deviation, s, that can be computed with the STDEV Excel function from the sample data

size = the sample size

The CONFIDENCE function will calculate

$$Z \times \frac{\sigma}{\sqrt{n}} \text{ or } Z \times \frac{s}{\sqrt{n}}$$

EXAMPLE 3

M&M Marrs uses a quality control technique based on periodic sampling to monitor the number of pieces of Skittles placed in each bag. If the "bagging" machine is working correctly, each candy bag should have a mean of 60 pieces per bag. Assume the process (population) standard deviation is 2 pieces per bag.

Every hour a random sample of 30 bags is selected. Data from one of these samplings of 30 bags is in Table 4.2.



© Emilio100/Shutterstock.com

56	58	58	57	60	57
62	55	59	62	61	57
59	59	61	63	61	63
58	62	59	60	57	60
57	58	61	61	56	59

Table 4.2 Data for Skittles

Source: Fred Hulme

The sample mean for the above data is 59.2 pieces per bag. The Excel function, = **CONFIDENCE(5%, 2, 30)**, will calculate $1.96 \times \frac{2}{\sqrt{30}} = 0.716$. Therefore, the sample mean of 59.2 is outside the "margin of error" that would be expected if the process mean equals 60 pieces per bag. This sample mean would be considered somewhat unusual and would possibly cast doubt on the assumption that the process mean equals 60.

You can use the **MegaStat** command shown in Figure 4.1, **Confidence Intervals/Sample Size** (Confidence Interval—mean), to calculate the "margin of error."

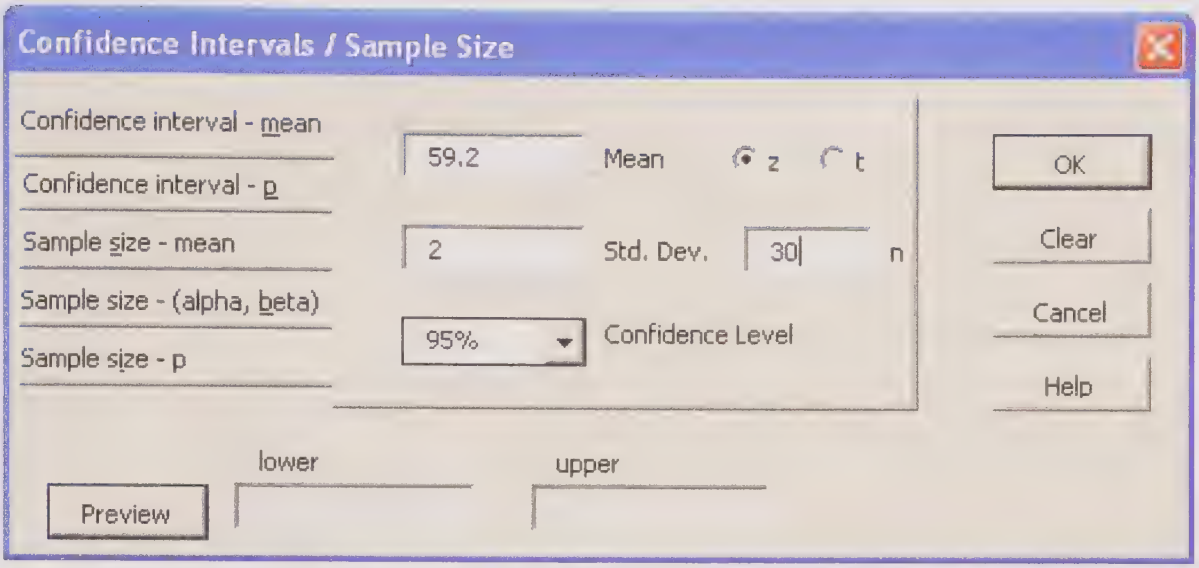


Figure 4.1 Dialog Box for a Confidence Interval
Source: Fred Hulme

MegaStat Results

Confidence Interval—Mean	
95%	Confidence level
59.2	Mean
2	Std. dev.
30	*n*
1.960	*z*
0.716	**Half-width**
59.916	Upper confidence limit
58.484	Lower confidence limit

Table 4.3 Skittles Output
Source: Fred Hulme

The "half-width" result is MegaStat's name for the "margin of error" calculation shown in Table 4.3.

The **MegaStat** command, **Descriptive Statistics** shown in Figure 4.2, can be used to estimate the population standard deviation if given the actual sample data. This estimate will be referred to as the "sample standard deviation."

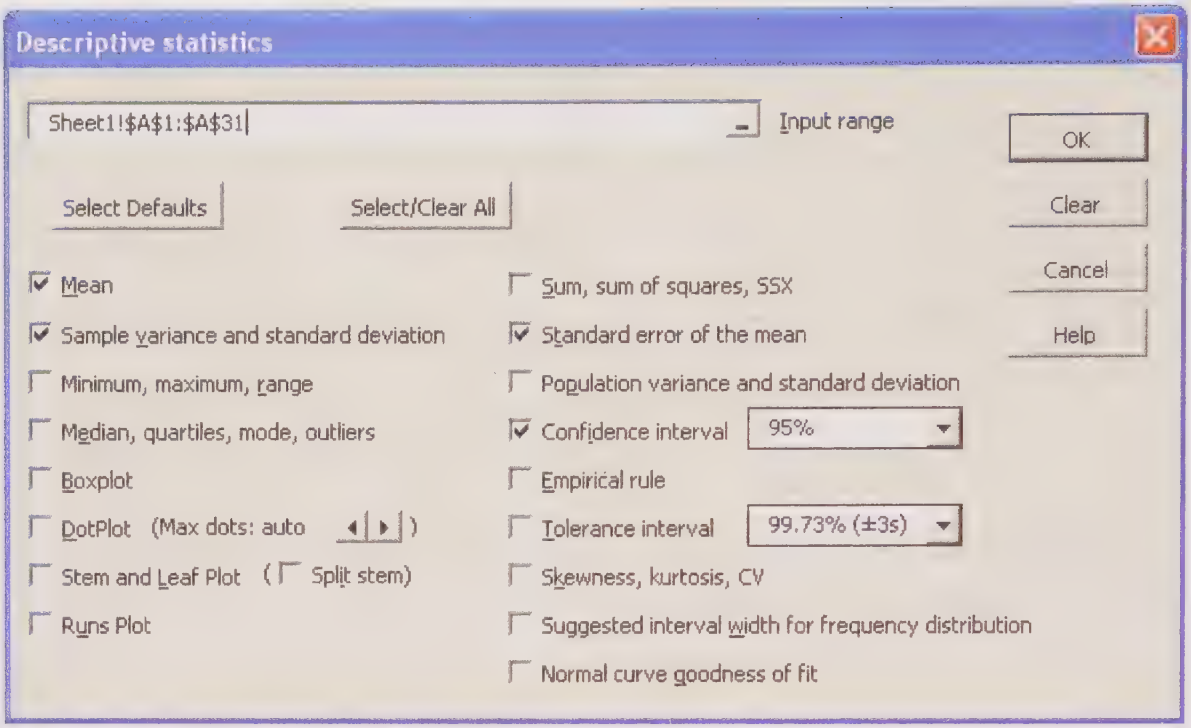


Figure 4.2 Descriptive Statistics Dialog Box
Source: Fred Hulme

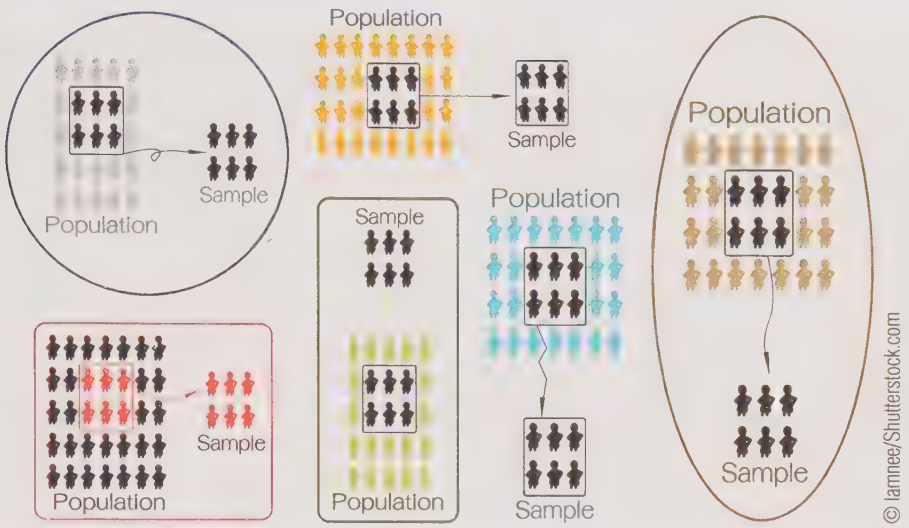
MegaStat Results

Descriptive Statistics	
	Pieces
Count	30
Mean	59.20
Sample variance	4.79
Sample standard deviation	2.19
Standard error of the mean	0.40
Confidence interval 95% lower	58.38
Confidence interval 95% upper	60.02
Half-width	0.82

Table 4.4 Skittles Descriptions
Source: Fred Hulme

When the sample size is 30 or more, the results from the Excel function, =**CONFIDENCE**, and the **MegaStat** command, **Descriptive Statistics** in Table 4.4, will give very similar results. For sample sizes less than 30, the results using the different procedures will differ. The difference will be greater for the smaller sample sizes. The MegaStat command, **Descriptive Statistics**, will use a confidence coefficient from the *t*-distribution rather than the normal distribution. Technically, if the population standard deviation is unknown (as it usually is), the *t*-distribution more accurately describes the correct values for the confidence coefficients.

Sample Size Determination



If you know before taking your sample how large a margin of error you desire, then you can calculate approximately how large a sample you will need for the desired accuracy. You will need an estimate of the population standard deviation, σ, in order to make the sample size determination. You could take a pilot sample in order to estimate σ, or you can use the following approximation to obtain a guess for σ:

$$\sigma \approx \frac{\text{Range}}{6}$$

To determine the necessary sample size such that the margin of error is no more than E units with an approximate 95% surety, you solve the following equation for n, the sample size.

$$E = 2 \times \left(\frac{\sigma}{\sqrt{n}} \right) \qquad n = \left(\frac{2 \times \sigma}{E} \right)^2$$

The more general formula that would allow for varying levels of surety would be:

$$n = \left(\frac{Z \times \sigma}{E} \right)^2$$

EXAMPLE 4



The calculated margin of error from Example 2 was approximately 154 shoppers. This was achieved with a sample size of 32 days. How much larger would the sample size have needed to be to reduce the margin of error to 100 shoppers? An estimate for σ was 436 shoppers.

$$n = \left(\frac{2 \times 436}{100}\right)^2 \approx 76 \text{ or } 77 \textbf{ days}$$

The MegaStat command, **Confidence Intervals/Sample Size – [Sample size—mean]** results in the following summary, which is shown in Table 4.5.

Sample Size—Mean	
100	*E*, error tolerance
436	Standard deviation
95.45%	Confidence level
2.000	*z*
76.039	Sample size
77	Rounded up

Table 4.5 Shopper Samples
Source: Fred Hulme

Sampling Distribution of the Sample Proportion, $\hat{p}$

If the random variable being observed is a qualitative variable, the appropriate data summary would be a proportion or fraction. For example, assume 300 Baylor students were asked if they were in favor of some new campus service. The response would simply be a "Yes" or a "No." There would not be a numeric response to average. The only summary for the data would be to find the proportion or fraction of the students in the sample that answered say "Yes" to the question. If you sampled a different set of 300 students, you would expect a different proportion of "Yes" answers. This type of data would have an underlying binomial distribution. The sampling error can be estimated by approximating a binomial distribution with a normal distribution. The sample proportion will be approximately normally distributed with the following mean and standard deviation (standard error) where p represents the population proportion.


© VGstockstudio/Shutterstock.com

$$\mu_{\hat{p}} = p$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Using this information about the distribution of the possible values for $\hat{p}$, an estimate of the "margin of error" can be made.

Inferential Calculations for the Proportion of a Qualitative Variable

Point Estimate: $\hat{p} = \frac{x}{n}$

Standard Error of the Proportion: $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Margin of Error: $Z \times \sigma_{\hat{p}}$

Sample size determination: $n = p(1-p)\left(\frac{Z}{\text{Desired margin of error}}\right)^2$

(Use $p = 0.5$ if no estimate of p is provided)

EXAMPLE 5

A manufacturer of perfume claims that 10% of all females of age 18 to 30 prefer its brand. A random sample of 1,200 females between 18 and 30 years old is selected, of which, 88 of the 1,200 females prefer the manufacturer's brand. What would you think about the manufacturer's claim?

Point Estimate: $\hat{p} = \frac{x}{n} = \frac{88}{1,200} = 0.0733$

Standard Error of the Proportion:

$$\sigma_{\hat{p}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.0733(1-.0733)}{1,200}} = 0.0075$$

Margin of Error: $=2(0.0075) = 0.015$

Sample size determination: $n = 0.50(1-0.50)\left(\frac{2}{0.04}\right)^2 = 625$ females (18–30 years old)

Looking at Table 4.6, since the sample proportion of 0.0733 is not within the margin of error of the claimed population proportion of 0.10, the manufacturer's claim would be in doubt.



© Matryoha/Shutterstock.com

If a margin of error of only 0.04 (4%) would have been acceptable and you had no prior information about the population proportion, a sample size of only 625 females (18–30 years old) would have required as seen in Table 4.7.

When you make conclusions based on incomplete data (i.e., a sample of only 300 students rather than all Baylor students), there is a chance of being wrong. However, when your decisions are based on random samples and statistics with known sampling distributions, you can estimate the risk of making an incorrect decision.

MEGASTAT SOLUTION FOR EXAMPLE 5

MegaStat Command: Confidence Intervals/Sample size—[Confidence interval—*p*] Confidence interval—proportion	
95%	Confidence level
0.073	Proportion
1,200	*n*
2.000	*z*
0.015	Half-width
0.088	Upper confidence limit
0.058	Lower confidence limit

Table 4.6 Perfume Proportions

Source: Fred Hulme

MegaStat Command: Confidence Intervals/Sample size—[Sample size—*p*] Sample size—proportion	
0.04	E, error tolerance
0.5	Estimated population proportion
95%	Confidence level
2.000	*z*
625.002	Sample size
626	Rounded up

Table 4.7 Sample Size

Source: Fred Hulme

Exercises

Short Answer/Discussion

1. Explain the reason why someone might choose to use only the incomplete information from a sample rather than conducting an exhaustive census of the data in question?

2. How does a parameter differ from a statistic?

3. What is meant by the term, "sampling error"?

4. What is meant by the term, "sampling distribution"?

5. What does the Central Limit Theorem say about the sampling distribution of the sample mean?

6. The standard deviation of the theoretical values that a statistic (i.e., the sample mean or sample proportion) can assume is known as a _____ _____.

7. If the observed value of a sample mean is more than the estimated "margin of error" from an assumed value for the population mean of a numeric variable, what conclusion can be made about the assumed value?

8. When is a proportion rather than a mean the appropriate data summary?

Problems



© Grzegorz Czapski/Shutterstock.com

1. A new car dealer sampled 100 car owners and found that the sample mean (average) number of miles driven last year was 15,000 miles with a sample standard deviation of 1,500 miles. Calculate the "margin of error."

2. Interpret the meaning of the "margin of error" measure from Problem (1).

3. Using the "margin of error" from Problem (1), what would you think about the claim that the population mean miles driven is 15,500?

4. How large a sample from Problem (1) would have been necessary to estimate the population average number of miles driven with a maximum error of 200 miles with an approximate 95% probability?

Problem 5 pertains to the following data.
The data represent the number of miles driven in a taxi fleet before a brake lining replacement is required. Table 4.8 represents the statistics from a random sample of 40 taxis from a finite population of 3,500 taxis.



© Leonard Zhukovsky/Shutterstock.com

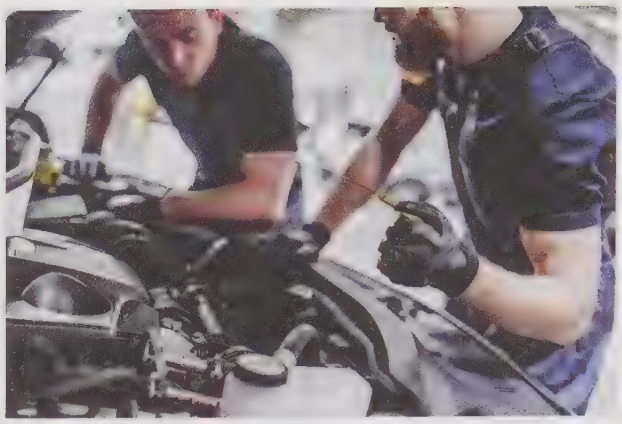
Mileage to Replacement (In thousands of miles)									
31	44	43	41	40	41	32	43	41	38
46	42	41	44	41	39	45	41	37	37
45	38	37	41	43	42	41	40	39	42
41	45	39	42	39	41	43	44	37	38

Descriptive Statistics	
	Miles
Count	40
Sample mean	40.60
Sample variance	10.35
Sample standard deviation	3.22
Minimum	31
Maximum	46
Range	15

Table 4.8 Taxi Statistics
Source: Fred Hulme

5. **a.** The mean mileage of all 3,500 taxis before a brake lining replacement is required would be considered a (a) sample, (b) population, (c) statistic, or (d) parameter.
   b. The mean mileage of the sample of 40 taxis before a brake lining replacement is required would be considered a (a) sample, (b) population, (c) statistic, or (d) parameter.
   c. Before taking the sample of 40 taxis, would the mean mileage of the 20 taxis be considered a (a) constant, (b) variable.
   d. Calculate the "standard error of the mean" using this data.
   e. Calculate the "margin of error" using this data.
   (f) Does it appear that the population mean mileage could be 42,000 miles? Why or why not? Use the margin of error in your assessment.
   (g) How large a sample would have been necessary to estimate the population mean within a maximum error of 1,200 miles? Use a 99% level of surety in your assessment?

Problem 6 pertains to the following data.
Most car manufacturers recommend changing a vehicle's oil every 3,000 miles or every 3 months whichever occurs first. A study was conducted to check how well car owners follow this suggestion. Thirty-six car owners are sampled and the mileage between oil changes was recorded in Table 4.9.29



© 4 PM production/Shutterstock.com

Mileage between Oil Changes					
3,903	3,722	3,729	3,874	3,690	3,727
3,566	3,798	3,853	3,855	3,532	3,719
3,732	3,835	3,544	3,529	3,674	3,769
3,852	3,917	3,898	3,947	3,775	3,918
3,735	3,843	3,898	3,718	3,774	3,799
3,682	3,560	3,777	3,781	3,782	3,770

Descriptive Statistics	Miles
Count	36
Mean	3,763.25
Sample standard deviation	114.08

Table 4.9 Proportion Oil Changes
Source: Fred Hulme

6. **a.** Would the mean mileage between oil changes of all car owners be considered a parameter or statistic?
   b. Would the sample mean be considered a parameter or statistic?
   c. One uses a _____ to estimate or guess a _____.
   d. Would the mean mileage of all the customers be considered a constant or variable prior to the sampling?
   e. Would the sample mean be considered a constant or variable prior to the sampling?
   f. Do we know anything about the possible values that the sample mean could assume? Do we know anything about the shape of the distribution of the possible values of the sample mean? What value would the possible sample means center around? What allows us to know this information?
   g. Calculate the "standard error of the mean." What does this number measure?

   (Help: The "standard error of the mean" $= \frac{\text{Standard Deviation}}{\sqrt{\text{SampleSize}}}$)

   h. With a 95% surety, how far away do you think any of the possible sample means would be from the population mean mileage between oil changes? This measure is sometimes referred to as "the bound on error" or "margin of error."

   Margin of Error = _____

   i. How large a sample would you have to take if you wanted the answer in Part (h) to be 25 miles?
   Sample Size = _____

7. A poll of 1,500 randomly selected union members was taken to estimate the proportion of the union that was in favor of a strike. 600 of the sampled members were in favor of a strike.



© bakdc/Shutterstock.com

a. Provide a "point estimate" of the proportion of all the union members who were in favor of a strike. A "point estimate" is just one single number.

b. Calculate the "standard error" for your estimate in (a).

c. What is the "margin of error" for the estimate in (a)? Use a surety level of approximately 95%.

d. How large a sample would be required to reduce the "margin of error" to only 0.01?

Self-Check

In questions 1 to 9, choose one of the two options in the parentheses.

1. The "sampling distribution of the mean," $\bar{x}$ describes the distribution of the possible values of the (sample, population) mean.

2. The sample mean, $\bar{x}$ is considered a (parameter, statistic).

3. A statistic is considered a (constant, variable) before the sample is taken.

4. We use a (parameter, statistic) to estimate a (parameter, statistic).

5. We would (expect, not expect) a point estimate to exactly equal the parameter being estimated.

6. The (sample standard deviation, standard error of the mean) measures how spread out the possible sample means, $\bar{x}$ could assume with repeated sampling.

7. The larger the inherited variation of the variable being sampled, the (smaller, larger) the standard error of the mean.

8. For the central limit theorem to apply, it is (necessary, not necessary) that all the population values being sampled have a normal (bell-shaped) distribution.

9. The central limit theorem (has, does not have) sample size requirements to be applicable.

10. If you were sampling the Baylor undergraduate population in order to estimate the mean numbers of hours that BU students work in part-time jobs during the school year, what advantage might there be by stratifying the sample on the students' classification (Freshmen, Sophomore, Junior, or Senior) in school?



© GaudiLab/Shutterstock.com

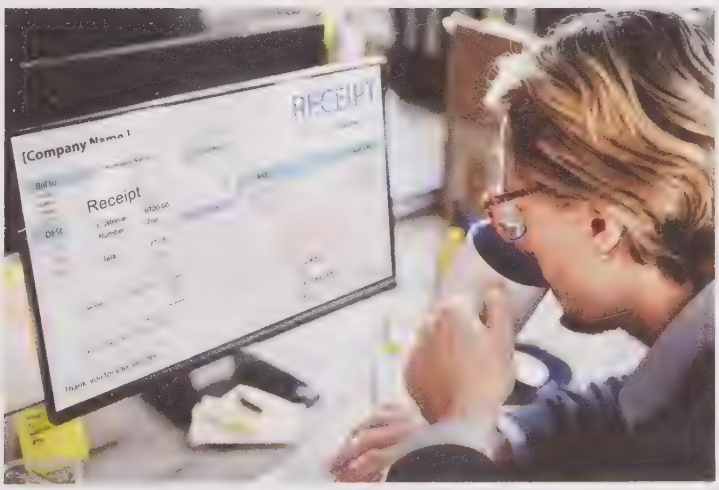
11. Explain how the sample in Question (10) would be obtained.

12. Assume you have a list of 10,000 names and telephone numbers. You have time and money to sample only 200 of the people from the list. You decide to use "systematic" sampling to obtain the names for your sample. Explain how "systematic" sampling works.

13. You have been given the task of sampling the households of a large city. You do not have a list of the addresses of all the households in the city. You do have a list of all the residential streets of the city. What method of sampling would be appropriate for this situation?

14. Explain how the sample would be obtained in the situation described in Question (13).

Questions 15 to 20 pertain to the following description.
Sarah Warner, comptroller of Beck's Inc., recently read an industry survey based on an analysis of 10,000 accounts receivable records collected from 75 firms. Among other things, the survey indicated that the average time between billing and payment was 15.7 days with a standard deviation of 4 days. Ms. Warner feels that this is substantially shorter than the time experienced by her company. In order to investigate if her company is performing below the national average, she takes a random sample of 49 accounts. The "descriptive statistics" results from the sample are given in Table 4.10.



© Rawpixel.com/Shutterstock.com

Descriptive Statistics					
Sample				Sample	
Variable	*n*	Mean	Std. Dev.	Minimum	Maximum
Days	49	16.6	7.0	9.0	44.0

Table 4.10 Beck's Inc Statistics
Source: Fred Hulme

15. Give a "point estimate" for the population mean days between billing and payment for Beck's Inc.

16. What do we know about the shape of the distribution of the possible values of the "point estimate" from Question (15)? How do we know this fact?

17. Calculate the "standard error of the mean."

18. What maximum difference would you expect between any sample mean and the population mean approximately 95% of the time?

19. Based on the "margin of error" estimate from Question (18), should Beck, Inc. be concerned that they are not doing as well as the industry overall ("smaller" means better for the number of days between billing and receipt of payment.) Why or why not?

20. How large a sample would have been required to decrease the "margin of error" to 1 day?

Chapter 4—End of Chapter Answers

Short Answer/Discussion

1. Exhaustive studies take too much time and money to complete; therefore, samples are used with a given level of confidence.

2. A parameter describes a population while a statistic describes a sample. Parameters are usually unknown but would be constant if a census were taken. Statistics are known but variable depending on the values chosen for the sample.

3. Sampling error is the difference between the parameter value minus the statistic. The error is due to the fact the data came from a sample rather than a census. (It is not human error.)

4. A sampling distribution results from taking repeated random samples of a population and is described by its shape, measure of central tendency, and standard error.

5. The Central Limit Theorem states that regardless of the population shape, the sampling distribution will be approximately normal for a sufficiently large sample size.

6. Standard error

7. It would be considered an outlier.

8. We use proportions to describe where qualitative variables tend to center and means to describe where quantitative variables center.

Problems

1. 150

2. Approximately 95% of the miles driven last year will be within ±2*150 or 300 miles.

3. 15,500 is over the upper limit for the estimated miles of 15,300. Thus, 15,500 is not a likely value.

4. n = 216.082 rounded up to 217

5. **a.** parameter
   b. statistic
   c. variable
   d. 3.22 / sqrt(40) = 0.509
   e. ±2*.509 = ±1.02
   f. 40.6 ± 1.02 = (39.58 to 41.62), the upper limit is 41,620 miles so 42,000 is not very likely.
   g. Take 1,200/1,000 to put it in the same units as the data for the error value. N = 47.77 rounded up to 48.

6. **a.** parameter
   b. statistic
   c. statistic to estimate a parameter
   d. constant
   e. variable
   f. Mean is approximately 3,763.25 with a normal distribution, and standard error of 114.08/6
   g. 114.08/6 = 19.013
   h. Bound on error = ±2*19.013 = ±38.026
   i. 80

7. **a.** 600/1,500
   b. 0.0126
   c. 0.0252
   d. 9,220

Self-Check

1. Population
2. Statistic
3. Variable
4. Statistic, Parameter
5. Not expect
6. Standard error of the mean
7. Larger
8. Not necessary
9. Has
10. The distribution of the stratified sample will have a greater chance of following the population distribution.
11. Obtain a list of BU students by classification, then randomly sample from each class the same number of students and record the hours of work for each student.
12. Survey every 50th household on the list.
13. Use cluster sampling.
14. Survey *n* number of houses from each neighborhood.
15. 16.6 days
16. It is normal *b/c* the sample size is ≥ 30.
17. 7/sqrt 49 = 1
18. ±2*1 = ±2 days
19. Beck's mean number of days between billing and payment will fall between 14.6 and 18.6 days approximately 95% of the time. Since the national average of 15.7 falls within this interval, Beck should not be concerned.
20. 189

CHAPTER

5

Statistical Estimation—Confidence Intervals



© Kendall Hunt Publishing Company

Introduction

This chapter applies the concepts from Chapter 4 using sample data to provide an estimate for a range of numbers that will include with a high chance, the value of an unknown population parameter. The unknown population parameter could be a mean or proportion of a random variable from one population or process. The unknown population parameter could also be the differences in the means or proportions of a random variable for two populations or processes. A distinction is made as to whether the population standard deviation is known or unknown in estimating a mean or the difference between two means. The unknown population standard deviation case will introduce a new statistical distribution—the ***t*-distribution**.

The resulting estimate from the sample data for the range of numbers will be referred to as an "**interval estimate**." Another name for this range is a "**confidence interval estimate**." Building contractors often give interval estimates on the length of time a project will require. They don't want to be tied to one particular finish date, so they give a range of time the project will probably take.

Confidence Interval for the Population Mean

EXAMPLE 1



© Kendall Hunt Publishing Company

A rental agent, anxious to rent a location for a restaurant in a shopping center, assures a restaurant owner that during an average business day at least 3,500 people pass through the shopping center. Being a cautious businessperson, the restaurant owner conducts his own study and obtains a sample mean (average) of 3,350 people for 32 randomly selected days. Is the sample mean result (3,350) too far away from the claimed process or population mean (3,500) to cause the restaurant owner to doubt the rental agent's claim?

In determining if the restaurant owner should doubt the rental agent, you would compare the sample average based on a random sample of 32 days against the "claimed" population average. The population average would be the average of all days. It is the population average or mean that the restaurant owner would be most concerned with and not the sample average value. You might hastily jump to the conclusion that since the sample average of 3,350 is less than the claimed population average that the claimed average is too high. This line of reasoning ignores the sampling variation of the sample average. If you took another sample of 32 randomly selected days, you would get a value different from 3,350 due to sampling error.

From the Central Limit Theorem, you can develop an upper and lower limit based on the sample mean that should contain the unknown population mean with a high probability. For this example, let us assume that the population standard deviation equals 436 persons per day. If you want to be approximately 95% confident that your interval contains the true population mean, μ, then your interval would be:

$$\text{Approximate 95\% Confidence Interval for } \mu\text{: } \bar{x} \pm \left(2 \times \frac{\sigma}{\sqrt{n}}\right)$$

In the above formula, $\bar{x}$ = the sample mean, σ = the population standard deviation, and n = the sample size.

The multiplier of the "2" in the formula comes from the fact that approximately 95% of the values of a normally distributed variable lies within 2 standard deviations of the mean. You can use other multipliers to alter the level of confidence of the interval. You can decrease the width of the interval by decreasing the level of confidence using Table 5.1. Narrower intervals are more precise, but you pay a price in that you are not as confident that you have captured the unknown population mean within your interval.

Confidence Level (%)	*Z* Factor from the Standard Normal Table
80	1.28
90	1.64
95	1.96
Approx. 95	2.00
98	2.33
99	2.58

Table 5.1 Common *Z* Factors

Source: Fred Hulme

An approximate 95% confidence interval for the population mean number of shoppers is:

$$3{,}350 \pm \left(2 \times \frac{436}{\sqrt{32}}\right)$$

$$(3{,}196,\ 3{,}504)$$

$$3{,}350 \pm (2 \times 77.07)$$

The interpretation of the above interval is: *Due to the sampling procedure used and using 95% confidence, the true population mean number of shoppers falls between 3,196 and 3,504 per day.* That means that 95 out of 100 times, when taking a sample from this population, the real average number of shoppers per day will be in this interval. Since the interval, (3,196, 3,504) is one of those intervals, you are approximately 95% confident that the true population mean number of shoppers is contained between 3,196 and 3,504. This sampling process would produce intervals only 5% of the time that would not contain the true, unknown population mean. Since the rental agent's claim that the true population mean number of shoppers is at least 3,500 **and** 3,500 is a possible value for the true population mean number of shoppers based on the confidence interval, you would conclude that the sample data is **not** in enough conflict with the claim to reject the claim. If the upper limit of the confidence interval had been less than 3,500, you would have concluded that the claim was probably too high.

Confidence Interval for the Population Mean—Population Standard Deviation Known

Point Estimate of μ:	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$	$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$
Standard Error of the Mean:	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$	
Confidence Interval for μ:	$\bar{x} \pm (z \times \sigma_{\bar{x}})$	

You can use the Excel statistical function,

= CONFIDENCE(alpha, standard_dev, size)

to calculate the quantity that you add to and subtract from the sample mean to give the upper and lower limits of the confidence interval. This value is called the **Margin of Error or Bound on Error**. In the CONFIDENCE function, alpha = 100% minus the level of confidence desired (i.e., use alpha = 5% for a 95% level of confidence) std. dev. = the assumed population standard deviation size = the sample size

The CONFIDENCE function will calculate: $Z \times \frac{\sigma}{\sqrt{n}}$.

You could then calculate:

$$\bar{x} \pm \left(Z \times \frac{\sigma}{\sqrt{n}}\right)$$

to give the lower and upper limits of the confidence interval.

EXAMPLE 2

A candy manufacturer uses a quality control technique based on periodic sampling to monitor the number of pieces of candy placed in each bag. If the "bagging" machine is working correctly, each candy bag should have an average of 60 pieces per bag. Assume the process (population) standard deviation is 2 pieces per bag.

Every hour a random sample of 30 bags is selected. The data below were found from one of these samplings of 30 bags in Table 5.2



© Kendall Hunt Publishing Company

56	58	58	57	60	57
62	55	59	62	61	57
59	59	61	63	61	63
58	62	59	60	57	60
57	58	61	61	56	59

Table 5.2 M&M Data
Source: Fred Hulme

The sample mean (average) for the above data is 59.2 pieces per bag. The Excel function, **=CONFIDENCE(5%, 2, 30),** will calculate 0.716. The 95% confidence interval for estimating the unknown process mean number of pieces per bag is shown in Table 5.3.

$$59.2 \pm 0.716 \text{ or } (58.484, 59.916)$$

95% Confidence Interval for the Mean Pieces per Bag	
Mean	59.2
1/2 Width of Confidence Interval	0.716
Upper Limit of the Confidence Interval	59.916
Lower Limit of the Confidence Interval	58.484

Table 5.3 Excel Output M&M Interval Estimate/Bag
Source: Fred Hulme

You can use the **MegaStat** command, **Confidence Intervals/Sample Size** (Confidence Interval—mean), to calculate the confidence interval limits shown in Figure 5.1.

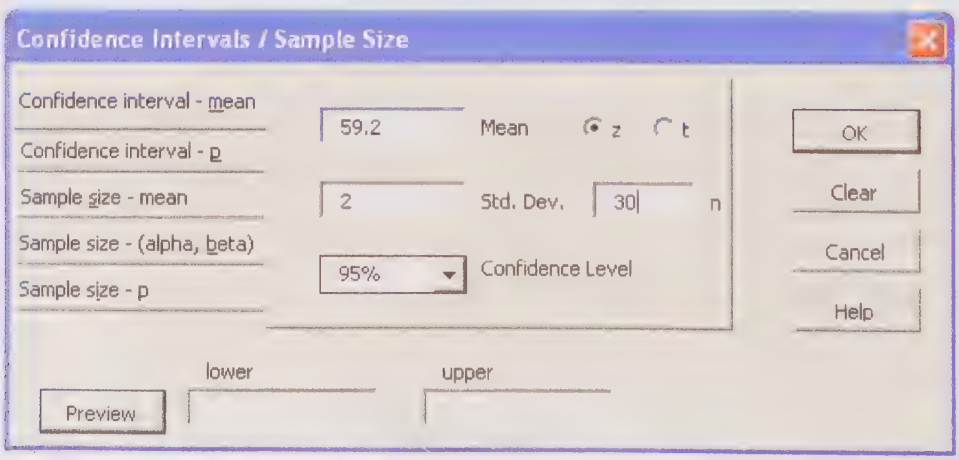


Figure 5.1 MegaStat Confidence Interval Dialog Box
Source: Fred Hulme

MegaStat Results

Confidence Interval—Mean	
95%	Confidence level
59.2	Mean
2	Std. dev.
30	*n*
1.960	*z*
0.716	Half-width
59.916	Upper confidence limit
58.484	Lower confidence limit

Table 5.4 MegaStat Output for M&Ms
Source: Fred Hulme

The interpretation of the interval in Table 5.4 is as follows: Due to the sampling procedure used, 95% of the time an interval containing the true, unknown population mean number of pieces of candy per bag will be between 58.48 and 59.92. If the goal is to average 60 pieces per bag, then this sample and corresponding confidence interval would cause concern that the process mean is too low.

Confidence Interval for the Population Mean—Population Standard Deviation Unknown

If the population or process standard deviation is unknown, you can still calculate a confidence interval for the unknown population mean of a numeric variable, but it will be based on the ***t*-distribution** rather than the normal distribution. The following ratio will have a *t*-distribution with degrees of freedom (df) $= n - 1$:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

After a little algebra, the following probability statement can be made:

$$\text{Prob}[\,\bar{x} - t_{(n-1,0\,.95)} \times \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{(n-1,0.95)} \times \frac{s}{\sqrt{n}}\,] = 0.95$$

95% Confidence Interval for μ with the population standard deviation unknown:

$$\bar{x} \pm (t_{(n-1,0\,.95)} \times \frac{s}{\sqrt{n}})$$

The confidence coefficient, $t_{(n-1,\,0.95)}$, can be found using the Excel function:

$$\textbf{=tinv(0.05, } n \textbf{ - 1).}$$

In general,

$(1 - \alpha) \times 100\%$ Confidence Interval for μ with the population standard deviation unknown:

$$\bar{x} \pm \left(t_{(n-1,\,1-\alpha)} \times \frac{s}{\sqrt{n}} \right)$$

The confidence coefficient, $t_{(n-1,\,1-\alpha)}$, can be found using the Excel function, **=tinv(α, n − 1)** or the *t*-Table 5.5.

		t-Distribution—Confidence Intervals				
df	**80%**	**90%**	**95%**	**98%**	**99%**	**99.90%**
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.599
3	1.638	2.353	3.182	4.541	5.841	12.924
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.869
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.408
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.768
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
.	.	.	.	.	.	.
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.358	2.617	3.373
Infinity	1.282	1.645	1.960	2.326	2.576	3.291

Table. 5.5 *t*-Distribution
Source: Fred Hulme

EXAMPLE 3

The sample statistics for the "candy bagging" data are in Table 5.6.

Sample Statistics	
n = Sample Size	30
$\bar{X}$ = Sample Mean	59.20
s = Sample Standard Deviation	2.19

Table 5.6 Descriptive Statistics for M&M
Source: Fred Hulme


© Kendall Hunt Publishing Company

The appropriate confidence coefficient from the t-distribution table for the "candy bagging" data would be:

$$t_{(30-1=29,0.95)} = 2.045$$

95% Confidence Interval for μ, the population mean number of pieces per bag (population standard deviation unknown):

$$59.20 \pm \left(2.045 \times \frac{2.19}{\sqrt{30}} \right)$$

$$(58.38, 60.02)$$

The interval, (58.38, 60.02), is the result of a sampling procedure that 95% of the time would lead to an interval that would contain the unknown population mean number of pieces of candy per bag.

The **MegaStat** command, **Descriptive Statistics,** is used to estimate the population standard deviation from the sample. This estimate will be referred to as the "sample standard deviation." With the following options, the sample size, mean, sample variance, sample standard deviation, standard error of the mean, a 95% confidence interval estimate will be calculated and displayed in Figure 5.2.

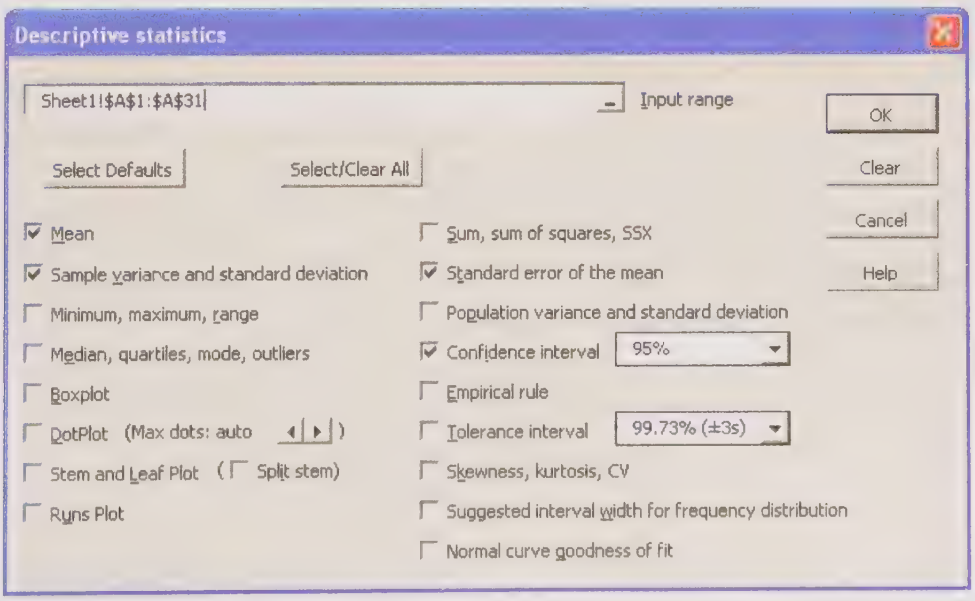


Figure 5.2 MegaStat Dialog Box
Source: Fred Hulme

MegaStat Results

Descriptive Statistics	Pieces
Count	30
Mean	59.20
Sample variance	4.79
Sample standard deviation	2.19
Standard error of the mean	0.40
Confidence interval 95% lower	58.38
Confidence interval 95% upper	60.02
Half-width	0.82

Table 5.7 Descriptive Statistics
Source: Fred Hulme

Looking at Table 5.7, when the sample size is 30 or more, the results from the Excel function, **=CONFIDENCE**, and the **MegaStat** command, **Descriptive Statistics**, will give very similar results. For sample sizes less than 30, the results using the different procedures will differ. The difference will be greater for the smaller sample sizes. The MegaStat command, **Descriptive Statistics**, will use a confidence coefficient from the *t*-distribution rather than the normal distribution. Technically, if the population standard deviation is unknown (as it usually is), the *t*-distribution more accurately describes the correct values for the confidence coefficients.

Use of the *t*-distribution for the confidence coefficient requires an additional assumption that the population that you are sampling is at least approximately normally distributed. The values for the confidence coefficients from the *t*-distribution will be slightly larger than the corresponding ones from the standard normal distribution. As the sample size increases, the *t*-distribution approaches the standard normal distribution.

Estimating Population Proportions

If the random variable being observed is a qualitative variable, the appropriate data summary would be a proportion or fraction. For example, assume 300 Baylor students were asked if they were in favor of some new campus service. The response would simply be a "Yes" or a "No." There would not be a numeric response to average. The only summary for the data would be to find the proportion or fraction of the students in the sample that answered say "Yes" to the question. If you sampled a different set of 300 students, you would expect a different proportion of "Yes" answers. This type of data would have an underlying binomial distribution. The sampling error can be estimated by approximating a binomial distribution with a normal distribution. The sample proportion is approximately normally distributed with the following mean and standard deviation (standard error) where p represents the population proportion.

$$\mu_{\hat{p}} = p \text{ (where } \hat{p} \text{ is the sample fraction)} = x/n$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Using this measurement of the sampling error, a confidence interval for estimating the population proportion or fraction can be made.

Confidence Interval for the Population Proportion, *p*

Point Estimate of *p*: $\hat{p} = \frac{x}{n}$

Standard Error of the Sample Proportion, $\hat{p}$: $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Confidence Interval for *p*: $\hat{p} \pm (z \times \sigma_{\hat{p}})$



© Kendall Hunt Publishing Company

EXAMPLE 4

A manufacturer of perfume claims that 10% of all females age 18 to 30 prefer its brand. A random sample of 1,200 females between 18 and 30 years old is selected, and 88 of the 1,200 females prefer the manufacturer's brand. What would you think about the manufacturer's claim?

Point Estimate: $\hat{p} = \frac{x}{n} = \frac{88}{1,200} = .0733$

Standard Error of the Sample Proportion:

$$\sigma_{\hat{p}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.0733(1-0.0733)}{1,200}} = 0.0075$$

Approximate 95% CI for p: $\hat{p} \pm (z \times \sigma_{\hat{p}})$

$0.0733 \pm [2 \times (0.0075)]$

0.0733 ± 0.0150

$(0.0583, 0.0883)$

or $(5.83\%, 8.83\%)$

The interval (5.83%, 8.83%) is the result of a sampling procedure that approximately 95% of the time will produce an interval that would contain the unknown population proportion of 18 to 30-year-old females that favor the manufacturer's brand. Since the manufacturer's claim of 10% is not contained in the confidence interval, the manufacturer's claim would be in doubt.

When you make conclusions based on incomplete data (i.e., a sample of only 300 students rather than all Baylor students), there is a chance of being wrong. However, when your decisions are based on random samples and statistics with known sampling distributions, you can estimate the risk of making an incorrect decision.

Confidence Interval—Difference Between Population Means (Assuming Independent Samples and Equal Population Standard Deviations)

You can use a confidence interval estimate to compare two population means using independent samples. The confidence interval produces an estimate for the difference between two unknown population means—$\mu_1 - \mu_2$. If a confidence interval for $\mu_1 - \mu_2$ does not contain zero, this is an indication that the unknown populations are probably different.

Sampling Distribution of $\bar{X}_1 - \bar{X}_2$

The following ratio will have a t-distribution with degrees of freedom $(df) = n_1 + n_2 - 2$.

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad \text{where } s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

After a little algebra, the following confidence interval statement can be made:

$(1 - \alpha) \times 100\%$ Confidence Interval for $(\mu_1 - \mu_2)$:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{(n_1+n_2-2,\,1-\alpha)} \times \sqrt{s_{Pooled}^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$\text{Where } s_{Pooled}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

EXAMPLE 5

The following data represent the number of miles driven in a taxi fleet before a brake lining replacement is required. Two different brands of brake linings are to be compared based on the mean mileage before a brake lining replacement is required.



© Kendall Hunt Publishing Company

Mileage to Replacement (in thousands of miles)

Brand A						Brand B				
37	43	34	51	33		41	41	43	40	35
42	45	39	36	39		45	34	35	33	34
36	43	39	38	38		40	36	40	38	32
39	36	38	39	39		35	30	37	45	48
49	47	39	31	42		40	37	36	34	38
33	31	42	38	40		27	36	37	41	35
36	35	39	43	43		34	34	36	43	41
37	42	45	47	36		39	49	32	41	35
38	42	47	42	38		42	39	40	45	40
42	43	34	36	39		31	41	42	30	38
41	35	39	41	46		37	41	36	36	37
39	41	33	28	39		36	32	39	38	38
37	37	42	37	38		38	37	34	40	44
31	44	43	41	40		41	32	43	41	38
46	42	41	44	41		39	45	41	37	37

Descriptive Statistics

Brand A	Brand B	
39.55	38.03	Mean
4.38	4.23	Std. dev.
75	75	*n*

Table 5.8 Descriptive Statistics for Taxis
Source: Fred Hulme

$$s^2_{Pooled} = \frac{(75-1)(4.38^2)+(75-1)(4.23^2)}{75+75-2} = 18.558; = \text{tinv}(.05, 75+75-2) = 1.976$$

95% Confidence Interval for $\mu_A - \mu_B$ using Table 5.8

$$(39.55 - 38.03) \pm 1.976\sqrt{(18.558)\left(\frac{1}{75}+\frac{1}{75}\right)}$$

$$(0.130, 2.910)$$

This interval is the result of a sampling procedure that 95% of the time would result in an interval that would contain the true, unknown difference of $\mu_A - \mu_B$. Since zero is not contained in this interval, you could conclude that the population means for the two brands are different. Since the Brand A sample mean was larger, you could infer that Brand A should be preferred to Brand B. The following MegaStat command could be used to find the confidence interval for the difference between two population means. You would need to select the "**Display Confidence Interval**" option.

MegaStat Command: Hypothesis Tests/Compare Two Independent Groups shown in Figure 5.3.

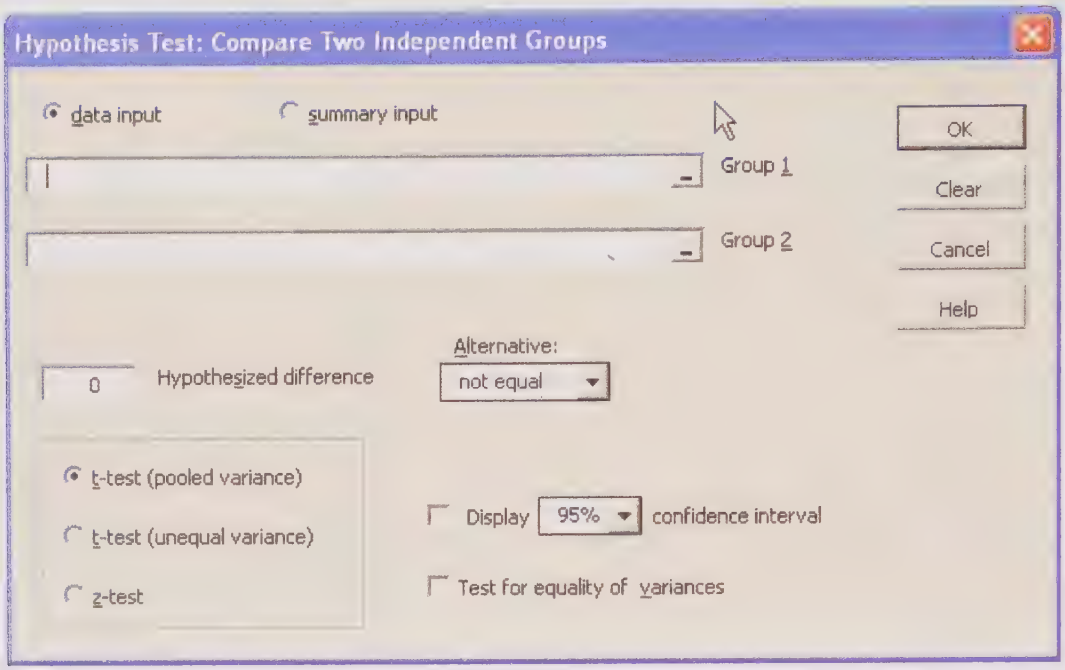


Figure 5.3 MegaStat Dialog Box Two Independent Means
Source: Fred Hulme

Results from *t*-test option (pooled variance) are in Table 5.9.

Hypothesis Test: Independent Groups (*t*-test, Pooled Variance)		
Brand A	**Brand B**	
39.55	38.03	mean
4.38	4.23	Std. dev.
75	75	*n*
	148	*df*
	1.520	Difference (Brand A − Brand B)
	18.558	Pooled variance
	4.308	Pooled std. dev.
	0.703	Standard error of difference
	0	Hypothesized difference
	2.16	*T*
	0.0323	*p*-value (two-tailed)
	0.130	Confidence interval 95.% lower
	2.910	Confidence interval 95.% upper
	1.390	Margin of error

Table 5.9 Output for Taxi Data
Source: Fred Hulme

Since zero is not in the interval estimate of 0.13 to 2.91, there is a significant difference in the two brands, 95% of the time, based on Table 5.9.

Confidence Interval—Difference Between Population Proportions

Sampling Distribution of $\bar{p}_1 - \bar{p}_2$

The values of the statistic, $\hat{p}_1 - \hat{p}_2$, will be approximately normally distributed with a mean equal to $p_1 - p_2$ and a standard deviation equal to $\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$. From this fact, a confidence interval for $p_1 - p_2$ can be developed.

Confidence Interval for $(p_1 - p_2)$:

$$(\hat{p}_1 - \hat{p}_2) \pm z\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

EXAMPLE 6

A telemarketing company gives a sales aptitude test to job applicants. The company wants to know if a particular question discriminates well between good salespersons and poor salespersons. The following table gives the relevant data in Table 5.10.

	"Good" salespersons	"Poor" salespersons
Correct answer	72	55
Incorrect answer	28	45

Table 5.10 Aptitude Test Data
Source: Fred Hulme



© Kendall Hunt Publishing Company

If the question acts as a good discriminator, then there should be a significantly higher fraction (proportion) of "good" salespersons answering the question correctly than the "poor" salespersons.

Descriptive Statistics

$$\hat{p}_1 = \frac{72}{100} = 0.72 \qquad \hat{p}_2 = \frac{55}{100} = 0.55$$

95% Confidence Interval for $(p_1 - p_2)$:

$$(0.72 - 0.55) \pm 1.96\sqrt{\frac{0.72(1-0.72)}{100} + \frac{0.55(1-0.55)}{100}}$$

(0.0387, 0.3013)

MegaStat Command: Hypothesis Tests/Compare Two Independent Proportions is in Figure 5.4.

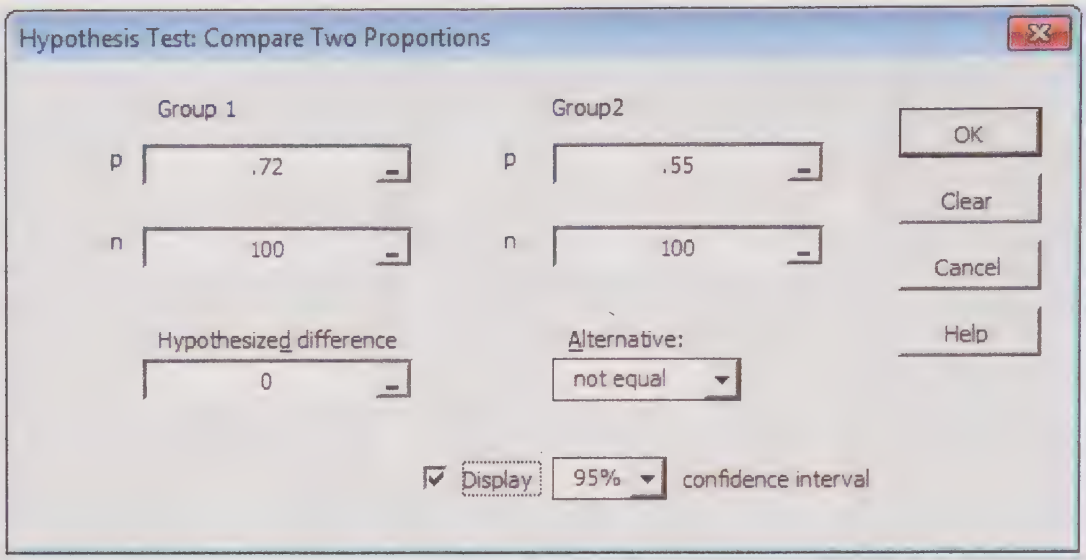


Figure 5.4 Dialog Box for Aptitude Data
Source: Fred Hulme

The interpretation of Table 5.11 should state that since the confidence interval does not contain zero, the population proportions are probably different, which implies that the question is a good discriminator of aptitude, 95% of the time.

Hypothesis Test for Two Independent Proportions			
p_1	p_2	p_c	
0.72	0.55	0.635	*p* (as decimal)
72/100	55/100	127/200	*p* (as fraction)
72	55	127	*X*
100	100	200	*n*
	0.17		Difference
	0.		Hypothesized difference
	0.0681		Std. error
	2.50		*z*
	0.0125		*p*-value (two-tailed)
	0.0387		Confidence interval 95% lower
	0.3013		Confidence interval 95% upper
	0.1313		Margin of error

Table 5.11 Aptitude Output
Source: Fred Hulme

Exercises

Short Answer/Discussion

1. What information does a confidence interval convey about a parameter?

2. What is the relationship between the width of a confidence interval and the size of the sample?

3. What is the relationship between the width of a confidence interval and the level of confidence?

4. If an assumed value for the population mean of a numeric variable does not fall within an observed confidence interval, what conclusion do you draw about the assumed value?

5. When is a proportion rather than a mean the appropriate data summary?

Problems

1. A new car dealer sampled 100 car owners and found that the sample mean (average) number of miles driven last year was 15,000 miles. Find a 95% confidence interval for estimating the population mean average number of miles driven. Assume the population standard deviation equals 1,500 miles. Use the exact z value rather than the empirical rule.



© Kendall Hunt Publishing Company

2. Using the sample results from Problem (1), change the level of confidence to 99%.

Problems 3 to 6 require the following information: The Evergreen Company operates retail pharmacies in 10 eastern states. Recently the company's internal audit department selected a random sample of 50 prescriptions issued throughout the system to customers under the age of 50. The objective of the sampling was to estimate the mean (average) dollar value of all prescriptions issued to such people by the company. The Excel output at right summarizes this data.



© Kendall Hunt Publishing Company

Descriptive Statistics	
Mean	56.92
Standard deviation	8.0
Minimum	16.38
Maximum	108.30
Count	50

Table 5.12 Evergreen Pharmacy
Source: Fred Hulme

3. Give a "point estimate" from the sample data in Table 5.12 to estimate the population mean (average) prescription cost.

4. Find a "confidence interval" estimate for the population average prescription cost. Assume a **95%** level of confidence for your interval.

5. Choose the correct interpretation of "confidence" in Problem (4).
   a. The range of the prescription costs will be in the interval in repeated sampling.
   b. For this sample only, we can say that the population mean prescription cost will be in the interval constructed with probability 0.95.
   c. The population mean prescription cost will be in intervals constructed this way 95% of the time in repeated sampling.
   d. The sample mean prescription cost will be in intervals constructed in this way 95% of the time in repeated sampling.
   e. none of the above.

6. Before the data displayed above was collected, it had been suggested that the mean (population) prescription cost equaled $58 during the last year. Use your confidence interval from Problem (4) to decide if you believe this claim or not claim. Explain your answer.

Problem 7 pertains to the following data.

The following data represent the number of miles driven in a taxi fleet before a brake lining replacement is required. The data represents a random sample of 20 taxis from a finite population of 3,500 taxis.

Mileage to Replacement (In thousands of miles)									
31	44	43	41	40	41	32	43	41	38
46	42	41	44	41	39	45	41	37	37

Descriptive Statistics	Miles
Count	20
Sample mean	40.35
Sample variance	15.08
Sample standard deviation	3.88
Minimum	31
Maximum	46
Range	15

Table 5.13 Taxi Statistics
Source: Fred Hulme



© Kendall Hunt Publishing Company

7. **a.** The mean mileage of all 3,500 taxis before a brake lining replacement is required would be considered a (a) sample, (b) population, (c) statistic, or (d) parameter.
   b. The mean mileage of the sample of 20 taxis before a brake lining replacement is required would be considered a (a) sample, (b) population, (c) statistic, or (d) parameter.
   c. Before taking the sample of 20 taxis, would the mean mileage of the 20 taxis be considered a (a) constant, (b) variable.
   d. The appropriate confidence coefficient used in the calculation for a confidence interval estimate should be based on the (a) normal probability table or (b) *t*-distribution table.
   e. Calculate the "standard error of the mean" using this data.
   f. Based on the above data, calculate the upper and lower limits of a 99% confidence interval for estimating the population mean mileage of all 3,500 taxis before a brake lining replacement is required.
   g. Choose one of the following that would provide an appropriate interpretation of the interval estimate in part (f). Assume the interval you found was (LL, UL), where LL = lower limit and UL = upper limit.
      i. 99% of all the taxis would have a mileage between (LL, UL).
      ii. 99% of the 20 sampled taxis would have a mileage between (LL, UL)
      iii. The interval (LL, UL) is the result of a sampling process that approximately 99 out of 100 times would produce an interval that would contain the true unknown population mean mileage.
      iv. None of the above accurately describes the meaning of the interval.
   h. Does it appear that the population mean mileage could be 42,000 miles? Why or why not? Use your interval estimate with a 99% level of confidence in your assessment.
   i. You decide that the interval in part (f) is too wide to be useful. What two options are available to you?

8. In a random sample of 600 television viewers contacted by phone in a certain suburban area, 210 were watching a movie on Netflix.
   1. Estimate the proportion of all the television viewers that were watching a movie on Netflix?
   2. Calculate the "standard error" of your estimate in part (a).
   3. With an approximate 95% surety, how much would your estimate in (a) differ from the proportion of all the television viewers that were watching a movie on Netflix?
   4. Calculate the approximate 95% confidence interval for estimating the proportion of all the television viewers that were a watching the movie on Netflix.
   5. What would you conclude from your survey results if the management of Netflix claimed that 45% off all viewers watched a movie?



© Kendall Hunt Publishing Company

9. A poll of 1,500 randomly selected union members was taken to estimate the proportion of the union that was in favor of a strike. 600 of the sampled members were in favor of a strike.
   a. Provide a "point estimate" of the proportion of all the union members who were in favor of a strike. A "point estimate" is just one single number.
   b. Calculate the "standard error" for your estimate in (a).
   c. Find the approximate 95% confidence interval estimate of the proportion of all the union members who were in favor of a strike.
   d. You decide that the interval in (c) is too wide to be useful. What 2 options are available to you?



© Kendall Hunt Publishing Company

Self-Check

In questions 1 to 13, choose one of the two options in the parentheses.

1. The sample mean, provides a (point, interval) estimate of the population mean,

2. A point estimate (does, does not) provide an estimate of its degree of accuracy.

3. As the sample size increases, the width of the 95% confidence interval for , the population mean, (increases, decreases) while holding all other factors constant.

4. If you increase the level of confidence of a confidence interval, the width of the interval (increases, decreases) while holding all other factors constant.

5. The sample mean, is considered a (parameter, statistic).

6. A statistic is considered a (constant, variable) before the sample is taken.

7. We use a (parameter, statistic) to estimate a (parameter, statistic).

8. We would (expect, not expect) a point estimate to exactly equal the parameter being estimated.

9. The (sample standard deviation, standard error of the mean) measures how spread out the possible sample means, could assume with repeated sampling.

10. The larger the inherited variation of the variable being sampled, the (smaller, larger) the standard error of the mean.

11. As the sample size increases, the difference between the z and t-distribution becomes (larger, smaller).

12. To find the appropriate t-value used in a confidence interval for based on a small sample, one enters the row with the degrees of freedom (df) equal to (n, $n - 1$).

13. You should use the t-distribution confidence coefficient in a confidence interval for when the sample size, n, is (smaller, larger) than (30, 50).

The bad debt ratio, the dollars of loans defaulted (i.e., loans that were paid back in full) divided by the total amount loaned, on new cars was measured on a random sample of 36 banks and the output is in Table 5.14.

Descriptive Statistics

Ratio	
Sample mean	4.65%
Sample standard deviation	0.60%
Range	1.87%
Minimum	3.48%
Maximum	5.35%
Count	36

Table 5.14 Defaulted Loans Ratio
Source: Fred Hulme



© Kendall Hunt Publishing Company

14. Using the sample results, give a "point estimate" for the population mean debt ratio.

15. What is known about the shape of the distribution of the possible values of the "point estimate" from (14)?

16. Calculate the "standard error of the sample mean."

17. Give the 95% confidence limits on the estimation for the population mean debt ratio based on the sample data.

18. Choose <u>one</u> of the following that would provide an appropriate interpretation of the interval estimate in (17). Assume the interval you found was (XX,YY), where XX = lower limit and YY = upper limit.
    - i. Approximately 95% of <u>all</u> the banks would have a bad debt ratio between (XX,YY)
    - ii. Approximately 95% of the 36 sampled banks would have a bad debt ratio between (XX,YY)
    - iii. The interval (XX,YY) is the result of a sampling process that approximately 95 out of 100 times would produce an interval that would contain the true unknown population mean bad debt ratio.
    - iv. None of the above accurately describes the meaning of the interval.

19. Based on the confidence interval estimate in (17), is there overwhelming evidence that the population mean debt ratio of all banks is equal to 5%? (i.e., would 5% be a reasonable answer for the population mean debt ratio.)

20. If you ask each of 400 owners of the newly designed VW Tiguan if they purchased the extended warranty that was available (Yes or No), the response would be considered:

a. qualitative data
b. discrete data
c. continuous data
d. ordinal data
e. none of the above



© Kendall Hunt Publishing Company

21. What type of data summary would make sense for the data described in (20)?

a. mean
b. average
c. standard deviation
d. proportion
e. median

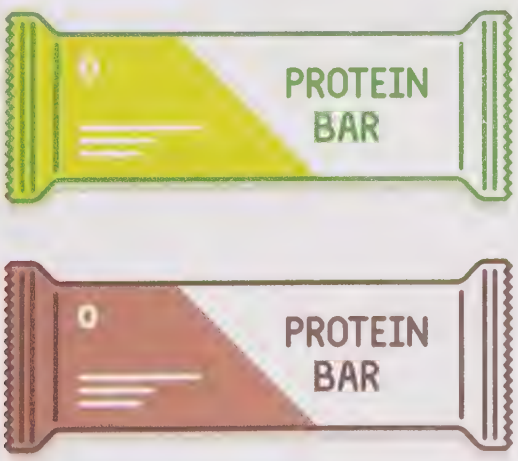
22. If 100 of the 400 owners responded that they purchased the extended warranty, your best point estimate of the proportion of all new VW Tiguan owners who purchased the extended warranty would be:

a. 0.75
b. 0.25
c. 0.50
d. 0.09
e. None of the above

23. A confidence interval with a 99% level of confidence for estimating the proportion of all VW Tiguan owners who purchased the extended warranty would be:

(_____ , _____)

24. If the management of VW claimed that 28% of all new Tiguan owners will purchase the extended warranty, what do you think about this claim?



© Kendall Hunt Publishing Company

To study the effectiveness of two different kinds (A vs. B) of packaging, a processor of breakfast foods puts a breakfast bar packaged in two different ways into 30 different supermarkets (60 supermarkets in all), obtaining the following summary of the sales (between 9 a.m. and noon on a given day):

25. Find a 95% confidence interval estimate for $\mu_A - \mu_B$ using the sample results in Table 5.15.

Package A	Package B	
31	25.8	Mean
7.38	5.76	Std. dev.
30	30	*n*

Table 5.15 Packaging
Source: Fred Hulme

26. Based on the confidence interval estimate found in (26), what would you conclude about the two types of packages?

A bank conducted a survey to compare the attitudes of young married customers (those having recently opened their first bank account) with the attitudes of older, established customers to their new automated teller system (ATM). A summary of the customers' responses follows.



© Kendall Hunt Publishing Company

	Younger Customers	Older Customers	Totals
Favor the ATM	87	65	152
Favor personal service	113	135	248
Totals	200	200	400

Table 5.16 ATM Survey
Source: Fred Hulme

27. Find a 95% confidence interval estimate for $p_{\text{Young}} - p_{\text{Old}}$ using the sample results in Table 5.16.

28. Based on the confidence interval estimate found in (27), what would you conclude about the attitudes toward ATM usage of younger vs. older customers?

Chapter 5—End of Chapter Answers

Short Answer/Discussion

1. % of the time the True population parameter falls in the interval.
2. N increases the width decreases.
3. More confidence = wider interval
4. It is not a likely value.
5. Use for qualitative data

Answers to Problems

1. 15,000 ± 1.96 (1,500/sqrt(100))
2. $Z = 2.58$

   15,000 ± 2.58(150)
3. Give a "point estimate" from the sample data to estimate the population mean (average) prescription cost. 56.92
4. Find a "confidence interval" estimate for the population average prescription cost. Assume a **95%** level of confidence for your interval.

   56.92 ± 1.96(8/sqrt(50)) = (54.7, 59.14)
5. The interpretation of "confidence" in Problem (4) is as follows.
   c. The population mean prescription cost will be in intervals constructed this way 95% of the time in repeated sampling.
6. 58 is in the interval; therefore, it is a possible cost.
7. **a.** The mean mileage of all 3,500 taxis before a brake lining replacement is required would be considered a (a) sample, (b) population, (c) statistic, or (d) parameter.
   b. The mean mileage of the sample of 20 taxis before a brake lining replacement is required would be considered a (a) sample, (b) population, (c) statistic, or (d) parameter.
   c. Before taking the sample of 20 taxis, would the mean mileage of the 20 taxis be considered a (a) constant, (b) variable.
   d. The appropriate confidence coefficient used in the calculation for a confidence interval estimate should be based on the (a) normal probability table or (b) t-distribution table.
   e. Calculate the "standard error of the mean" using this data.

   3.88/sqrt(20) = 0.867

f. Based on the above data, calculate the upper and lower limits of a 99% confidence interval for estimating the population mean mileage of all 3,500 taxis before a brake lining replacement is required.

$40.35 \pm 2.58(0.867) = (37.87 \text{ to } 42.83)$

g. Choose one of the following that would provide an appropriate interpretation of the interval estimate in part (f). Assume the interval you found was (LL, UL), where LL = lower limit and UL = upper limit.

i. 99% of all the taxis would have a mileage between (LL, UL).

ii. 99% of the 20 sampled taxis would have a mileage between (LL, UL)

iii. The interval, (LL, UL) is the result of a sampling process that approximately 99 out of 100 times would produce an interval that would contain the true unknown population mean mileage.

iv. None of the above accurately describes the meaning of the interval.

h. Does it appear that the population mean mileage could be 42,000 miles? Why or why not? Use your interval estimate with a 99% level of confidence in your assessment.

Yes, 42 is in the interval.

i. You decide that the interval in part (f) is too wide to be useful. What 2 options are available to you?

Take a larger sample or use less confidence.

8. In a random sample of 600 television viewers contacted by phone in a certain suburban area, 210 were watching the movie on Channel 12.

1. Estimate the proportion of all the television viewers that were watching the movie on Channel 12?

$P = 210/600 = 0.35$

2. Calculate the "standard error" of your estimate in part (a).

Sqrt((.35*.65)/600) = 0.0195

3. With an approximate 95% surety, how much would your estimate in (a) differ from the proportion of all the television viewers that were watching the movie on Channel 12?

Margin of Error = 1.96 * 0.0195 = ±0.038

4. Calculate the approximate 95% confidence interval for estimating the proportion of all the television viewers that were watching the movie on Channel 12.

$0.35 \pm 0.038 = (0.31, 0.39)$

5. What would you conclude from your survey results if the management of Channel 12 claimed that 45% off all viewers watched the movie?

Not likely, since 0.45 is not in the interval.

9. A poll of 1,500 randomly selected union members was taken to estimate the proportion of the union that was in favor of a strike. 600 of the sampled members were in favor of a strike.
   a. Provide a "point estimate" of the proportion of all the union members who were in favor of a strike. A "point estimate" is just one single number.

   600/1,500 = 0.40

   b. Calculate the "standard error" for your estimate in (a).

   0.013

   c. Find the approximate 95% confidence interval estimate of the proportion of all the union members who were in favor of a strike.

   $0.4 \pm 0.025 = (0.375, 0.425)$

   d. You decide that the interval in (c) is too wide to be useful. What 2 options are available to you? Take a larger sample or use less confidence.

Self-Check

In questions 1 to 13, choose one of the two options in the parentheses.

1. The sample mean, $\bar{x}$, provides a (point, interval) estimate of the population mean,
2. A point estimate (does, does not) provide an estimate of its degree of accuracy.
3. As the sample size increases, the width of the 95% confidence interval for, the population mean (increases, decreases), while holding all other factors constant.
4. If you increase the level of confidence of a confidence interval, the width of the interval (increases, decreases) while holding all other factors constant.
5. The sample mean, $\bar{x}$, is considered a (parameter, statistic).
6. A statistic is considered a (constant, variable) before the sample is taken.
7. We use a (parameter, statistic) to estimate a (parameter, statistic).
8. We would (expect, not expect) a point estimate to exactly equal the parameter being estimated.
9. The (sample standard deviation, standard error of the mean) measures how spread out the possible sample means, $\bar{x}$, could assume with repeated sampling.
10. The larger the inherited variation of the variable being sampled, the (smaller, larger) the standard error of the mean.
11. As the sample size increases, the difference between the z and t distribution becomes (larger, smaller).

12. To find the appropriate t-value used in a confidence interval for based on a small sample, one enters the row with the degrees of freedom (df) equal to ($n, n - 1$).

13. You should use the t-distribution confidence coefficient in a confidence interval forwhen the sample size, n, is (smaller, larger) than (30, 50).

The bad debt ratio, the dollars of loans defaulted (i.e., loans that were paid back in full) divided by the total amount loaned, on new cars was measured on a random sample of 36 banks.

Descriptive Statistics

Ratio	
Sample Mean	4.65%
Sample Standard Deviation	0.60%
Range	1.87%
Minimum	3.48%
Maximum	5.35%
Count	36

14. Using the sample results, give a "point estimate" for the population mean debt ratio.

4.65

15. What do we know about the shape of the distribution of the possible values of the "point estimate" from (14)?

Normal

16. Calculate the "standard error of the sample mean".

0.6/sqrt(36) = 0.1

17. Give the 95% confidence limits on the estimation for the population mean debt ratio based on the sample data.

$4.65 \pm 1.96(0.1) = (4.45, 4.85)$

18. Choose <u>one</u> of the following that would provide an appropriate interpretation of the interval estimate in (17). Assume the interval you found was (XX,YY), where XX = lower limit and

YY = upper limit.

iv. Approximately 95% of <u>all</u> the banks would have a bad debt ratio between (XX,YY)
v. Approximately 95% of the 36 sampled banks would have a bad debt ratio between (XX,YY)
vi. The interval (XX,YY) is the result of a sampling process that approximately 95 out of 100 times would produce an interval that would contain the true unknown population mean bad debt ratio.
vii. None of the above accurately describes the meaning of the interval.

19. Based on the confidence interval estimate in (17), is there overwhelming evidence that the population mean debt ratio of all banks is equal to 5%? (i.e., would 5% be a reasonable answer for the population mean debt ratio.) No, 5% is beyond the interval's range.

20. If you ask each of 400 owners of the newly designed 2012 Honda CR-V if they purchased the extended warranty that was available (Yes or No), the response would be considered:
    a. qualitative data
    b. discrete data
    c. continuous data
    d. ordinal data
    e. none of the above

21. What type of data summary would make sense for the data described in (20)?
    a. mean
    b. average
    c. standard deviation
    d. proportion
    e. median

22. If 100 of the 400 owners responded that they purchased the extended warranty, your best point estimate of the proportion of all 2012 Honda CR-V's owners who purchased the extended warranty would be:
    a. 0.75
    b. 0.25
    c. 0.50
    d. 0.09
    e. None of the above

23. A confidence interval with a 99% level of confidence for estimating the proportion of all 2012 Honda CR-V's owners who purchased the extended warranty would be:

    0.25 ± 2.58* sqrt((0.25*0.75)/400)

    (0.194, 0.306)

24. If the management of Honda claimed that 28% of all 2012 Honda CR-V's owners will purchase the extended warranty, what do you think about this claim?

    28% is a possible value.

To study the effectiveness of two different kinds (*A* vs. *B*) of packaging, a processor of breakfast foods puts a breakfast food packaged in the two different ways into 30 different supermarkets (60 supermarkets in all), obtaining the following summary of the sales (between 9 a.m. and noon on a given day):

Packaging A	Packaging B	
31	25.8	Mean
7.38	5.76	Std. dev.
30	30	*n*

25. Find a 95% confidence interval estimate for $\mu_A - \mu_B$ using the sample results.

Given from Excel Output.

1.8500	Confidence interval 95% lower
8.5500	Confidence interval 95% upper

26. Based on the confidence interval estimate found in (26), what would you conclude about the two types of packages?

Due to the sampling procedure used, I'm 95% confident the mean difference in sales is between 1.85 and 8.55. (It is a significant difference.)

A bank conducted a survey to compare the attitudes of young married customers (those having recently opened their first bank account) with the attitudes of older, established customers to their new automated teller system (ATM). A summary of the customers' responses is shown below.

	Younger customers	Older customers	Totals
Favor the ATM	87	65	152
Favor personal service	113	135	248
Totals	200	200	400

27. Find a 95% confidence interval estimate for $p_{\text{Young}} - p_{\text{Old}}$ using the sample results.

P(Young for ATM) = 87/200 = 0.435

p(Old for ATM) = 65/200 = 0.325

0.0155	Confidence interval 95.% lower
0.2045	Confidence interval 95.% upper

28. Based on the confidence interval estimate found in (27), what would you conclude about the attitudes toward ATM usage of younger vs. older customers?

I'm 95% confident, due to the sampling procedure used there is a significant diff. in the proportion of young and old customers favoring the ATM. It's btw 1.55% and 20.45%

CHAPTER

6

Hypothesis Testing



© Kendall Hunt Publishing Company

Introduction

Another approach to drawing conclusions about the unknown parameters of populations from sample data is a procedure known as **hypothesis testing** or **significance testing**. In a hypothesis testing situation, a claim about a specific value for a population parameter is made. A sample from the population provides a point estimate for the unknown population parameter. Using the sampling distribution of the point estimate, you answer how likely would the sample result have occurred if the claim about the population parameter were true. If the chance is low, then the sample result casts doubt about the claim being made on the value of the unknown population parameter.

Steps in Setting Up a Hypothesis

1. The symbol H_0 represents the null hypothesis, which would indicate the sample value does not statistically differ from a given or known value. The symbol H_a or in some texts H_1 represents the alternative hypothesis, which shows the claimed or theorized value is true.

2. Parameter symbols are always used in the hypothesis setup. μ, P or Π, $\mu_1 - \mu_2$, and $P_1 - P_2$ will be used in this chapter.

3. The equal sign is typically used in H_0, and either $<$, $>$, or $\neq$ is placed in H_a.

4. The right-hand side of the hypothesis is a value based on past information, claims, or "yardstick" values we are testing the sample against.

EXAMPLE 1—ONE SAMPLE MEAN CASE

A candy manufacturer uses a quality control technique based on periodic sampling to monitor the number of pieces placed in each bag. If the "bagging" machine is working correctly, each bag should have an average or mean of 60 pieces per bag. The manufacturer is most concerned about the process average falling below 60 per bag. They want to minimize the risk of allowing a defective machine to produce bags that would "short" the customer.

a. Give the null hypothesis, H_0, and the alternative hypothesis, H_a, for this situation.

H_0: $\mu = 60$ (the "null" hypothesis)

H_a: $\mu < 60$ (the "alternative" hypothesis)



If H_0 appears more believable than H_a, then you conclude that the machine is "in control" and working properly. If H_a appears more believable than H_0, then you conclude that the machine is "out of control" and not working properly.

This test is referred to as a **"one-tail" or "one-sided"** test in that you are interested in determining if the process mean, μ, is less than 60. You are checking for a difference in a specific direction. More specifically, the test would be considered a "lower, one-tail" test because of the "$<$" sign in the alternative hypothesis, H_a.

b. The quality control personnel take a random sample of the finished bags of candy. Prior to drawing the sample and deciding from the sample results if the machine is "in control" or "out of control," there is a chance of making a wrong decision or conclusion. For your hypotheses above, discuss what it would mean if a **Type I error** were made. **Type II error**?

	Fail to Reject H_0	**Reject H_0**
H_0 is true	Correct decision	**Type I Error**
H_0 is false	**Type II Error**	Correct decision

Table 6.1 Type I and Type II Errors
Source: Fred Hulme

© Kendall Hunt Publishing Company

From Table 6.1, a **Type I error** is made when you reject a true null hypothesis (H_0). For this situation, a Type I error would be made if the sample data caused the candy manufacturer to conclude that the machine was "out of control" when it was actually "in control." For quality control testing, the chance of a Type I error is sometimes referred to as the **"Producer's Risk."**

A **Type II error** is made when you fail to reject a false null hypothesis (H_0). For this situation, a Type II error would be made if the sample data caused the candy manufacturer to conclude that the machine was "in control" when it was actually "out of control." For quality control testing, the chance of a Type II error is sometimes referred to as the **"Consumer's Risk."**

C. Every hour a random sample of 20 bags is selected. The data below were found from one of these samplings of 20 bags:

56	58	58	57
57	55	59	62
59	59	61	63
58	62	59	60
57	58	61	61

Table 6.2 Candy Sample
Source: Fred Hulme

Based on the sample in Table 6.2, you can decide if there is sufficient evidence that the machine is "out of control." In making this decision, you assume that the machine is "in control" (i.e., that H_0 is true) and calculate the chance or likelihood that your particular sample mean or more unusual could occur. This chance is referred to as the "***p*-value**" of the test. If the calculated p-value is small, say less than 0.05, you would conclude that it is unlikely that H_0 is true and lead you to "reject H_0." This would mean you think the machine is "out of control." If the "p-value" is small, then it is probably not chance alone as to why the sample mean is being considered unusual. The more likely reason for the value of an unusual sample mean is that the null hypothesis, H_0, is false. The calculation of the "p-value" for the test can be found with the following:

Use the Excel command, **Data/Data Analysis/Descriptive Statistics**, to calculate the sample mean and sample standard deviation of the sample data. You could also use the MegaStat command, **Descriptive Statistics**, to get the sample statistics in Table 6.3.

Pieces	
Mean	59
Standard error	0.481227
Median	59
Mode	59
Standard deviation	2.15211
Sample variance	4.631579
Kurtosis	−0.65658
Skewness	0.140806
Range	8
Minimum	55
Maximum	63
999,999	1,180
Count	20

Table 6.3 Candy Excel Output
Source: Fred Hulme

Using these results, calculate the "**test statistic**" with the following formula:

$$t = \frac{(\bar{x} - \mu_0)}{\frac{s}{\sqrt{n}}} = \frac{(59 - 60)}{\frac{2.152}{\sqrt{20}}} = -2.08$$

The Excel function, =**TDIST(*x*, deg_freedom, tails)** can be used to calculate the *p*-value. The inputs to =**TDIST** are:

x = the absolute value of the "t" you calculated from above

Deg_freedom = $n - 1$, where n = the sample size

tails = 1 if using a one-tail test; 2 if using a two-tail test

***p*-value: =TDIST(2.08, 19, 1) = 0.0257**

Conclusion: Reject H_0 since the calculated *p*-value of 0.0257 is less than 0.05. The sample mean equal to 59 pieces per bag would be considered unusual or rare if the process or population mean is 60 pieces per bag. Therefore, from the sample data results, it appears that the machine should be considered "out of control."

d. Explain what the *p*-value found in c. represents. Why are we interested in the *p*-value?

The *p*-value = 0.0257 means that if the machine had been "in control" (i.e., $\mu = 60$), then there is only a 2.57% chance of getting a sample mean based for 20 bags of 59 pieces or less. If the *p*-value for a test is small (say less than 0.05), you conclude that H_0 is probably false. The comparison number used in deciding if the *p*-value is sufficiently small to conclude that H_0 is false is known as the "**level of significance**" of the test. The common values for the level of significance of a hypothesis test are 0.10, 0.05, or 0.01.

e. In your decision from part c., what type of error might you be making?

Since the decision was to "reject H_0," a Type I error may have possibly occurred. This would mean the candy manufacturer was misled by the sample data in concluding that the machine was "out of control" when it was actually "in control."

f. If this process is repeated many, many times, what is the chance or probability of the error that might have been made in part e.?

The probability of a **Type I error** is equal to the **level of significance** of the test. A level of significance of 0.05 was used in this example. Therefore, there would be a chance of 0.05 of declaring that the machine was "out of control" when it was actually "in control."

g. Compare your answers to the **MegaStat** results using the command, **Hypothesis Tests/Mean vs. Hypothesized Value**.

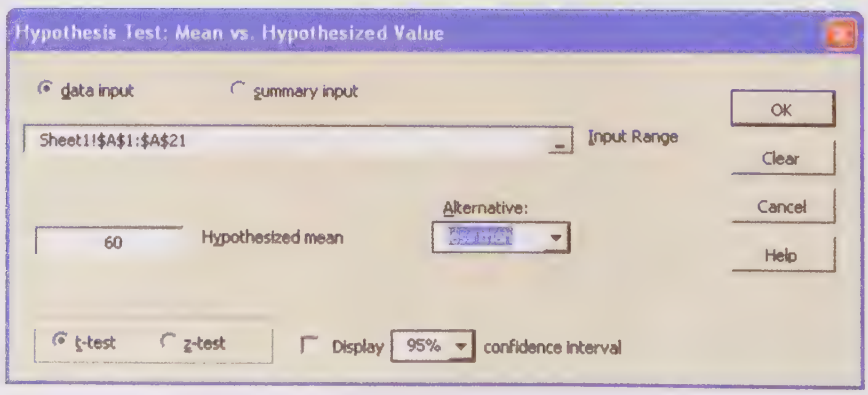


Figure 6.1 Hypothesis Test for a Single Mean Value

Source: Fred Hulme

The input screen assumes that the 20 data values have been entered into an Excel spreadsheet in cells A1 through A21 (with cell A1 containing a title or label for the data—"Pieces") is shown in Figure 6.1 and the output in Table 6.4.

MegaStat Results

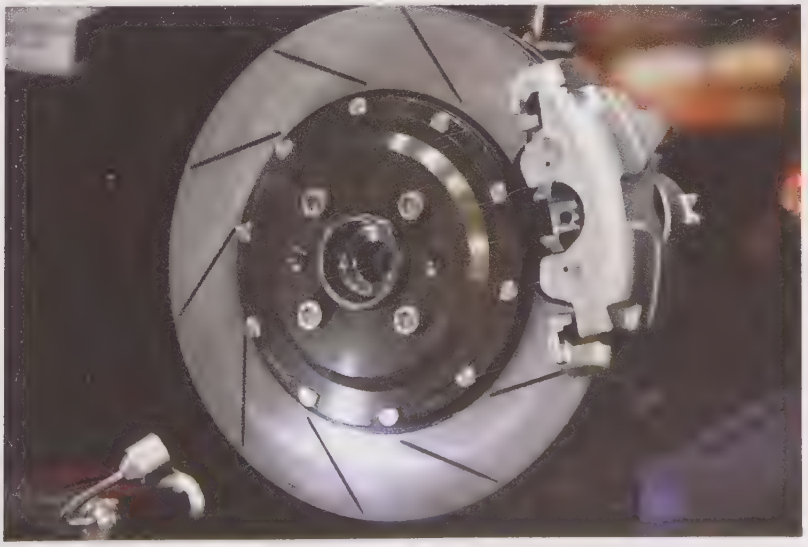
Hypothesis Test: Mean vs. Hypothesized Value	
60.000	Hypothesized value
59.000	Mean pieces
2.152	Std. dev.
0.481	Std. error
20	*n*
19	*df*
−2.08	*t*
0.0258	*p*-value (one-tailed, lower)

Table 6.4 Candy MegaStat Output
Source: Fred Hulme

You can conduct a hypothesis test to compare the means of two samples as to whether they come from populations with equal means. If the standard deviations for the two populations are known and you have sufficiently large sample sizes (both samples sizes ≥ 30), then you can base the hypothesis test on a normally distributed, Z, statistic. If the standard deviations for the two populations are unknown but can be assumed approximately equal, then you can base the hypothesis test on a test statistic with an assumed t-distribution. An alternative procedure exists in MegaStat for comparing two population means when the population standard deviations can't be assumed equal.

EXAMPLE 2—TWO-SAMPLE TEST—DIFFERENCE BETWEEN POPULATION MEANS (ASSUMING INDEPENDENT SAMPLES AND EQUAL VARIANCES)

The following data represent the number of miles driven in a taxi fleet before a brake lining replacement is required. Two different brands of brake linings are to be compared.



© Kendall Hunt Publishing Company

Mileage to Replacement (in thousands of miles)

Brand A						Brand B				
37	43	34	51	33		41	41	43	40	35
42	45	39	36	39		45	34	35	33	34
36	43	39	38	38		40	36	40	38	32
39	36	38	39	39		35	30	37	45	48
49	47	39	31	42		40	37	36	34	38
33	31	42	38	40		27	36	37	41	35
36	35	39	43	43		34	34	36	43	41
37	42	45	47	36		39	49	32	41	35
38	42	47	42	38		42	39	40	45	40
42	43	34	36	39		31	41	42	30	38
41	35	39	41	46		37	41	36	36	37
39	41	33	28	39		36	32	39	38	38
37	37	42	37	38		38	37	34	40	44
31	44	43	41	40		41	32	43	41	38
46	42	41	44	41		39	45	41	37	37

Should the two brands be considered different based on the mean mileage before a brake lining replacement is required assuming no cost difference? Perform the appropriate hypothesis test to answer this question. Use a level of significance of $\alpha = 0.05$.

a. Give the null hypothesis, H_0, and the alternative hypothesis, H_a, for this situation.

H_0: $\mu_A = \mu_B$ or $\mu_A - \mu_B = 0$ (the "null" hypothesis)
H_a: $\mu_A \neq \mu_B$ or $\mu_A - \mu_B \neq 0$ (the "alternative" hypothesis)

This is an example of a "two-tail" or "two-sided" test because of the "≠" sign in the alternate hypothesis, H_a. You are just testing if the means are equal or not. You aren't specifying in the alternate hypothesis a specific direction for the difference—simply stating that there IS a difference.

If H_0 appears more believable than H_a, then you conclude that the brand doesn't matter in comparing the mean mileage between brake linings. If Ha appears more believable than H_0, then you conclude that the brake lining brand matters.

MegaStat Command: Hypothesis Tests/Compare Two Independent Groups

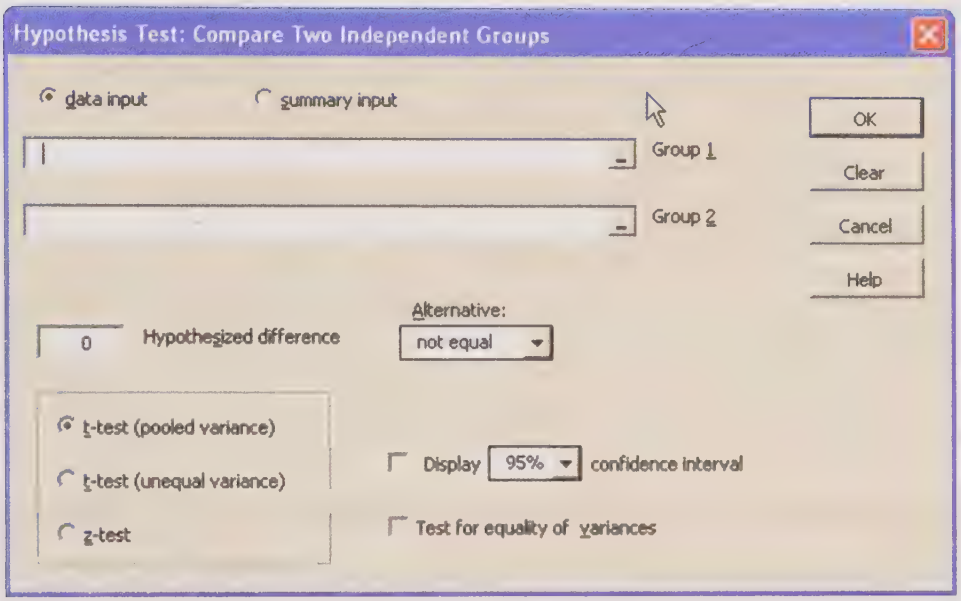


Figure 6.2 MegaStat Hypothesis Dialog Box Two Independent Means
Source: Fred Hulme

Results from *t*-test option (pooled variance)

Hypothesis Test: Independent Groups (*t*-Test, Pooled Variance)		
Brand A	**Brand B**	
39.55	38.03	Mean
4.38	4.23	Std. Dev.
75	75	*n*
	148	*df*
	1.520	Difference (Brand A – Brand B)
	18.558	Pooled variance
	4.308	Pooled std. dev.
	0.703	Standard error of difference
	0	Hypothesized difference
	2.16	*T*
	0.0323	*p*-value (two-tailed)

Table 6.5 Brake Lining Output
Source: Fred Hulme

b. Using Figure 6.2 and Table 6.5, what is the value of the "test statistic" determined from the *t*-test and give the appropriate *p*-value associated with each test statistic.

$$t = +2.16 \qquad p\text{-value} = 0.0323$$

c. Using a level of significance of 5%, should you reject H_0 or fail to reject H_0?

Using the results from the *t*-test, you should reject H_0 since the calculated *p*-value = 0.0323 is less than the stated level of significance = 0.05.

d. What does your conclusion from c. mean?

The entry in the MegaStat summary, *p*-value (two-tailed) = 0.0323, represents the probability or chance of finding as much difference or more between the sample means observed in the data if the true process means for Brand A and Brand B are equal. You typically compare this calculated probability to say 5% (0.05). If the calculated *p*-value is less than 5%, you would conclude that the observed difference between the two-sample means is not due to chance but due to some "assignable cause." For our specific example, the calculated *p*-value, 0.0323, is less than the stated level of significance of 0.05. You would conclude that there is a "significant" difference between the two sample means for the brands. You would choose Brand A as the preferred brake lining since its sample mean (39.55) is greater than Brand B's (38.03).

EXAMPLE 3—COMPARING THE MEAN OF DIFFERENCES FOR DEPENDENT (PAIRED) SAMPLES

To determine the effectiveness of an industrial safety program, the following data was collected over a period of one year on the average weekly loss of man-hours due to accidents in 12 plants "before and after" the program was put into operation. Using a level of significance of 5%, test if the program has been effective. The "before" value followed by the "after" value for the 12 plants is given in Table 6.6.

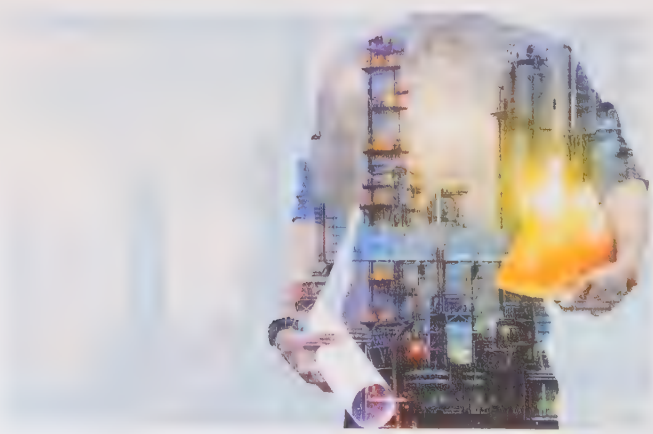
38 and 29	75 and 63	25 and 23	129 and 117
47 and 48	53 and 41	12 and 14	76 and 73
13 and 17	24 and 19	38 and 36	23 and 25

Table 6.6 Lost Man-hours Paired Data
Source: Fred Hulme

a. State the null and alternate hypotheses to be tested.

H_0: $\mu_{(\text{Before} - \text{After})} = 0$ (Safety program was not effective)
H_a: $\mu_{(\text{Before} - \text{After})} > 0$ (Safety program was effective)

The following summary was obtained by the **MegaStat** command, **Hypothesis Testing/Paired Observations**.



© Kendall Hunt Publishing Company

Hypothesis Test: Paired Observations	
0.000	Hypothesized value
46.083	Mean before
42.083	Mean after
4.000	Mean difference (Before − After)
5.939	Std. dev.
1.714	Std. error
12	*N*
11	*df*
2.33	*T*
0.0198	*p*-value (one-tailed, upper)

Table 6.7 Lost Man-Hours Output
Source: Fred Hulme

b. Using Table 6.7, what is the value of the "test statistic" and its associated p-value?

$t = 2.33$
p-value $= 0.0198$

c. Using a 5% level of significance, should you reject H_0 or fail to reject H_0?
Conclusion: Reject H_0, since the calculated p-value $= 0.0198$ is less than the stated level of significance of 0.05.

d. What does your conclusion from c. mean?
The safety program appears to be effective in reducing the loss of man-hours due to accidents in the plants.

EXAMPLE 4—ONE SAMPLE PROPORTION CASE

In a random sample of 600 television viewers contacted by phone in a certain suburban area, 210 were watching a certain movie on Netflix. What would you conclude from your survey results if the management of Netflix claimed that at least 40% of all viewers watched the movie? Use a level of significance of 5% in your decision. Use Table 6.8 to answer the questions.



© Kendall Hunt Publishing Company

a. H_0: $p \geq 0.40$
H_a: $p < 0.40$

b. Test statistic:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}} = \frac{0.35 - 0.40}{\sqrt{\frac{0.40(1 - .040)}{600}}} = -2.50$$

c. *p*-Value: 0.0062 (Found using the standard normal probability table or from the Excel calculator, NORMSDIST(−2.50) = 0.0062.)

d. Decision: Fail to reject H_0 or Reject H_0

e. Conclusion in words: The sample proportion of 0.35 would indicate that the population proportion is probably some value less than 0.40. The management's claim is probably too high.

Results from **MegaStat/Hypothesis Tests/Proportion vs. Hypothesized Value**

Hypothesis Test for Proportion vs. Hypothesized Value		
Observed	Hypothesized	
0.35	0.4	*p* (as decimal)
210/600	240/600	*p* (as fraction)
210.	240.	*X*
600	600	*n*
	0.02	Std. error
	−2.50	*z*
	0.0062	*p*-value (one-tailed, lower)

Table 6.8 Netflix Output
Source: Fred Hulme

EXAMPLE 5—COMPARING THE DIFFERENCE IN TWO INDEPENDENT PROPORTIONS

A sample of 300 customers using store credit and under age 30 showed that 50 did not pay their entire balance at the end of the month. A sample of 200 department store credit customers 50 years or older showed that 20 did not pay their entire balance at the end of the month. At the 0.05 significance level can we conclude that a larger percent of the younger customers do not pay their entire balance at the end of the month? Use Table 6.9 to answer the questions.



© Kendall Hunt Publishing Company

a. State the null and alternate hypotheses to be tested.

H_0: $p_{<30} = p_{>50}$ (Age doesn't matter as related to paying off debt)

H_a: $p_{<30} > p_{>50}$ (Age does matter as related to paying off debt; there is a higher % of younger credit customers who fail to pay off their entire balance.)

Results from **MegaStat/Hypothesis Tests/Compare Two Independent Proportions**

Hypothesis Test for Two Independent Proportions			
p_1	p_2	p_c	
0.1667	0.1	0.14	p (as decimal)
50/300	20/200	70/500	p (as fraction)
50.	20.	70.	X
300	200	500	n
	0.0667		Difference
	0.		Hypothesized difference
	0.0317		Std. error
	2.10		z
	0.0177		p-value (one-tailed, upper)

Table 6.9 Store Credit Card Payment Output
Source: Fred Hulme

b. What is the value of the "test statistic" from the data determined from the MegaStat summary and give the appropriate p-value associated with the test statistic.

$Z = +2.10$ (this is the value of the "test statistic")

p-value $= 0.0177$

c. Using a 5% level of significance, should you reject H_0 or fail to reject H_0?
Conclusion: Reject H_0, since the calculated p-value $= 0.0177$ is less than the stated level of significance of 0.05.

d. What does your conclusion from c. mean?
The difference in the two sample proportions (0.167 vs. 0.10) is large enough to conclude that the population proportion of younger customers who fail to pay off their entire credit balance is greater than that of the older customers.

Exercises

Short Answer/Discussion

1. A hypothesis test involves testing a claim about the value of *a* a. statistic or b. parameter.

2. Using the following hypotheses, H_0: $\mu = 100$ H_a: $\mu \neq 100$, H_0 is considered the a. null hypothesis or b. alternative hypothesis.

3. Using the following hypotheses, H_0: $\mu = 100$ H_a: $\mu \neq 100$, H_a is considered the a. null hypothesis or b. alternative hypothesis.

4. Using the following hypotheses, H_0: $\mu = 100$ H_a: $\mu \neq 100$, H_0, Ha is considered a a. two-sided alternative hypothesis or b. one-sided alternative hypothesis.

5. Using the following hypotheses, H_0: $\mu = 100$ H_a: $\mu \neq 100$, H_0, H_a is considered *a*(*n*) a. lower, one-sided alternative hypothesis or b. upper, one-sided alternative hypothesis.

6. Using the following hypotheses, H_0: $\mu = 100$ H_a: $\mu \neq 100$, H_0, H_a is considered *a*(*n*) a. lower, one-sided alternative hypothesis or b. upper, one-sided alternative hypothesis.

7. If your sample data leads you to the decision to reject H_0, then you might be making *a* a. Type I error or b. Type II error.

8. If your sample data leads you to the decision to fail to reject H_0, then you might be making a a. Type I error or b. Type II error.

9. Which type of error is more serious? a. Type I, b. Type II, or c. It varies.

10. If the null hypothesis, H_0, is true, the likelihood or chance of the sample results is known as the a. level of significance, b. test statistic, or c. *p*-value.

11. The probability of making a Type I error is known as the a. level of significance, b. test statistic, or c. *p*-value.

12. In a hypothesis test, if the *p*-value associated with the sample is less than the stated level of significance, you should a. fail to reject H_0, b. reject H_0, or c. take another sample.

13. In a hypothesis test, if you want to protect against making a Type I error, you should choose a relatively (a) small level of significance or b. large level of significance.

14. In a hypothesis test, if you want to protect against making a Type II error, you should choose a relatively a. small level of significance or b. large level of significance.

15. For a fixed sample size, as you decrease the chance of making a Type I error, the chance of making a Type II error a. increases or b. decreases.

16. The only way to decrease the chance of making both a Type I and Type II error is by _____ the sample size. a. increasing or b. decreasing

Problems

1. A national study found that the mean stopping distance for a school bus traveling 50 miles per hour is 264 feet. A school district wants to check if the population mean stopping distance for its busses is greater than the national mean distance. For a sample of ten busses, the mean stopping distance was 270 feet and the standard deviation was 15 feet. Use Table 6.10 to answer the questions concerning the bus.



© Kendall Hunt Publishing Company

a. State the hypotheses to be tested.

H_0: _____

H_a: _____

MegaStat Results

Hypothesis Test: Mean vs. Hypothesized Value	
264.00	Hypothesized value
270.00	Mean stopping distance
15.00	Std. dev.
4.74	Std. error
10	*n*
9	*df*
1.26	*t*
0.1188	*p*-value (one-tailed, upper)

Table 6.10 School Bus Output

Source: Fred Hulme

b. What is the value of the "test statistic" that can be used in deciding H_0 or H_a?

c. What is the "*p*-value" associated with the test statistic?

d. Using a level of significance of 0.05, what is your decision about H_0? Fail to reject H_0 **or** Reject H_0

e. What is your conclusion in words?

f. What kind of error, Type I or Type II, is possible with your decision in part d.?

2. An appliance store reports in its commercials that more than 70% of our customers have purchased an appliance with us before. The management of the company hired a research firm to independently validate this claim. In a sample of 200 recent buyers, 160 reported that they had, in fact, purchased and an appliance from the store before. Use Table 6.11 to answer the questions.



© Kendall Hunt Publishing Company

a. State the hypotheses to be tested.

H_0: _____

H_a: _____

24

MegaStat Results

Hypothesis Test for Proportion vs. Hypothesized Value		
Observed	Hypothesized	
0.8	0.7	*p* (as decimal)
160/200	140/200	*p* (as fraction)
160.	140.	*X*
200	200	*n*
	0.0324	Std. error
	3.09	*z*
	0.0010	*p*-value (one-tailed, upper)

Table 6.11 Appliance Output

Source: Fred Hulme

b. What is the value of the "test statistic" that can be used in deciding H_0 or H_a?

c. What is the "*p*-value" associated with the test statistic?

d. Using a level of significance of 0.05, what is your decision about H_0? Fail to reject H_0 **or** Reject H_0

e. What is your conclusion in words?

f. What kind of error, Type I or Type II, is possible with your decision in part d.?

3. A bank wants to study the use of its ATMs. Of particular interest is whether younger customers (under 25 years of age) use the ATMs more on average than older customers (over 60 years of age). The following data was collected on the number of ATM transactions last month by the two groups. The data is given in Table 6.12 and descriptive statistics in Table 6.13.



© Kendall Hunt Publishing Company

Data

Under 25	10	10	11	15	7	11	10	9			
Over 60	4	8	7	7	4	5	1	7	4	10	5

Table 6.12 ATM Data
Source: Fred Hulme

Descriptive Statistics

	Under 25		Over 60
Mean	10.38	Mean	5.67
Sample variance	5.13	Sample Variance	6.05
Count	8	Count	11

Table 6.13 ATM Descriptive Statistics
Source: Fred Hulme

a. Should the samples be considered independent or paired?

b. Give the hypotheses to be tested in order to answer the bank's question.

H_0: _____

H_a: _____

c. Should the test statistic be based on the "pooled variance" or "unequal variance" procedure?

(Rule of Thumb: Form the ratio of the two sample variances—the larger variance divided by the smaller variance. If this ratio is larger than 2, you should use the "unequal variance" procedure.)

MegaStat Results

Hypothesis Test: Independent Groups (*t*-test, Pooled Variance)		
Under 25	Over 60	
10.38	5.64	Mean
2.26	2.46	Std. dev.
8	11	*n*
	17	*df*
	4.739	Difference (Under 25 – Over 60)
	5.672	Pooled variance
	2.382	Pooled std. dev.
	1.107	Standard error of difference
	0	Hypothesized difference
	4.28	*t*
	0.0003	*p*-value (one-tailed, upper)

Table 6.14 ATM Output Two Sample Means
Source: Fred Hulme

d. Using Table 6.14, what is the value of the "test statistic" that can be used in deciding H_0 or H_a?

e. What is the "*p*-value" associated with the test statistic?

f. Using a level of significance of 0.05, what is your decision about H_0?

Fail to reject H_0 **or** Reject H_0

g. What is your conclusion in words?

h. What kind of error, Type I or Type II, is possible with your decision in part d.?

4. An egg farmer wanted to determine if increasing the time the lights were on in his hen house would increase egg production. For a sample of eight chickens he determined their production before and after increasing the amount of time the lights were on. The data are reported in Table 6.15. At the 0.01 significance level, has there been an increase in production?



© Kendall Hunt Publishing Company

Hen	1	2	3	4	5	6	7	8
Before	10	8	5	2	3	7	3	3
After	7	5	6	8	8	8	10	2

Table 6.15 Chicken Ranch Data
Source: Fred Hulme

a. Should the samples be considered independent or dependent (paired)?

MegaStat Results

Hypothesis Test: Paired Observations	
0.000	Hypothesized value
6.750	Mean After
5.125	Mean Before
1.625	Mean difference (After − Before)
3.962	Std. dev.
1.401	Std. error
8	*n*
7	*df*
1.16	*t*
0.1420	*p*-value (one-tailed, upper)

Table 6.16 Chicken Ranch Output
Source: Fred Hulme

b. State the hypotheses to be tested.

H_0: _____

H_a: _____

c. Using Table 6.16, what is the value of the "test statistic" that can be used in deciding H_0 or H_a?

d. What is the "*p*-value" associated with the test statistic?

e. Using a level of significance of 0.05, what is your decision about H_0?

Fail to reject H_0 or Reject H_0

f. What is your conclusion in words?

g. What kind of error, Type I or Type II, is possible with your decision in part d.?

5. A telemarketing company gives a sales aptitude test to job applicants. The company wants to know if a particular question discriminates well between good salespersons and poor salespersons. Table 6.17 gives the relevant data:

If the question acts as a good discriminator, then there should be a significantly higher percentage of "good" salespersons answering the question correctly than the "poor" salespersons.



© Kendall Hunt Publishing Company

	"Good" salespersons	"Poor" salespersons
Correct answer	72	55
Incorrect answer	28	45

Table 6.17 Salesperson Aptitude Test
Source: Fred Hulme

a. State the hypotheses to be tested.

H_0: _____

H_a: _____

Hypothesis Test for Two Independent Proportions			
p_1	p_2	p_c	
0.72	0.55	0.635	p (as decimal)
72/100	55/100	127/200	p (as fraction)
72.	55.	127.	X
100	100	200	n
	0.17		Difference
	0.		Hypothesized difference
	0.0681		Std. error
	2.50		z
	??????		p-value (one-tailed, upper)

Table 6.18 Aptitude Test Output
Source: Fred Hulme

b. Using Table 6.18, what is the value of the "test statistic" that can be used in deciding H_0 or H_a.

c. Calculate the "p-value" associated with the test statistic.

d. Using a level of significance of 0.05, what is your decision about H_0?

Fail to reject H_0 or Reject H_0

e. What is your conclusion in words?

f. What kind of error, Type I or Type II, is possible with your decision in part d.?

Self-Check

Questions 1 to 9 involves the following: Prior to the institution of a new safety program, the mean (average) number of on-the-job accidents per day at a factory was 4.5. To determine if a safety program has been effective, a random sample of 30 days is taken after the institution of the new safety program and the number of accidents recorded. The sample mean, and sample standard deviation, s, are:

$$X = 3.9 \quad \text{and} \quad s = 1.3$$

At the 1% level of significance, does it appear that the new safety program has been successful at decreasing the mean number of accidents per day?



© Kendall Hunt Publishing Company

1. State the hypotheses to be tested.

   H_0: _____

   H_a: _____

2. The appropriate "test statistic" for deciding whether H_0 or H_a is more believable based on the sample results should be based on which statistical distribution?

   a. t b. Z

3. Calculate the "test statistic" based on the sample data that can be used to determine which appears to be more believable H_0 or H_a.

4. Calculate the "p-value" associated with the test statistic from (3) and decide about rejecting H_0 or failing to reject H_0. In your decision, use a level of significance (α) = 0.01.

   p-value = _____

5. Decision: Fail to Reject H_0 or Reject H_0

6. What is your conclusion in words? (Don't just say Fail to Reject H_0 or Reject H_0)

7. What type of error might you be making in your decision in question (5)?

8. Explain what a Type I error would mean specific to this problem.

9. Before collecting the sample and making the decision, what would the chance of a Type I error be?36

Questions 10 to 17 involve the following: A random sample of 50 consumers taste a new snack food. Their responses were coded:

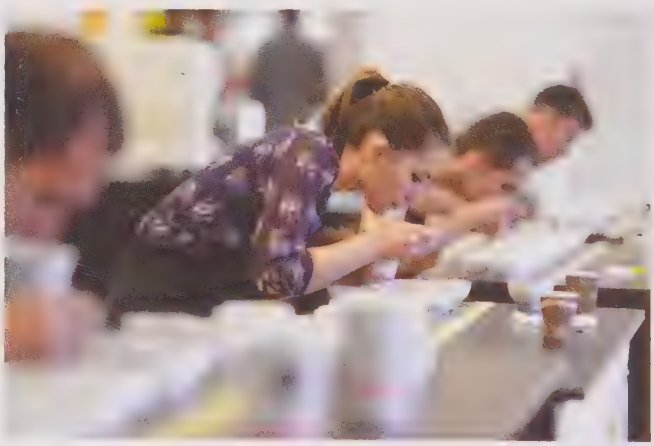
(0 = do not like; 1 = like; and 2 = indifferent)

Use Table 6.19 and put the data in Excel in a single column.

1	0	0	1	2	0	1	1	0	0
0	1	0	2	0	2	2	0	0	1
1	0	0	0	0	1	0	2	0	0
0	1	0	0	1	0	0	1	0	1
0	2	0	0	1	1	0	0	0	1

Table 6.19 Taste Testing Data
Source: Fred Hulme

From past studies, it has been shown when a new snack food product is tested that 50% of the taste testers usually do not like it. At the 10% level of significance, do the sample results indicate that the new snack food that was tested is more disliked than normal for an introductory product?



© Kendall Hunt Publishing Company

10. State the hypotheses to be tested.

H_0: _____

H_a: _____

11. The appropriate "test statistic" for deciding whether H_0 or H_a is more believable based on the sample results should be based on which statistical distribution?

a. t b. Z

12. Calculate the "test statistic" based on the sample data that can be used to determine which appears to be more believable H_0 or H_a.

Calculate the "p-value" associated with the test statistic from (12) and decide about rejecting H_0 or failing to reject H_0. In your decision, use a level of significance (α) = 0.10.

13. p-value = _____

14. Decision: Fail to Reject H_0 or Reject H_0

15. What is your conclusion in words? (Don't just say Fail to Reject H_0 or Reject H_0)

16. What type of error might you be making in your decision in question (5)?

17. Explain what a Type I error would mean specific to this problem.

Questions 18 to 25 involve the following: Some homeowners view having a swimming pool as a positive; others, however, view it as a negative. Using the MegaStat results below, perform a hypothesis test to see if the mean selling price is different for homes with a pool versus homes without a pool. The units are in thousands of dollars. Use a level of significance of 0.05.

Hypothesis Test		
W/Pool	W/O Pool	
231.48	202.79	Mean
50.58	33.71	Std. dev.
67	38	n
	28.69000	Difference (W/Pool − W/O Pool)
	8.25075	Standard error of difference
	0	Hypothesized difference
	3.48	z
	0.0005	p-value (two-tailed)

Table 6.20 Pool Survey
Source: Fred Hulme

18. Should the two samples be considered "independent" or "paired"?

19. State the hypotheses being tested.

H_0: _____

H_a: _____

20. Using Table 6.20, how different are the sample means?

21. Give the value of "test statistic" based on the sample data that can be used to determine which appears to be more believable H_0 or H_a.



© Kendall Hunt Publishing Company

22. Give the corresponding p-value associated with the value of the test statistic:

p-value = _____

23. In words, explain what the p-value measures.

24. Decision: Fail to reject H_0 or Reject H_0 (circle one)

25. Conclusion:

Questions 26 to 32 involve the following: In planning the construction of a new manufacturing facility, the purchasing department has narrowed its choice of paints to that of two brands. Brand A is less expensive than Brand B, but the department is concerned that Brand A's quality is inferior. To aid in this decision, several chips are painted and exposed to weather conditions for a period of six months. Each chip is then judged as to several qualities and a score determined (where the larger the score, the better the quality of the paint.) Use the paint samples given in Table 6.21. Use a 0.05 level of significance

Chip	1	2	3	4	5	6
Brand A	84	86	91	93	84	88
Brand B	90	88	92	94	84	85

Table 6.21 Paint Brands Data
Source: Fred Hulme

26. Should the two samples be considered "independent" or "paired"?

27. Set up the hypothesis to see if Brand A's score is inferior to Brand B's.

H_0: _____

H_a: _____

28. Give the value of "test statistic" based on the sample data that can be used to determine which appears to be more believable H_0 or H_a.



© Sebastian Duda

29. Give the corresponding p-value associated with the value of the test statistic:

p-value = _____

30. In words, explain what the p-value measures.

31. Decision: Fail to reject H_0 or Reject H_0 (circle one)

32. Conclusion:

Questions 33 to 38 involve the following: A marketing study conducted in a large city shows that of 100 married women who work full time, 43 eat dinner at a restaurant at least one night during a typical workweek. Twenty-seven out of a sample of 100 married women who do not work fulltime eat dinner out at least once a week. Using a 5% level of significance, is there evidence of a significant difference between the two groups?



© Kendall Hunt Publishing Company

33. State the hypotheses being tested.

H_0: _____

H_a: _____

34. Give the value of "test statistic" based on the sample data that can be used to determine which appears to be more believable H_0 or H_a.

35. Give the corresponding p-value associated with the value of the test statistic:

p-value = _____

36. In words, explain what the p-value measures.

37. Decision: Fail to reject H_0 or Reject H_0 (circle one)

38. Conclusion:

Chapter 6—End of Chapter Answers

Short Answer/Discussion

1. B
2. A
3. B
4. A
5. B
6. A
7. A
8. B
9. C
10. C
11. A
12. B
13. A
14. B
15. A
16. A

Problems

1. **a.** H_0: $\mu = 264$ H_a: $\mu > 264$
   b. $t = 1.26$
   c. 0.1188
   d. Fail to Reject H_0
   e. Using 95% confidence the mean stopping distance is not significantly different from 264 feet.
   f. Type II

2. **a.** H_0: $p = 0.7$ H_a: $p > 0.7$
   b. $z = 3.09$
   c. 0.0010
   d. Reject H0
   e. Using 95% confidence, more than 70% of the customers have purchased an appliance with store before.
   f. Type I

3. **a.** Independent
   b. H_0: $\mu_{under} = \mu_{over\ 60}$ H_a: $\mu_{under} > \mu_{over\ 60}$
   c. 6.05/5.13 = 1.18 use pooled variance
   d. $t = 4.28$
   e. 0.0003
   f. Reject H_0
   g. Using 95% confidence, the mean number of ATM uses by the under 25 age group is significantly greater than the over 60 age group.
   h. Type I

4. **a.** Dependent
   b. H_0: $\mu_{After} = \mu_{Before}$ H_a: $\mu_{After} > \mu_{Before}$
   c. $t = 1.16$
   d. 0.1420
   e. Fail to Reject H_0
   f. Using 99% confidence there is no significant difference in the mean egg production after the light time was increased than before.
   g. Type II

5. **a.** H_0: $P_{good} = P_{poor}$ H_a: $P_{good} > P_{poor}$
   b. 2.50
   c. 0.0063
   d. Reject H_0
   e. Using 95% confidence, significantly more "good salespeople" answered correctly than "poor salespeople."
   f. Type I

Self Check

1. H_0: $\mu_{After} = \mu_{Before}$ H_a: $\mu_{After} < \mu_{Before}$

2. a. t

3. $t = -2.528$

4. p-value $= 0.0086 < 0.01$

5. Reject H_0

6. Using 99% confidence, the mean number of accidents significantly decreased after the safety program.

7. Type I

8. We decide the program does work to reduce accidents based on the sample, but in reality it does not.

9. 1%

10. $H_0: p = 0.5$ $H_a: p > 0.05$

11. b. z

12. $z = 1.13$

13. p-value $= 0.1289$

14. Fail to Reject H_0

15. Using 90% confidence, the proportion of taste testers disliking the new snack is not significantly different from 50%.

16. Type II

17. Type I error results in believing the proportion of taste testers disliking the snack is over 50% when it really is not.

18. Independent

19. $H_0: \mu_{\text{pool}} = \mu_{\text{without}}$ $H_a: \mu_{\text{pool}} \neq \mu_{\text{without}}$

20. 28.69 ($000)

21. $Z = 3.48$

22. p-value $= 0.0005$

23. The chance of obtaining a test statistic of 3.48 if the true mean difference in price is 0 is 0.0005.

24. Reject Ho

25. Using 95% confidence, there is a significant difference in the mean price of homes with pools and homes without pools.

26. Samples are paired.

27. $H_0: \mu_A = \mu_B$ $H_a: \mu_A < \mu_B$

28. $t = -4.238$

29. p-value $= 0.1869$

30. There is a 0.1869 chance that a t statistic of -4.238 would be obtained if the true mean difference in the paints were zero. This is a fairly large probability.

31. Fail to Reject H_0

32. Using 95% confidence, there is no significant difference in the mean scores for the two paints.

33. H_0: $p_{FT} = p_{not}$ H_a: $p_{ft} \neq p_{not}$

34. $z = 2.37$

35. p-value $= 0.0177$

36. There is only a 1.77% chance of obtaining a test statistic of 2.37 if the true difference in the proportions was 0.

37. Reject H_0

38. Using 95% confidence, there is sufficient evidence to conclude that the proportion of full-time married women eat out more than the proportion who are not full time.

CHAPTER

7

Analysis of Count Data—Chi-Square Applications

Business data often involves non-numeric, qualitative outcomes. Customer satisfaction levels (low, moderate, or high) and athletic shoe brand preference (Nike, New Balance, Adidas, or Reebok) would be examples of qualitative data. The outcomes of qualitative data are just categories. Frequency tables are used to summarize qualitative data. The table below summarizes the responses of a randomly selected sample of 240 customers as to the brand of athletic shoes they purchased and their level of satisfaction with the shoes. The frequencies are in Table 7.1



© Kendall Hunt Publishing Company

		Nike	New Balance	Adidas	Reebok
Satisfaction Level	Low	25	7	7	2
	Moderate	41	14	8	6
	High	76	27	21	6

Table 7.1 Sneaker Satisfaction
Source: Fred Hulme

The above summary is referred to as a "two factor frequency table" or a "cross tabulation." This cross tabulation of the data can be used to decide if customers' athletic shoe preferences are related to their level of satisfaction. The relationship between customer satisfaction and shoe brand could also be described by the following column chart in Figure 7.1.

Visually it appears that the patterns of the satisfaction levels are not too different for the different shoe brands indicating that the two factors are probably not related. Later in this chapter we will calculate how likely the pattern in the frequency table would be if customer satisfaction and shoe brand preference were not related. If this pattern is not very likely (say less than 0.05), we would conclude that the two factors are related. If the pattern is not considered unusual, then we conclude that there is not enough evidence to determine that the two are related.



Figure 7.1 Chart of Sneaker Preferences
Source: Fred Hulme

Goodness of Fit Analysis

A "Goodness of Fit" analysis involves comparing a set of "observed" frequencies or counts formed from a sample of data to a set of "expected" frequencies assuming some underlying theoretical pattern. If there is a great deal of difference between the "observed" and the "expected" frequencies, it makes you think that the sample does not come from a population with the underlying theoretical pattern. How large a difference must exist between the "observed" and the "expected" frequencies? The following statistic assesses how likely the sample frequencies would occur if the population follows some underlying pattern. In other words, did we observe what we expected based on past knowledge or not?

$$X^2 = \sum \frac{(Obs - Exp)^2}{Exp}$$

EXAMPLE 1

A certain town has only three gasoline Stations (A, B, C). When the three Stations were all self-serving, their shares of the total market were equal. Two months ago, Station A converted to a full-service concept in an attempt to improve its market share. A recent survey of a sample of 180 residents found that 76 now use Station A; 54 now use Station B; and 50 now use Station C.

If the three stations still have the same market share of all possible customers in the town after Station A switched to a full-service concept, then you would expect each station to have an equal number of 60 customers. Some variation around the expected equal number usually occurs but not a vast amount.



© Kendall Hunt Publishing Company

How much variation could you allow in the sample and still think there was an equal market share for the three stations across all possible customers? This depends on how large the value of the "chi-square test statistic" is. When the value of χ^2 is large, it implies that there is too large a difference between the "observed" and "expected" frequencies to conclude that the market shares are still equal. So, the question becomes, how large is large? If the value of the chi-square test statistic is large enough that it would occur with less than say a 5% chance assuming an equal market share, then you conclude that there probably isn't an equal market share any longer.

The Expected Value for a Goodness of Fit test is either calculated by averaging the observed values and each group has an equal chance of occurring or by taking n^*p if there are known proportions attached to each group.

For our data, $X^2 = \sum \frac{(Obs - Exp)^2}{Exp} = 6.53$ (see Table 7.2).

	Observed	Expected	$O - E$	$(O - E)^2 \div E$
A	76	60	16	4.27
B	54	60	-6	0.60
C	50	60	-10	1.67
Total	180	180	0	6.53

Table 7.2 Gas Station Data
Source: Fred Hulme

Percentage Points of the X^2 Distribution				
Degrees of Freedom	**Probability That a Chi-Square Value Will Be Exceeded**			
	0.25	0.10	0.05	0.01
1	1.32	2.71	3.84	6.63
2	2.77	4.61	5.99	9.21
3	4.11	6.25	7.81	11.34
4	5.39	7.78	9.49	13.28
5	6.63	9.24	11.07	15.09
6	7.84	10.64	12.5	16.81

Table 7.3 Chi-Square Table
Source: Fred Hulme

The χ^2 distribution in Table 7.3 can be used to estimate the probability of getting a value larger than our observed chi-square test statistic. You refer to this probability as a "p-value." Using the table, the best you can do is estimate a range for the "p-value."

The χ^2 distribution is a family of distributions indexed on a factor known as the associated degrees of freedom of the test statistic. With a "Goodness of Fit" analysis, the degrees of freedom (df) equals # of categories minus one. For our problem, $df = 3 - 1 = 2$. This leads us to use the second row of the χ^2 distribution table. Our chi-square test statistic value, 6.53, falls between 5.99 and 9.21, the 0.05 and 0.01 columns, respectively. This means that the p-value is between 0.01 and 0.05 or $0.01 < p\text{-value} < 0.05$. This indicates that a χ^2 value of 6.53 would be considered unusual if the three Stations had an equal market share of all possible customers in the town. This leads you to conclude that the market share for the three Stations is probably not equal.

H_0: The market share is equal for all three Stations

Ha: The market share is not equal for all three Stations

Value of the Chi-Square "Test Statistic": 6.53

p-value: $0.01 < p\text{-value} < 0.05$

Decision: Fail to reject H_0 or Reject H_0 (assuming a level of significance = 0.05)

Conclusion: Using 95% confidence, the market share for the gasoline Stations appears not to be equal.

MegaStat Solution to Example 1

MegaStat command: Chi-Square/Crosstab – Goodness of Fit Test

Goodness of Fit Test				
Observed	**Expected**	***O* – *E***	**$(O-E)^2/E$**	**% of chi-square**
76	60.000	16.000	4.267	65.31
54	60.000	-6.000	0.600	9.18
50	60.000	-10.000	1.667	25.51
180	180.000	0.000	6.533	100.00
6.53	Chi-square			
2	*df*			
.0381	*p*-value			

Table 7.4 MegaStat Chi-Square Output for Gas Stations
Source: Fred Hulme

H_0: The market share is equal for all three Stations

Ha: The market share is not equal for all three Stations

Value of the Chi-Square "Test Statistic": 6.53 (from Table 7.4)

p-value: 0.0381

Decision: Fail to reject H_0 or Reject H_0 (assuming a level of significance = 0.05)

Conclusion: On using 95% confidence, the market share for the gasoline Stations no longer appear to be equal.

EXAMPLE 2

The owner of an online jewelry business would like to compare her sales with the geographic distribution of the population. According to the United States Census Bureau, 21% of the population lives in the Northeast, 24% in the Midwest, 35 percent in the South, and 20% in the West. Listed on the next page is a breakdown of a sample of 400 orders randomly selected from those shipped last month as shown in Table 7.5. At the 0.05 significance level, does the distribution of the orders reflect the population?



© Kendall Hunt Publishing Company

Region	Frequency
Northeast	68
Midwest	104
South	155
West	73
Total	400

Table 7.5 Jewelry Customers by Region
Source: Fred Hulme

Goodness of Fit Test				
Observed	**Expected**	***O* – *E***	**$(O-E)^2/E$**	**% of chi-sq**
68	84.000	–16.000	3.048	51.36
104	96.000	8.000	0.667	11.23
155	140.000	15.000	1.607	27.08
73	80.000	–7.000	0.613	10.32
400	400.000	0.000	5.934	100.00
5.93	Chi-square			
3	*df*			
0.1149	*p*-value			

Table 7.6 Regional Customers Output
Source: Fred Hulme

H_0: The origins of the orders follow the Census Bureau population pattern

Ha: The origins of the orders do not follow the Census Bureau population pattern

Value of the Chi-Square "Test Statistic": 5.93 (from Table 7.6)

p-value: 0.1149

Decision: Fail to reject H_0 or Reject H_0 (assuming a level of significance = 0.05)

Conclusion: Using 95% confidence, there is not enough evidence to conclude that the origins of the orders do not follow the Census Bureau population pattern.

Contingency Tables—Testing for the Independence of Classifications

EXAMPLE 3

Table 7.7 is a 2-factor frequency table giving the cross tabulations of the gender and rank of college attended for 60 employees. Figure 7.2 gives a visualization of the data.



© Kendall Hunt Publishing Company

	A	B	C	D	E
1	Count of Gender	Rank			
2	Gender	Above average	Average	Below average	Grand Total
3	F	11	16	3	30
4	M	7	14	9	30
5	Grand Total	18	30	12	60

Table 7.7 Frequencies of College Rank by Gender
Source: Fred Hulme

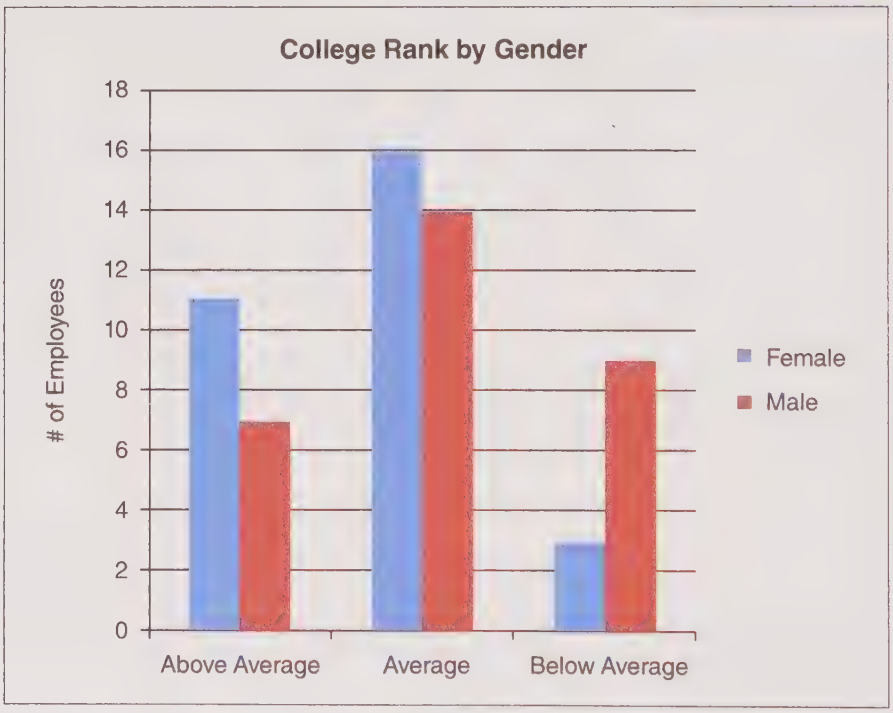


Figure 7.2 College Rank by Gender
Source: Fred Hulme

Is the ranking of college attended by the employees related to their gender? In order to quantify the answer to this question—you would have to consider what is known as **sampling error**. From your visual comparison of the frequencies, it appears there is a difference in the distributions between female and male. However, this data represents a sample of only 60 employees. If the study were conducted with a different group of employees,

would you get similar results? The results would be at least somewhat different because the results would be based on a different set of people. This difference is called sampling error. The word "error" does not imply that a mistake was made in collecting the original data. The word "error" refers to the fact that if you repeated the study with a different set of people that you would not get exactly the same results.

As the number of subjects in your sample increases, the differences due to sampling error decreases. Most studies are based on data from samples as opposed to taking a complete census because there is limited money and time available. You usually will have to accept the uncertainty in your decision due to the sampling error. Sampling error cannot be eliminated in a study based on a sample of the complete data, but the size of the sampling error can be estimated. In order to estimate the sampling error in the two-factor frequency table example above, you would have to assume that there is no relationship between the variables, gender and ranking of college attended. By making this assumption, you can calculate how many employees you would expect for each gender and college ranking combination. Since there is an equal number of females and males in the sample, you would expect 50% of the 18 employees that graduated from a "1" ranked college to be female and 50% to be male. Therefore, you would expect 9 females and 9 males to have graduated from "1" ranked colleges if there is no relationship between the classifications of gender and ranking of college attended. If there had been 1/3 females and 2/3 males in the sample, you would expect 1/3 of 18 or 6 females and 2/3 of 18 or 12 males to have attended a "1" ranked college. If there were no difference between the genders with respect to which type of college attended, you would expect the same proportion between females and males attending "1" ranked colleges as the **overall** proportion of females and males. Assuming no relationship between the classifications of the two variables, the calculation for the expected number for each cell in the two-dimensional frequency table can be represented by:

Expected Frequency for a Given Cell = (Row Total*Column Total)÷(Grand Total)

Using this formula, the expected frequency out of the 60 employees who are female and graduated from a college ranked as a "1" would be:

$$(30 \cdot 18) \div 60 = 9.$$

Using this formula, the expected frequency out of the 60 employees who are male and graduated from a college ranked as a "3" would be: $(30 \cdot 12) \div 60 = 6$.

If you calculate the expected frequency for each cell, you would get the values in Table 7.8.

Table 7.8 displays the Expected Frequencies Assuming Independence of Gender and College Ranking

		Ranking of College			
		1	2	3	Total
Gender	Female	9	15	6	30
	Male	9	15	6	30
Total		18	30	12	60

Table 7.8 Expected Frequencies
Source: Fred Hulme

As you would expect there is some discrepancy between what you observed in your sample of 60 employees and what you would expect to occur over repeated samplings if there is no relationship between gender and ranking of college attended. How much discrepancy can be allowed and you still conclude there is no relationship between the factors? Obviously the more the discrepancy, the less likely it appears that the

assumption of independence is true. But how much is too much? The question is answered by calculating a chi-square, X^2, statistic, given by the formula:

$$X^2 = \sum \frac{(Obs - Exp)^2}{Exp}$$

$$X^2 = \frac{(11-9)^2}{9} + \frac{(16-15)^2}{15} + \frac{(3-6)^2}{6} + \frac{(7-9)^2}{9} + \frac{(14-15)^2}{15} + \frac{(9-6)^2}{6}$$

$$X^2 = 0.44 + 0.07 + 1.5 + 0.44 + 0.07 + 1.5$$

$$X^2 = 4.02$$

The larger the calculated X^2 statistic is, the less likely it appears that the observed frequencies follow the assumption used in calculating the expected frequencies. This assumption was that the factors of gender and the ranking of college attended are independent of one another. The question becomes—"How large is large?" This is answered by comparing the calculated X^2 statistic to a value from a X^2 distribution in Table 7.9 with a certain number of degrees of freedom. The degrees of freedom involve the fact that since the row and column totals are fixed for the expected frequencies only two cells are free to vary. The degrees of freedom can be found by calculating (the number of rows − 1) times the (the number of columns − 1). For this example that would be:

$$(2-1) \cdot (3-1) = 2 \text{ df}$$

Percentage Points of the χ^2 Distribution				
Degrees of freedom	**Probability that a Chi-Square Value will be Exceeded**			
	0.25	0.10	0.05	0.01
1	1.32	2.71	3.84	6.63
2	2.77	4.61	5.99	9.21
3	4.11	6.25	7.81	11.34
4	5.39	7.78	9.49	13.28
5	6.63	9.24	11.07	15.09
6	7.84	10.64	12.5	16.81

Table 7.9 Chi-Square Table
Source: Fred Hulme

From the above table you could then find the probability of the occurrence of your calculated X^2 value assuming the claim of independence or no relationship. Since the calculated X^2 value of 4.02 is between $X^2_{(2, 0.25)} = 2.77$ and $X^2_{(2, 0.10)} = 4.61$, you would say that the chance of your sample observed values occurring if the classifications are not related is somewhere between 10% and 25%. You call this percentage the probability value or *p*-value of the test. Before the test results are considered significant, you typically require the *p*-value to be less than 5%. Based on the fact that $10\% \leq p\text{-value} \leq 25\%$, you would fail to reject the hypothesis that the classifications of gender and the ranking of college attended are independent. Another way of stating the conclusion of the statistical test is that the observed frequencies of the data were not in sufficient enough conflict with the expected frequencies assuming independence to say that the classifications are dependent on one another.

You can have Microsoft Excel calculate the *p*-value by first summarizing the observed frequencies into a two-dimensional table. You would then calculate the expected frequencies using the formula:

Expected Frequency for a Given Cell = (Row Total) · (Column Total)÷(Grand Total).

The Microsoft Excel function,

=CHITEST(range for observed frequencies, range for expected frequencies), in Table 7.10,

can be used to calculate the *p*-value of the statistical test.

	A	B	C	D	E
1	Count of Gender	Index			
2	Gender	1	2	3	Grand Total
3	1	11	16	3	30
4	2	7	14	9	30
5	Grand Total	18	30	12	60
6					
7	Expected frequencies				
8		9	15	6	
9		9	15	6	
10					
11	P-value for test of independence	0.13			

Table 7.10 Excel Chitest Output
Source: Fred Hulme

The entry for cell C11 would be: **=CHITEST(B3:D4, B8:D9)**. The value of 0.13 represents the *p*-value for the statistical test of independence. Since 0.13 would not be considered a rare event if you use as your yardstick of what is a rare event as 0.05, then you would fail to reject the claim that the classifications are independent. This wording of the conclusion may sound like double talk, but in decision-making involving sample data you never conclude that you accept a claim as being true. You either reject the claim or fail to reject the claim. This is similar to a decision in a court of law in which a defendant is found not guilty. A jury decision of not guilty is not saying the defendant is innocent, but that the evidence presented did not "prove" beyond a shadow of doubt that the defendant was guilty.

The MegaStat command, **Chi-Square/Crosstab**, can be used to calculate the expected frequencies and corresponding *p*-value necessary to test if the classifications should be considered independent or not.

You would use the option, **Chi-Square/Crosstab/Contingency Table,** if you have a table of the observed frequencies for different combinations of the classifications. You would use the option, **Chi-Square/Crosstab/Crosstabulation,** if you have the raw data.

MegaStat Results Table 7.11					
Chi-square Contingency Table Test for Independence					
		1	2	3	Total
Female	Observed	**11**	**16**	**3**	30
	Expected	9.00	15.00	6.00	30.00
Male	Observed	**7**	**14**	**9**	30
	Expected	9.00	15.00	6.00	30.00
Total	Observed	18	30	12	60
	Expected	18.00	30.00	12.00	60.00
		4.02	Chi-square		
		2	*df*		
		0.1338	*p*-value		

Table 7.11 MegaStat Chi-Square for Gender by College Ranking
Source: Fred Hulme

H_0: Rank of college attended and gender are not related

Ha: Rank of college attended and gender are related

Value of the Chi-Square "Test Statistic": 4.02

p-value: 0.1338

Decision: Fail to reject H_0 or Reject H_0 (Assuming a level of significance = 0.05)

Conclusion: Using 95% confidence, there is not sufficient evidence to conclude that an employee's college ranking and gender are related.

EXAMPLE 4

A bank conducted a survey to compare the attitudes of young married customers (those having recently opened their first bank account) with the attitudes of older, established customers to their new automated teller system (ATM). A summary of the customers' responses is shown in Table 7.12.

	Younger customers	Older customers	Totals
Favor the ATM	87	65	152
Favor personal service	113	135	248
Totals	200	200	400

Table 7.12 ATM Survey Data
Source: Fred Hulme

Conduct the appropriate "significance test" in order to decide if there is a significant difference between the proportion of younger customers who favor using an ATM versus the older customers who favor using an ATM. Use a level of significance of 5%. Give H_0 and *Ha*, the value of the chi-square test statistic, the *p*-value of the test, and your conclusion using Table 7.13.

Chi-square Contingency Table Test for Independence				
		Younger Customers	**Older Customers**	**Total**
Favor the ATM	Observed	**87**	**65**	152
	Expected	76.00	76.00	152.00
	$(O-E)^2/E$	1.59	1.59	3.18
Favor personal service	Observed	**113**	**135**	248
	Expected	124.00	124.00	248.00
	$(O-E)^2/E$	0.98	0.98	1.95
Total	Observed	200	200	400
	Expected	200.00	200.00	400.00
	$(O-E)^2/E$	2.57	2.57	5.14
		5.14	Chi-square	
		1	*df*	
		0.0234	*p*-value	

Table 7.13 MegaStat Output for ATM Survey
Source: Fred Hulme

H_0: Customer's age and banking service preference are not related

Ha: Customer's age and banking service preference are related

Value of the Chi-Square "Test Statistic": 5.14 from Table 7.13

p-value: 0.0234

Decision: Fail to reject H_0 or Reject H_0 (Assuming a level of significance = 0.05)

Conclusion: Using 95% confidence, the customer's age and banking service preference are related

EXAMPLE 5

The following data is from a chain of 50 convenience stores. Each row or record of data contains the following variables:

1. Satisfaction rating from customer survey (L = low, M = moderate, H = high)
2. Yes/No response if the manager had received TQM (Total Quality Management) training

   (Y = yes, N = no)

   The first three data records are in Table 7.14



© Kendall Hunt Publishing Company

	A	B
1	Satisfaction	Training
2	M	Y
3	L	N
4	H	Y

Table 7.14 Partial Data
Source: Fred Hulme

The MegaStat command, **Chi-Square/Crosstab/Crosstabulation** is used below to conduct the appropriate "significance test" in order to decide if there is a relationship between customer satisfaction and whether the manager had received the TQM training or not. A level of significance of 5% is used in drawing the conclusion.

Cross tabulation					
		Satisfaction			
		L	M	H	Total
Training	N	13	5	3	21
	Y	4	15	10	29
	Total	17	20	13	50
		12.58	Chi-square		
		2	*df*		
		.0019	*p*-value		

Table 7.15 Convenience Store Output
Source: Fred Hulme

H_0: Customer satisfaction and manager's training are not related

Ha: Customer satisfaction and manager's training are related

Value of the Chi-Square test statistic: 12.58

p-value: 0.0019 from Table 7.15.

Decision: Fail to reject H_0 or Reject H_0 (Assuming a level of significance = 0.05)

Conclusion: Using 95% confidence, customer satisfaction and manager's training are related.

EXAMPLE 6

A credit score measures a person's creditworthiness. The following histogram summarizes the credit scores for 100 randomly selected adults. A person's credit score is a number that ranges between 300 and 950 where larger means better. A credit score is based on a person's ability to handle debt relationships. Actions such as not defaulting on loan payments and not making late payments on credit card debt will help a person's credit rating.



© Kendall Hunt Publishing Company

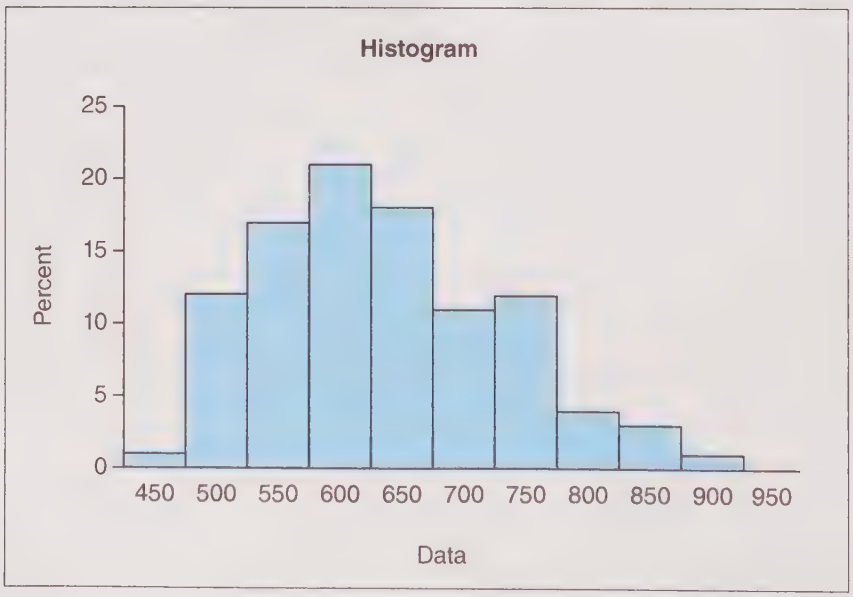


Figure 7.3 Visualization of Credit Scores

Source: Fred Hulme

Descriptive statistics	Credit_Score
Count	100
Mean	660.32
Sample variance	9,084.00
Sample standard deviation	95.31
Minimum	468
Maximum	934
Range	466
Normal curve GOF	
p-value	.4614
Chi-square(*df*=5)	4.64
E	12.50
O(−1.15)	13
O(−0.67)	14
O(−0.32)	13
O(+0.00)	16
O(+0.32)	9
O(+0.67)	9
O(+1.15)	10
O(inf.)	16

Table 7.16 Chi-Square Test for Normal Distribution Source: Fred Hulme

Results of the MegaStat command: **Descriptive Statistics/Normal Curve Goodness of Fit** are shown in Figure 7.3 and Table 7.16.

Use the MegaStat results to perform a hypothesis test to determine if the sample of the 100 credit scores indicates the underlying population distribution is normally distributed. Use a level of significance of 0.05.

H_0: The underlying distribution of credit scores is a normal distribution

Ha: The underlying distribution of credit scores is not a normal distribution

Value of the Chi-Square "Test Statistic": 4.64

p-value: 0.4614

Decision: Fail to reject H_0 or Reject H_0

Conclusion: Using 95% confidence, one cannot conclude the underlying distribution of credit scores is not normal.

Results of the MegaStat command are shown in Figure 7.4 using the command **Descriptive Statistics/ Normal Curve Plot**

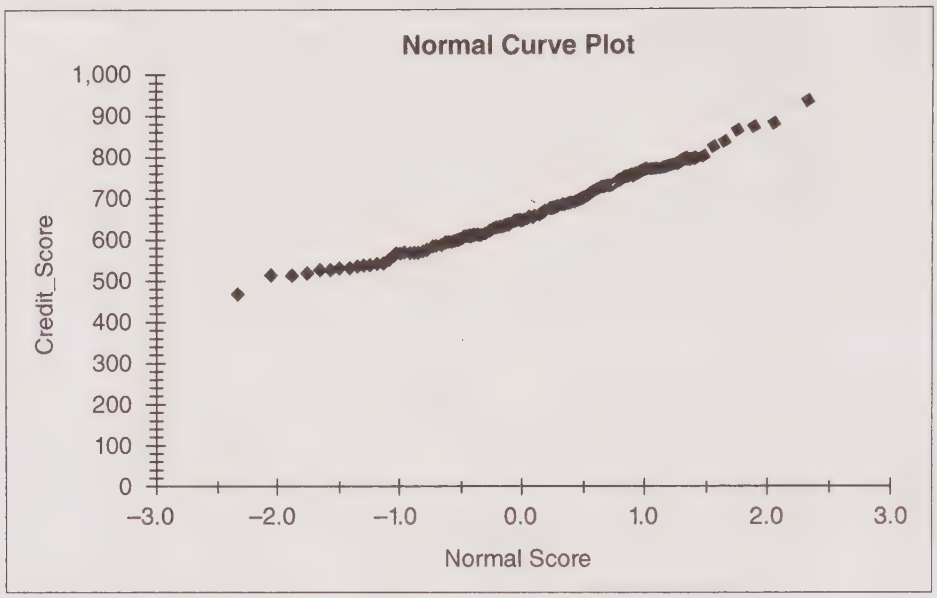


Figure 7.4 Normal Curve Plot
Source: Fred Hulme

When a set of sample values comes from a normal distribution, the "normal curve plot" will be approximately a straight line. The "normal curve plot" above for the credit score data indicates an underlying normal distribution.

Problems

1. A manufacturer of frosting for cake mixes wishes to determine which of the flavors is the most popular among customers. A sample of 200 customers revealed the following:



© Kendall Hunt Publishing Company

Flavor	Frequency
Light chocolate	62
Dark chocolate	55
White	45
Strawberry	38.

Is it reasonable to conclude that the flavors are equally preferred? Use a 0.05 significance level in reaching your conclusion.

H_0: _____

Ha: _____

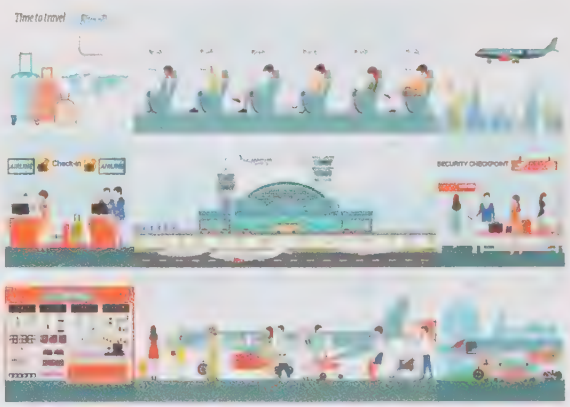
Value of the Chi-square Test Statistic: _____

p-value: _____

Decision: Fail to reject H_0 or Reject H_0

Conclusion:

2. An airline is interested in determining whether passenger satisfaction is related to the level of on-board service the passenger receives. 112 passengers were asked to rate their level of satisfaction and to give the level of service received during their flight. Given Table 7.17 and a 5% level of significance, does it appear that the customer satisfaction is related to the on-board service the passenger receives?



© Kendall Hunt Publishing Company

		Level of Service		
		Snacks	Snacks + Meals	Snacks + Meals + Entertainment
Passenger Satisfaction	High	2	15	17
	Moderate	16	11	10
	Low	21	15	5

Table 7.17 Service /Satisfaction Survey

Source: Fred Hulme

Chi-square Contingency Table Test for Independence					
		Snacks	Meals	Entertainment	Total
High	Observed	2	15	17	34
	Expected	11.84	12.45	9.71	34.00
Moderate	Observed	16	11	10	37
	Expected	12.88	13.54	10.57	37.00
Low	Observed	21	15	5	41
	Expected	14.28	15.01	11.71	41.00
Total		39.00	41.00	32.00	112.00
		22.44	Chi-square		
		4	*df*		
		.0002	*p*-value		

Table 7.18 Contingency Table for Airline Survey Source: Fred Hulme

H_0: _____

Ha: _____

Value of the Chi-square "Test Statistic": _____ See Table 7.18.

p-value: _____

Decision: Fail to reject H_0 or Reject H_0

Conclusion:

3. Digital Circuits is marketing a new battery for operating the flash unit of a digital camera using the slogan, "Why not use the best? On the average our batteries give 475 flashes." A consumer group has challenged this claim as being too high. In order to determine who is right, an independent product-testing lab tested 50 randomly selected batteries with the following results.



© Kendall Hunt Publishing Company

395	454	440	428	453	531	371	399	490	432
360	456	486	508	465	409	473	468	437	423
428	463	468	506	405	413	447	429	440	419
490	470	399	403	430	428	395	413	412	452
464	442	422	491	418	451	427	511	485	411

Table 7.19 Camera Battery Data
Source: Fred Hulme

In deciding about XYZ's claim, you must be able to assume that the lifetimes of the batteries are at least approximately normally distributed. **Copy the data into Excel from Table 7.19. Using a level of significance of 0.05, test the normality assumption.**

H_0: _____

Ha: _____

Value of the Chi-square Test Statistic: _____

p-value: _____

Decision: Fail to reject H_0 or Reject H_0

Conclusion:

Answers to Chapter 7 Problems

1. H_0: The popularity of flavors is equal

   Ha: The popularity of flavors is not equal.

   Chi-square = 6.76

   p-value = 0.08 > 0.05

   Fail to reject H_0

   Using 95% confidence, there is not enough evidence to conclude there is a difference in the flavors' popularity.

2. H_0: Passenger satisfaction and level of service are not related

   Ha: Passenger satisfaction and level of service are related

   Chi-square = 22.44

   p-value = 0.0002 < 0.05

   Reject H_0

   Using 95% confidence, there is significant evidence to indicate a relationship between passenger satisfaction and level of service on flights.

3. H_0: The underlying distribution of battery flashes is normal.

   Ha: The underlying distribution of battery flashes is not normally distributed.

   Chi-square = 2.64

   p-value = 0.6198

   Fail to reject H_0.

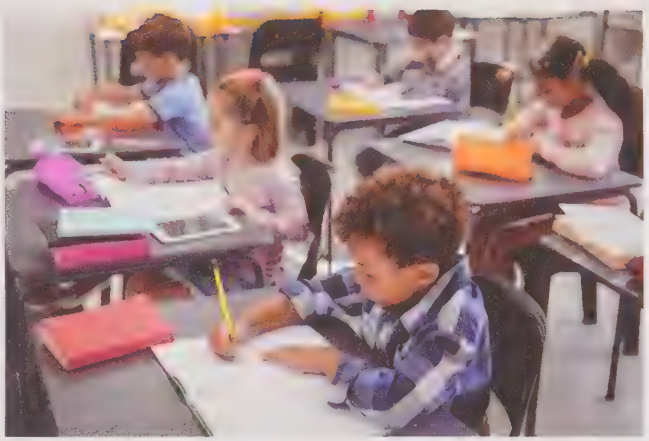
   Using 95% confidence, one cannot conclude the distribution of flashes is not normal.

CHAPTER

8

Analysis of Variance ANOVA

Introduction



© Kendall Hunt Publishing Company

Data, whether from business transactions, manufacturing processes, or tests results, is either gathered through simple observation or through designed experiments. Analysis of variance makes pairwise comparisons of more than two groups of means by literally controlling the variation within groups while measuring the variation between groups.

Educational Psychologists gave three second grade classes a spelling test. They then returned the graded exams and praised one class with the words, "You did so well; you are so smart." The next class was told, "You did so well, you must have worked very hard." Then the control group received no praise. The word lists and classrooms were also controlled, in that they did not vary. The students were randomly assigned to the classrooms.

The next spelling test resulted in the following sample grades:

Smart	Worked	No Praise
88	98	89
85	95	88
90	97	89
80	95	82

Note there is variation "within" each group. ANOVA will control for this and then determine whether there is a significant difference "between" the three groups.



© Kendall Hunt Publishing Company

is typically quick, easy, and inexpensive to gather as opposed to carrying out experiments. However, it does have some drawbacks. We cannot always create a control group, balance the sample sizes, or prevent other factors from influencing the outcomes when we just observe things. Large amounts of data in business come to us "second-hand" and not from primary sources. This also can

introduce less accurate measurements. An example of primary observational data could be recording detergent sales in a grocery store in conjunction with coupons offered or location on the aisles. Secondary observational data could be gathered from the Internet comparing P&G detergent sales to Colgate-Palmolive, and Purex.

Observational data coming to us from past studies is considered retrospective. If HEB sent data to us from sales from 2019 to track buying behaviors based on coupon usage, it would be retrospective. However, if we followed shoppers on the detergent aisle noting their purchases and uses of coupons, this would be prospective. We would be actively obtaining the data as the subjects chose the items.



© Kendall Hunt Publishing Company

When afforded the time, resources, and equipment needed for an experiment, we prefer experimental data. Researchers start by determining what type of data they are seeking. In order to use Analysis of Variance, the response variable (the value being recorded) needs to be quantitative and contain more than two groups of values. The experiment involves manipulating attributes of participants or objects and observing the outcomes. The act of measuring or recording these outcomes is the experiment. ANOVA is the methodology making these pairwise comparisons of more than two means.

The attributes that contribute to the outcomes are called factors. Factors can be quantitative or qualitative. For instance, the detergent sales may be influenced by the amount of a coupon (quantitative) or simply the presence of a coupon offer (qualitative). The factors then may be set to various levels. Coupons can be for 0 cents off, 25 cents, 50, or a $1. Finally, the combination of all factors and levels is called a treatment effect.

The difference between an experiment and other forms of investigations hinge upon two components: 1. the deliberate manipulation of factors to create treatment effects and 2. the random assignment of those treatments to the subjects.

The Field of Experimental Design

Four elements are basic to every experiment.

1. Random assignment of subjects to the groups. Researchers try to eliminate any bias by using things like random numbers to generate lists of subjects to be chosen for the experiment.

2. Control of the sources of variation between the groups while measuring the variation within them. We can also include a control group that acts as a baseline. When placebos (fake treatments) are used in pharmaceutical trials, patients receiving them become the control group. If the technicians conducting the experiment do not know which treatments each group is receiving, it is called a blind experiment. Then if the data analysts are also unaware of the treatments each group had, it is considered a double blind experiment.

3. Replication of the treatments within each group is necessary. If the sample sizes are the same for each level of factors, it is balanced. This is preferred. However, if the sample sizes differ, which is often the case with observational data, it is unbalanced. When proving a scientific theory, researchers endeavor to replicate entire experiments to hopefully find the same outcomes. They would have different subjects at another point in time, possibly under different circumstances but with similar results.

4. Blocking: Grouping the subjects by a second factor is called blocking. If a researcher feels that another factor such as gender, ethnicity, or age could be affecting the outcome beyond the main factor, a block may be added.

Designs

There are three types of designs used in Analysis of Variance: Completely Randomized, Randomized Block, and Factorial.

Continuing with the "Praise" problem, note how the data is arranged differently and whether there are one or two factors involved.

1. *Completely Randomized*—One Factor ANOVA

Smart	Worked	No Praise
88	98	89
85	95	88
90	97	89
80	95	82

2. *Randomized Block*—Two Factor Without Replication

	Smart	Worked	No Praise
Male	88	98	89
Female	90	97	89

3. *Factorial Design (2 replications)*—Two Factor with Replication

	Smart	Worked	No Praise
Male	88	98	89
	85	95	88
Female	90	97	89
	80	95	82

Completely Randomized Design

EXAMPLE 1

Note there is one factor, the type of praise the children received. The children had been randomly assigned the group (level of praise) they would get. This is the simplest type of experiment. It is considered balanced since each group had a sample size of four. The response variable is the test score. The MegaStat command is Analysis of Variance / One Factor.

One factor ANOVA

	Mean	*n*	*Std. Dev*	
	85.8	4	4.35	Smart
	96.3	4	1.50	Worked
	87.0	4	3.37	No Praise
	89.7	12	5.73	Total

ANOVA table

Source	*SS*	*df*	*MS*	*F*	*p-value*
Treatment	263.17	2	131.583	12.15	.0028
Error	97.50	9	10.833		
Total	360.67	11			

Post hoc analysis

p-values for pairwise *t*-tests

		Smart	No Praise	Worked
		85.8	87.0	96.3
Smart	85.8			
No Praise	87.0	0.6042		
Worked	96.3	0.0015	0.0032	

Tukey simultaneous comparison *t*-values (df = 9)

		Smart	No Praise	Worked
		85.8	87.0	96.3
Smart	85.8			
No Praise	87.0	0.54		
Worked	96.3	4.51	3.97	

Critical values for experimentwise error rate:

	0.05	2.79
	0.01	3.84

Figure 8.1 Example 1 Output

Source: Carolyn Turner

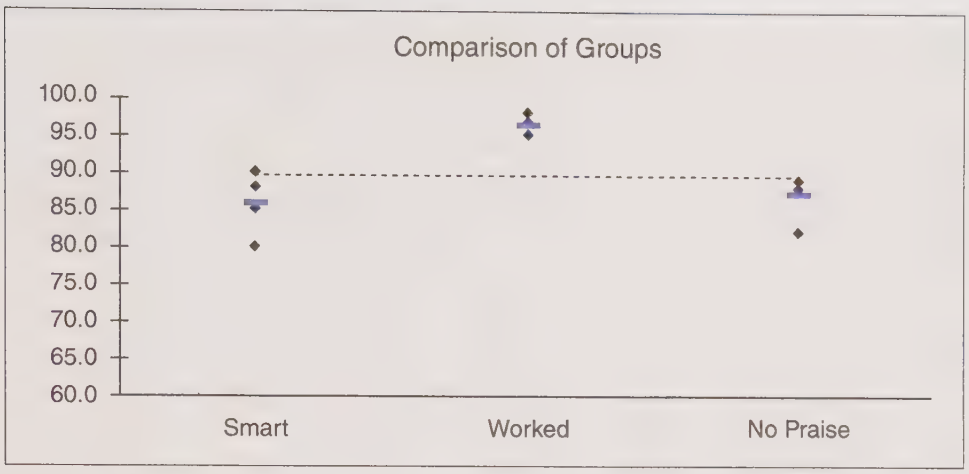


Figure 8.2 Example 1 Output
Source: Carolyn Turner

The hypothesis: H_0: $u_1 = u_2 = u_3$

H_a: At least 1 mean differs

In Figure 8.1, the *p*-value of .0028 indicates there is only a 0.28% chance that we would get a test statistic of $F = 12.15$ if there was really no difference between the groups. Therefore, we will reject H_0 and conclude with 95% certainty that at least one of the groups has a different test score than the others.

Multiple Comparison Tests

Now you should ask, "Which group is different?"

The *Post Hoc Analysis* contains three areas to help with this answer. The *p*-values that are below 0.05 are highlighted in yellow. When pairing the groups, the highlighted values indicate differences between them. Tukey's simultaneous comparison *t*-values are shown in blue indicating differences between the groups. The plot in Figure 8.2 indicates differences when the intervals do not overlap.

So our final analysis would state that the "Worked Hard" group received significantly higher scores than the other two. There was no significant difference between the "Smart" and "No Praise" groups. You might also note that there is more variation within the "Smart" group.

EXAMPLE 2

Randomized Block Design

The Randomized Block Design randomly assigns the subjects to treatments within each block. In this example, researchers **blocked** on gender to see if it had an effect on the test scores. Praise type will still be the main factor. The MegaStat Command is Analysis of Variance / Randomized Block. The results are below in Figure 8.3.

Randomized blocks ANOVA

Mean	*n*	*Std. dev*	
89.000	2	1.414	Smart
97.500	2	0.707	Worked
89.000	2	0.000	No Praise
91.667	3	5.508	Male
92.000	3	4.359	Female
91.833	6	4.446	Total

ANOVA table

Source	*SS*	*df*	*MS*	*F*	*p-value*
Treatments	96.33	2	48.167	41.29	.0236
Blocks	0.17	1	0.167	0.14	.7418
Error	2.33	2	1.167		
Total	98.83	5			

Post hoc analysis

p-values for pairwise *t*-tests

		Smart	No Praise	Worked
		89.000	89.000	97.500
Smart	89.000			
No Praise	89.000	1.0000		
Worked	97.500	0.0158	0.0158	

Tukey simultaneous comparison *t*-values (df = 2)

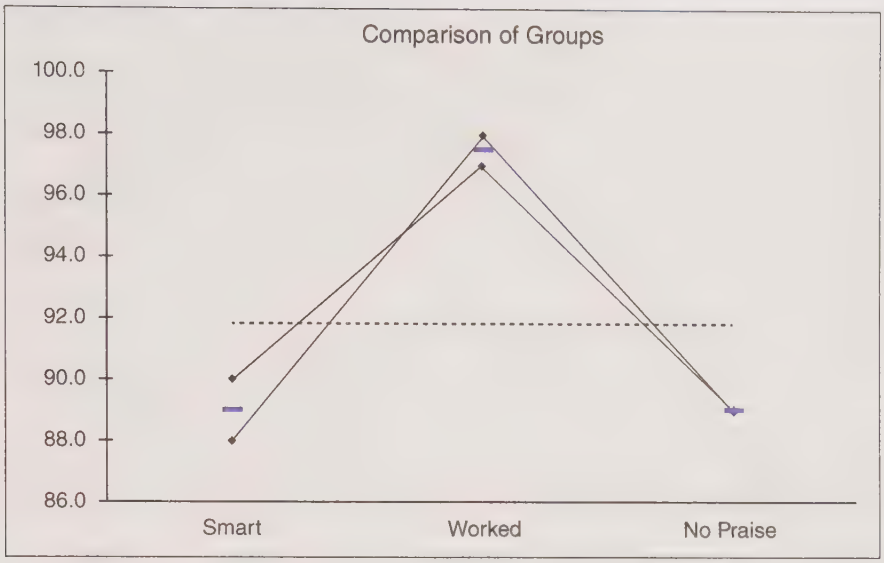
		Smart	No Praise	Worked
		89.000	89.000	97.500
Smart	89.000			
No Praise	89.000	0.00		
Worked	97.500	7.87	7.87	

Critical values for experimentwise error rate:

	0.05	5.89
	0.01	13.44

Figure 8.3 Example 2 Output

Source: Carolyn Turner



We set up two hypotheses: H_0: $u_S = u_W = u_{NP}$ H_a: **At least one mean differs**

H_0: $u_M = u_F$ H_a: **At least one mean differs**

In Figure 8.3, the *p*-value for the Treatments (Praise type) is 0.0236; therefore, at least one group of tests has significantly different scores. The *p*-value for the Block (Gender) is 0.7418, so it made no significant difference in test scores.

Only on a look at the post hoc analysis for those groups were they found to have significant differences. Again, we are 95% certain that the Worked Hard group is significantly different from the other two.

EXAMPLE 3

Factorial Design

Now look at the third group of data. Notice how the samples have been taken more than one time in each pairing of Gender and Praise type. Figure 8.4 contains the MegaStat output.

Two factor ANOVA

		Factor 2			
	Means:	Smart	Worked	No Praise	
	Male	86.5	96.5	88.5	90.5
Factor 1	Female	85.0	96.0	85.5	88.8
		85.8	96.3	87.0	89.7

Replications per cell 2

ANOVA table

Source	*SS*	*df*	*MS*	*F*	*p-value*
Factor 1	8.33	1	8.333	0.58	0.4747
Factor 2	263.17	2	131.583	9.18	0.0149
Interaction	3.17	2	1.583	0.11	0.8972
Error	86.00	6	14.333		
Total	360.67	11			

Post hoc analysis
p-values for pairwise *t*-tests for Factor 2

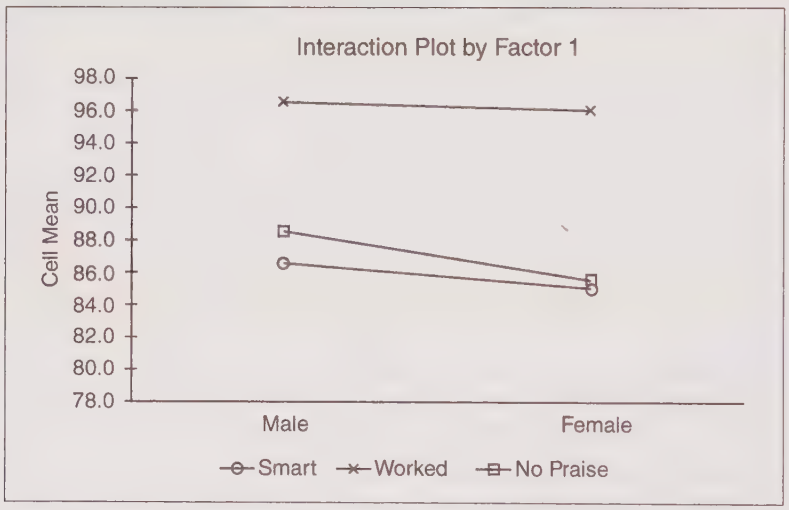
		Smart	No Praise	Worked
		85.8	87.0	96.3
Smart	85.8			
No Praise	87.0	0.6570		
Worked	96.3	0.0078	0.0135	

Tukey simultaneous comparison *t*-values (df = 6)

		Smart	No Praise	Worked
		85.8	87.0	96.3
Smart	85.8			
No Praise	87.0	0.47		
Worked	96.3	3.92	3.46	

Critical values for experimentwise error rate:

0.05	3.07
0.01	4.48



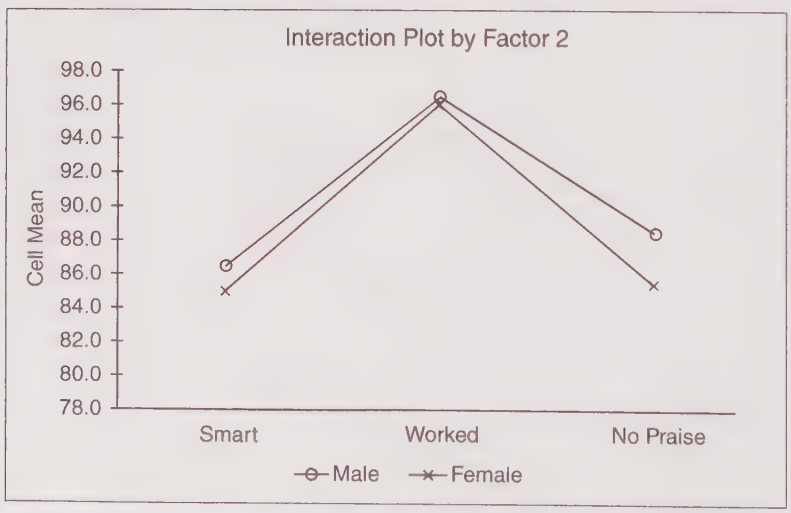


Figure 8.4 Two-Factor ANOVA Output
Source: Carolyn Turner

In a factorial design, there are *treatments* that represent all possible combination of level and factors.

Factorial designs test for something called Interaction. This appears significant when two factors combine at different levels to yield different effects than would be expected when multiplying the effects of two factors together. Many researchers begin their analysis with a factorial design, so they don't inadvertently leave out the test for interaction.

The first hypothesis tests for interaction. If it is present, there is no need to test the individual factors. The plot will display intersecting lines and detail how the factors combine at the different levels.

H_0: There is No Interaction

H_a: There is Interaction

In this example there is no interaction. p-value for Interaction is 0.8972. Do Not Reject H_0.

H^0: $u_m = u_F$ H_a: At least one mean differs p-value = 0.4747 > 0.05 Do Not Reject H_0.

H_0: $u_s = u_w = u_{NP}$ H_a: At least one mean differs p-value = 0.0149 < 0.05 Reject H_0.

When the lines on the plot fail to intersect, there is no interaction. So the only factor contributing the variation in test scores is the type of praise.

Beware!

Secondary or observational data often contain confounded factors. These are two separate items that were surveyed in unison and cannot be broken down after the fact. For instance, if I had First Grade Males, First Grade Females, Second Grade Males, and Second Grade Females in the study: The grade and gender are confounded. The levels contain more than one factor. Many times, questionnaires are poorly designed and confounding occurs. An example follows:

√ Married with Children vs. √ Married without Children
√ Single with Children vs. √ Single without Children.

Two factors have been combined—marital status and children. It would be preferable to know how many children were in each household and more detail about their marital status in separate questions.

Another problem that can occur is called a lurking variable. Something other than the factors used in the experiment is driving the outcomes. If the students had teachers with very different teaching methods this could make the praise type appear to be causing differences in scores when it really wasn't. Lurking variables may not be in the actual model but driving two other variables to make a causal relationship appear between the two variables. The average number of housing starts may appear to be "caused by" the cost of building materials, but another variable may be *lurking* or driving the housing starts such as interest rates.

The statistic calculated in ANOVA is F.

$$F_{k-1,N-k} = \frac{MST}{MSE}$$

What if a researcher doesn't use ANOVA, but does single t tests for each pair of factors? The chance of making a Type I Error increases exponentially with each wrong pairing. The chance of this is called the Effective Error Rate. It is calculated by taking

$1-(1-\alpha)_k$, **k** is the number of pairwise comparisons that can be made.

So in the Praise problem, if we had simply made comparisons with *t* tests, the Effective Error Rate would equal $1-(1-0.05)^3 = 0.1426$.

Observational data is often unbalanced, not randomized, and contains no control group. In these cases avoid making any statements about causality.

Hints for Reliable Data Analysis

- Look for Outliers and changing Variances.
- Choose proper multiple comparison tests. Tukey is best for balanced experiments. Other statistical programs such as SAS can run Bonferroni and Scheffe when data is unbalanced.
- Begin with replication so that an Interaction test can be run, then simplify the design if there is none.
- Do not draw causal conclusions with observational data.



© Kendall Hunt Publishing Company

Problems

1. *Champoo* offers several lines of hair care products in the following categories: Anti-Frizz, Volume, and Repair. A marketing VP has noticed there appears to be some kind of relationship between the mean sales of the various products and summer and fall months. Data was obtained from the months of June through October from three types of distributors in Table 8.1. Use the output in Figure 8.5 to answer questions concerning the interaction of the variables.

Month	Anti Frizz	Volume	Repair
June	7.5	9	11.1
	8.05	8.8	10.7
	8.25	8.2	10.2
July	10.75	7.8	12.4
	10.2	8.2	11.75
	9.8	7.5	11.5
August	13.2	10	8.2
	12.5	9	7.2
	11.45	8.2	6.1
September	8.7	9.15	13.2
	6.4	9.4	14.4
	7.2	8.7	15
October	9.1	11.2	9.4
	7.8	9.8	9.5
	11	10	10.2

Table 8.1 *Champoo* Data
Source: Carolyn Turner

Define the following:
a. Experiment
b. Factors
c. Levels
d. Design
e. Response Variable
f. Treatments
g. Replications
h. What MegaStat command should be run?
i. Set up the necessary hypotheses.

Two factor ANOVA

Means:

		Factor 2			
		Anti Frizz	Volume	Repair	
Factor 1	June	7.93	8.67	10.67	9.089
	July	10.25	7.83	11.88	9.989
	August	12.38	9.07	7.17	9.539
	September	7.43	9.08	14.20	10.239
	October	9.30	10.33	9.70	9.778
		9.46	9.00	10.72	9.727

3 Replications per cell

ANOVA table

Source	*SS*	*df*	*MS*	*F*	*p-value*
Factor 1	6.98	4	1.745	2.76	0.0455
Factor 2	23.96	2	11.980	18.98	4.71E-06
Interaction	131.10	8	16.388	25.96	1.84E-11
Error	18.94	30	0.631		
Total	180.98	44			

Post hoc analysis

p-values for pairwise *t*-tests for Factor 1

		June	August	October	July	September
		9.089	9.539	9.778	9.989	10.239
June	9.089					
August	9.539	0.2390				
October	9.778	0.0758	0.5284			
July	9.989	0.0227	0.2390	0.5772		
September	10.239	0.0045	0.0714	0.2278	0.5096	

p-values for pairwise *t*-tests for Factor 2

		Volume	Anti Frizz	Repair
		8.997	9.460	10.723
Volume	8.997			
Anti Frizz	9.460	0.1207		
Repair	10.723	1.60E-06	.0001	

j. State your conclusion. Alpha = 0.05 Use the plot in Figure 8.5.

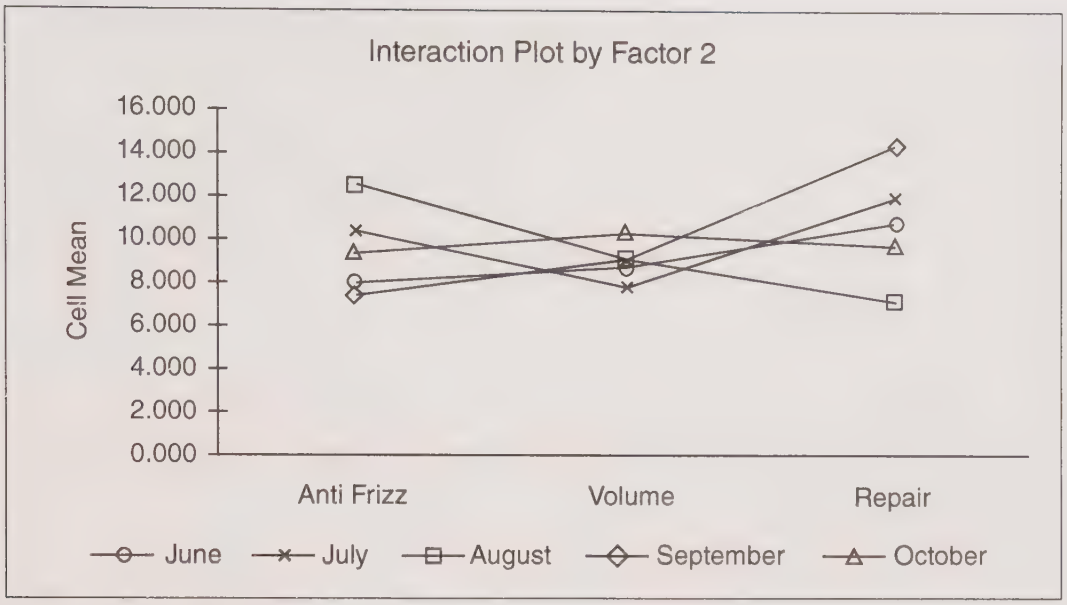


Figure 8.5 Champoo Output
Source: Carolyn Turner

2. Good Housekeeping Labs measured the brightness level of 16 loads of laundry. The researchers manipulated the brand of detergent and amount of time each load was washed. Test to see if a difference exists in the mean brightness levels. Alpha = 0.05 Use the output in

Time (min)	Tide	Cheer	Gain	All
18	0.120	0.120	0.165	0.110
20	0.115	0.145	0.145	0.125
22	0.180	0.175	0.185	0.160
24	0.165	0.190	0.200	0.160

Table 8.2 Detergent Data
Source: Carolyn Turner

Transposed data to test the blocks.				
Time	**18 min**	**20 min**	**22 min**	**24 min**
Tide	0.120	0.115	0.180	.165
Cheer	0.120	0.145	0.175	.190
Gain	0.165	0.145	0.185	.200
All	0.110	0.125	0.160	.160

Table 8.3 Transposed Data
Source: Carolyn Turner

Define the following:

a. Experiment
b. Factors
c. Levels
d. Design
e. Response Variable
f. Treatments
g. Replications
h. Set up the necessary hypotheses.
i. What MegaStat command should be run?

Randomized Blocks ANOVA

Mean	*n*	*Std. dev*	
0.145000	4	0.032404	Tide
0.157500	4	0.031225	Cheer
0.173750	4	0.023936	Gain
0.138750	4	0.025290	All
0.128750	4	0.024622	18 min
0.132500	4	0.015000	20 min
0.175000	4	0.010801	22 min
0.178750	4	0.019311	24 min
0.153750	16	0.028954	Total

ANOVA table

Source	*SS*	*df*	*MS*	*F*	*p-value*
Treatments	0.00286	3	0.000954	7.81	0.0071
Blocks	0.00861	3	0.002871	23.49	0.0001
Error	0.00110	9	0.000122		
Total	0.01258	15			

Post hoc analysis

p-values for pairwise *t*-tests

		All	Tide	Cheer	Gain
		0.138750	0.145000	0.157500	0.173750
All	0.138750				
Tide	0.145000	0.4446			
Cheer	0.157500	0.0400	0.1443		
Gain	0.173750	0.0015	0.0051	0.0674	

Tukey simultaneous comparison *t*-values (df = 9)

		All	Tide	Cheer	Gain
		0.138750	0.145000	0.157500	0.173750
All	0.138750				
Tide	0.145000	0.80			
Cheer	0.157500	2.40	1.60		
Gain	0.173750	4.48	3.68	2.08	

Critical values for experimentwise error rate:

	0.05	3.12
	0.01	4.21

p-values for pairwise *t*-tests

		18 min	20 min	22 min	24 min
		0.128750	0.132500	0.175000	0.178750
18 min	0.128750				
20 min	0.132500	0.6429			
22 min	0.175000	0.0002	0.0004		
24 min	0.178750	0.0001	0.0002	0.6429	

Tukey simultaneous comparison *t*-values (df = 9)

		18 min	20 min	22 min	24 min
		0.128750	0.132500	0.175000	0.178750
18 min	0.128750				
20 min	0.132500	0.48			
22 min	0.175000	5.92	5.44		
24 min	0.178750	6.40	5.92	0.48	

Table 8.3 Detergent Output
Source: Carolyn Turner

j. State your conclusions. Alpha = 0.05

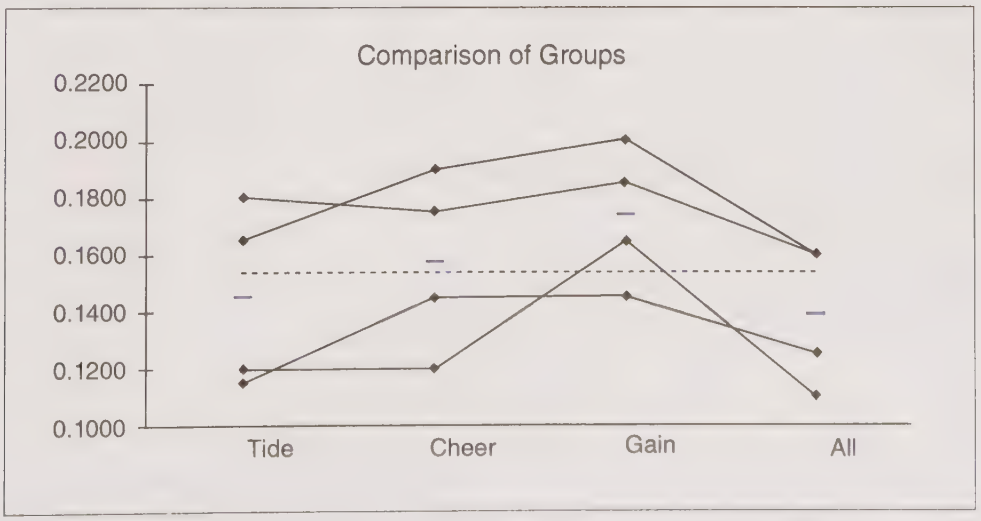


Figure 8.6 Detergent / Time Chart
Source: Carolyn Turner

k. What do the intersecting lines on Figure 8.6 indicate?
l. Was it tested in the current experiment?
m. What would be needed to test for this condition?

3. Three neighborhoods competed to collect the most recycling material over four months. Test to see if there is a difference in the mean amount of material using a 0.05 level of significance.

Neighborhood Recycling		
Mtnvw.	**Lake Shore**	**Cedar Rdg.**
64	74	75
68	71	80
70	69	76
60	70	78

Table 8.4 Recycling Data
Source: Carolyn Turner

Define the following:
a. Experiment
b. Factors
c. Levels
d. Design
e. Response Variable
f. Replications
g. Set up the necessary hypothesis.
h. What MegaStat command should be run?
i. State your conclusion. Alpha = 0.05 Use Table 8.5 and Figure 8.7.

One factor ANOVA

Mean	*n*	*Std. dev*	
65.5	4	4.43	Mtnvw.
71.0	4	2.16	Lake Shore
77.3	4	2.22	Cedar Rdg
71.3	12	5.75	Total

ANOVA table

Source	*SS*	*df*	*MS*	*F*	*p-value*
Treatment	276.50	2	138.250	14.18	0.0017
Error	87.75	9	9.750		
Total	364.25	11			

Post hoc analysis
p-values for pairwise *t*-tests

		Mtnvw.	Lake Shore	Cedar Rdg
		65.5	71.0	77.3
Mtnvw.	65.5			
Lake Shore	71.0	0.0344		
Cedar Rdg	77.3	0.0005	.0197	

Table 8.5 Recycling Output
Source: Carolyn Turner

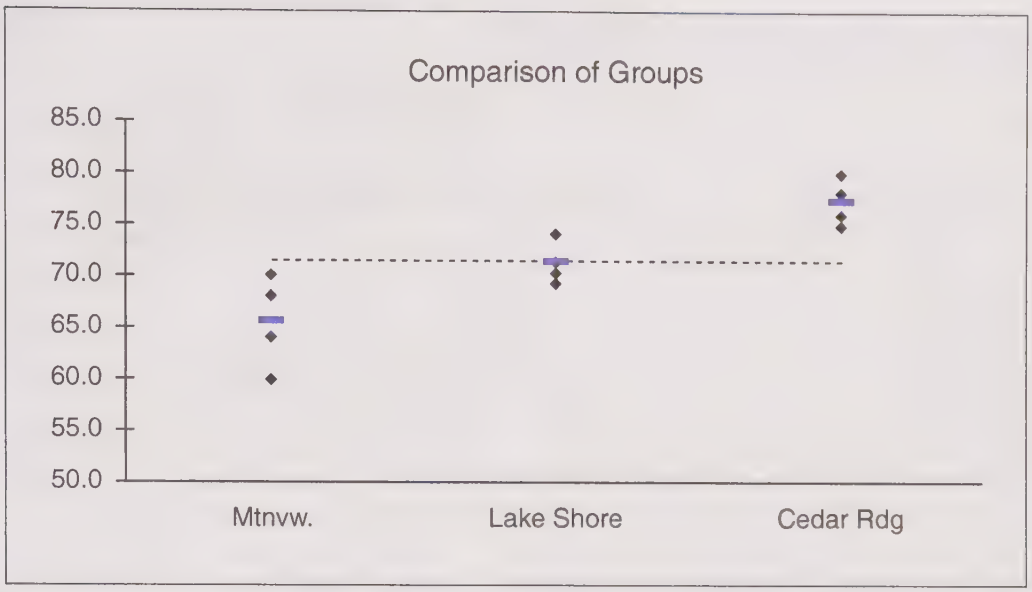


Figure 8.7 Recycling Chart
Source: Carolyn Turner

American Council on Exercise Report on Reebok, Sketchers, and MBT shoes

In the early 2000s, Skechers, Reebok, and MBT ran advertisements claiming to "tone up" women's bodies 23% by simply walking around in the shoes doing normal activities.

The American Council on Exercise study revealed that there is no significant health benefit to wearing toning shoes. "Across the board, none of the toning shoes showed statistically significant increases in either exercise response or muscle activation during any of the treadmill trials," said ACE. "There is simply no evidence to support the claims that these shoes will help wearers exercise more intensely, burn more calories or improve muscle strength and tone."[1]

In response, Skechers has indicated that they have competent and reliable scientific evidence that supports certain benefits to wearing their Shape-Ups shoes. Several lawsuits were filed against all three manufacturers for making false claims resulting in Skechers paying $40 million in damages and Reebok $25 million.

Go to the website listed below and look at the output.
What were the findings of the council?

[1] Porcari, John, Greany, John, Tepper, Stephanie, Edmonson, Brian, Foster, Carl, Anders, Mark, "Will Toning Shoes Really Give You a Better Body? Accessed August 2014. http://www.acefitness.org/getfit/studies/toningshoes072010.pdf

Answers to Chapter 8 Problems

Problem 1

a. Experiment—Recording Champoo sales based on type of shampoo and month sold.
b. Factors—Type of shampoo and month
c. Levels—Anti Frizz, Volume, Repair
d. Design—Factorial
e. Response Variable— Average Sales
f. Treatments—All combinations of types of shampoos and months at different levels.
g. Replications—3
h. MegatStat: Analysis of Variance / Two Factor
i. H_0: There is no interaction between types of shampoo and months
H_a: There is interaction between types of shampoo and months
$F = 25.96$ p-value = 1.84E-11 Reject H_0
j. Using 95% confidence, there is interaction between the two factors.
Anti Frizz sells best in August
Volumizing sells best in October
Repair type sells best in September

Problem 2

a. Experiment—Measuring brightness levels of laundry based on brand of detergent and number of minutes washed.
b. Factors—Brands and Minutes
c. Levels—Tide, Cheer, Gain, All, 18, 20, 22, 24 minutes
d. Design—Randomized Block
e. Response Variable—Average brightness level
f. Treatments—All combinations of brands and minutes at different levels.
g. Replications—1per combination of brand and minutes
h. H_0: $\mu T=\mu C=\mu G=\mu A$ H_a: At least 2 means differ $F = 7.81$ p-value = 0.0071 Reject H_0
H_0: $\mu_{18}=\mu_{20}=\mu_{22}=\mu_{24}$ H_a: At least 2 means differ $F = 23.49$ p-value = 0.0001 Reject H_0
i. MegaStat: Analysis of Variance/Randomized Block
j. Conclusion: Using 95% confidence, at least two means differ based on brand and at least two means differ based on minutes washed.
k. Based on the Post hoc Analysis, Gain, and Cheer are brighter than All and Tide, and washing 22 or 24 minutes yields brighter levels than 18 or 22 minutes.
l. The intersecting lines indicate the two factors may interact.
m. The experiment would need to be replicated and a factorial design set up.

Problem 3

a. Experiment—Recording recycling amounts based neighborhood.
b. Factor—Neighborhood
c. Levels—Mtnvw., Lake Shore, Cedar Rdg.
d. Design—Completely Randomized
e. Response Variable—Average amount recycled.
f. No replications
g. H_0: $\mu_M = \mu_{LS} = \mu_{CR}$ H_a: At least 2 means differ
h. Run Analysis of Variance / One Factor ANOVA
i. $F = 14.18$ p-value $= 0.0017$ Reject H_0.
Using 95% confidence at least 2 neighborhoods differ in recycling amounts. Cedar Ridge recycles more than Lake Shore or Mountainview.



CPSIA information can be obtained
at www.ICGtesting.com
Printed in the USA
LVHW060306041221
705097LV00006B/6



INTRODUCTION TO BUSINESS ANALYSIS

FRED HULME • CAROLYN HULME TURNER

Fred Hulme wrote this textbook to introduce the fundamental concepts of quantitative business analysis to students. It is written with the knowledge that most business school students have not had a full course in statistics and especially not one focusing on business applications. The problems lend themselves to the analytical side more than the mathematical side with the idea that the course is a tool for business students and not an end unto itself. An emphasis is placed on using computer software to solve realistic problems in terms of complexity and size.

Chapters 1 and 2 involve summarizing data through tables, charts, and numerical calculations. These summaries provide information used in decision making. Techniques for condensing large amounts of data into a meaningful summary are presented. Both Excel and MegaStat are used within the text, but the same concepts can easily be understood using JMP, SAS, and R. Proficiency in Microsoft Excel is often requested by organizations that recruit university business school interns and graduates.

Business decisions often involve choosing from several alternatives. In order to decide among several alternatives, one must be able to estimate the consequences of each alternative. The consequences often depend upon factors that are subject to chance. The uncertainty cannot be eliminated, but risk analysis methods exist which allow you to estimate the risk in your decisions. These tools are discussed in Chapter 3.

The concepts of gathering samples, analyzing the data, and making inferences about the populations are covered in Chapters 4 through 6. The topics of sampling procedures, probability, interval estimation, and hypothesis testing are all covered.

The last two chapters cover the analysis of count data using Chi Square techniques and comparing more than two means with Analysis of Variance designs. Understanding both methodologies boost the analytical skills of business students.

09-AKF-879

ISBN 978-1-7924-2293-5
90000 >
9 781792 422935

Kendall Hunt
publishing company