Nonparametric Statistics

A Step-by-Step Approach

Second Edition

Gregory W. Corder • Dale I. Foreman

WWW. LINK AVAILABLE

WILEY

NONPARAMETRIC STATISTICS

SECOND EDITION

NONPARAMETRIC STATISTICS

A Step-by-Step Approach

GREGORY W. CORDER

DALE I. FOREMAN

WILEY

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Corder, Gregory W., 1972– author.
Nonparametric statistics : a step-by-step approach / Gregory W. Corder. Dale I. Foreman – Second edition.
pages cm
Includes bibliographical references and index.
ISBN 978-1-118-84031-3 (hardback)
1. Nonparametric statistics. I. Foreman, Dale I., author II. Title.
QA278.8.C67 2014
519.5′4–dc23

2013042040

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CONTENTS

PREFACE

The social, behavioral, and health sciences have a need for the ability to use nonparametric statistics in research. Many studies in these areas involve data that are classified in the nominal or ordinal scale. At times, interval data from these fields lack parameters for classification as normal. Nonparametric statistical tests are useful tools for analyzing such data.

Purpose of This Book

This book is intended to provide a conceptual and procedural approach for nonparametric statistics. It is written so that someone who does not have an extensive mathematical background may work through the process necessary to conduct the given statistical tests presented. In addition, the outcome includes a discussion of the final decision for each statistical test. Each chapter takes the reader through an example from the beginning hypotheses, through the statistical calculations, to the final decision as compared with the hypothesis. The examples are then followed by a detailed, step-by-step analysis using the computer program SPSS®. Finally, research literature is identified which uses the respective nonparametric statistical tests.

Intended Audience

While not limited to such, this book is written for graduate and undergraduate students in social science programs. As stated earlier, it is targeted toward the student who does not have an especially strong mathematical background, but can be used effectively with a mixed group of students that includes students who have both strong and weak mathematical background.

Special Features of This Book

There are currently few books available that provide a practical and applied approach to teaching nonparametric statistics. Many books take a more theoretical approach to the instructional process that can leave students disconnected and frustrated, in need of supplementary material to give them the ability to apply the statistics taught.

It is our hope and expectation that this book provides students with a concrete approach to performing the nonparametric statistical procedures, along with their application and interpretation. We chose these particular nonparametric procedures since they represent a breadth of the typical types of analyses found in social science research. It is our hope that students will confidently learn the content presented with the promise of future successful applications.

In addition, each statistical test includes a section that explains how to use the computer program SPSS. However, the organization of the book provides effective instruction of the nonparametric statistical procedures for those individuals with or without the software. Therefore, instructors (and students) can focus on learning the tests with a calculator, SPSS, or both.

A Note to the Student

We have written this book with you in mind. Each of us has had a great deal of experience working with students just like you. Over the course of that time, it has been our experience that most people outside of the fields of mathematics or hard sciences struggle with and are intimidated by statistics. Moreover, we have found that when statistical procedures are explicitly communicated in a step-by-step manner, almost anyone can use them.

This book begins with a brief introduction (Chapter 1) and is followed with an explanation of how to perform the crucial step of checking your data for normality (Chapter 2). The chapters that follow (Chapters 3–9) highlight several nonparametric statistical procedures. Each of those chapters focuses on a particular type of variable and/or sample condition.

Chapters 3–9 each have a similar organization. They each explain the statistical methods included in their respective chapters. At least one sample problem is included for each test using a step-by-step approach. (In some cases, we provide additional sample problems when procedures differ between large and small samples.) Then, those same sample problems are demonstrated using the statistical software package SPSS. Whether or not your instructor incorporates SPSS, this section will give you the opportunity to learn how to use the program. Toward the end of each chapter, we identify examples of the tests in published research. Finally, we present sample problems with solutions.

As you seek to learn nonparametric statistics, we strongly encourage you to work through the sample problems. Then, using the sample problems as a reference, work through the problems at the end of the chapters and additional data sets provided.

New to the Second Edition

Given an opportunity to write a second edition of this book, we revised and expanded several portions. Our changes are based on feedback from users and reviewers.

We asked several undergraduate and graduate students for feedback on Chapters 1 and 2. Based on their suggestions, we made several minor changes to Chapter 1 with a goal to improve understanding. In Chapter 2, we expanded the section that describes and demonstrates the Kolmogorov–Smirnov (K-S) one-sample test.

After examining current statistics textbooks and emerging research paper, we decided to include two additional tests. We added the sign test to Chapter 3 and the Kolmogorov–Smirnov (K-S) two-sample test to Chapter 4. We also added a discussion on statistical power to Chapter 3 as requested by instructors who had adopted our book for their courses.

Since our book's first edition, SPSS has undergone several version updates. Our new edition of the book also has updated directions and screen captures for images of SPSS. Specifically, these changes reflect SPSS version 21.

We have included web-based tools to support our book's new edition. If you visit the publisher's book support website, you will find a link to a Youtube channel that includes narrated screen casts. The screen casts demonstrate how to use SPSS to perform the tests included in this book. The publisher's book support website also includes a link to a decision tree that helps the user determine an appropriate type of statistical test. The decision tree is organized using Prezi. The branches terminate with links to the screen casts on YouTube.

Gregory W. Corder
Dale I. Foreman

LIST OF VARIABLES

English Symbols

C	number of columns in a contingency table; number of categories
C_F	tie correction for the Friedman test
C_H	tie correction for the Kruskal–Wallis *H*-test
D, $\tilde{D}$	divergence between values from cumulative frequency distributions
D_i	difference between a ranked pair
df	degrees of freedom
f_e	expected frequency
f_o	observed frequency
$\hat{f}_r$	empirical frequency value
F_r	Friedman test statistic
g	number of tied groups in a variable
h	correction for continuity
H	Kruskal–Wallis test statistic
H_A	alternate hypothesis
H_O	null hypothesis
k	number of groups
K	kurtosis
M	midpoint of a sample
n	sample size
N	total number of values in a contingency table
p	probability
P_i	a category's proportion with respect to all categories
r_b	biserial correlation coefficient for a sample
r_{pb}	point-biserial correlation coefficient for a sample
r_s	Spearman rank-order correlation coefficient for a sample
R	number of runs; number of rows in a contingency table
R_i	sum of the ranks from a particular sample
s	standard deviation of a sample
S_k	skewness
SE	standard error
t	*t* statistic
t_i	number of tied values in a tie group
T	Wilcoxon signed rank test statistic
U	Mann–Whitney test statistic
$\overline{x}$	sample mean
y	height of the unit normal curve ordinate at the point dividing two proportions

z	the number of standard deviations away from the mean
Z	Kolmogorov–Smirnov test statistic

Greek Symbols

α	alpha, probability of making a type I error
α_B	adjusted level of risk using the Bonferroni procedure
β	beta, probability of making a type II error
θ	theta, median of a population
μ	mu, mean value for a population
ρ	rho, correlation coefficient for a population
σ	sigma, standard deviation of a population
Σ	sigma, summation
χ^2	chi-square test statistic

CHAPTER 1

NONPARAMETRIC STATISTICS: AN INTRODUCTION

1.1 OBJECTIVES

In this chapter, you will learn the following items:

- The difference between parametric and nonparametric statistics.
- How to rank data.
- How to determine counts of observations.

1.2 INTRODUCTION

If you are using this book, it is possible that you have taken some type of introductory statistics class in the past. Most likely, your class began with a discussion about probability and later focused on particular methods of dealing with populations and samples. Correlations, z-scores, and t-tests were just some of the tools you might have used to describe populations and/or make inferences about a population using a simple random sample.

Many of the tests in a traditional, introductory statistics text are based on samples that follow certain assumptions called parameters. Such tests are called *parametric tests*. Specifically, parametric assumptions include samples that

- are randomly drawn from a normally distributed population,
- consist of independent observations, except for paired values,
- consist of values on an interval or ratio measurement scale,
- have respective populations of approximately equal variances,
- are adequately large,* and
- approximately resemble a normal distribution.

*The minimum sample size for using a parametric statistical test varies among texts. For example, Pett (1997) and Salkind (2004) noted that most researchers suggest $n > 30$. Warner (2008) encouraged considering $n > 20$ as a minimum and $n > 10$ per group as an absolute minimum.

Nonparametric Statistics: A Step-by-Step Approach, Second Edition. Gregory W. Corder and
Dale I. Foreman.

If any of your samples breaks one of these rules, you violate the assumptions of a parametric test. You do have some options, however.

You might change the nature of your study so that your data meet the needed parameters. For instance, if you are using an ordinal or nominal measurement scale, you might redesign your study to use an interval or ratio scale. (See Box 1.1 for a description of measurement scales.) Also, you might seek additional participants to enlarge your sample sizes. Unfortunately, there are times when one or neither of these changes is appropriate or even possible.

BOX 1.1

MEASUREMENT SCALES.

We can measure and convey variables in several ways. *Nominal* data, also called categorical data, are represented by counting the number of times a particular event or condition occurs. For example, you might categorize the political alignment of a group of voters. Group members could either be labeled democratic, republican, independent, undecided, or other. No single person should fall into more than one category.

A *dichotomous* variable is a special classification of nominal data; it is simply a measure of two conditions. A dichotomous variable is either discrete or continuous. A *discrete dichotomous* variable has no particular order and might include such examples as gender (male vs. female) or a coin toss (heads vs. tails). A *continuous dichotomous* variable has some type of order to the two conditions and might include measurements such as pass/fail or young/old.

Ordinal scale data describe values that occur in some order of rank. However, distance between any two ordinal values holds no particular meaning. For example, imagine lining up a group of people according to height. It would be very unlikely that the individual heights would increase evenly. Another example of an ordinal scale is a Likert-type scale. This scale asks the respondent to make a judgment using a scale of three, five, or seven items. The range of such a scale might use a 1 to represent *strongly disagree* while a 5 might represent *strongly agree*. This type of scale can be considered an ordinal measurement since any two respondents will vary in their interpretation of scale values.

An *interval* scale is a measure in which the relative distances between any two sequential values are the same. To borrow an example from the physical sciences, we consider the Celsius scale for measuring temperature. An increase from −8 to −7°C degrees is identical to an increase from 55 to 56°C.

A *ratio* scale is slightly different from an interval scale. Unlike an interval scale, a ratio scale has an absolute zero value. In such a case, the zero value indicates a measurement limit or a complete absence of a particular condition. To borrow another example from the physical sciences, it would be appropriate to measure light intensity with a ratio scale. Total darkness is a complete absence of light and would receive a value of zero.

On a general note, we have presented a classification of measurement scales similar to those used in many introductory statistics texts. To the best of our knowledge, this hierarchy of scales was first made popular by Stevens (1946). While Stevens has received agreement (Stake, 1960; Townsend & Ashby, 1984) and criticism (Anderson, 1961; Gaito, 1980; Velleman & Wilkinson, 1993), we believe the scale classification we present suits the nature and organization of this book. We direct anyone seeking additional information on this subject to the preceding citations.

If your samples do not resemble a normal distribution, you might have learned a strategy that modifies your data for use with a parametric test. First, if you can justify your reasons, you might remove extreme values from your samples called outliers. For example, imagine that you test a group of children and you wish to generalize the findings to typical children in a normal state of mind. After you collect the test results, most children earn scores around 80% with some scoring above and below the average. Suppose, however, that one child scored a 5%. If you find that this child speaks no English because he arrived in your country just yesterday, it would be reasonable to exclude his score from your analysis. Unfortunately, outlier removal is rarely this straightforward and deserves a much more lengthy discussion than we offer here.* Second, you might utilize a parametric test by applying a mathematical transformation to the sample values. For example, you might square every value in a sample. However, some researchers argue that transformations are a form of data tampering or can distort the results. In addition, transformations do not always work, such as circumstances when data sets have particularly long tails. Third, there are more complicated methods for analyzing data that are beyond the scope of most introductory statistics texts. In such a case, you would be referred to a statistician.

Fortunately, there is a family of statistical tests that do not demand all the parameters, or rules, that we listed earlier. They are called *nonparametric tests*, and this book will focus on several such tests.

1.3 THE NONPARAMETRIC STATISTICAL PROCEDURES PRESENTED IN THIS BOOK

This book describes several popular nonparametric statistical procedures used in research today. Table 1.1 identifies an overview of the types of tests presented in this book and their parametric counterparts.

TABLE 1.1

Type of analysis	Nonparametric test	Parametric equivalent
Comparing two related samples	Wilcoxon signed ranks test and sign test	*t*-Test for dependent samples
Comparing two unrelated samples	Mann–Whitney *U*-test and Kolmogorov–Smirnov two-sample test	*t*-Test for independent samples
Comparing three or more related samples	Friedman test	Repeated measures, analysis of variance (ANOVA)
Comparing three or more unrelated samples	Kruskal–Wallis *H*-test	One-way ANOVA

(*Continued*)

*Malthouse (2001) and Osborne and Overbay (2004) presented discussions about the removal of outliers.

TABLE 1.1 (***Continued***)

Type of analysis	Nonparametric test	Parametric equivalent
Comparing categorical data	Chi square (χ^2) tests and Fisher exact test	None
Comparing two rank-ordered variables	Spearman rank-order correlation	Pearson product–moment correlation
Comparing two variables when one variable is discrete dichotomous	Point-biserial correlation	Pearson product–moment correlation
Comparing two variables when one variable is continuous dichotomous	Biserial correlation	Pearson product–moment correlation
Examining a sample for randomness	Runs test	None

When demonstrating each nonparametric procedure, we will use a particular step-by-step method.

1.3.1 State the Null and Research Hypotheses

First, we state the hypotheses for performing the test. The two types of hypotheses are null and alternate. The *null hypothesis* (H_O) is a statement that indicates no difference exists between conditions, groups, or variables. The *alternate hypothesis* (H_A), also called a research hypothesis, is the statement that predicts a difference or relationship between conditions, groups, or variables.

The alternate hypothesis may be directional or nondirectional, depending on the context of the research. A directional, or one-tailed, hypothesis predicts a statistically significant change in a particular direction. For example, a treatment that predicts an improvement would be directional. A nondirectional, or two-tailed, hypothesis predicts a statistically significant change, but in no particular direction. For example, a researcher may compare two new conditions and predict a difference between them. However, he or she would not predict which condition would show the largest result.

1.3.2 Set the Level of Risk (or the Level of Significance) Associated with the Null Hypothesis

When we perform a particular statistical test, there is always a chance that our result is due to chance instead of any real difference. For example, we might find that two samples are significantly different. Imagine, however, that no real difference exists. Our results would have led us to reject the null hypothesis when it was actually true. In this situation, we made a type I error. Therefore, statistical tests assume some level of risk that we call alpha, or α.

There is also a chance that our statistical results would lead us to not reject the null hypothesis. However, if a real difference actually does exist, then we made a type II error. We use the Greek letter beta, β, to represent a type II error. See Table 1.2 for a summary of type I and type II errors.

TABLE 1.2

	We do not reject the null hypothesis	We reject the null hypothesis
The null hypothesis is actually true	No error	Type-I error, α
The null hypothesis is actually false	Type-II error, β	No error

After the hypotheses are stated, we choose the level of risk (or the level of significance) associated with the null hypothesis. We use the commonly accepted value of $\alpha = 0.05$. By using this value, there is a 95% chance that our statistical findings are real and not due to chance.

1.3.3 Choose the Appropriate Test Statistic

We choose a particular type of test statistic based on characteristics of the data. For example, the number of samples or groups should be considered. Some tests are appropriate for two samples, while other tests are appropriate for three or more samples.

Measurement scale also plays an important role in choosing an appropriate test statistic. We might select one set of tests for nominal data and a different set for ordinal variables. A common ordinal measure used in social and behavioral science research is the Likert scale. Nanna and Sawilowsky (1998) suggested that nonparametric tests are more appropriate for analyses involving Likert scales.

1.3.4 Compute the Test Statistic

The test statistic, or obtained value, is a computed value based on the particular test you need. Moreover, the method for determining the obtained value is described in each chapter and varies from test to test. For small samples, we use a procedure specific to a particular statistical test. For large samples, we approximate our data to a normal distribution and calculate a z-score for our data.

1.3.5 Determine the Value Needed for Rejection of the Null Hypothesis Using the Appropriate Table of Critical Values for the Particular Statistic

For small samples, we reference a table of critical values located in Appendix B. Each table provides a critical value to which we compare a computed test statistic. Finding a critical value using a table may require you to use such data characteristics as the degrees of freedom, number of samples, and/or number of groups. In addition, you may need the desired level of risk, or alpha (α).

For large samples, we determine a critical region based on the level of risk (or the level of significance) associated with the null hypothesis, α. We will determine if the computed z-score falls within a critical region of the distribution.

1.3.6 Compare the Obtained Value with the Critical Value

Comparing the obtained value with the critical value allows us to identify a difference or relationship based on a particular level of risk. Once this is accomplished, we can state whether we must reject or must not reject the null hypothesis. While this type of phrasing may seem unusual, the standard practice in research is to state results in terms of the null hypothesis.

Some of the critical value tables are limited to particular sample or group size(s). When a sample size exceeds a table's range of value(s), we approximate our data to a normal distribution. In such cases, we use Table B.1 in Appendix B to establish a critical region of z-scores. Then, we calculate a z-score for our data and compare it with a critical region of z-scores. For example, if we use a two-tailed test with $\alpha = 0.05$, we do not reject the null hypothesis if the z-score is between -1.96 and $+1.96$. In other words, we do not reject if the null hypothesis if $-1.96 \leq z \leq 1.96$.

1.3.7 Interpret the Results

We can now give meaning to the numbers and values from our analysis based on our context. If sample differences were observed, we can comment on the strength of those differences. We can compare the observed results with the expected results. We might examine a relationship between two variables for its relative strength or search a series of events for patterns.

1.3.8 Reporting the Results

Communicating results in a meaningful and comprehensible manner makes our research useful to others. There is a fair amount of agreement in the research literature for reporting statistical results from parametric tests. Unfortunately, there is less agreement for nonparametric tests. We have attempted to use the more common reporting techniques found in the research literature.

1.4 RANKING DATA

Many of the nonparametric procedures involve ranking data values. Ranking values is really quite simple. Suppose that you are a math teacher and wanted to find out if students score higher after eating a healthy breakfast. You give a test and compare the scores of four students who ate a healthy breakfast with four students who did not. Table 1.3 shows the results.

TABLE 1.3

Students who ate breakfast	Students who skipped breakfast
87	93
96	83
92	79
84	73

To rank all of the values from Table 1.3 together, place them all in order in a new table from smallest to largest (see Table 1.4). The first value receives a rank of 1, the second value receives a rank of 2, and so on.

TABLE 1.4

Value	Rank
73	1
79	2
83	3
84	4
87	5
92	6
93	7
96	8

Notice that the values for the students who ate breakfast are in bold type. On the surface, it would appear that they scored higher. However, if you are seeking statistical significance, you need some type of procedure. The following chapters will offer those procedures.

1.5 RANKING DATA WITH TIED VALUES

The aforementioned ranking method should seem straightforward. In many cases, however, two or more of the data values may be repeated. We call repeated values *ties*, or tied values. Say, for instance, that you repeat the preceding ranking with a different group of students. This time, you collected new values shown in Table 1.5.

TABLE 1.5

Students who ate breakfast	Students who skipped breakfast
90	75
85	80
95	55
70	**90**

Rank the values as in the previous example. Notice that the value of 90 is repeated. This means that the value of 90 is a tie. If these two student scores were different, they would be ranked 6 and 7. In the case of a tie, give all of the tied values the average of their rank values. In this example, the average of 6 and 7 is 6.5 (see Table 1.6).

TABLE 1.6

Value	Rank ignoring tied values	Rank accounting for tied values
55	1	1
70	2	2
75	3	3
80	4	4
85	5	5
90	**6**	**6.5**
90	**7**	**6.5**
95	8	8

Most nonparametric statistical tests require a different formula when a sample of data contains ties. It is important to note that the formulas for ties are more algebraically complex. What is more, formulas for ties typically produce a test statistic that is only slightly different from the test statistic formulas for data without ties. It is probably for this reason that most statistics texts omit the formulas for tied values. As you will see, however, we include the formulas for ties along with examples where applicable.

When the statistical tests in this book are explained using the computer program SPSS® (Statistical Package for Social Scientists), there is no mention of any special treatment for ties. That is because SPSS automatically detects the presence of ties in any data sets and applies the appropriate procedure for calculating the test statistic.

1.6 COUNTS OF OBSERVATIONS

Some nonparametric tests require *counts* (or frequencies) of observations. Determining the count is fairly straightforward and simply involves counting the total number of times a particular observations is made. For example, suppose you ask several children to pick their favorite ice cream flavor given three choices: vanilla, chocolate, and strawberry. Their preferences are shown in Table 1.7.

To find the counts for each ice cream flavor, list the choices and tally the total number of children who picked each flavor. In other words, count the number of children who picked chocolate. Then, repeat for the other choices, vanilla and strawberry. Table 1.8 reveals the counts from Table 1.7.

TABLE 1.7

Participant	Flavor
1	Chocolate
2	Chocolate
3	Vanilla
4	Vanilla
5	Strawberry
6	Chocolate
7	Chocolate
8	Vanilla

TABLE 1.8

Flavor	Count
Chocolate	4
Vanilla	3
Strawberry	1

To check your accuracy, you can add all the counts and compare them with the number of participants. The two numbers should be the same.

1.7 SUMMARY

In this chapter, we described differences between parametric and nonparametric tests. We also addressed assumptions by which nonparametric tests would be favorable over parametric tests. Then, we presented an overview of the nonparametric procedures included in this book. We also described the step-by-step approach we use to explain each test. Finally, we included explanations and examples of ranking and counting data, which are two tools for managing data when performing particular nonparametric tests.

The chapters that follow will present step-by-step directions for performing these statistical procedures both by manual, computational methods and by computer analysis using SPSS. In the next chapter, we address procedures for comparing data samples with a normal distribution.

1.8 PRACTICE QUESTIONS

1. Male high school students completed the 1-mile run at the end of their 9th grade and the beginning of their 10th grade. The following values represent the differences between the recorded times. Notice that only one student's time improved ($-2{:}08$). Rank the values in Table 1.9 beginning with the student's time difference that displayed improvement.

TABLE 1.9

Participant	Value	Rank
1	0:36	
2	0:28	
3	1:41	
4	0:37	
5	1:01	
6	2:30	
7	0:44	
8	0:47	
9	0:13	
10	0:24	
11	0:51	
12	0:09	
13	−2:08	
14	0:12	
15	0:56	

2. The values in Table 1.10 represent weekly quiz scores on math. Rank the quiz scores.

TABLE 1.10

Participant	Score	Rank
1	100	
2	60	
3	70	
4	90	
5	80	
6	100	
7	80	
8	20	
9	100	
10	50	

3. Using the data from the previous example, what are the counts (or frequencies) of passing scores and failing scores if a 70 is a passing score?

1.9 SOLUTIONS TO PRACTICE QUESTIONS

1. The value ranks are listed in Table 1.11. Notice that there are no ties.

2. The value ranks are listed in Table 1.12. Notice the tied values. The value of 80 occurred twice and required averaging the rank values of 5 and 6.

TABLE 1.11

Participant	Value	Rank
1	0:36	7
2	0:28	6
3	1:41	14
4	0:37	8
5	1:01	13
6	2:30	15
7	0:44	9
8	0:47	10
9	0:13	4
10	0:24	5
11	0:51	11
12	0:09	2
13	−2:08	1
14	0:12	3
15	0:56	12

TABLE 1.12

Participant	Score	Rank
1	**100**	**9**
2	60	3
3	70	4
4	90	7
5	**80**	**5.5**
6	**100**	**9**
7	**80**	**5.5**
8	20	1
9	**100**	**9**
10	50	2

$$(5+6)\div 2=5.5$$

The value of 100 occurred three times and required averaging the rank values of 8, 9, and 10.

$$(8+9+10)\div 3=9$$

3. Table 1.13 shows the passing scores and failing scores using 70 as a passing score. The counts (or frequencies) of passing scores is $n_{\text{passing}}=7$. The counts of failing scores is $n_{\text{failing}}=3$.

TABLE 1.13

Participant	Score	Pass/Fail
1	100	Pass
2	60	Fail
3	70	Pass
4	90	Pass
5	80	Pass
6	100	Pass
7	80	Pass
8	20	Fail
9	100	Pass
10	50	Fail

CHAPTER 2

TESTING DATA FOR NORMALITY

2.1 OBJECTIVES

In this chapter, you will learn the following items:

- How to find a data sample's kurtosis and skewness and determine if the sample meets acceptable levels of normality.
- How to use SPSS® to find a data sample's kurtosis and skewness and determine if the sample meets acceptable levels of normality.
- How to perform a Kolmogorov–Smirnov one-sample test to determine if a data sample meets acceptable levels of normality.
- How to use SPSS to perform a Kolmogorov–Smirnov one-sample test to determine if a data sample meets acceptable levels of normality.

2.2 INTRODUCTION

Parametric statistical tests, such as the *t*-test and one-way analysis of variance, are based on particular assumptions or parameters. The data samples meeting those parameters are randomly drawn from a normal population, based on independent observations, measured with an interval or ratio scale, possess an adequate sample size (see Chapter 1), and approximately resemble a normal distribution. Moreover, comparisons of samples or variables should have approximately equal variances. If data samples violate one or more of these assumptions, you should consider using a nonparametric test.

Examining the data gathering method, scale type, and size of a sample are fairly straightforward. However, examining a data sample's resemblance to a normal distribution, or its normality, requires a more involved analysis. Visually inspecting a graphical representation of a sample, such as a stem and leaf plot or a box and whisker plot, might be the most simplistic examination of normality. Statisticians advocate this technique in beginning statistics; however, this measure of normality does not suffice for strict levels of defensible analyses.

Nonparametric Statistics: A Step-by-Step Approach, Second Edition. Gregory W. Corder and Dale I. Foreman.

In this chapter, we present three quantitative measures of sample normality. First, we discuss the properties of the normal distribution. Then, we describe how to examine a sample's kurtosis and skewness. Next, we describe how to perform and interpret a Kolmogorov–Smirnov one-sample test. In addition, we describe how to perform each of these procedures using SPSS.

2.3 DESCRIBING DATA AND THE NORMAL DISTRIBUTION

An entire chapter could easily be devoted to the description of data and the normal distribution and many books do so. However, we will attempt to summarize the concept and begin with a practical approach as it applies to data collection.

In research, we often identify some population we wish to study. Then, we strive to collect several independent, random measurements of a particular variable associated with our population. We call this set of measurements a *sample*. If we used good experimental technique and our sample adequately represents our population, we can study the sample to make inferences about our population. For example, during a routine checkup, your physician draws a sample of your blood instead of all of your blood. This blood sample allows your physician to evaluate all of your blood even though he or she only tested the sample. Therefore, all of your body's blood cells represent the population about which your physician makes an inference using only the sample.

While a blood sample leads to the collection of a very large number of blood cells, other fields of study are limited to small sample sizes. It is not uncommon to collect less than 30 measurements for some studies in the behavioral and social sciences. Moreover, the measurements lie on some scale over which the measurements vary about the mean value. This notion is called *variance*. For example, a researcher uses some instrument to measure the intelligence of 25 children in a math class. It is highly unlikely that every child will have the same intelligence level. In fact, a good instrument for measuring intelligence should be sensitive enough to measure differences in the levels of the children.

The variance s^2 can be expressed quantitatively. It can be calculated using Formula 2.1:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \tag{2.1}$$

where x_i is an individual value in the distribution, $\bar{x}$ is the distribution's mean, and n is the number of values in the distribution

As mentioned in Chapter 1, parametric tests assume that the variances of samples being compared are approximately the same. This idea is called homogeneity of variance. To compare sample variances, Field (2005) suggested that we obtain a variance ratio by taking the largest sample variance and dividing it by the smallest sample variance. The variance ratio should be less than 2. Similarly, Pett (1997) indicated that no sample's variance be twice as large as any other sample's variance. If the homogeneity of variance assumption cannot be met, one would use a nonparametric test.

A more common way of expressing a sample's variability is with its standard deviation, s. Standard deviation is the square root of variance where $s = \sqrt{s^2}$. In other words, standard deviation is calculated using Formula 2.2:

$$s = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n-1}} \tag{2.2}$$

As illustrated in Figure 2.1, a small standard deviation indicates that a sample's values are fairly concentrated about its mean, whereas a large standard deviation indicates that a sample's values are fairly spread out.

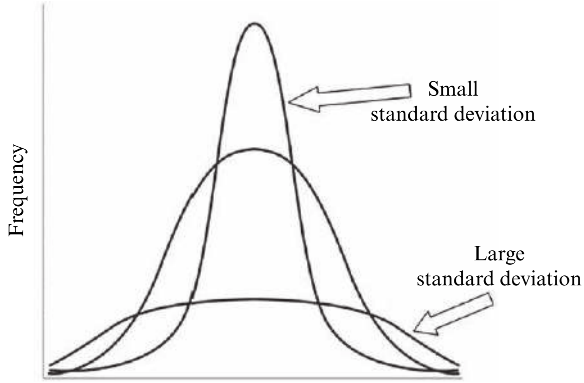


FIGURE 2.1

A histogram is a useful tool for graphically illustrating a sample's frequency distribution and variability (see Fig. 2.2). This graph plots the value of the measurements horizontally and the frequency of each particular value vertically. The middle value is called the median and the greatest frequency is called the mode.

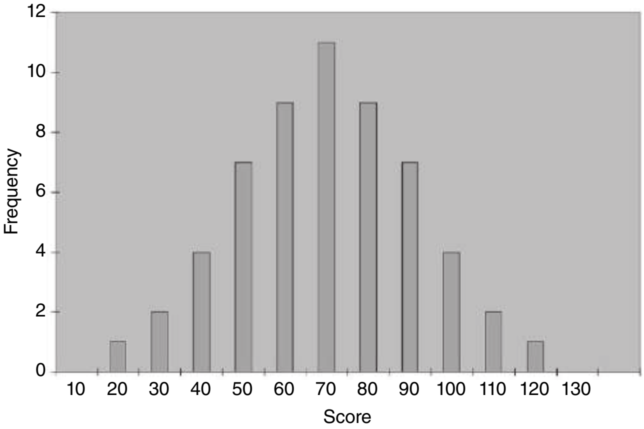


FIGURE 2.2

The mean and standard deviation of one distribution differ from the next. If we want to compare two or more samples, then we need some type of standard. A standard score is a way we can compare multiple distributions. The standard score that we use is called a z-score, and it can be calculated using Formula 2.3:

$$z=\frac{x_i-\overline{x}}{s} \tag{2.3}$$

where x_i is an individual value in the distribution, $\overline{x}$ is the distribution's mean, and s is the distribution's standard deviation.

There is a useful relationship between the standard deviation and z-score. We can think of the standard deviation as a unit of horizontal distance away from the mean on the histogram. One standard deviation from the mean is the same as $z = 1.0$. Two standard deviations from the mean are the same as $z = 2.0$. For example, if $s = 10$ and $\overline{x} = 70$ for a distribution, then $z = 1.0$ at $x = 80$ and $z = 2.0$ at $x = 90$. What is more, z-scores that lie below the mean have negative values. Using our example, $z = -1.0$ at $x = 60$ and $z = -2.0$ at $x = 50$. Moreover, $z = 0.0$ at the mean value, $x = 70$. These z-scores can be used to compare our distribution with another distribution, even if the mean and standard deviation are different. In other words, we can compare multiple distributions in terms of z-scores.

To this point, we have been focused on distributions with finite numbers of values, n. As more data values are collected for a given distribution, the histogram begins to resemble a bell shape called the normal curve. Figure 2.3 shows the relationship among the raw values, standard deviation, and z-scores of a population. Since we are describing a population, we use sigma, σ, to represent standard deviation and mu, μ, to represent the mean.

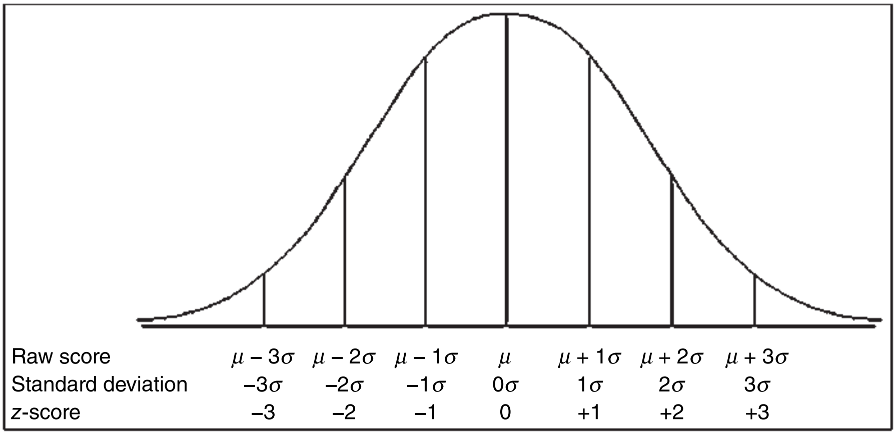


FIGURE 2.3

The normal curve has three particular properties (see Fig. 2.4). First, the mean, median, and mode are equal. Thus, most of the values lie in the center of the distribution. Second, the curve displays perfect symmetry about the mean. Third, the left and right sides of the curve, called the tails, are asymptotic. This means that they approach the horizontal axis, but never touch it.

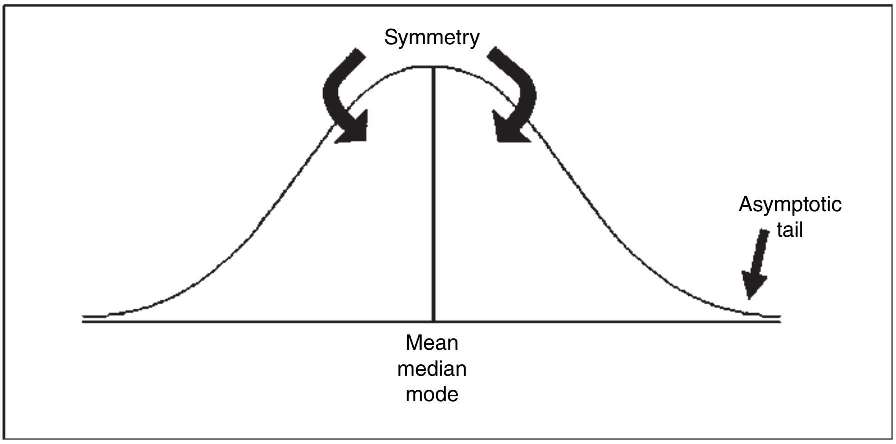


FIGURE 2.4

When we use a normal curve to represent probabilities p, we refer to it as the normal distribution. We set the area under the curve equal to $p = 1.0$. Since the distribution is symmetrical about the mean, $p = 0.50$ on the left side of the mean and $p = 0.50$ on the right. In addition, the ordinate of the normal curve, y, is the height of the curve at a particular point. The ordinate is tallest at the curve's center and decreases as you move away from the center. Table B.1 in Appendix B provides the z-scores, probabilities, and ordinates for the normal distribution.

2.4 COMPUTING AND TESTING KURTOSIS AND SKEWNESS FOR SAMPLE NORMALITY

A frequency distribution that resembles a normal curve is approximately normal. However, not all frequency distributions have the approximate shape of a normal curve. The values might be densely concentrated in the center or substantially spread out. The shape of the curve may lack symmetry with many values concentrated on one side of the distribution. We use the terms kurtosis and skewness to describe these conditions, respectively.

Kurtosis is a measure of a sample or population that identifies how flat or peaked it is with respect to a normal distribution. Stated another way, kurtosis refers to how concentrated the values are in the center of the distribution. As shown in Figure 2.5, a peaked distribution is said to be leptokurtic. A leptokurtic distribution has a positive kurtosis. If a distribution is flat, it is said to be platykurtic. A platykurtic distribution has a negative kurtosis.

The skewness of a sample can be described as a measure of horizontal symmetry with respect to a normal distribution. As shown in Figure 2.6, if a distribution's scores are concentrated on the right side of the curve, it is said to be left skewed. A left skewed distribution has a negative skewness. If a distribution's scores are concentrated on the left side of the curve, it is said to be right skewed. A right skewed distribution has a positive skewness.

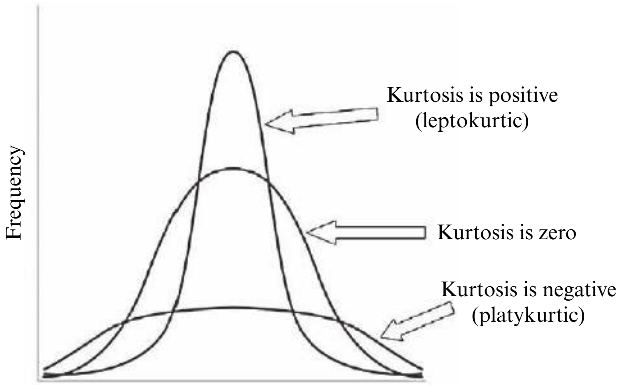


FIGURE 2.5

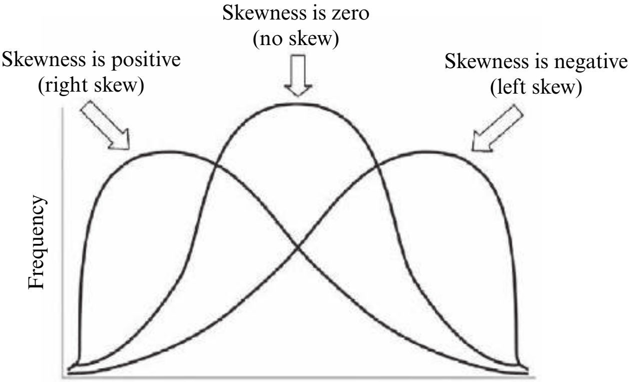


FIGURE 2.6

The kurtosis and skewness can be used to determine if a sample approximately resembles a normal distribution. There are five steps for examining sample normality in terms of kurtosis and skewness.

1. Determine the sample's mean and standard deviation.
2. Determine the sample's kurtosis and skewness.
3. Calculate the standard error of the kurtosis and the standard error of the skewness.
4. Calculate the z-score for the kurtosis and the z-score for the skewness.
5. Compare the z-scores with the critical region obtained from the normal distribution.

The calculations to find the values for a distribution's kurtosis and skewness require you to first find the sample mean $\bar{x}$ and the sample standard deviation s. Recall that standard deviation is found using Formula 2.2. The mean is found using Formula 2.4:

$$\overline{x} = \frac{\sum x_i}{n} \tag{2.4}$$

where $\sum x_i$ is the sum of the values in the sample and n is the number of values in the sample.

The kurtosis K and standard error of the kurtosis, SE_K, are found using Formula 2.5 and Formula 2.6:

$$K = \left[\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum\left(\frac{x_i-\overline{x}}{s}\right)^4\right] - \frac{3(n-1)^2}{(n-2)(n-3)} \tag{2.5}$$

and

$$SE_K = \sqrt{\frac{24n(n-1)^2}{(n-2)(n-3)(n+5)(n+3)}} \tag{2.6}$$

The skewness S_k and standard error of the skewness, SE_{S_k}, are found using Formula 2.7 and Formula 2.8:

$$S_k = \frac{n}{(n-1)(n-2)}\sum\left(\frac{x_i-\overline{x}}{s}\right)^3 \tag{2.7}$$

$$SE_{S_k} = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}} \tag{2.8}$$

Normality can be evaluated using the z-score for the kurtosis, z_K, and the z-score for the skewness, z_{S_k}. Use Formula 2.9 and Formula 2.10 to find those z-scores:

$$z_K = \frac{K-0}{SE_K} \tag{2.9}$$

$$z_{S_k} = \frac{S_k-0}{SE_{S_k}} \tag{2.10}$$

Compare these z-scores with the values of the normal distribution (see Table B.1 in Appendix B) for a desired level of confidence α. For example, if you set $\alpha = 0.05$, then the calculated z-scores for an approximately normal distribution must fall between -1.96 and $+1.96$.

2.4.1 Sample Problem for Examining Kurtosis

The scores in Table 2.1 represent students' quiz performance during the first week of class. Use $\alpha = 0.05$ for your desired level of confidence. Determine if the samples of week 1 quiz scores are approximately normal in terms of its kurtosis.

TABLE 2.1

Week 1 quiz scores		
90	72	90
64	95	89
74	88	100
77	57	35
100	64	95
65	80	84
90	100	76

First, find the mean of the sample:

$$\overline{x} = \frac{\sum x_i}{n} = \frac{1706}{21}$$
$$\overline{x} = 80.24$$

Next, find the standard deviation. It is helpful to set up Table 2.2 to manage the summation when computing the standard deviation (see Formula 2.2):

$$s = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n-1}} = \sqrt{\frac{5525.81}{21-1}} = \sqrt{276.29}$$
$$s = 16.62$$

Use the values for the mean and standard deviation to find the kurtosis. Again, it is helpful to set up Table 2.3 to manage the summation when computing the kurtosis (see Formula 2.5).

TABLE 2.2

x_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
90	9.76	95.29
72	−8.24	67.87
90	9.76	95.29
64	−16.24	263.68
95	14.76	217.91
89	8.76	76.77
74	−6.24	38.91
88	7.76	60.25
100	19.76	390.53
77	−3.24	10.49
57	−23.24	540.01

TABLE 2.2 (*Continued*)

x_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
35	−45.24	2046.49
100	19.76	390.53
64	−16.24	263.68
95	14.76	217.91
65	−15.24	232.20
80	−0.24	0.06
84	3.76	14.15
90	9.76	95.29
100	19.76	390.53
76	−4.24	17.96
		$\sum(x_i - \overline{x})^2 = 5525.81$

TABLE 2.3

x_i	$\dfrac{x_i - \overline{x}}{s}$	$\left(\dfrac{x_i - \overline{x}}{s}\right)^4$
90	0.587	0.119
72	−0.496	0.060
90	0.587	0.119
64	−0.977	0.911
95	0.888	0.622
89	0.527	0.077
74	−0.375	0.020
88	0.467	0.048
100	1.189	1.998
77	−0.195	0.001
57	−1.398	3.820
35	−2.722	54.864
100	1.189	1.998
64	−0.977	0.911
95	0.888	0.622
65	−0.917	0.706
80	−0.014	0.000
84	0.226	0.003
90	0.587	0.119
100	1.189	1.998
76	−0.255	0.004
		$\sum\left(\dfrac{x_i - \overline{x}}{s}\right)^4 = 69.020$

Compute the kurtosis:

$$
\begin{aligned}
K &= \left[\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum\left(\frac{x_i-\bar{x}}{s}\right)^4\right]-\frac{3(n-1)^2}{(n-2)(n-3)}\\
&= \left[\frac{21(21+1)}{(21-1)(21-2)(21-3)}(69.020)\right]-\frac{3(21-1)^2}{(21-2)(21-3)}\\
&= \left[\frac{21(22)}{(20)(19)(18)}(69.020)\right]-\frac{3(20)^2}{(19)(18)}\\
&= [0.0675(69.020)]-3.509=4.662-3.509\\
K &= 1.153
\end{aligned}
$$

Next, find the standard error of the kurtosis:

$$
\begin{aligned}
SE_K &= \sqrt{\frac{24n(n-1)^2}{(n-2)(n-3)(n+5)(n+3)}}\\
&= \sqrt{\frac{24(21)(21-1)^2}{(21-2)(21-3)(21+5)(21+3)}}\\
&= \sqrt{\frac{24(21)(20)^2}{(19)(18)(26)(24)}}=\sqrt{\frac{201{,}600}{213{,}408}}=\sqrt{0.945}\\
SE_K &= 0.972
\end{aligned}
$$

Finally, use the kurtosis and the standard error of the kurtosis to find a z-score:

$$
\begin{aligned}
z_K &= \frac{K-0}{SE_K}=\frac{1.153-0}{0.972}\\
z_K &= 1.186
\end{aligned}
$$

Use the z-score to examine the sample's approximation to a normal distribution. This value must fall between -1.96 and $+1.96$ to pass the normality assumption for $\alpha = 0.05$. Since this z-score value does fall within that range, the sample has passed our normality assumption for kurtosis. Next, the sample's skewness must be checked for normality.

2.4.2 Sample Problem for Examining Skewness

Based on the same values from the example listed earlier, determine if the samples of week 1 quiz scores are approximately normal in terms of its skewness.

Use the mean and standard deviation from the previous example to find the skewness. Set up Table 2.4 to manage the summation in the skewness formula.

Compute the skewness:

TABLE 2.4

x_i	$\frac{x_i - \overline{x}}{s}$	$\left(\frac{x_i - \overline{x}}{s}\right)^3$
90	0.587	0.203
72	−0.496	−0.122
90	0.587	0.203
64	−0.977	−0.932
95	0.888	0.700
89	0.527	0.146
74	−0.375	−0.053
88	0.467	0.102
100	1.189	1.680
77	−0.195	−0.007
57	−1.398	−2.732
35	−2.722	−20.159
100	1.189	1.680
64	−0.977	−0.932
95	0.888	0.700
65	−0.917	−0.770
80	−0.014	0.000
84	0.226	0.012
90	0.587	0.203
100	1.189	1.680
76	−0.255	−0.017
		$\sum\left(\frac{x_i - \overline{x}}{s}\right)^3 = -18.415$

$$S_k = \frac{n}{(n-1)(n-2)}\sum\left(\frac{x_i - \overline{x}}{s}\right)^3 = \frac{21}{(21-1)(21-2)}(-18.415)$$

$$= \frac{21}{(20)(19)}(-18.415)$$

$$S_k = -1.018$$

Next, find the standard error of the skewness:

$$SE_{S_k} = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}} = \sqrt{\frac{6(21)(21-1)}{(21-2)(21+1)(21+3)}}$$

$$= \sqrt{\frac{6(21)(20)}{(19)(22)(24)}} = \sqrt{\frac{2520}{10{,}032}} = \sqrt{0.251}$$

$$SE_{S_k} = 0.501$$

Finally, use the skewness and the standard error of the skewness to find a *z*-score:

$$z_{S_k} = \frac{S_k - 0}{SE_{S_k}} = \frac{-1.018}{0.501}$$

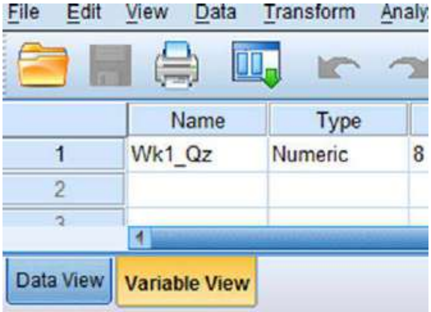
$$z_{S_k} = -2.032$$

Use the *z*-score to examine the sample's approximation to a normal distribution. This value must fall between −1.96 and +1.96 to pass the normality assumption for $\alpha = 0.05$. Since this *z*-score value does not fall within that range, the sample has failed our normality assumption for skewness. Therefore, either the sample must be modified and rechecked or you must use a nonparametric statistical test.

2.4.3 Examining Skewness and Kurtosis for Normality Using SPSS

We will analyze the examples earlier using SPSS.

2.4.3.1 Define Your Variables First, click the "Variable View" tab at the bottom of your screen. Then, type the name of your variable(s) in the "Name" column. As shown in Figure 2.7, we have named our variable "Wk1_Qz."



	Name	Type	
1	Wk1_Qz	Numeric	8
2			

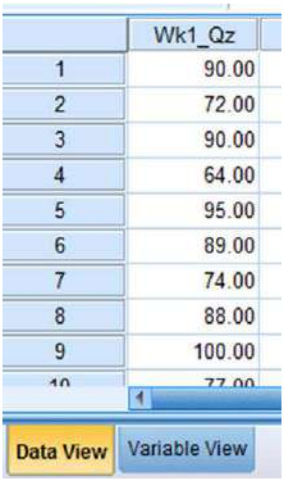
FIGURE 2.7

2.4.3.2 Type in Your Values Click the "Data View" tab at the bottom of your screen and type your data under the variable names. As shown in Figure 2.8, we have typed the values for the "Wk1_Qz" sample.

2.4.3.3 Analyze Your Data As shown in Figure 2.9, use the pull-down menus to choose "Analyze," "Descriptive Statistics," and "Descriptives . . ."

Choose the variable(s) that you want to examine. Then, click the button in the middle to move the variable to the "Variable(s)" box, as shown in Figure 2.10. Next, click the "Options . . ." button to open the "Descriptives: Options" window shown in Figure 2.11. In the "Distribution" section, check the boxes next to "Kurtosis" and "Skewness." Then, click "Continue."

Finally, once you have returned to the "Descriptives" window, as shown in Figure 2.12, click "OK" to perform the analysis.



	Wk1_Qz
1	90.00
2	72.00
3	90.00
4	64.00
5	95.00
6	89.00
7	74.00
8	88.00
9	100.00
10	77.00

FIGURE 2.8

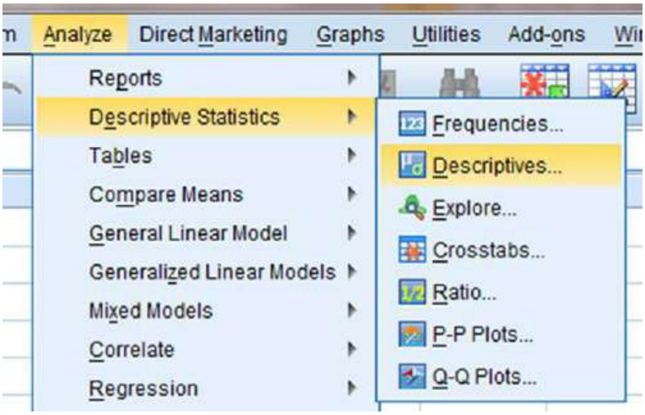
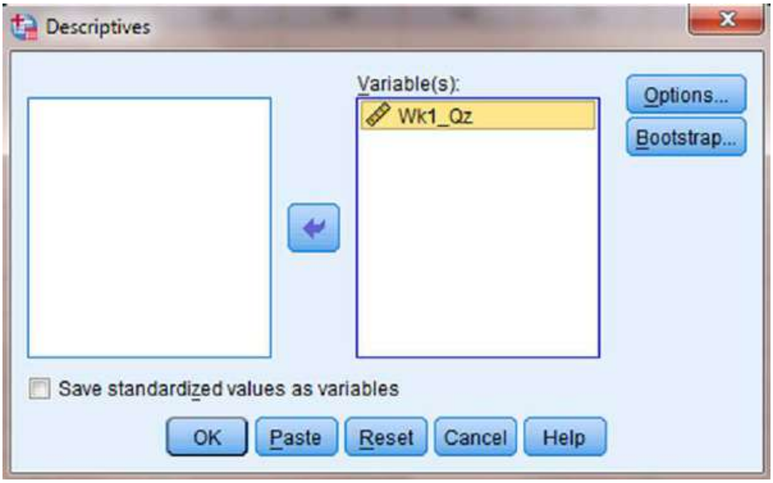


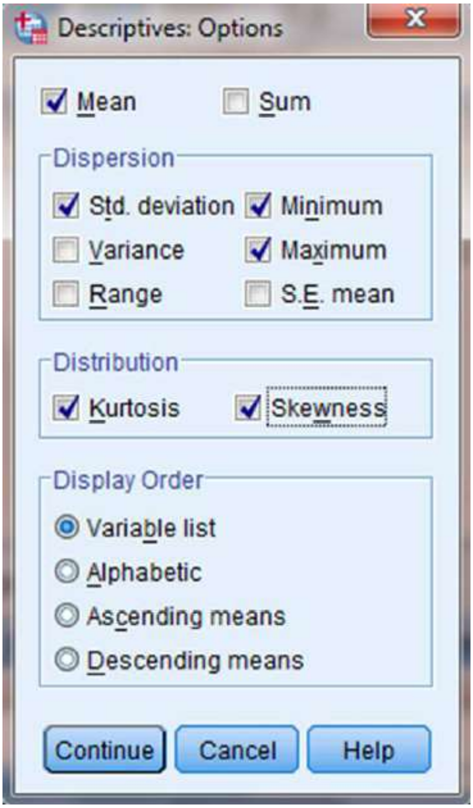
FIGURE 2.9



Variable(s): Wk1_Qz

☐ Save standardized values as variables

FIGURE 2.10



☑ Mean □ Sum

Dispersion

☑ Std. deviation ☑ Minimum
□ Variance ☑ Maximum
□ Range □ S.E. mean

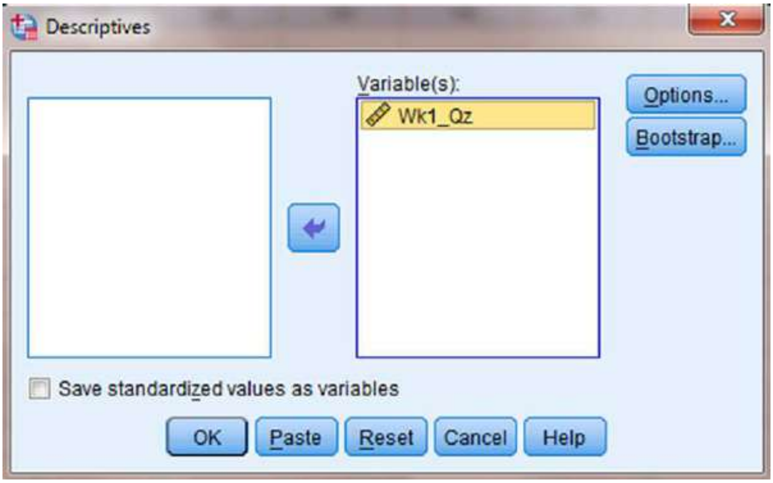
Distribution

☑ Kurtosis ☑ Skewness

Display Order

● Variable list
○ Alphabetic
○ Ascending means
○ Descending means

FIGURE 2.11

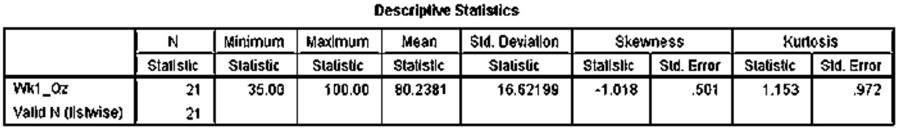


Variable(s):
Wk1_Qz

□ Save standardized values as variables

FIGURE 2.12

2.4.3.4 Interpret the Results from the SPSS Output Window The SPSS Output 2.1 provides the kurtosis and the skewness, along with their associated standard errors. In our example, the skewness is −1.018 and its standard error is 0.501. The kurtosis is 1.153 and its standard error is 0.972.



Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Wk1_Qz	21	35.00	100.00	80.2381	16.62199	-1.018	.501	1.153	.972
Valid N (listwise)	21								

SPSS OUTPUT 2.1

At this stage, we need to manually compute the z-scores for the skewness and kurtosis as we did in the previous examples. First, compute the z-score for kurtosis:

$$z_K = \frac{K-0}{SE_K} = \frac{1.153-0}{0.972}$$
$$z_K = 1.186$$

Next, we compute the z-score for skewness:

$$z_{S_k} = \frac{S_k-0}{SE_{S_k}} = \frac{-1.018}{0.501}$$
$$z_{S_k} = -2.032$$

Both of these values must fall between −1.96 and +1.96 to pass the normality assumption for $\alpha = 0.05$. The z-score for kurtosis falls within the desired range, but the z-score for skewness does not. Using $\alpha = 0.05$, the sample has passed the normality assumption for kurtosis, yet failed the normality assumption for skewness. Therefore, either the sample must be modified and rechecked or you must use a nonparametric statistical test.

2.5 COMPUTING THE KOLMOGOROV–SMIRNOV ONE-SAMPLE TEST

The Kolmogorov–Smirnov one-sample test is a procedure to examine the agreement between two sets of values. For our purposes, the two sets of values compared are an observed frequency distribution based on a randomly collected sample and an empirical frequency distribution based on the sample's population. Furthermore, the observed sample is examined for normality when the empirical frequency distribution is based on a normal distribution.

The Kolmogorov–Smirnov one-sample test compares two cumulative frequency distributions. A cumulative frequency distribution is useful for finding the number of observations above or below a particular value in a data sample. It is calculated by taking a given frequency and adding all the preceding frequencies in the list. In other words, it is like making a running total of the frequencies in a distribution. Creating cumulative frequency distributions of the observed and empirical frequency distributions allow us to find the point at which these two distributions show the largest divergence. Then, the test uses the largest divergence to identify a two-tailed probability estimate p to determine if the samples are statistically similar or different.

To perform the Kolmogorov–Smirnov one-sample test, we begin by determining the relative empirical frequency distribution $\hat{f}_{x_i}$ based on the observed sample. This relative empirical frequency distribution will approximate a normal distribution since we are examining our observed values for sample normality. First, calculate the observed frequency distribution's midpoint M and standard deviation s. The midpoint and standard deviation are found using Formula 2.11 and Formula 2.12:

$$M = (x_{\text{max}} + x_{\text{min}}) \div 2 \tag{2.11}$$

where x_{max} is the largest value in the sample and x_{min} is the smallest value in the sample, and

$$s = \sqrt{\frac{\sum (f_i x_i^2) - \frac{\left(\sum f_i x_i\right)^2}{n}}{n-1}} \tag{2.12}$$

where x_i is a given value in the observed sample, f_i is the frequency of a given value in the observed sample, and n is the number of values in the observed sample.

Next, use the midpoint and standard deviation to calculate the z-scores (see Formula 2.13) for the sample values x_i,

$$z = \left| \frac{x_i - M}{s} \right| \tag{2.13}$$

Use those z-scores and Table B.1 in Appendix B to determine the probability associated with each sample value, $\hat{p}_{x_i}$. These p-values are the relative frequencies of the empirical frequency distribution $\hat{f}_r$.

Now, we find the relative values of the observed frequency distribution f_r. Use Formula 2.14:

$$f_r = \frac{f_i}{n} \tag{2.14}$$

where f_i is the frequency of a given value in the observed sample and n is the number of values in the observed sample.

Since the Kolmogorov–Smirnov test uses cumulative frequency distributions, both the relative empirical frequency distribution and relative observed frequency distribution must be converted into cumulative frequency distributions $\hat{F}_{x_i}$ and S_{x_i}, respectively. Use Formula 2.15 and Formula 2.16 to find the absolute value divergence $\tilde{D}$ and D between the cumulative frequency distributions:

$$\tilde{D} = \left| \hat{F}_{x_i} - S_{x_i} \right| \tag{2.15}$$

$$D = \left| \hat{F}_{x_i} - S_{x_{i-1}} \right| \tag{2.16}$$

Use the largest divergence with Formula 2.17 to calculate the Kolmogorov–Smirnov test statistic Z:

$$Z = \sqrt{n}\max\left(|D|, |\tilde{D}|\right) \tag{2.17}$$

Then, use the Kolmogorov–Smirnov test statistic Z and the Smirnov (1948) formula (see Formula 2.18, Formula 2.19, Formula 2.20, Formula 2.21, Formula 2.22, and Formula 2.23) to find the two-tailed probability estimate p:

$$\text{if } 0 \le Z < 0.27, \text{then } p = 1 \tag{2.18}$$

$$\text{if } 0.27 \le Z < 1, \text{then } p = 1 - \frac{2.506628}{Z}(Q + Q^9 + Q^{25}) \tag{2.19}$$

where

$$Q = e^{-1.233701Z^{-2}} \tag{2.20}$$

$$\text{if } 1 \le Z < 3.1, \text{then } p = 2(Q - Q^4 + Q^9 - Q^{16}) \tag{2.21}$$

where

$$Q = e^{-2Z^2} \tag{2.22}$$

$$\text{if } Z \ge 3.1, \text{then } p = 0 \tag{2.23}$$

A p-value that exceeds the level of risk associated with the null hypothesis indicates that the observed sample approximates the empirical sample. Since our empirical distributions approximated a normal distribution, we can state that our observed sample is sufficiently normal for parametric statistics. Conversely, a p-value that is smaller than the level of risk indicates an observed sample that is not sufficiently normal for parametric statistics. The nonparametric statistical tests in this book are useful if a sample lacks normality.

2.5.1 Sample Kolmogorov–Smirnov One-Sample Test

A department store has decided to evaluate customer satisfaction. As part of a pilot study, the store provides customers with a survey to rate employee friendliness. The survey uses a scale of 1–10 and its developer indicates that the scores should conform to a normal distribution. Use the Kolmogorov–Smirnov one-sample test to decide if the sample of customers surveyed responded with scores approximately matching a normal distribution. The survey results are shown in Table 2.5.

TABLE 2.5

Survey results			
7	3	3	6
4	4	4	5
5	5	8	9
5	5	5	7
6	8	6	2

2.5.1.1 State the Null and Research Hypotheses The null hypothesis states that the observed sample has an approximately normal distribution. The research hypothesis states that the observed sample does not approximately resemble a normal distribution.

The null hypothesis is

> H_O: There is no difference between the observed distribution of survey scores and a normally distributed empirical sample.

The research hypothesis is

> H_A: There is a difference between the observed distribution of survey scores and a normally distributed empirical sample.

2.5.1.2 Set the Level of Risk (or the Level of Significance) Associated with the Null Hypothesis The level of risk, also called an alpha (α), is frequently set at 0.05. We will use an $\alpha = 0.05$ in our example. In other words, there is a 95% chance that any observed statistical difference will be real and not due to chance.

2.5.1.3 Choose the Appropriate Test Statistic We are seeking to compare our observed sample against a normally distributed empirical sample. The Kolmogorov–Smirnov one-sample test will provide this comparison.

2.5.1.4 Compute the Test Statistic First, determine the midpoint and standard deviation for the observed sample. Table 2.6 helps to manage the summations for this process.

TABLE 2.6

Survey score	Score frequency		
x_i	f_i	$f_i x_i$	$f_i x_i^2$
1	0	0	0
2	1	2	4
3	2	6	18
4	3	12	48
5	6	30	150
6	3	18	108
7	2	14	98
8	2	16	128
9	1	9	81
10	0	0	0
	$n = 20$	$\sum f_i x_i = 107$	$\sum f_i x_i^2 = 635$

Use Formula 2.11 to find the midpoint:

$$\begin{aligned} M &= (x_{\max} + x_{\min}) \div 2 \\ &= (9+2) \div 2 \\ M &= 5.5 \end{aligned}$$

Then, use Formula 2.12 to find the standard deviation:

$$\begin{aligned} s &= \sqrt{\frac{\sum (f_i x_i^2) - \frac{\left(\sum f_i x_i\right)^2}{n}}{n-1}} \\ &= \sqrt{\frac{635 - \frac{107^2}{20}}{20-1}} \\ s &= 1.81 \end{aligned}$$

Now, determine the z-scores, empirical relative frequencies, and observed relative frequencies for each score value (see Table 2.7).

TABLE 2.7

Survey score	Score frequency			Empirical frequency	Observed frequency
x_i	f_i	z-score	$\hat{p}_{x_i}$	$\hat{f}_r$	f_r
1	0	2.49	0.0064	0.006	0.000
2	1	1.93	0.0266	0.020	0.050
3	2	1.38	0.0838	0.064	0.100
4	**3**	**0.83**	**0.2033**	**0.140**	**0.150**
5	6	0.28	0.3897	0.250	0.300
6	3	0.28	0.3897	0.250	0.150
7	2	0.83	0.2033	0.140	0.100
8	2	1.38	0.0838	0.064	0.100
9	1	1.93	0.0266	0.020	0.050
10	0	2.49	0.0064	0.006	0.000

We will provide a sample calculation for survey score = 4 as seen in Table 2.7. Use Formula 2.13 to calculate the z-scores:

$$\begin{aligned} z &= \left| \frac{x_i - M}{s} \right| \\ &= \left| \frac{4 - 5.5}{1.81} \right| \\ z &= 0.83 \end{aligned}$$

Use each z-score and Table B.1 in Appendix B to determine the probability associated with the each value, $\hat{p}_{x_i}$:

$$\hat{p}_4 = 0.2033$$

To find the empirical frequency value $\hat{f}_r$ for each value, subtract its preceding value, $\hat{f}_{r-1}$, from the associated probability value $\hat{p}_{x_i}$. In other words,

$$\hat{f}_r = \hat{p}_{x_i} - \hat{f}_{r-1}$$

We establish our empirical frequency distribution beginning at the tail, $x_i = 1$, and work to the midpoint, $x_i = 5$:

$$\begin{aligned}
\hat{f}_{r1} &= \hat{p}_1 - \hat{f}_{r0} = 0.0064 - 0.000 = 0.006 \\
\hat{f}_{r2} &= \hat{p}_2 - \hat{f}_{r1} = 0.0266 - 0.006 = 0.020 \\
\hat{f}_{r3} &= \hat{p}_3 - \hat{f}_{r2} = 0.0838 - 0.020 = 0.064 \\
\hat{f}_{r4} &= \hat{p}_4 - \hat{f}_{r3} = 0.2033 - 0.064 = 0.140 \\
\hat{f}_{r5} &= \hat{p}_5 - \hat{f}_{r4} = 0.3897 - 0.140 = 0.250
\end{aligned}$$

Our empirical frequency distribution is based on a normal distribution, which is symmetrical. Therefore, we can complete our empirical frequency distribution by basing the remaining values on a symmetrical distribution. Those values are in Table 2.7.

Now, we find the values of the observed frequency distribution f_r with Formula 2.14. We provide a sample calculation with survey result = 4. That survey value occurs three times:

$$\begin{aligned}
f_{r4} &= \frac{f_{x_i=4}}{n} = \frac{3}{20} \\
f_r &= 0.150
\end{aligned}$$

Next, we create cumulative frequency distributions using the empirical and observed frequency distributions. A cumulative frequency distribution is created by taking a frequency and adding all the preceding values. We demonstrate this in Table 2.8.

Now, we find the absolute value divergence $\tilde{D}$ and D between the cumulative frequency distributions. Use Formula 2.15 and Formula 2.16. See the sample calculation for survey score = 4 as seen in bold in Table 2.9.

$$\begin{aligned}
\tilde{D}_4 &= \left|\hat{F}_4 - S_4\right| = |0.230 - 0.300| \\
\tilde{D}_4 &= 0.070
\end{aligned}$$

and

$$\begin{aligned}
D_4 &= \left|\hat{F}_4 - S_3\right| = |0.230 - 0.150| \\
D_4 &= 0.080
\end{aligned}$$

TABLE 2.8

	Relative frequency		Cumulative frequency	
Survey score	Empirical	Observed	Empirical	Observed
x_i	$\hat{f}_r$	f_r	$\hat{F}_{x_i}$	S_{x_i}
1	0.006	0.000	0.006	0.000
2	0.020	0.050	0.020 + 0.006 = 0.026	0.050 + 0.000 = 0.050
3	0.064	0.100	0.064 + 0.026 = 0.090	0.100 + 0.050 = 0.150
4	0.140	0.150	0.140 + 0.090 = 0.230	0.150 + 0.150 = 0.300
5	0.250	0.300	0.250 + 0.230 = 0.480	0.300 + 0.300 = 0.600
6	0.250	0.150	0.250 + 0.480 = 0.730	0.150 + 0.600 = 0.750
7	0.140	0.100	0.140 + 0.730 = 0.870	0.100 + 0.750 = 0.850
8	0.064	0.100	0.064 + 0.870 = 0.934	0.100 + 0.850 = 0.950
9	0.020	0.050	0.020 + 0.934 = 0.954	0.050 + 0.950 = 1.000
10	0.006	0.000	0.006 + 0.954 = 0.960	0.000 + 1.000 = 1.000

TABLE 2.9

	Cumulative frequency		Cumulative frequency	
Survey score	Empirical	Observed	Divergence	
x_i	$\hat{F}_{x_i}$	S_{x_i}	$\tilde{D}$	D
1	0.006	0.000	0.006	
2	0.026	0.050	0.024	0.026
3	0.090	0.150	0.060	0.040
4	**0.230**	**0.300**	**0.070**	**0.080**
*5	0.480	0.600	0.120	*0.180
6	0.730	0.750	0.020	0.130
7	0.870	0.850	0.020	0.120
8	0.934	0.950	0.016	0.084
9	0.954	1.000	0.046	0.004
10	0.960	1.000	0.040	0.040

To find the test statistic Z, use the largest value from $\tilde{D}$ and D in Formula 2.17. Table 2.9 has an asterisk next to the largest divergence. That value is located at survey value = 5. It is $\max\left(|D|, |\tilde{D}|\right) = 0.180$:

$$\begin{aligned} Z &= \sqrt{n}\max\left(|D|, |\tilde{D}|\right) \\ &= \sqrt{20}(0.180) \\ Z &= 0.805 \end{aligned}$$

2.5.1.5 Determine the* p*-Value Associated with the Test Statistic The Kolmogorov–Smirnov test statistic Z and the Smirnov (1948) formula (see Formula 2.18, Formula 2.19, Formula 2.20, Formula 2.21, Formula 2.22, and Formula 2.23) are used to find the two-tailed probability estimate p. Since $0.27 \leq Z < 1$, we use Formula 2.19 and Formula 2.20:

$$\begin{aligned} Q &= e^{-1.233701Z^{-2}} \\ &= e^{-(1.233701)(0.805)^{-2}} \\ Q &= 0.149 \end{aligned}$$

and

$$\begin{aligned} p &= 1 - \frac{2.506628}{Z}(Q + Q^9 + Q^{25}) \\ &= 1 - \frac{2.506628}{0.805}(0.149 + 0.149^9 + 0.149^{25}) \\ p &= 0.536 \end{aligned}$$

2.5.1.6 Compare the* p*-Value with the Level of Risk (or the Level of Significance) Associated with the Null Hypothesis The critical value for rejecting the null hypothesis is $\alpha = 0.05$ and the obtained p-value is $p = 0.536$. If the critical value is greater than the obtained value, we must reject the null hypothesis. If the critical value is less than the obtained p-value, we must not reject the null hypothesis. Since the critical value is less than the obtained value ($0.05 < 0.536$), we do not reject the null hypothesis.

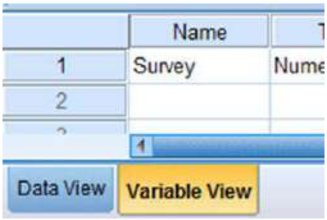
2.5.1.7 Interpret the Results We did not reject the null hypothesis, suggesting the customers' survey ratings of employee friendliness sufficiently resembled a normal distribution. This means that a parametric statistical procedure may be used with this sample.

2.5.1.8 Reporting the Results When reporting the results from the Kolmogorov–Smirnov one-sample test, we include the test statistic (D), the degrees of freedom (which equals the sample size), and the p-value in terms of the level of risk α. Based on our analysis, the sample of customers is approximately normal, where $D_{(20)} = 0.180$, $p > 0.05$.

2.5.2 Performing the Kolmogorov–Smirnov One-Sample Test Using SPSS

We will analyze the data from the example earlier using SPSS.

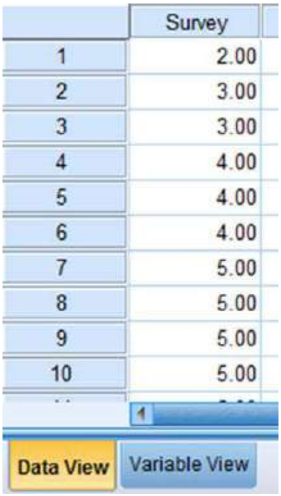
2.5.2.1 Define Your Variables First, click the "Variable View" tab at the bottom of your screen. Then, type the names of your variables in the "Name" column. As shown in Figure 2.13, the variable is called "Survey."



	Name	T
1	Survey	Nume
2		

FIGURE 2.13

2.5.2.2 Type in Your Values Click the “Data View” tab at the bottom of your screen. Type your sample values in the “Survey” column as shown in Figure 2.14.



	Survey
1	2.00
2	3.00
3	3.00
4	4.00
5	4.00
6	4.00
7	5.00
8	5.00
9	5.00
10	5.00

FIGURE 2.14

2.5.2.3 Analyze Your Data As shown in Figure 2.15, use the pull-down menus to choose “Analyze,” “Nonparametric Tests,” “Legacy Dialogs,” and “1-Sample K-S . . .”

Use the arrow button to place your variable with your data values in the box labeled “Test Variable List:” as shown in Figure 2.16. Finally, click “OK” to perform the analysis.

2.5.2.4 Interpret the Results from the SPSS Output Window SPSS Output 2.2 provides the most extreme difference ($D = 0.176$), Kolmogorov–Smirnov Z-test statistic ($Z = 0.789$), and the significance ($p = 0.562$). Based on the results from SPSS, the p-value exceeds the level of risk associated with the null hypothesis ($\alpha = 0.05$). Therefore, we do not reject the null hypothesis. In other words, the sample distribution is sufficiently normal.

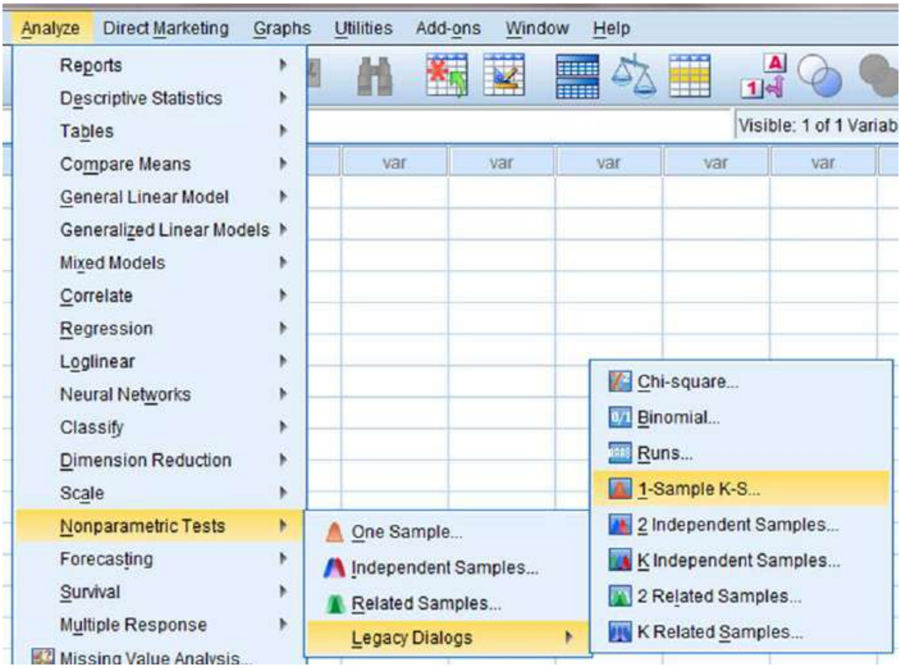
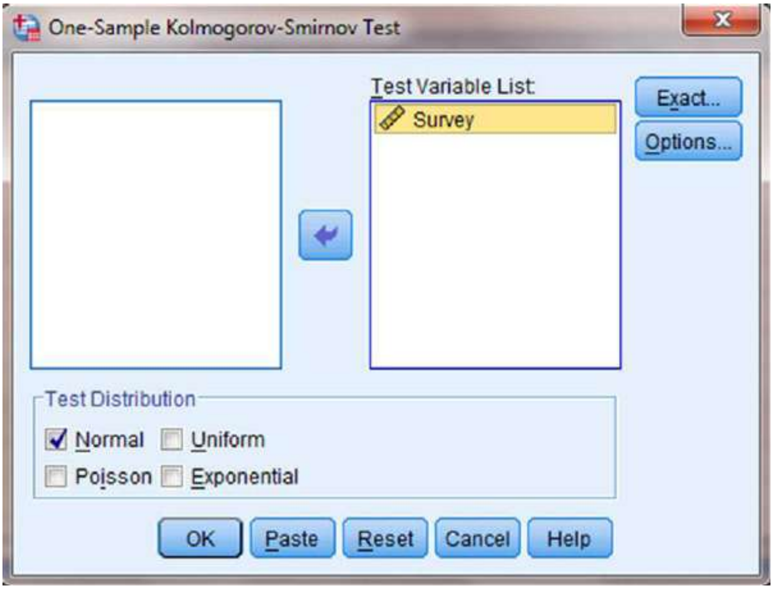


FIGURE 2.15


One-Sample Kolmogorov-Smirnov Test

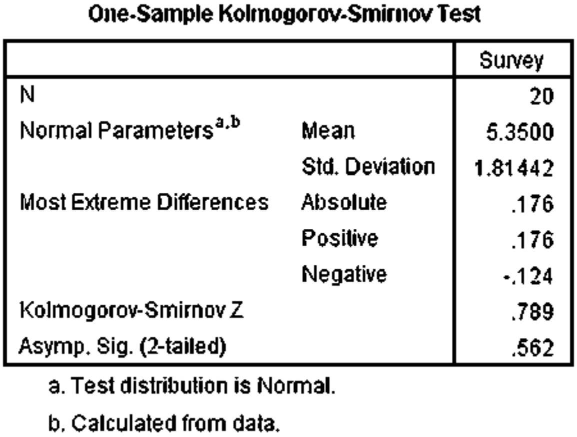
Test Variable List: Survey

Test Distribution

☑ Normal □ Uniform

□ Poisson □ Exponential

FIGURE 2.16



One-Sample Kolmogorov-Smirnov Test

		Survey
N		20
Normal Parameters[a,b]	Mean	5.3500
	Std. Deviation	1.81442
Most Extreme Differences	Absolute	.176
	Positive	.176
	Negative	-.124
Kolmogorov-Smirnov Z		.789
Asymp. Sig. (2-tailed)		.562

a. Test distribution is Normal.

b. Calculated from data.

SPSS OUTPUT 2.2

On an added note, differences between the values from the sample problem earlier and the SPSS output are likely due to value precision and computational round off errors.

2.6 SUMMARY

Parametric statistical tests, such as the *t*-test and one-way analysis of variance, are based on particular assumptions or parameters. Therefore, it is important that you examine collected data for its approximation to a normal distribution. Upon doing that, you can consider whether you will use a parametric or nonparametric test for analyzing your data.

In this chapter, we presented three quantitative measures of sample normality. First, we described how to examine a sample's kurtosis and skewness. Then, we described how to perform and interpret a Kolmogorov–Smirnov one-sample test. In the following chapters, we will describe several nonparametric procedures for analyzing data samples that do not meet the assumptions needed for parametric statistical tests. In the chapter that follows, we will begin by describing a test for comparing two unrelated samples.

2.7 PRACTICE QUESTIONS

1. The values in Table 2.10 are a sample of reading-level score for a 9th-grade class. They are measured on a ratio scale. Examine the sample's skewness and kurtosis for normality for $\alpha = 0.05$. Report your findings.

2. Using a Kolmogorov–Smirnov one-sample test, examine the sample of values from Table 2.10. Report your findings.

TABLE 2.10

Ninth-grade reading-level score									
8.10	8.20	8.20	8.70	8.70	8.80	8.80	8.90	8.90	8.90
9.20	9.20	9.20	9.30	9.30	9.30	9.40	9.40	9.40	9.40
9.50	9.50	9.50	9.50	9.60	9.60	9.60	9.70	9.70	9.90

2.8 SOLUTIONS TO PRACTICE QUESTIONS

1. SPSS returned the following values:

skewness = −0.904
standard error of the skewness = 0.427
kurtosis = 0.188
standard error of the kurtosis = 0.833

The computed *z*-scores are as follows:

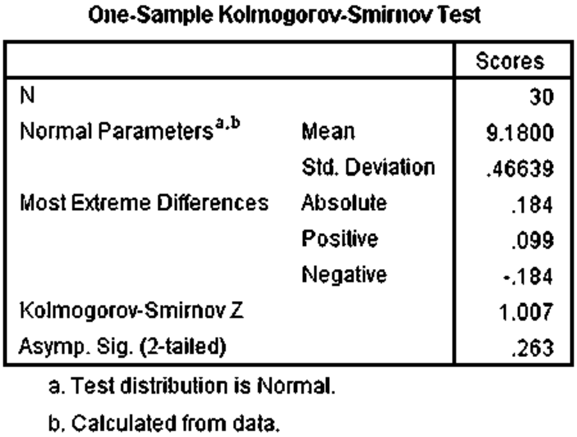
$$z_{S_k} = -2.117$$

and

$$z_K = 0.226$$

At $\alpha = 0.05$, the sample's skewness fails the normality test, while the kurtosis passes the normality test. Based on our standard of $\alpha = 0.05$, this sample of reading levels for 9th-grade students is not sufficiently normal.

2. SPSS Output 2.3 shows the results from the Kolmogorov–Smirnov one-sample test.
Kolmogorov–Smirnov obtained value = 1.007
Two-Tailed significance = 0.263



One-Sample Kolmogorov-Smirnov Test

		Scores
N		30
Normal Parameters[a,b]	Mean	9.1800
	Std. Deviation	.46639
Most Extreme Differences	Absolute	.184
	Positive	.099
	Negative	-.184
Kolmogorov-Smirnov Z		1.007
Asymp. Sig. (2-tailed)		.263

a. Test distribution is Normal.

b. Calculated from data.

SPSS OUTPUT 2.3

According to the Kolmogorov–Smirnov one-sample test with $\alpha = 0.05$, this sample of reading levels for 9th-grade students is sufficiently normal.

CHAPTER 3

COMPARING TWO RELATED SAMPLES: THE WILCOXON SIGNED RANK AND THE SIGN TEST

3.1 OBJECTIVES

In this chapter, you will learn the following items:

- How to compute the Wilcoxon signed rank test.
- How to perform the Wilcoxon signed rank test using SPSS®.
- How to construct a median confidence interval based on the Wilcoxon signed rank test for matched pairs.
- How to compute the sign test.
- How to perform the sign test using SPSS.

3.2 INTRODUCTION

Imagine that you give an attitude test to a small group of people. After you deliver some type of treatment, say, a daily vitamin C supplement for several weeks, you give that same group of people another attitude test. Finally, you compare the two measures of attitude to see if there is any type of difference between the two sets of scores.

The two sets of test scores in the previous scenario are related or paired. This is because each person was tested twice. In other words, each test score in one group of scores has another test score counterpart. The Wilcoxon signed rank test and the sign test are nonparametric statistical procedures for comparing two samples that are paired or related. The parametric equivalent to these tests goes by names such as the Student's *t*-test, *t*-test for matched pairs, *t*-test for paired samples, or *t*-test for dependent samples.

In this chapter, we will describe how to perform and interpret a Wilcoxon signed rank test and a sign test, using both small samples and large samples. In addition, we demonstrate the procedures for performing both tests using SPSS. Finally, we offer varied examples of these nonparametric statistics from the literature.

Nonparametric Statistics: A Step-by-Step Approach, Second Edition. Gregory W. Corder and Dale I. Foreman.

3.3 COMPUTING THE WILCOXON SIGNED RANK TEST STATISTIC

The formula for computing the Wilcoxon T for small samples is shown in Formula 3.1. The signed ranks are the values that are used to compute the positive and negative values in the formula:

$$T = \text{smaller of } \Sigma R_+ \text{ and } \Sigma R_- \tag{3.1}$$

where ΣR_+ is the sum of the ranks with positive differences and ΣR_- is the sum of the ranks with negative differences.

After the T statistic is computed, it must be examined for significance. We may use a table of critical values (see Table B.3 in Appendix B). However, if the numbers of pairs n exceeds those available from the table, then a large sample approximation may be performed. For large samples, compute a z-score and use a table with the normal distribution (see Table B.1 in Appendix B) to obtain a critical region of z-scores. Formula 3.2, Formula 3.3, and Formula 3.4 are used to find the z-score of a Wilcoxon signed rank test for large samples:

$$\overline{x}_T = \frac{n(n+1)}{4} \tag{3.2}$$

where $\overline{x}_T$ is the mean and n is the number of matched pairs included in the analysis,

$$s_T = \sqrt{\frac{n(n+1)(2n+1)}{24}} \tag{3.3}$$

where s_T is the standard deviation,

$$z^* = \frac{T - \overline{x}_T}{s_T} \tag{3.4}$$

where z^* is the z-score for an approximation of the data to the normal distribution and T is the T statistic.

At this point, the analysis is limited to identifying the presence or absence of a significant difference between the groups and does not describe the strength of the treatment. We can consider the effect size (ES) to determine the degree of association between the groups. We use Formula 3.5 to calculate the ES:

$$ES = \frac{|z|}{\sqrt{n}} \tag{3.5}$$

where $|z|$ is the absolute value of the z-score and n is the number of matched pairs included in the analysis.

The ES ranges from 0 to 1. Cohen (1988) defined the conventions for ES as *small* = 0.10, *medium* = 0.30, and *large* = 0.50. (Correlation coefficient and ES are both measures of association. See Chapter 7 concerning correlation for more information on Cohen's assignment of ES's relative strength.)

3.3.1 Sample Wilcoxon Signed Rank Test (Small Data Samples)

The counseling staff of Clear Creek County School District has implemented a new program this year to reduce bullying in their elementary schools. The school district does not know if the new program resulted in improvement or deterioration. In order to evaluate the program's effectiveness, the school district has decided to compare the percentage of successful interventions last year before the program began with the percentage of successful interventions this year with the program in place. In Table 3.1, the 12 elementary school counselors, or participants, reported the percentage of successful interventions last year and the percentage this year.

TABLE 3.1

	Percentage of successful interventions	
Participants	Last year	This year
1	31	31
2	14	14
3	53	50
4	18	30
5	21	28
6	44	48
7	12	35
8	36	32
9	22	23
10	29	34
11	17	27
12	40	42

The samples are relatively small, so we need a nonparametric procedure. Since we are comparing two related, or paired, samples, we will use the Wilcoxon signed rank test.

3.3.1.1 State the Null and Research Hypotheses The null hypothesis states that the counselors reported no difference in the percentages last year and this year. The research hypothesis states that the counselors observed some differences between this year and last year. Our research hypothesis is a two-tailed, nondirectional hypothesis because it indicates a difference, but in no particular direction.

The null hypothesis is

$$H_O\colon \mu_D = 0$$

The research hypothesis is

$$H_A\colon \mu_D \neq 0$$

3.3.1.2 Set the Level of Risk (or the Level of Significance) Associated with the Null Hypothesis The level of risk, also called an alpha (α), is frequently set at 0.05. We will use $\alpha = 0.05$ in our example. In other words, there is a 95% chance that any observed statistical difference will be real and not due to chance.

3.3.1.3 Choose the Appropriate Test Statistic The data are obtained from 12 counselors, or participants, who are using a new program designed to reduce bullying among students in the elementary schools. The participants reported the percentage of successful interventions last year and the percentage this year. We are comparing last year's percentages with this year's percentages. Therefore, the data samples are related or paired. In addition, sample sizes are relatively small. Since we are comparing two related samples, we will use the Wilcoxon signed rank test.

3.3.1.4 Compute the Test Statistic First, compute the difference between each sample pair. Then, rank the absolute value of those computed differences. Using this method, the differences of zero are ignored when ranking. We have done this in Table 3.2.

TABLE 3.2

	Percentage of successful interventions			Rank	
Participant	Last year	This year	Difference	Without zero	Sign
1	31	31	0	Exclude	
2	14	14	0	Exclude	
3	53	50	–3	3	–
4	18	30	+12	9	+
5	21	28	+7	7	+
6	44	48	+4	4.5	+
7	12	35	+23	10	+
8	36	32	–4	4.5	–
9	22	23	+1	1	+
10	29	34	+5	6	+
11	17	27	+10	8	+
12	40	42	+2	2	+

Compute the sum of ranks with positive differences. Using Table 3.2, the ranks with positive differences are 9, 7, 4.5, 10, 1, 6, 8, and 2. When we add all of the ranks with positive difference we get $\Sigma R_+ = 47.5$.

Compute the sum of ranks with negative differences. The ranks with negative differences are 3 and 4.5. The sum of ranks with negative difference is $\Sigma R_- = 7.5$.

The obtained value is the smaller of the two rank sums. Therefore, the Wilcoxon is $T = 7.5$.

3.3.1.5 Determine the Value Needed for Rejection of the Null Hypothesis Using the Appropriate Table of Critical Values for the Particular Statistic Since the sample sizes are small, we use Table B.3 in Appendix B, which lists the critical values for the Wilcoxon T. As noted earlier in Table 3.2, the two counselors with score differences of zero were discarded. This reduces our sample size to $n = 10$. In this case, we look for the critical value under the two-tailed test for $n = 10$ and $\alpha = 0.05$. Table B.3 returns a critical value for the Wilcoxon test of $T = 8$. An obtained value that is less than or equal to 8 will lead us to reject our null hypothesis.

3.3.1.6 Compare the Obtained Value with the Critical Value The critical value for rejecting the null hypothesis is 8 and the obtained value is $T = 7.5$. If the critical value equals or exceeds the obtained value, we must reject the null hypothesis. If instead, the critical value is less than the obtained value, we must not reject the null hypothesis. Since the critical value exceeds the obtained value, we must reject the null hypothesis.

3.3.1.7 Interpret the Results We rejected the null hypothesis, suggesting that a real difference exists between last year's percentages and this year's percentages. In addition, since the sum of the positive difference ranks (ΣR_+) was larger than the negative difference ranks (ΣR_-), the difference is positive, showing a positive impact of the program. Therefore, our analysis provides evidence that the new bullying program is providing positive benefits toward the improvement of student behavior as perceived by the school counselors.

3.3.1.8 Reporting the Results When reporting the findings, include the T statistic, sample size, and p-value's relation to α. The directionality of the difference should be expressed using the sum of the positive difference ranks (ΣR_+) and sum of the negative difference ranks (ΣR_-).

For this example, the Wilcoxon signed rank test ($T = 7.5$, $n = 12$, $p < 0.05$) indicated that the percentage of successful interventions was significantly different. In addition, the sum of the positive difference ranks ($\Sigma R_+ = 47.5$) was larger than the sum of the negative difference ranks ($\Sigma R_- = 7.5$), showing a positive impact from the program. Therefore, our analysis provides evidence that the new bullying program is providing positive benefits toward the improvement of student behavior as perceived by the school counselors.

3.3.2 Confidence Interval for the Wilcoxon Signed Rank Test

The American Psychological Association (2001) has suggested that researchers report the *confidence interval* for research data. A confidence interval is an inference to a population in terms of an estimation of sampling error. More specifically, it provides a range of values that fall within the population with a level of confidence of $100(1 - \alpha)\%$.

A median confidence interval can be constructed based on the Wilcoxon signed rank test for matched pairs. In order to create this confidence interval, all of the possible matched pairs (X_i,X_j) are used to compute the differences $D_i = X_i - X_j$. Then, compute all of the averages u_{ij} of two difference scores using Formula 3.6. There will be a total of $[n(n - 1)/2] + n$ averages.

$$u_{ij} = (D_i + D_j)/2 \quad 1 \le i \le j \le n \tag{3.6}$$

We will perform a 95% confidence interval using the sample Wilcoxon signed rank test with a small data sample (as stated earlier). Table 3.1 provides the values for obtaining our confidence interval. We begin by using Formula 3.6 to compute all of the averages u_{ij} of two difference scores. For example,

$$u_{11} = (D_1 + D_1)/2 = (-3 + -3)/2$$

$$u_{11} = -3$$

$$u_{12} = (D_1 + D_2)/2 = (-3 + 12)/2$$

$$u_{12} = 4.5$$

$$u_{13} = (D_1 + D_3)/2 = (-3 + 7)/2$$

$$u_{13} = 2$$

Table 3.3 shows each value of u_{ij}.

TABLE 3.3

	−3	12	7	4	23	−4	1	5	10	2
−3	−3	4.5	2	0.5	10	−3.5	−1	1	3.5	−0.5
12		12	9.5	8	17.5	4	6.5	8.5	11	7
7			7	5.5	15	1.5	4	6	8.5	4.5
4				4	13.5	0	2.5	4.5	7	3
23					23	9.5	12	14	16.5	12.5
−4						−4	−1.5	0.5	3	−1
1							1	3	5.5	1.5
5								5	7.5	3.5
10									10	6
2										2

Next, arrange all of the averages in order from smallest to largest. We have arranged all of the values for u_{ij} in Table 3.4.

The median of the ordered averages gives a point estimate of the population median difference. The median of this distribution is 4.5, which is the point estimate of the population.

Use Table B.3 in Appendix B to find the endpoints of the confidence interval. First, determine T from the table that corresponds with the sample size and desired

TABLE 3.4

1	–4.0	12	1.0	22	4.0	34	6.5	45	10.0
2	−3.5	13	1.5	23	4.0	35	7.0	46	11.0
3	−3.0	14	1.5	24	4.0	36	7.0	47	12.0
4	−1.5	15	2.0	25	4.5	37	7.0	48	12.0
5	−1.0	15	2.0	26	4.5	38	7.5	49	12.5
6	−1.0	16	2.5	27	4.5	39	8.0	50	13.5
7	−0.5	17	3.0	28	5.0	40	8.5	51	14.0
8	0.0	18	3.0	29	5.5	41	8.5	52	15.0
9	0.5	19	3.0	30	5.5	42	9.5	53	16.5
10	0.5	20	3.5	31	6.0	43	9.5	54	17.5
11	1.0	21	3.5	32	6.0	44	10.0	55	23.0

confidence such that $p = \alpha/2$. We seek to find a 95% confidence interval. For our example, $n = 10$ and $p = 0.05/2$. The table provides $T = 8$.

The endpoints of the confidence interval are the Kth smallest and the Kth largest values of u_{ij}, where $K = T + 1$. For our example, $K = 8 + 1 = 9$. The ninth value from the bottom is 0.5 and the ninth value from the top is 12.0. Based on these findings, it is estimated with 95% confident that the difference of successful interventions due to the new bullying programs lies between 0.5 and 12.0.

3.3.3 Sample Wilcoxon Signed Rank Test (Large Data Samples)

Hearing of Clear Creek School District's success with their antibullying program, Jonestown School District has implemented the program this year to reduce bullying in their own elementary schools. The Jonestown School District evaluates their program's effectiveness by comparing the percentage of successful interventions last year before the program began with the percentage of successful interventions this year with the program in place. In Table 3.5, the 25 elementary school counselors, or participants, reported the percentage of successful interventions last year and the percentage this year.

TABLE 3.5

	Percentage of successful interventions	
Participant	Last year	This year
1	53	50
2	18	43
3	21	28
4	44	48
5	12	35
6	36	32

(*Continued*)

TABLE 3.5 (*Continued*)

	Percentage of successful interventions	
Participant	Last year	This year
7	22	23
8	29	34
9	17	27
10	10	42
11	38	44
12	37	16
13	19	33
14	37	50
15	28	20
16	15	27
17	25	27
18	38	30
19	40	51
20	30	50
21	23	45
22	41	20
23	31	49
24	28	43
25	14	30

We will use the same nonparametric procedure to analyze the data. However, use a large sample ($n \geq 20$) approximation.

3.3.3.1 State the Null and Research Hypotheses

The null hypothesis states that the counselors reported no difference in the percentages last year and this year. The research hypothesis states that the counselors observed some differences between this year and last year. Our research hypothesis is a two-tailed, nondirectional hypothesis because it indicates a difference, but in no particular direction.

The null hypothesis is

$$H_O\text{: } \mu_D = 0$$

The research hypothesis is

$$H_A\text{: } \mu_D \neq 0$$

3.3.3.2 Set the Level of Risk (or the Level of Significance) Associated with the Null Hypothesis

The level of risk, also called an alpha (α), is frequently set at 0.05. We will use $\alpha = 0.05$ in our example. In other words, there is a 95% chance that any observed statistical difference will be real and not due to chance.

3.3.3.3 Choose the Appropriate Test Statistic The data are obtained from 25 counselors, or participants, who are using a new program designed to reduce bullying among students in the elementary schools. The participants reported the percentage of successful interventions last year and the percentage this year. We are comparing last year's percentages with this year's percentages. Therefore, the data samples are related or paired. Since we are comparing two related samples, we will use the Wilcoxon signed rank test.

3.3.3.4 Compute the Test Statistic First, compute the difference between each sample pair. Then, rank the absolute value of those computed differences. We have done this in Table 3.6.

TABLE 3.6

	Percentage of successful interventions				
Participant	Last year	This year	Difference	Rank	Sign
1	53	50	−3	3	−
2	18	43	+25	24	+
3	21	28	+7	8	+
4	44	48	+4	4.5	+
5	12	35	+23	23	+
6	36	32	−4	4.5	−
7	22	23	+1	1	+
8	29	34	+5	6	+
9	17	27	+10	11	+
10	10	42	+32	25	+
11	38	44	+6	7	+
12	37	16	−21	20.5	−
13	19	33	+14	15	+
14	37	50	+13	14	+
15	28	20	−8	9.5	−
16	15	27	+12	13	+
17	25	27	+2	2	+
18	38	30	−8	9.5	−
19	40	51	+11	12	+
20	30	50	+20	19	+
21	23	45	+22	22	+
22	41	20	−21	20.5	−
23	31	49	+18	18	+
24	28	43	+15	16	+
25	14	30	+16	17	+

Compute the sum of ranks with positive differences. Using Table 3.6, when we add all of the ranks with positive difference, we get $\Sigma R_+ = 257.5$.

Compute the sum of ranks with negative differences. The ranks with negative differences are 3, 4.5, 9.5, 9.5, 20.5, and 20.5. The sum of ranks with negative difference is $\Sigma R_- = 67.5$.

The obtained value is the smaller of these two rank sums. Thus, the Wilcoxon $T = 67.5$.

Since our sample size is larger than 20, we will approximate it to a normal distribution. Therefore, we will find a z-score for our data using a normal approximation. We must find the mean $\bar{x}_T$ and the standard deviation s_T for the data:

$$\bar{x}_T = \frac{n(n+1)}{4} = \frac{25(25+1)}{4}$$

$$\bar{x}_T = 162.5$$

and

$$s_T = \sqrt{\frac{n(n+1)(2n+1)}{24}} = \sqrt{\frac{25(25+1)(50+1)}{24}} = \sqrt{\frac{33{,}150}{24}}$$

$$s_T = 37.17$$

Next, we use the mean, standard deviation, and the T-test statistic to calculate a z-score. Remember, we are testing the hypothesis that there is no difference in ranks of percentages of successful interventions between last year and this year:

$$z^* = \frac{T - \bar{x}_T}{s_T} = \frac{67.5 - 162.5}{37.17}$$

$$z^* = -2.56$$

3.3.3.5 Determine the Value Needed for Rejection of the Null Hypothesis Using the Appropriate Table of Critical Values for the Particular Statistic Table B.1 in Appendix B is used to establish the critical region of z-scores. For a two-tailed test with $\alpha = 0.05$, we must not reject the null hypothesis if $-1.96 \le z^* \le 1.96$.

3.3.3.6 Compare the Obtained Value to the Critical Value We find that z^* is not within the critical region of the distribution, $-2.56 < -1.96$. Therefore, we reject the null hypothesis. This suggests a difference in the percentage of successful interventions after the program was implemented.

3.3.3.7 Interpret the Results We rejected the null hypothesis, suggesting that a real difference exists between last year's percentages and this year's percentages. In addition, since the sum of the positive difference ranks (ΣR_+) was larger than the negative difference ranks (ΣR_-), the difference is positive, showing a positive impact of the program. Therefore, our analysis provides evidence that the new bullying program is providing positive benefits toward the improvement of student behavior as perceived by the school counselors.

At this point, the analysis is limited to identifying the presence or absence of a significant difference between the groups. In other words, the statistical test's level of significance does not describe the strength of the treatment. The American Psychological Association (2001), however, has called for a measure of the strength called the ES.

We can consider the ES for this large sample test to determine the degree of association between the groups. We use Formula 3.5 to calculate the ES. For the example, $|z| = 2.56$ and $n = 25$:

$$ES = \frac{|z|}{\sqrt{n}} = \frac{|-2.56|}{\sqrt{25}}$$

$$ES = 0.51$$

Our ES for the matched-pair samples is 0.51. This value indicates a high level of association between the percentage of successful interventions before and after the implementation of the new bullying program.

3.3.3.8 Reporting the Results For this example, the Wilcoxon signed rank test ($T = 67.5, n = 25, p < 0.05$) indicated that the percentage of successful interventions was significantly different. In addition, the sum of the positive difference ranks ($\Sigma R_+ = 257.5$) was larger than the sum of the negative difference ranks ($\Sigma R_- = 67.5$), showing a positive impact from the program. Moreover, the ES for the matched-pair samples was 0.51. Therefore, our analysis provides evidence that the new bullying program is providing positive benefits toward the improvement of student behavior as perceived by the school counselors.

3.4 COMPUTING THE SIGN TEST

You can analyze related samples more efficiently by reducing values to dichotomous results ("yes" or "no") or ("+" or "−"). The sign test allows you to perform that analysis. Our procedure for performing the sign test is based on the method described by Gibbons and Chakraborti (2010).

We begin the procedure for performing a sign test by identifying whether each set from the related data samples demonstrates a positive difference, a negative difference, or no difference at all. Then, we find the sum of the positive differences n_p and the sum of negative differences n_n. Cases with no difference are ignored.

We perform the next part of the analysis based on the sum of differences. If $n_p + n_n = 0$, then the one-sided probability is $p = 0.5$. If $0 < n_p + n_n < 25$, then p is calculated recursively from the binomial probability function using Formula 3.7. Table B.9 in Appendix B includes several factorials to simplify computation:

$$P(X) = \frac{n!}{(n-X)!X!} \cdot p^X \cdot (1-p)^{n-X} \tag{3.7}$$

where $n = n_p + n_n$ and p is the probability of event occurrence.

If $n_p + n_n \geq 25$, we use Formula 3.8:

$$z_c = \frac{\max(n_p, n_n) - 0.5(n_p + n_n) - 0.5}{0.5\sqrt{n_p + n_n}} \tag{3.8}$$

Formula 3.8 approximates a binomial distribution to the normal distribution. However, the binomial distribution is a discrete distribution, while the normal distribution is continuous. More to the point, discrete values deal with heights but not widths, while the continuous distribution deals with both heights and widths. The correction adds or subtracts 0.5 of a unit from each discrete X-value to fill the gaps and make it continuous.

The one sided p-value is $p_1 = 1 - \Phi|z_c|$, where $\Phi|z_c|$ is the area under the respective tail of the normal distribution at z_c. The two-sided p-value is $p = 2p_1$.

3.4.1 Sample Sign Test (Small Data Samples)

To present the process for performing the sign test, we are going to use the data from Section 3.3.1, which used the Wilcoxon signed rank test. Recall that the sample involves 12 members of the counseling staff from Clear Creek County School District who are working on a program to improve response to bullying in the schools. The data from Table 3.1 are being reduced to a binomial distribution for use with the sign test. The relatively small sample size warrants a nonparametric procedure.

3.4.1.1 State the Null and Research Hypotheses The null hypothesis states that the counselors reported no difference between positive or negative interventions between last year and this year. In other words, the changes in responses produce a balanced number of positive and negative differences. The research hypothesis states that the counselors observed some differences between this year and last year. Our research hypothesis is a two-tailed, nondirectional hypothesis because it indicates a difference, but in no particular direction.

The null hypothesis is

H_O: $p = 0.5$

The research hypothesis is

H_A: $p \neq 0.5$

3.4.1.2 Set the Level of Risk (or the Level of Significance) Associated with the Null Hypothesis The level of risk, also called an alpha (α), is frequently set at 0.05. We will use $\alpha = 0.05$ in our example. In other words, there is a 95% chance that any observed statistical difference will be real and not due to chance.

3.4.1.3 Choose the Appropriate Test Statistic Recall from Section 3.3.1 that the data are obtained from 12 counselors, or participants, who are using a new program designed to reduce bullying among students in the elementary schools. The participants reported the percentage of successful interventions last year and the percentage this year. We are comparing last year's percentages with this year's percentages. Therefore, the data samples are related or paired. In addition, sample sizes are relatively small. Since we are comparing two related samples, we will use the sign test.

3.4.1.4 Compute the Test Statistic First, decide if there is a difference in intervention score from year 1 to year 2. Determine if the difference is positive or negative and put the sign of the difference in the sign column. If we count the number of ties or "0" differences among the group, we find only two with no difference from last year to this year. Ties are discarded.

Now, we count the number of positive and negative differences between last year and this year. Count the number of "+" or positive differences. When we look at Table 3.7, we see that eight participants showed positive differences, $n_p = 8$. Count the number of "−" or negative differences. When we look at Table 3.7, we see only two negative differences, $n_n = 2$.

TABLE 3.7

	Percentage of successful intervention		
Participant	Last year	This year	Sign of difference
1	31	31	0
2	14	14	0
3	53	50	−
4	18	30	+
5	21	28	+
6	44	48	+
7	12	35	+
8	36	32	−
9	22	23	+
10	29	34	+
11	17	27	+
12	40	42	+

Next, we find the X-score at and beyond where the area under our binomial probability function is $\alpha = 0.05$. Since we are performing a two-tailed test, we use 0.025 for each tail. We will calculate the probabilities associated with the binomial distribution for $p = 0.5$ and $n = 10$. We will demonstrate one of the calculations, but list the results for each value. To simplify calculation, use the table of factorials in Appendix B, Table B.9:

$$P(X) = \frac{n!}{(n-X)!X!} \cdot p^X \cdot (1-p)^{n-X}$$

$$P(0) = \frac{10!}{(10-0)!0!} \cdot 0.5^0 \cdot (1-0.5)^{10-0}$$

$$P(0) = \frac{3{,}628{,}800}{(3{,}628{,}800)(0)} \cdot 1 \cdot 0.000977$$

$$P(0) = 0.0010$$

$$P(1) = 0.0098$$

$$P(2) = 0.0439$$

$$P(3) = 0.1172$$

$$P(4) = 0.2051$$

$$P(5) = 0.2461$$

$$P(6) = 0.2051$$

$$P(7) = 0.1172$$

$$P(8) = 0.0439$$

$$P(9) = 0.0098$$

$$P(10) = 0.0010$$

Notice that the values form a symmetric distribution with the median at $P(5)$, as shown in Figure 3.1. Using this distribution, we find the p-values for each tail. To do that, we sum the probabilities for each tail until we find a probability equal to or greater than $\alpha/2 = 0.025$. First, calculate P for pluses:

$$P(8, 9, \text{or } 10) = 0.0439 + 0.0098 + 0.0010 = 0.0547$$

Second, calculate P for minuses:

$$P(0, 1, \text{or } 2) = 0.0010 + 0.0098 + 0.0439 = 0.0547$$

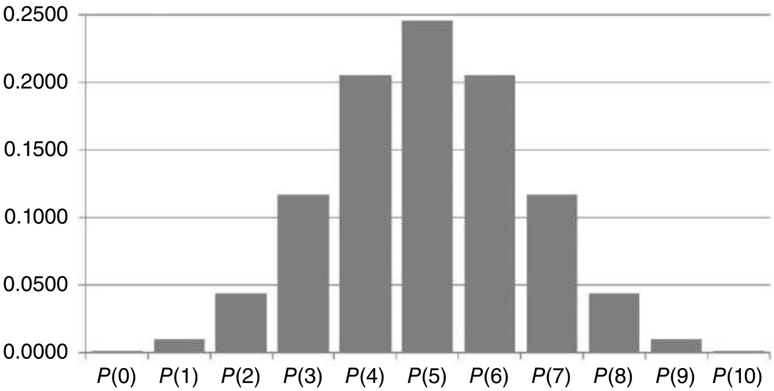


FIGURE 3.1

Finally, calculate the obtained value p by combining the two tails:

$$p = P(8, 9, \text{or } 10) + P(0, 1, \text{or } 2) = 0.0547 + 0.0547$$

$$p = 0.1094$$

3.4.1.5 Determine the Critical Value Needed for Rejection of the Null Hypothesis In the example in this chapter, the two-tailed probability was computed and is compared with the level of risk specified earlier, $\alpha = 0.05$.

3.4.1.6 Compare the Obtained Value with the Critical Value The critical value for rejecting the null hypothesis is $\alpha = 0.05$ and the obtained p-value is $p = 0.1094$. If the critical value is greater than the obtained value, we must reject the null hypothesis. If the critical value is less than the obtained value, we do not reject the null hypothesis. Since the critical value is less than the obtained value ($p > \alpha$), we do not reject the null hypothesis.

3.4.1.7 Interpret the Results We did not reject the null hypothesis, suggesting that no real difference exists between last year's and this year's percentages. There was no evidence of positive or negative intervention by counselors. These results differ from the data's analysis using the Wilcoxon signed rank test. A discussion about statistical power addresses those differences toward the end of this chapter.

3.4.1.8 Reporting the Results When reporting the findings for the sign test, you should include the sample size, the number of pluses, minuses, and ties, and the probability of getting the obtained number of pluses and minuses.

For this example, the obtained value, $p = 0.1094$, was greater than the critical value, $\alpha = 0.05$. Therefore, we did not reject the null hypothesis, suggesting that the new bullying program is not providing evidence of a change in student behavior as perceived by the school counselors.

3.4.2 Sample Sign Test (Large Data Samples)

We are going to demonstrate a sign test with large samples using the data from the Wilcoxon signed rank test for large samples in Section 3.3.3. The data from the implementation of the bullying program in the Jonestown School District are presented in Table 3.8. The data are used to determine the effect of the bullying program from year 1 to year 2. If there is an increase in successful intervention, we will use a "+" to identify the positive difference in response. If there is a decrease in successful intervention in the response, we will identify a negative difference with a "−." There are 25 participants in this study.

3.4.2.1 State the Null and Alternate Hypotheses The null hypothesis states that there was no positive or negative effect of the bullying program on successful intervention. The research hypothesis states that either a positive or negative effect exists from the bullying program.

TABLE 3.8

	Percentage of successful interventions	
Participant	Last year	This year
1	53	50
2	18	43
3	21	28
4	44	48
5	12	35
6	36	32
7	22	23
8	29	34
9	17	27
10	10	42
11	38	44
12	37	16
13	19	33
14	37	50
15	28	20
16	15	27
17	25	27
18	38	30
19	40	51
20	30	50
21	23	45
22	41	20
23	31	49
24	28	43
25	14	30

The null hypothesis is

$$H_O\colon p = 0.5$$

The research hypothesis is

$$H_A\colon p \neq 0.5$$

3.4.2.2 Set the Level of Risk (or the Level of Significance) Associated with the Null Hypothesis The level of risk, also called an alpha (α), is frequently set at 0.05. We will use $\alpha = 0.05$ in our example. In other words, there is a 95% chance that any observed statistical difference will be real and not due to chance.

3.4.2.3 Choose the Appropriate Test Statistic Recall from Section 3.3.3 that the data were obtained from 25 counselors, or participants, who were using a new program designed to reduce bullying among students in the elementary schools. The

TABLE 3.9

	Percentage of successful interventions		
Participant	Last year	This year	Sign of difference
1	53	50	−
2	18	43	+
3	21	28	+
4	44	48	+
5	12	35	+
6	36	32	−
7	22	23	+
8	29	34	+
9	17	27	+
10	10	42	+
11	38	44	+
12	37	16	−
13	19	33	+
14	37	50	+
15	28	20	−
16	15	27	+
17	25	27	+
18	38	30	−
19	40	51	+
20	30	50	+
21	23	45	+
22	41	20	−
23	31	49	+
24	28	43	+
25	14	30	+

participants reported the percentage of successful interventions last year and the percentage this year. We are comparing last year's percentages with this year's percentages. Therefore, the data samples are related or paired. Since we are making dichotomous comparisons of two related samples, we will use the sign test.

3.4.2.4 Compute the Test Statistic First, we determine the sign of the differences between last year and this year. Table 3.9 includes the column for the sign of the difference for each participant. Next, we count the numbers of positive and negative differences. We find six negative differences, $n_n = 6$, and 19 positive differences, $n_p = 19$.

Since the sample size is $n \geq 25$, we will use a z-score approximation of the binomial distribution. The binomial distribution becomes an approximation of the normal distribution as n becomes large and p is not too close to the 0 or 1 values. If this approximation is used, $P(Y \leq k)$ is obtained by computing the corrected z-score for the given data that are as extreme or more extreme than the data given:

$$z_c = \frac{\max(n_p, n_n) - 0.5(n_p + n_n) - 0.5}{0.5\sqrt{n_p + n_n}} = \frac{19 - (0.5)(19+6) - 0.5}{(0.5)(\sqrt{19+6})}$$

$$= \frac{19 - 12.5 - 0.5}{(0.5)(5)} = \frac{6}{2.5}$$

$$z_c = 2.4$$

Next, we find the one-sided p-value. Table B.1 is used to establish $\Phi|z_c|$.

$$p_1 = 1 - \Phi|z_c| = 1 - 0.9918$$

$$p_1 = 0.0082$$

We now multiply two times the one-sided p-value to find the two-sided p-value:

$$p = 2p_1 = (2)(0.0082)$$

$$p = 0.016$$

3.4.2.5 Determine the Critical Value Needed for Rejection of the Null Hypothesis In the example in this chapter, the two-tailed probability was computed and compared with the level of risk specified earlier, $\alpha = 0.05$.

3.4.2.6 Compare the Obtained Value with the Critical Value The critical value for rejecting the null hypothesis is $\alpha = 0.05$ and the obtained p-value is $p = 0.016$. If the critical value is greater than the obtained value, we must reject the null hypothesis. If the critical value is less than the obtained value, we do not reject the null hypothesis. Since the critical value is greater than the obtained value ($p < \alpha$), we reject the null hypothesis.

3.4.2.7 Interpret the Results We rejected the null hypothesis, suggesting that there is a real difference between last year's and this year's degree of successful intervention for the 25 counselors who were in the study.

Analysis was limited to the identification of the presence of positive "+" or negative "−" differences between year 1 and year 2 for each participant. The level of significance does not describe the strength of the test's level of significance.

3.4.2.8 Reporting the Results When reporting the findings for the sign test, you should include the sample size, the number of pluses, minuses, and ties, and the probability of getting the obtained number of pluses and minuses.

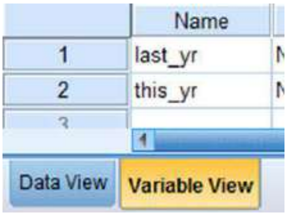
For this example, the obtained significance, $p = 0.016$, was less than the critical value, $\alpha = 0.05$. Therefore, we rejected the null hypothesis, suggesting that the number of successful interventions was significantly different from year 1 to year 2.

3.5 PERFORMING THE WILCOXON SIGNED RANK TEST AND THE SIGN TEST USING SPSS

We will analyze the small sample examples for the Wilcoxon signed rank test and the sign test using SPSS.

3.5.1 Define Your Variables

First, click the "Variable View" tab at the bottom of your screen. Then, type the names of your variables in the "Name" column. As shown in Figure 3.2, we have named our variables "last_yr" and "this_yr."

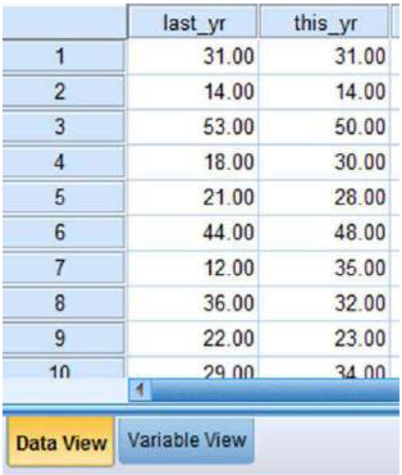


	Name
1	last_yr
2	this_yr

FIGURE 3.2

3.5.2 Type in Your Values

Click the "Data View" tab at the bottom of your screen and type your data under the variable names. As shown in Figure 3.3, we are comparing "last_yr" with "this_yr."



	last_yr	this_yr
1	31.00	31.00
2	14.00	14.00
3	53.00	50.00
4	18.00	30.00
5	21.00	28.00
6	44.00	48.00
7	12.00	35.00
8	36.00	32.00
9	22.00	23.00
10	29.00	34.00

FIGURE 3.3

3.5.3 Analyze Your Data

As shown in Figure 3.4, use the pull-down menus to choose “Analyze,” “Nonparametric Tests,” “Legacy Dialogs,” and “2 Related Samples . . .”

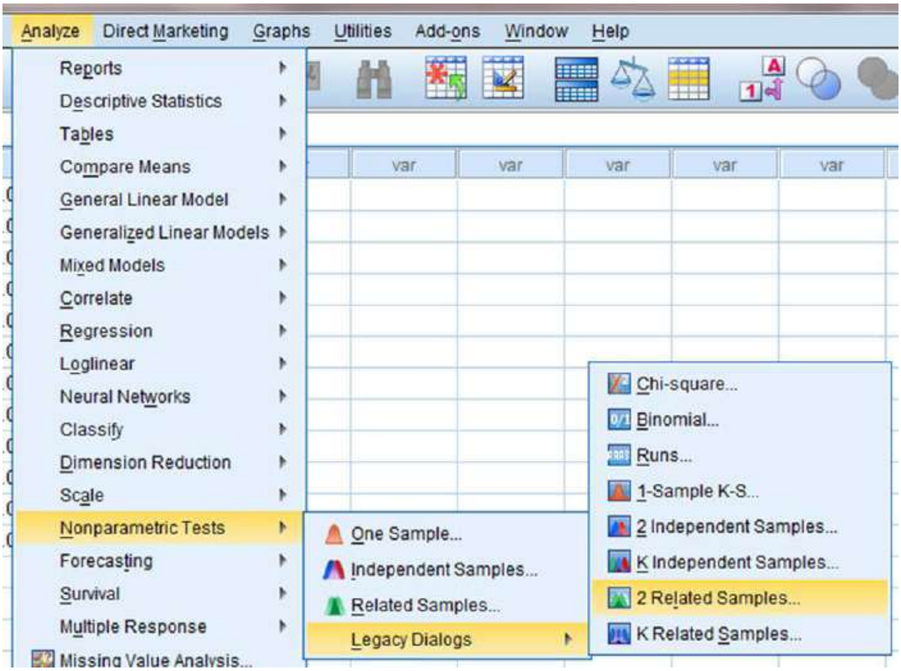


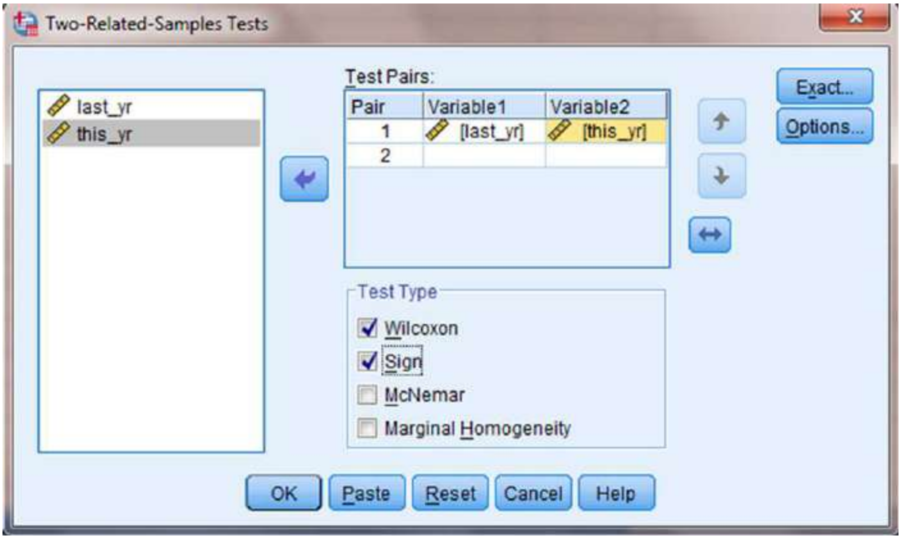
FIGURE 3.4

In the upper left box, select both variables that you want to compare. Then, use the arrow button to place your variable pair in the box labeled “Test Pairs:”. Next, check the “Test Type” you wish to perform. In Figure 3.5, we have checked “Wilcoxon” and “Sign” to perform both tests. Finally, click “OK” to perform the analysis.

3.5.4 Interpret the Results from the SPSS Output Window

SPSS Output 3.1 begins by reporting the results from the Wilcoxon signed rank test. The first output table (called “Ranks”) provides the Wilcoxon T or obtained value. From the “Sum of Ranks” column, we select the smaller of the two values. In our example, $T = 7.5$. The second output table (called “Test Statistics”) returns the critical z-score for large samples. In addition, SPSS calculates the two-tailed significance ($p = 0.041$).

Based on the results from SPSS, the number of successful interventions was significantly different ($T = 7.5$, $n = 12$, $p < 0.05$). In addition, the sum of the positive difference ranks ($\Sigma R_+ = 47.5$) was larger than the sum of the negative difference ranks ($\Sigma R_- = 7.5$), demonstrating a positive impact from the program.



Two-Related-Samples Tests

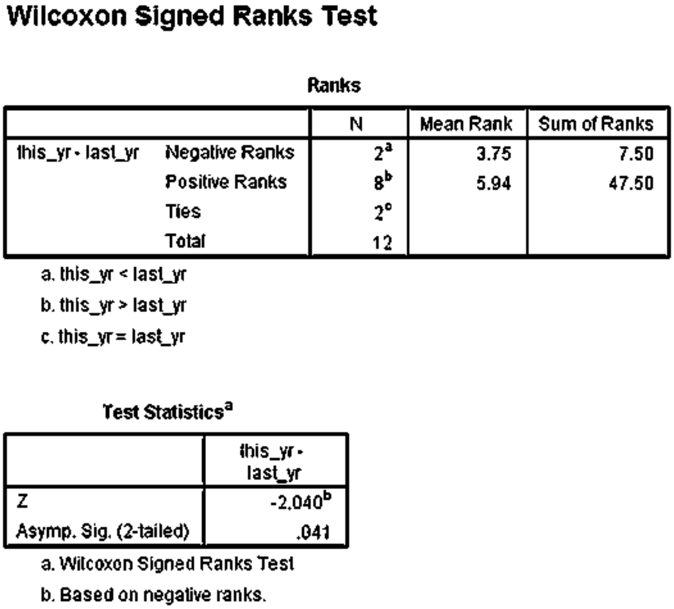
Test Pairs:

Pair	Variable1	Variable2
1	[last_yr]	[this_yr]
2		

Test Type

☑ Wilcoxon
☑ Sign
□ McNemar
□ Marginal Homogeneity

FIGURE 3.5



Wilcoxon Signed Ranks Test

Ranks

		N	Mean Rank	Sum of Ranks
this_yr - last_yr	Negative Ranks	2^a	3.75	7.50
	Positive Ranks	8^b	5.94	47.50
	Ties	2^c		
	Total	12		

a. this_yr < last_yr

b. this_yr > last_yr

c. this_yr = last_yr

Test Statistics[a]

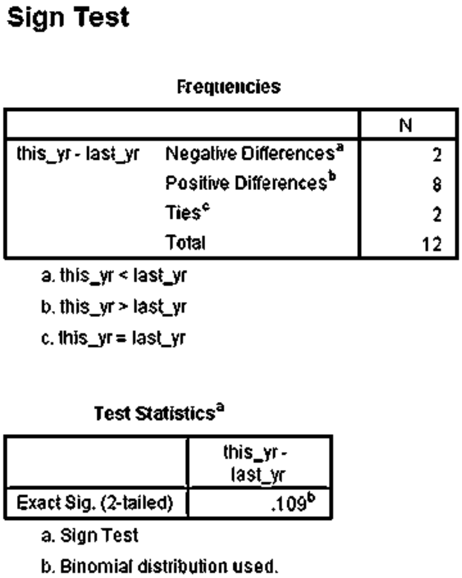
	this_yr - last_yr
Z	-2.040^b
Asymp. Sig. (2-tailed)	.041

a. Wilcoxon Signed Ranks Test

b. Based on negative ranks.

SPSS OUTPUT 3.1

Next, SPSS Output 3.2 reports the results from the sign test. The first output table (called “Frequencies”) provides the negative differences, positive differences, ties, and total comparisons. The second output table (called “Test Statistics”) returns the two-tailed significance ($p = 0.109$). Based on the results of the sign test using SPSS, the number of successful interventions was not significantly different ($0.109 > 0.05$).



Sign Test

Frequencies

		N
this_yr - last_yr	Negative Differences[a]	2
	Positive Differences[b]	8
	Ties[c]	2
	Total	12

a. this_yr < last_yr

b. this_yr > last_yr

c. this_yr = last_yr

Test Statistics[a]

	this_yr - last_yr
Exact Sig. (2-tailed)	.109[b]

a. Sign Test

b. Binomial distribution used.

SPSS OUTPUT 3.2

The notion that the Wilcoxon signed rank test produced significant results while the sign test did not is addressed next in a brief discussion about statistical power.

3.6 STATISTICAL POWER

Comparing our conflicting results from the small sample Wilcoxon signed rank test with the sign test presents an opportunity to discuss statistical power. That difference is especially visible when comparing the results from the sample problems in Sections 3.3.1 and 3.4.1 of this chapter. Both sections analyzed the same data; however, one section demonstrated a Wilcoxon signed rank test and the other demonstrated the sign test.

Notice that the result from the Wilcoxon signed rank test was significant, yet the result from the sign test was not significant. In other words, one test produced significant results and the other test did not. The reason involves differences in statistical power.

Nonparametric methods generally have less statistical power compared with their parametric equivalents, especially when used in small samples. For instance, a test with less statistical power has a smaller chance of detecting a true effect where one might actually exist. This difference in statistical power is especially true for the sign test (Siegel and Castellan, 1988).

A statistical test's power depends on several factors: the size of the effect (discussed later), level of desired significance (α), and sample size. Researchers use this information to perform a statistical power analysis before performing the experiment. This allows the researcher to determine the needed sample size. A quick search returns a variety of online power analysis tools. Currently, *G*Power* is a free tool. In addition, Cohen (1988) has provided several tables for finding sample sizes based on level of power.

3.7 EXAMPLES FROM THE LITERATURE

To be shown are varied examples of the nonparametric procedures described in this chapter. We have summarized each study's research problem and the researchers' rationale(s) for choosing a nonparametric approach. We encourage you to obtain these studies if you are interested in their results.

Boser and Poppen (1978) sought to determine which verbal responses by teacher held the greatest potential for improving student–teacher relationships. The seven verbal responses were feelings, thoughts, motives, behaviors, encounter/encouragement, confrontation, and sharing. They used a Wilcoxon signed rank test to examine 101 9th-grader responses because the student participants rank ordered their responses.

Vaughn et al. (1999) investigated kindergarten teachers' perceptions of practices identified to improve outcomes for children with disabilities transitioning from prekindergarten to kindergarten. The researchers compared the paired ratings of teachers' desirability to employ the identified practices with feasibility using a Wilcoxon signed rank test. This nonparametric procedure was considered the most appropriate because the study's measure was a Likert-type scale (1 = *low*, 5 = *high*).

Rinderknecht and Smith (2004) used a 7-month nutrition intervention to improve the dietary self-efficacy of Native American children (5–10 years) and adolescents (11–18 years). Wilcoxon signed rank tests were used to determine whether fat and sugar intake changed significantly between pre- and postintervention among adolescents. The researchers chose nonparametric tests for their data that were not normally distributed.

Seiver and Hatfield (2002) asked environmental health professionals about their willingness to dine in certain restaurants based on the method and history of health code evaluations. A paired-sample sign test was used to determine which health code evaluation method and history that participants preferred. The researchers chose a nonparametric test since they administered questionnaires with rank ordered scales (0 = *never*, 10 = *always*).

3.8 SUMMARY

Two samples that are paired, or related, may be compared using a nonparametric procedure called the Wilcoxon signed rank test or the sign test. The parametric equivalent to this test is known as the Student's t-test, t-test for matched pairs, or t-test for dependent samples.

In this chapter, we described how to perform and interpret a Wilcoxon signed rank test and a sign test, using both small samples and large samples. We also explained how to perform the procedure for both tests using SPSS. Finally, we offered varied examples of these nonparametric statistics from the literature. The next chapter will involve comparing two samples that are not related.

3.9 PRACTICE QUESTIONS

1. A teacher wished to determine if providing a bilingual dictionary to students with limited English proficiency improves math test scores. A small class of students ($n = 10$) was selected. Students were given two math tests. Each test covered the same type of math content; however, students were provided a bilingual dictionary on the second test. The data in Table 3.10 represent the students' performance on each math test.

TABLE 3.10

Student	Math test without a bilingual dictionary	Math test with a bilingual dictionary
1	30	39
2	56	46
3	48	37
4	47	44
5	43	32
6	45	39
7	36	41
8	44	40
9	44	38
10	40	46

Use a one-tailed Wilcoxon signed rank test and a one-tailed sign test to determine which testing condition resulted in higher scores. Use $\alpha = 0.05$. Report your findings.

2. A research study was done to investigate the influence of being alone at night on the human male heart rate. Ten men were sent into a wooded area, one at a time, at night, for 20 min. They had a heart monitor to record their pulse rate. The second night, the same men were sent into a similar wooded area accompanied by a companion. Their pulse rate was recorded again. The researcher wanted to see if having a companion would change their pulse rate. The median rates are reported in Table 3.11.

   Use a two-tailed Wilcoxon signed rank test and a two-tailed sign test to determine which condition produced a higher pulse rate. Use $\alpha = 0.05$. Report your findings.

TABLE 3.11

Participant	Median rate alone	Median rate with companion
A	88	72
B	77	74
C	91	80
D	70	77
E	80	71
F	85	83
G	90	80
H	82	91
I	93	86
J	75	69

3. A researcher conducts a pilot study to compare two treatments to help obese female teenagers lose weight. She tests each individual in two different treatment conditions. The data in Table 3.12 provide the number of pounds that each participant lost.

TABLE 3.12

	Pounds lost	
Participant	Treatment 1	Treatment 2
1	10	18
2	20	12
3	15	16
4	9	7
5	18	21
6	11	17
7	6	13
8	12	14

Use a two-tailed Wilcoxon signed rank test and a two-tailed sign test to determine which treatment resulted in greater weight loss. Use $\alpha = 0.05$. Report your findings.

4. Twenty participants in an exercise program were measured on the number of sit-ups they could do before other physical exercise (first count) and the number they could do after they had done at least 45 min of other physical exercise (second count). Table 3.13 shows the results for 20 participants obtained during two separate physical exercise sessions. Determine the ES for a calculated z-score.

TABLE 3.13

Participant	First count	Second count
1	18	28
2	19	18
3	20	28
4	29	20
5	15	30
6	22	25
7	21	28
8	30	18
9	22	27
10	11	30
11	20	24
12	21	27
13	21	10
14	20	40
15	18	20
16	27	14
17	24	29
18	13	30
19	10	24
20	10	36

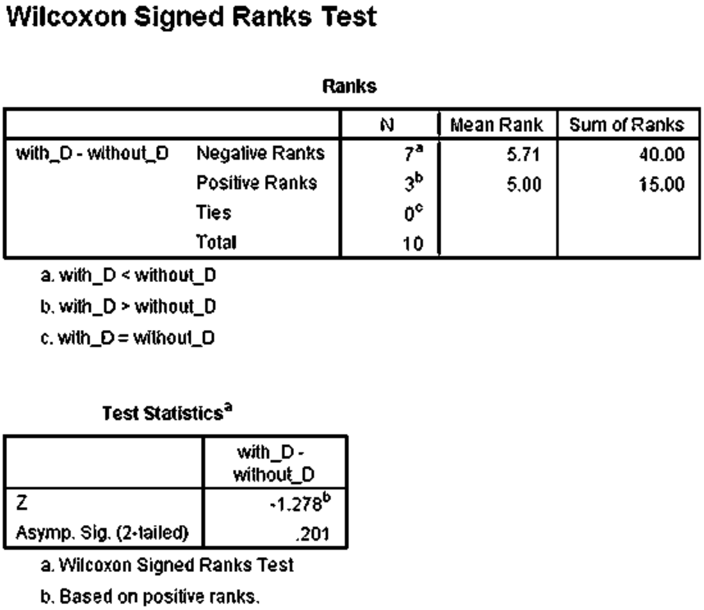
5. A school is trying to get more students to participate in activities that will make learning more desirable. Table 3.14 shows the number of activities that each of the 10 students in one class participated in last year before a new activity program was implemented and this year after it was implemented. Construct a 95% median confidence interval based on the Wilcoxon signed rank test to determine whether the new activity program had a significant positive effect on the student participation.

TABLE 3.14

Participants	Last year	This year
1	18	20
2	22	28
3	10	18
4	25	23
5	16	20
6	14	21
7	21	17
8	13	18
9	28	22
10	12	21

3.10 SOLUTIONS TO PRACTICE QUESTIONS

1. The results from the analysis are displayed in SPSS Outputs 3.3 and 3.4. Both tests report the two-tailed significance, but the question asked for the one-tailed significance. Therefore, divide the two-tailed significance by 2 to find the one-tailed significance.



Wilcoxon Signed Ranks Test

Ranks

		N	Mean Rank	Sum of Ranks
with_D - without_D	Negative Ranks	7[a]	5.71	40.00
	Positive Ranks	3[b]	5.00	15.00
	Ties	0[c]		
	Total	10		

a. with_D < without_D

b. with_D > without_D

c. with_D = without_D

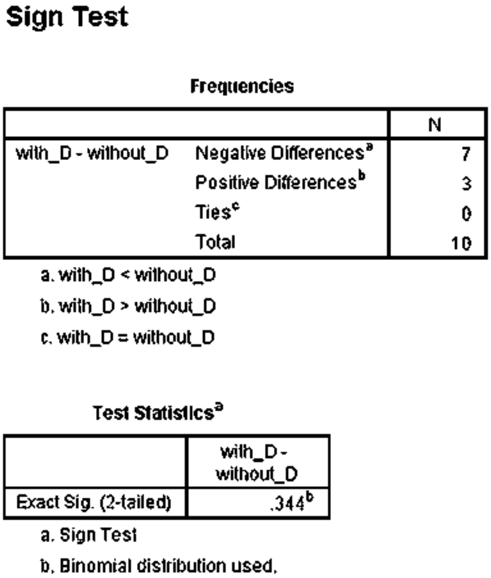
Test Statistics[a]

	with_D - without_D
Z	-1.278[b]
Asymp. Sig. (2-tailed)	.201

a. Wilcoxon Signed Ranks Test

b. Based on positive ranks.

SPSS OUTPUT 3.3



Sign Test

Frequencies

		N
with_D - without_D	Negative Differences[a]	7
	Positive Differences[b]	3
	Ties[c]	0
	Total	10

a. with_D < without_D

b. with_D > without_D

c. with_D = without_D

Test Statistics[a]

	with_D - without_D
Exact Sig. (2-tailed)	.344[b]

a. Sign Test

b. Binomial distribution used.

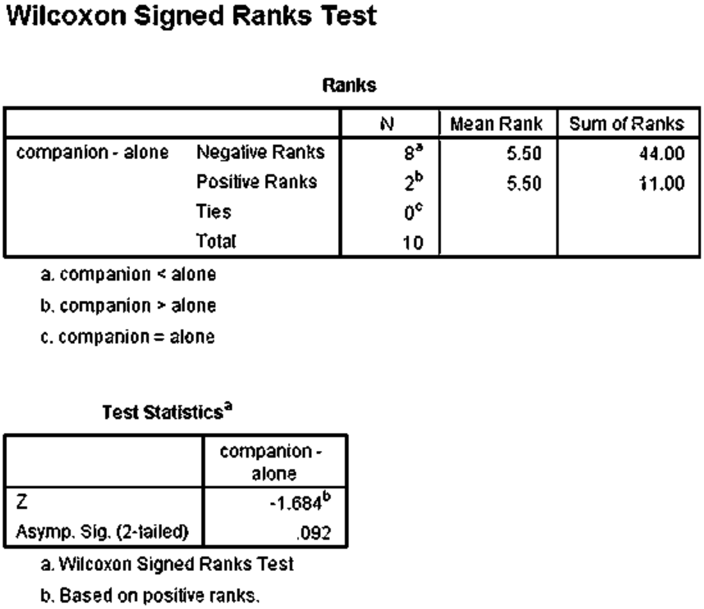
SPSS OUTPUT 3.4

The results from the Wilcoxon signed rank test reported a one-tailed significance of $p = 0.201/2 = 0.101$. The test results ($T = 15.0$, $n = 10$, $p > 0.05$) indicated that the two testing conditions were not significantly different.

The results from the sign test reported a one-tailed significance of $p = 0.344/2 = 0.172$. These test results ($p > 0.05$) also indicated that the two testing conditions were not significantly different.

Therefore, based on this study, the use of bilingual dictionaries on a math test did not significantly improve scores among limited English proficient students.

2. The results from the analysis are displayed in SPSS Outputs 3.5 and 3.6.



Wilcoxon Signed Ranks Test

Ranks

		N	Mean Rank	Sum of Ranks
companion - alone	Negative Ranks	8[a]	5.50	44.00
	Positive Ranks	2[b]	5.50	11.00
	Ties	0[c]		
	Total	10		

a. companion < alone

b. companion > alone

c. companion = alone

Test Statistics[a]

	companion - alone
Z	-1.684[b]
Asymp. Sig. (2-tailed)	.092

a. Wilcoxon Signed Ranks Test

b. Based on positive ranks.

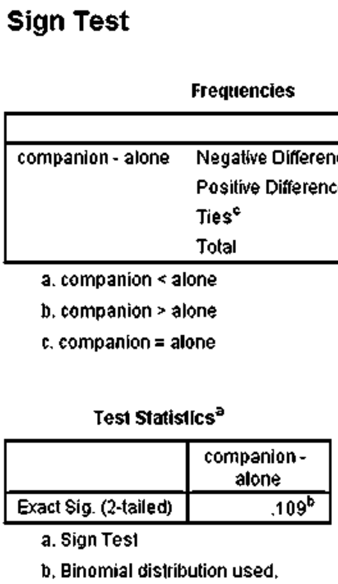
SPSS OUTPUT 3.5

The results from the Wilcoxon signed rank test reported a two-tailed significance of $p = 0.092$. The test results ($T = 11.0$, $n = 10$, $p > 0.05$) indicated that the two conditions were not significantly different.

The results from the sign test reported a two-tailed significance of $p = 0.109$. These test results ($p > 0.05$) also indicated that the two testing conditions were not significantly different.

Therefore, based on this study, the presence of a companion in the woods at night did not significantly influence the males' pulse rates.

3. The results from the analysis are displayed in SPSS Outputs 3.7 and 3.8.
The results from the Wilcoxon signed rank test ($T = 10.0$, $n = 8$, $p > 0.05$) indicated that the two treatments were not significantly different.



Sign Test

Frequencies

		N
companion - alone	Negative Differences[a]	8
	Positive Differences[b]	2
	Ties[c]	0
	Total	10

a. companion < alone

b. companion > alone

c. companion = alone

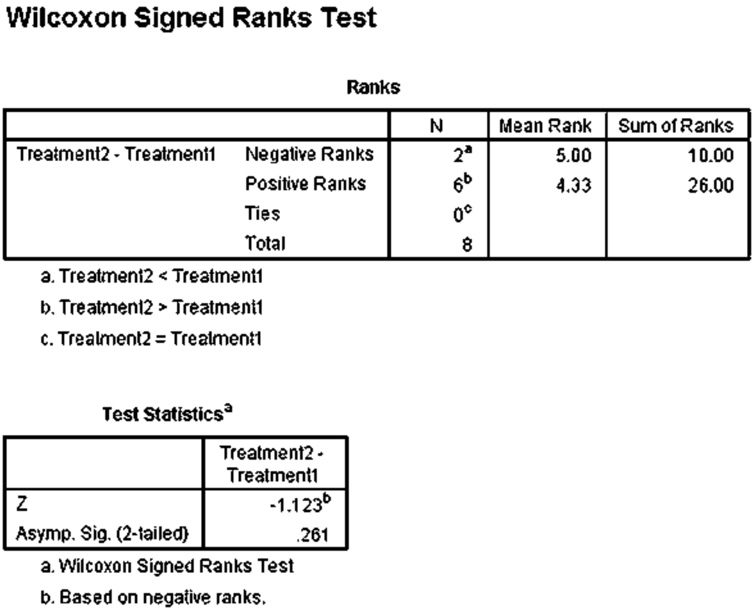
Test Statistics[a]

	companion - alone
Exact Sig. (2-tailed)	.109[b]

a. Sign Test

b. Binomial distribution used.

SPSS OUTPUT 3.6



Wilcoxon Signed Ranks Test

Ranks

		N	Mean Rank	Sum of Ranks
Treatment2 - Treatment1	Negative Ranks	2[a]	5.00	10.00
	Positive Ranks	6[b]	4.33	26.00
	Ties	0[c]		
	Total	8		

a. Treatment2 < Treatment1

b. Treatment2 > Treatment1

c. Treatment2 = Treatment1

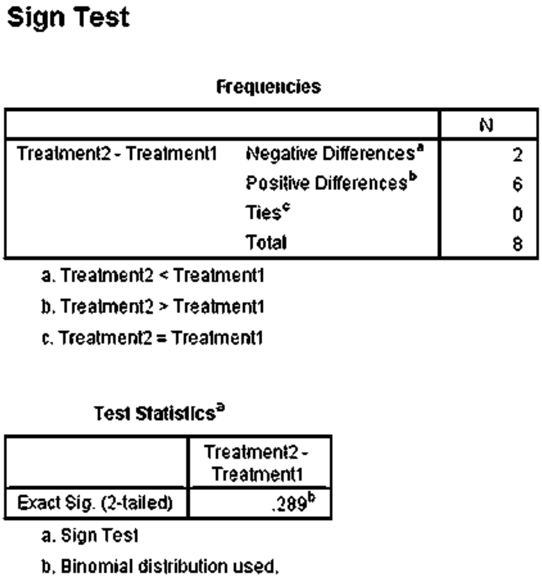
Test Statistics[a]

	Treatment2 - Treatment1
Z	-1.123[b]
Asymp. Sig. (2-tailed)	.261

a. Wilcoxon Signed Ranks Test

b. Based on negative ranks.

SPSS OUTPUT 3.7



Sign Test

Frequencies

		N
Treatment2 - Treatment1	Negative Differences[a]	2
	Positive Differences[b]	6
	Ties[c]	0
	Total	8

a. Treatment2 < Treatment1

b. Treatment2 > Treatment1

c. Treatment2 = Treatment1

Test Statistics[a]

	Treatment2 - Treatment1
Exact Sig. (2-tailed)	.289[b]

a. Sign Test

b. Binomial distribution used.

SPSS OUTPUT 3.8

The results from the sign test ($p > 0.05$) also indicated that the two testing conditions were not significantly different.

Therefore, based on this study, neither treatment program resulted in a significantly higher weight loss among obese female teenagers.

4. The results from the analysis are as follows:

$$T = 50$$

$$x_r = 105 \text{ and } s_r = 26.79$$

$$z^* = -2.05$$

$$ES = 0.46$$

This is a reasonably high ES which indicates a strong measure of association.

5. For our example, $n = 10$ and $p = 0.05/2$. Thus, $T = 8$ and $K = 9$. The ninth value from the bottom is -1.0 and the ninth value from the top is 7.0. Based on these findings, it is estimated with 95% confidence that the difference in students' number of activities before and after the new program lies between -1.0 and 7.0.

CHAPTER 4

COMPARING TWO UNRELATED SAMPLES: THE MANN–WHITNEY U-TEST AND THE KOLMOGOROV–SMIRNOV TWO-SAMPLE TEST

4.1 OBJECTIVES

In this chapter, you will learn the following items:

- How to perform the Mann–Whitney *U*-test.
- How to construct a median confidence interval based on the difference between two independent samples.
- How to perform the Kolmogorov–Smirnov two-sample test.
- How to perform the Mann–Whitney *U*-test and the Kolmogorov–Smirnov two-sample test using SPSS®.

4.2 INTRODUCTION

Suppose a teacher wants to know if his first-period's early class time has been reducing student performance. To test his idea, he compares the final exam scores of students in his first-period class with those in his fourth-period class. In this example, each score from one class period is independent, or unrelated, to the other class period.

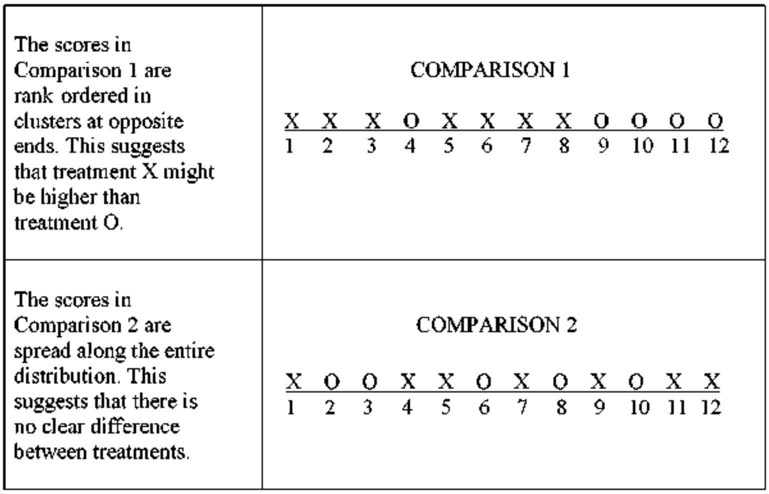
The Mann–Whitney *U*-test and the Kolmogorov–Smirnov two-sample test are nonparametric statistical procedures for comparing two samples that are independent, or not related. The parametric equivalent to these tests is the *t*-test for independent samples.

Nonparametric Statistics: A Step-by-Step Approach, Second Edition. Gregory W. Corder and Dale I. Foreman.

In this chapter, we will describe how to perform and interpret a Mann–Whitney U-test and a Kolmogorov–Smirnov two-sample test. We will demonstrate both small samples and large samples for each test. We will also explain how to perform the procedure using SPSS. Finally, we offer varied examples of these nonparametric statistics from the literature.

4.3 COMPUTING THE MANN–WHITNEY *U*-TEST STATISTIC

The Mann–Whitney U-test is used to compare two unrelated, or independent, samples. The two samples are combined and rank ordered together. The strategy is to determine if the values from the two samples are randomly mixed in the rank ordering or if they are clustered at opposite ends when combined. A random rank ordered would mean that the two samples are not different, while a cluster of one sample's values would indicate a difference between them. In Figure 4.1, two sample comparisons illustrate this concept.



The scores in Comparison 1 are rank ordered in clusters at opposite ends. This suggests that treatment X might be higher than treatment O.	COMPARISON 1 X X X O X X X X O O O O 1 2 3 4 5 6 7 8 9 10 11 12
The scores in Comparison 2 are spread along the entire distribution. This suggests that there is no clear difference between treatments.	COMPARISON 2 X O O X X O X O X O X X 1 2 3 4 5 6 7 8 9 10 11 12

FIGURE 4.1

Use Formula 4.1 to determine a Mann–Whitney U-test statistic for each of the two samples. The smaller of the two U statistics is the obtained value:

$$U_i = n_1 n_2 + \frac{n_i(n_i+1)}{2} - \sum R_i \tag{4.1}$$

where U_i is the test statistic for the sample of interest, n_i is the number of values from the sample of interest, n_1 is the number of values from the first sample, n_2 is the number of values from the second sample, and ΣR_i is the sum of the ranks from the sample of interest.

After the U statistic is computed, it must be examined for significance. We may use a table of critical values (see Table B.4 in Appendix B). However, if the numbers of values in each sample, n_i, exceeds those available from the table, then a large sample approximation may be performed. For large samples, compute a z-score and use a table with the normal distribution (see Table B.1 in Appendix B) to obtain a critical region of z-scores. Formula 4.2, Formula 4.3, and Formula 4.4 are used to find the z-score of a Mann–Whitney U-test for large samples:

$$\bar{x}_U = \frac{n_1 n_2}{2} \tag{4.2}$$

where $\bar{x}_U$ is the mean, n_1 is the number of values from the first sample, and n_2 is the number of values from the second sample;

$$s_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} \tag{4.3}$$

where s_U is the standard deviation;

$$z^* = \frac{U_i - \bar{x}_U}{s_U} \tag{4.4}$$

where z^* is the z-score for a normal approximation of the data and U_i is the U statistic from the sample of interest.

At this point, the analysis is limited to identifying the presence or absence of a significant difference between the groups and does not describe the strength of the treatment. We can consider the effect size (ES) to determine the degree of association between the groups. We use Formula 4.5 to calculate the ES:

$$ES = \frac{|z|}{\sqrt{n}} \tag{4.5}$$

where $|z|$ is the absolute value of the z-score and n is the total number of observations.

The ES ranges from 0 to 1. Cohen (1988) defined the conventions for ES as *small* = 0.10, *medium* = 0.30, and *large* = 0.50. (Correlation coefficient and ES are both measures of association. See Chapter 7 concerning correlation for more information on Cohen's assignment of ES's relative strength.)

4.3.1 Sample Mann–Whitney *U*-Test (Small Data Samples)

The following data were collected from a study comparing two methods being used to teach reading recovery in the 4th grade. Method 1 was a pull-out program in which the children were taken out of the classroom for 30 min a day, 4 days a week. Method 2 was a small group program in which children were taught in groups of four or five for 45 min a day in the classroom, 4 days a week. The students were tested using a reading comprehension test after 4 weeks of the program. The test results are shown in Table 4.1.

TABLE 4.1

Method 1	Method 2
48	14
40	18
39	20
50	10
41	12
38	102
53	17

4.3.1.1 State the Null and Research Hypotheses

The null hypothesis states that there is no tendency of the ranks of one method to be systematically higher or lower than the other. The hypothesis is stated in terms of comparison of distributions, not means. The research hypothesis states that the ranks of one method are systematically higher or lower than the other. Our research hypothesis is a two-tailed, nondirectional hypothesis because it indicates a difference, but in no particular direction.

The null hypothesis is

> H_O: There is no tendency for ranks of one method to be significantly higher (or lower) than the other.

The research hypothesis is

> H_A: The ranks of one method are systematically higher (or lower) than the other.

4.3.1.2 Set the Level of Risk (or the Level of Significance) Associated with the Null Hypothesis

The level of risk, also called an alpha (α), is frequently set at 0.05. We will use $\alpha = 0.05$ in our example. In other words, there is a 95% chance that any observed statistical difference will be real and not due to chance.

4.3.1.3 Choose the Appropriate Test Statistic

The data are obtained from two independent, or unrelated, samples of 4th-grade children being taught reading. Both the small sample sizes and an existing outlier in the second sample violate our assumptions of normality. Since we are comparing two unrelated, or independent, samples, we will use the Mann–Whitney U-test.

4.3.1.4 Compute the Test Statistic

First, combine and rank both data samples together (see Table 4.2).

Next, compute the sum of ranks for each method. Method 1 is ΣR_1 and method 2 is ΣR_2. Using Table 4.2,

$$\sum R_1 = 7+8+9+10+11+12+13$$

$$\sum R_1 = 70$$

TABLE 4.2

Ordered scores

Rank	Score	Sample
1	10	Method 2
2	12	Method 2
3	14	Method 2
4	17	Method 2
5	18	Method 2
6	20	Method 2
7	38	Method 1
8	39	Method 1
9	40	Method 1
10	41	Method 1
11	48	Method 1
12	50	Method 1
13	53	Method 1
14	102	Method 2

and

$$\sum R_2 = 1+2+3+4+5+6+14$$
$$\sum R_2 = 35$$

Now, compute the U-value for each sample. For sample 1,

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - \sum R_1 = 7(7) + \frac{7(7+1)}{2} - 70 = 49 + 28 - 70$$
$$U_1 = 7$$

and for sample 2,

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - \sum R_2 = 7(7) + \frac{7(7+1)}{2} - 35 = 49 + 28 - 35$$
$$U_2 = 42$$

The Mann–Whitney U-test statistic is the smaller of U_1 and U_2. Therefore, $U = 7$.

4.3.1.5 Determine the Value Needed for Rejection of the Null Hypothesis Using the Appropriate Table of Critical Values for the Particular Statistic

Since the sample sizes are small ($n < 20$), we use Table B.4 in Appendix B, which lists the critical values for the Mann–Whitney U. The critical values are found on the table at the point for $n_1 = 7$ and $n_2 = 7$. We set $\alpha = 0.05$. The critical value for the Mann–Whitney U is 8. A calculated value that is less than or equal to 8 will lead us to reject our null hypothesis.

4.3.1.6 Compare the Obtained Value with the Critical Value The critical value for rejecting the null hypothesis is 8 and the obtained value is $U = 7$. If the critical value equals or exceeds the obtained value, we must reject the null hypothesis. If instead, the critical value is less than the obtained value, we must not reject the null hypothesis. Since the critical value exceeds the obtained value, we must reject the null hypothesis.

4.3.1.7 Interpret the Results We rejected the null hypothesis, suggesting that a real difference exists between the two methods. In addition, since the sum of the ranks for method 1 (ΣR_1) was larger than method 2 (ΣR_2), we see that method 1 had significantly higher scores.

4.3.1.8 Reporting the Results The reporting of results for the Mann–Whitney U-test should include such information as the sample sizes for each group, the U statistic, the p-value's relation to α, and the sums of ranks for each group.

For this example, two methods were used to provide students with reading instruction. Method 1 involved a pull-out program and method 2 involved a small group program. Using the ranked reading comprehension test scores, the results indicated a significant difference between the two methods ($U = 7$, $n_1 = 7$, $n_2 = 7$, $p < 0.05$). The sum of ranks for method 1 ($\Sigma R_1 = 70$) was larger than the sum of ranks for method 2 ($\Sigma R_2 = 35$). Therefore, we can state that the data support the pull-out program as a more effective reading program for teaching comprehension to 4th-grade children at this school.

4.3.2 Confidence Interval for the Difference between Two Location Parameters

The American Psychological Association (2001) has suggested that researchers report the *confidence interval* for research data. A confidence interval is an inference to a population in terms of an estimation of sampling error. More specifically, it provides a range of values that fall within the population with a level of confidence of $100(1 - \alpha)\%$.

A median confidence interval can be constructed based on the difference between two independent samples. It consists of possible values of differences for which we do not reject the null hypothesis at a defined significance level of α.

The test depends on the following assumptions:

1. Data consist of two independent random samples: $X_1, X_2, \ldots, X_n$ from one population and $Y_1, Y_2, \ldots, Y_n$ from the second population.
2. The distribution functions of the two populations are identical except for possible location parameters.

To perform the analysis, set up a table that identifies all possible differences for each possible sample pair such that $D_{ij} = X_i - Y_j$ for (X_i, Y_j). Placing the values for X from smallest to largest across the top and the values for Y from smallest to largest down the side will eliminate the need to order the values of D_{ij} later.

TABLE 4.3

	X_i						
Y_j	38	39	40	41	48	50	53
10	28	29	30	31	38	40	43
12	26	27	28	29	36	38	41
14	24	25	26	27	34	36	39
17	21	22	23	24	31	33	36
18	20	21	22	23	30	32	35
20	18	19	20	21	28	30	33
102	−64	−63	−62	−61	−54	−52	−49

The sample procedure to be presented later is based on the data from Table 4.2 (small data sample Mann–Whitney *U*-test) near the beginning of this chapter.

The values from Table 4.2 are arranged in Table 4.3 so that the method 1 (X) scores are placed in order across the top and the method 2 (Y) scores are placed in order down the side. Then, the n_1n_2 differences are calculated by subtracting each Y value from each X value. The differences are shown in Table 4.3. Notice that the values of D_{ij} are ordered in the table from highest to lowest starting at the top right and ending at the bottom left.

We use Table B.4 in Appendix B to find the lower limit of the confidence interval, L, and the upper limit U. For a two-tailed test, L is the $w_{\alpha/2}$th smallest difference and U is the $w_{\alpha/2}$th largest difference that correspond to $\alpha/2$ for n_1 and n_2 for a confidence interval of $(1 - \alpha)$.

For our example, $n_1 = 7$ and $n_2 = 7$. For $\alpha/2 = 0.05/2 = 0.025$, Table B.4 returns $w_{\alpha/2} = 9$. This means that the ninth values from the top and bottom mark the limits of the 95% confidence interval on both ends. Therefore, $L = 19$ and $U = 36$. Based on these results, we are 95% certain that the difference in population median is between 18 and 36.

4.3.3 Sample Mann–Whitney *U*-Test (Large Data Samples)

The previous comparison of teaching methods for reading recovery was repeated with 5th-grade students. The 5th-grade used the same two methods. Method 1 was a pull-out program in which the children were taken out of the classroom for 30 min a day, 4 days a week. Method 2 was a small group program in which children were taught in groups of four or five for 45 min a day in the classroom, 4 days a week. The students were tested using the same reading comprehension test after 4 weeks of the program. The test results are shown in Table 4.4.

4.3.3.1 State the Null and Research Hypotheses The null hypothesis states that there is no tendency of the ranks of one method to be systematically higher or lower than the other. The hypothesis is stated in terms of comparison of distributions, not means. The research hypothesis states that the ranks of one method are

TABLE 4.4

Method 1	Method 2
48	14
40	18
39	20
50	10
41	12
38	102
71	21
30	19
15	100
33	23
47	16
51	82
60	13
59	25
58	24
42	97
11	28
46	9
36	34
27	52
93	70
72	22
57	26
45	8
53	17

systematically higher or lower than the other. Our research hypothesis is a two-tailed, nondirectional hypothesis because it indicates a difference, but in no particular direction.

The null hypothesis is

> H_O: There is no tendency for ranks of one method to be significantly higher (or lower) than the other.

The research hypothesis is

> H_A: The ranks of one method are systematically higher (or lower) than the other.

4.3.3.2 Set the Level of Risk (or the Level of Significance) Associated with the Null Hypothesis The level of risk, also called an alpha (α), is frequently set at 0.05. We will use $\alpha = 0.05$ in our example. In other words, there is a 95% chance that any observed statistical difference will be real and not due to chance.

4.3.3.3 Choose the Appropriate Test Statistic The data are obtained from two independent, or unrelated, samples of 5th-grade children being taught reading. Since we are comparing two unrelated, or independent, samples, we will use the Mann−Whitney *U*-test.

4.3.3.4 Compute the Test Statistic First, combine and rank both data samples together (see Table 4.5). Next, compute the sum of ranks for each method. Method 1 is ΣR_1 and method 2 is ΣR_2. Using Table 4.5,

TABLE 4.5

Ordered scores

Rank	Score	Sample
1	8	Method 2
2	9	Method 2
3	10	Method 2
4	11	Method 1
5	12	Method 2
6	13	Method 2
7	14	Method 2
8	15	Method 1
9	16	Method 2
10	17	Method 2
11	18	Method 2
12	19	Method 2
13	20	Method 2
14	21	Method 2
15	22	Method 2
16	23	Method 2
17	24	Method 2
18	25	Method 2
19	26	Method 2
20	27	Method 1
21	28	Method 2
22	30	Method 1
23	33	Method 1
24	34	Method 2
25	36	Method 1
26	38	Method 1
27	39	Method 1
28	40	Method 1
29	41	Method 1
30	42	Method 1
31	45	Method 1

(*Continued*)

TABLE 4.5 (*Continued*)

Ordered scores		
Rank	Score	Sample
32	46	Method 1
33	47	Method 1
34	48	Method 1
35	50	Method 1
36	51	Method 1
37	52	Method 2
38	53	Method 1
39	57	Method 1
40	58	Method 1
41	59	Method 1
42	60	Method 1
43	70	Method 2
44	71	Method 1
45	72	Method 1
46	82	Method 2
47	93	Method 1
48	97	Method 2
49	100	Method 2
50	102	Method 2

$$\sum R_1 = 779$$

and

$$\sum R_2 = 496$$

Now, compute the U-value for each sample. For sample 1,

$$\begin{aligned} U_1 &= n_1 n_2 + \frac{n_1(n_1+1)}{2} - \sum R_1 \\ &= 25(25) + \frac{25(25+1)}{2} - 779 = 625 + 325 - 779 \\ U_1 &= 171 \end{aligned}$$

and for sample 2,

$$\begin{aligned} U_2 &= n_1 n_2 + \frac{n_2(n_2+1)}{2} - \sum R_2 \\ &= 25(25) + \frac{25(25+1)}{2} - 496 = 625 + 325 - 496 \\ U_2 &= 454 \end{aligned}$$

The Mann−Whitney U-test statistic is the smaller of U_1 and U_2. Therefore, $U = 171$.

Since our sample sizes are large, we will approximate them to a normal distribution. Therefore, we will find a z-score for our data using a normal approximation. We must find the mean $\overline{x}_U$ and the standard deviation s_U for the data:

$$\overline{x}_U = \frac{n_1 n_2}{2} = \frac{(25)(25)}{2}$$
$$\overline{x}_U = 312.5$$

and

$$s_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = \sqrt{\frac{(25)(25)(25+25+1)}{12}} = \sqrt{\frac{31{,}875}{12}}$$
$$s_U = 51.54$$

Next, we use the mean, standard deviation, and the U-test statistic to calculate a z-score. Remember, we are testing the hypothesis that there is no difference in the ranks of the scores for two different methods of reading instruction for 5th-grade students:

$$z^* = \frac{U_i - \overline{x}_U}{s_U} = \frac{171 - 312.5}{51.54}$$
$$z^* = -2.75$$

4.3.3.5 Determine the Value Needed for Rejection of the Null Hypothesis Using the Appropriate Table of Critical Values for the Particular Statistic Table B.1 in Appendix B is used to establish the critical region of z-scores. For a two-tailed test with $\alpha = 0.05$, we must not reject the null hypothesis if $-1.96 \leq z^* \leq 1.96$.

4.3.3.6 Compare the Obtained Value with the Critical Value We find that z^* is not within the critical region of the distribution, $-2.75 < -1.96$. Therefore, we reject the null hypothesis. This suggests a difference between method 1 and method 2.

4.3.3.7 Interpret the Results We rejected the null hypothesis, suggesting that a real difference exists between the two methods. In addition, since the sum of the ranks for method 1 (ΣR_1) was larger than method 2 (ΣR_2), we see that method 1 had significantly higher scores.

At this point, the analysis is limited to identifying the presence or absence of a significant difference between the groups. In other words, the statistical test's level of significance does not describe the strength of the treatment. The American Psychological Association (2001), however, has called for a measure of the strength called the *effect size*.

We can consider the ES for this large sample test to determine the degree of association between the groups. We can use Formula 4.5 to calculate the ES. For the example, $z = -2.75$ and $n = 50$:

$$ES = \frac{|z|}{\sqrt{n}} = \frac{|-2.75|}{\sqrt{50}}$$
$$ES = 0.39$$

Our ES for the sample difference is 0.39. This value indicates a medium−high level of association between the teaching methods for the reading recovery program with 5th graders.

4.3.3.8 Reporting the Results For this example, two methods were used to provide 5th-grade students with reading instruction. Method 1 involved a pull-out program and method 2 involved a small group program. Using the ranked reading comprehension test scores, the results indicated a significant difference between the two methods ($U = 171$, $n_1 = 25$, $n_2 = 25$, $p < 0.05$). The sum of ranks for method 1 ($\Sigma R_1 = 779$) was larger than the sum of ranks for method 2 ($\Sigma R_2 = 496$). Moreover, the ES for the sample difference was 0.39. Therefore, we can state that the data support the pull-out program as a more effective reading program for teaching comprehension to 5th-grade children at this school.

4.4 COMPUTING THE KOLMOGOROV–SMIRNOV TWO-SAMPLE TEST STATISTIC

In Chapter 2, we used the Kolmogorov–Smirnov one-sample test to compare a sample with the normal distribution. We can use the Kolmogorov–Smirnov two-sample test to analyze two different data samples for independence. Our data must meet two assumptions.

1. Observations $X_1, \ldots, X_m$ are a random sample from a continuous population 1, where the X-values are mutually independent and identically distributed. Likewise, observations $Y_1, \ldots, Y_n$ are a random sample from a continuous population 2, where the Y-values are mutually independent and identically distributed.
2. The two samples are independent.

We begin by placing the data in a form that will permit us to compute the two-sided Kolmogorov–Smirnov test statistic Z. The first step in this procedure is to find the empirical distribution functions $F_m(t)$ and $G_n(t)$ for the samples of X and Y, respectively. Combine and rank order both sets of values. For every real number t, let

$$F_m(t) = \frac{\text{number of observed } X\text{'s} \leq t}{m}$$

and

$$G_n(t) = \frac{\text{number of observed } Y\text{'s} \leq t}{n}$$

where m is the sample size of X and n the sample size of Y.

Next, use Formula 4.6 to find each absolute value divergence D between the empirical distributions functions:

$$D=|F_m(t)-G_n(t)| \tag{4.6}$$

Use the largest divergence $D_{\max}$ with Formula 4.7 to calculate the Kolmogorov–Smirnov test statistic Z:

$$Z=D_{\max}\sqrt{\frac{mn}{m+n}} \tag{4.7}$$

Then, use the Kolmogorov–Smirnov test statistic, Z, and the Smirnov (1948) formula (see Formula 4.8, Formula 4.9, Formula 4.10, Formula 4.11, Formula 4.12, and Formula 4.13) to find the two-tailed probability estimate p. This is the same procedure shown in Chapter 2 when we performed the Kolmogorov–Smirnov one-sample test:

$$\text{if } 0\le Z<0.27, \text{ then } p=1 \tag{4.8}$$

$$\text{if } 0.27\le Z<1, \text{ then } p=1-\frac{2.506628}{Z}(Q+Q^9+Q^{25}) \tag{4.9}$$

where

$$Q=e^{-1.233701Z^{-2}} \tag{4.10}$$

$$\text{if } 1\le Z<3.1, \text{ then } p=2(Q-Q^4+Q^9-Q^{16}) \tag{4.11}$$

where

$$Q=e^{-2Z^2} \tag{4.12}$$

$$\text{if } Z\ge 3.1, \text{ then } p=0 \tag{4.13}$$

Once we have our p-value, we can compare it against our level of risk α to determine if the two samples are significantly different.

4.4.1 Sample Kolmogorov–Smirnov Two-Sample Test

We will use the data from Section 4.3.1 to demonstrate the Kolmogorov–Smirnov two-sample test. Table 4.6 recalls the data from the study involving reading recovery in the 4th grade. Method 1 was a program in which children were taken out of the

TABLE 4.6

Method 1	Method 2
48	14
40	18
39	20
50	10
41	12
38	102
53	17

classroom for 30 min a day, Monday through Thursday each week. Method 2 was a small group program in which the children were taught in groups of no more than five for 45 min a day in the classroom. These small classes were taught Monday through Thursday, also. The students were tested using a reading comprehension test after 4 weeks of instruction.

4.4.1.1 State the Null and Alternate Hypotheses

Let $X_1, \ldots, X_m$, and $Y_1, \ldots, Y_n$ be independent random samples. The null hypothesis indicates that there is no difference between the reading groups X and Y. Our research hypothesis is a two-tailed, nondirectional hypothesis because it indicates a difference, but in no particular direction.

The null hypothesis is

H_O: $[F(t) = G(t)$, for every $t]$

The research hypothesis is

H_A: $[F(t) \neq G(t)$ for at least one value of $t]$

4.4.1.2 Set the Level of Risk (or the Level of Significance) Associated with the Null Hypothesis

We will use $\alpha = 0.05$ in our example. In other words, there is a 95% chance that any observed statistical difference will be real and not due to chance.

4.4.1.3 Choose the Appropriate Test Statistic

We are seeking to compare two random samples, X and Y. Each sample is mutually independent and identically distributed. The X's and Y's are mutually independent. The Kolmogorov–Smirnov two-sample test will provide this comparison.

4.4.1.4 Compute the Test Statistic

Begin by computing the empirical distribution functions for the X and Y samples:

$$F_m(t) = \frac{\text{number of observed } X\text{'s} \leq t}{m}$$

and

$$G_n(t) = \frac{\text{number of observed } Y\text{'s} \leq t}{n}$$

where $m = 7$ and $n = 7$.

We use the data in Table 4.6 and Formula 4.6 to find each divergence and generate Table 4.7.

Next, we find the largest divergence $D_{\max}$. Table 4.7 shows that $D_{\max} = 6/7 = 0.86$. Now, we use Formula 4.7 to calculate the Kolmogorov–Smirnov test statistic Z:

$$Z = D_{\max}\sqrt{\frac{mn}{m+n}} = (0.86)\cdot\sqrt{\frac{(7)(7)}{7+7}} = (0.86)\cdot\sqrt{3.5} = (0.86)(1.87)$$

$$Z = 1.604$$

TABLE 4.7

	Z_i	$F_7(Z_i)$	$G_7(Z_i)$	$\lvert F_7(Z_i) - G_7(Z_i)\rvert$
1	10	0/7	1/7	1/7
2	12	0/7	2/7	2/7
3	14	0/7	3/7	3/7
4	17	0/7	4/7	4/7
5	18	0/7	5/7	5/7
6	20	0/7	6/7	6/7
7	38	1/7	6/7	5/7
8	39	2/7	6/7	4/7
9	40	3/7	6/7	3/7
10	41	4/7	6/7	2/7
11	48	5/7	6/7	1/7
12	50	6/7	6/7	0/7
13	53	7/7	6/7	1/7
14	102	7/7	7/7	0/7

4.4.1.5 *Determine the* p*-Value Associated with the Test Statistic*

Now, we find the p-value using Formula 4.11 since they satisfy the condition that $1 \leq Z < 3.1$. We first need Q using Formula 4.12:

$$Q = e^{-2Z^2} = e^{(-2)(1.604)^2} = e^{-5.146}$$
$$Q = 0.0058$$

Now, we can use Formula 4.11:

$$\begin{aligned} p = 2(Q - Q^4 + Q^9 - Q^{16}) &= (2)(0.0058 - 0.0058^4 + 0.0058^9 - 0.0058^{16}) \\ &= (2)(0.0058) \end{aligned}$$
$$p = 0.012$$

4.4.1.6 *Compare the Obtained Value with the Critical Value Needed for Rejection of the Null Hypothesis*

The two-tailed probability, $p = 0.012$, was computed and is now compared with the level of risk specified earlier, $\alpha = 0.05$. If α is greater than the p-value, we must reject the null hypothesis. If α is less than the p-value, we must not reject the null hypothesis. Since α is greater than the p-value ($0.05 > 0.012$), we reject the null hypothesis.

4.4.1.7 *Interpret the Results*

We rejected the null hypothesis, suggesting that the two methods for teaching reading recovery have significantly different effects on the learning of students. In studying the results, it appears that method 1 was more effective than method 2, in general.

4.4.1.8 Reporting the Results When reporting the results from the Kolmogorov–Smirnov two-sample test, include such information as the sample sizes for each group, the D statistic, and the p-value's relation to α.

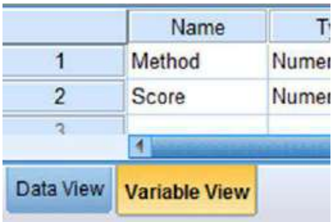
For this example, two methods were used to provide students with reading instruction. Method 1 involved a pull-out program and method 2 involved a small group program. Both methods include seven participants. The results from the Kolmogorov–Smirnov two-sample test ($D = 0.857$, $p < 0.05$) indicate a significant difference between the two methods. Therefore, we can state that the data support the pull-out program as a more effective reading program for teaching comprehension to 4th-grade children at this school.

4.5 PERFORMING THE MANN–WHITNEY *U*-TEST AND THE KOLMOGOROV–SMIRNOV TWO-SAMPLE TEST USING SPSS

We will analyze the data from the example in Sections 4.3.1 and 4.4.1 using SPSS.

4.5.1 Define Your Variables

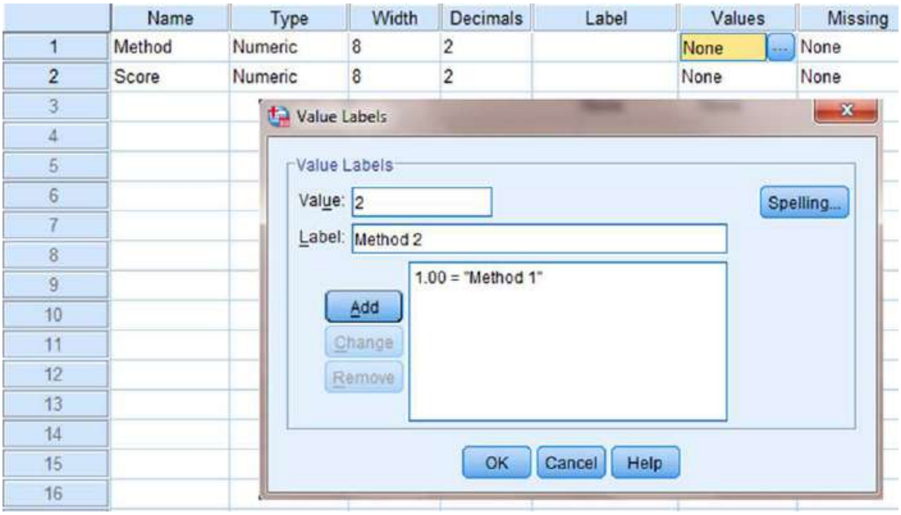
First, click the "Variable View" tab at the bottom of your screen. Then, type the names of your variables in the "Name" column. Unlike the related samples described in Chapter 2, you cannot simply enter each unrelated samples into a separate column to execute the Mann–Whitney U-test or Kolmogorov–Smirnov two-sample test. You must use a grouping variable to distinguish each sample. As shown in Figure 4.2, the first variable is the grouping variable that we called "Method." The second variable that we called "Score" will have our actual values.



	Name	T
1	Method	Numer
2	Score	Numer
3		

FIGURE 4.2

When establishing a grouping variable, it is often easiest to assign each group a whole number value. In our example, our groups are "Method 1" and "Method 2." Therefore, we must set our grouping variables for the variable "Method." First, we selected the "Values" column and clicked the gray square, as shown in Figure 4.3. Then, we set a value of 1 to equal "Method 1." Now, as soon as we click the "Add" button, we will have set "Method 2" equal to 2 based on the values we inserted above.



	Name	Type	Width	Decimals	Label	Values	Missing
1	Method	Numeric	8	2		None	None
2	Score	Numeric	8	2		None	None

Value Labels

Value Labels

Value: 2

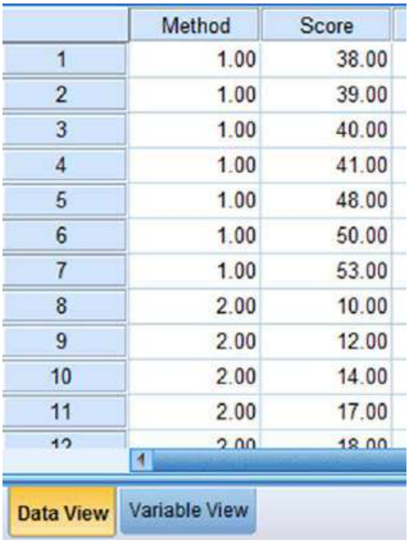
Label: Method 2

1.00 = "Method 1"

FIGURE 4.3

4.5.2 Type in Your Values

Click the “Data View” tab at the bottom of your screen as shown in Figure 4.4. Type in the values for both sets of data in the “Score” column. As you do so, type in the corresponding grouping variable in the “Method” column. For example, all of the values for “Method 2” are signified by a value of 2 in the grouping variable column that we called “Method.”



	Method	Score
1	1.00	38.00
2	1.00	39.00
3	1.00	40.00
4	1.00	41.00
5	1.00	48.00
6	1.00	50.00
7	1.00	53.00
8	2.00	10.00
9	2.00	12.00
10	2.00	14.00
11	2.00	17.00
12	2.00	18.00

FIGURE 4.4

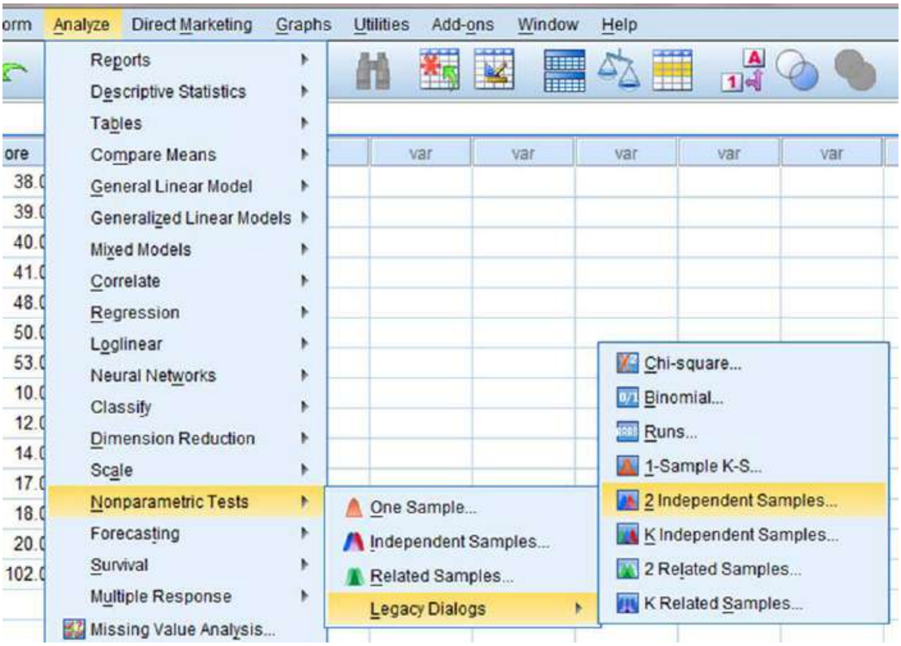


FIGURE 4.5

4.5.3 Analyze Your Data

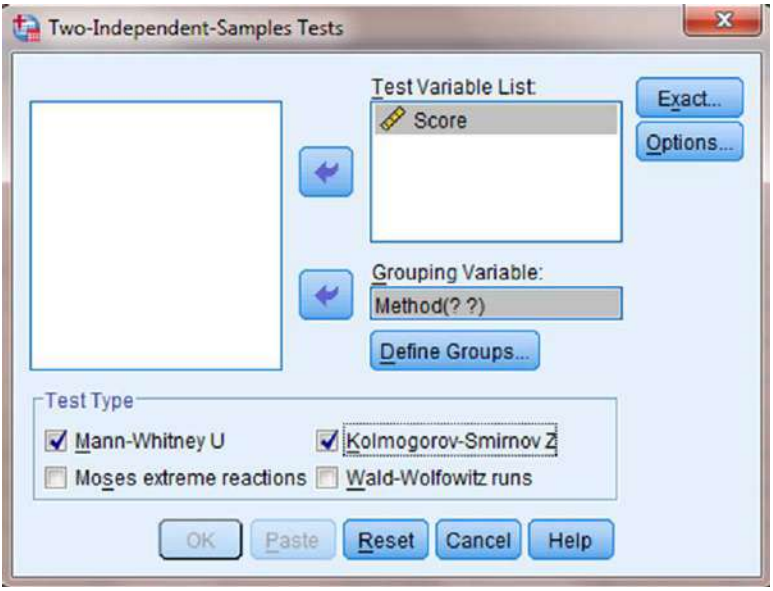
As shown in Figure 4.5, use the pull-down menus to choose “Analyze,” “Nonparametric Tests,” “Legacy Dialogs,” and “2 Independent Samples. . . .”

Use the top arrow button to place your variable with your data values, or dependent variable (DV), in the box labeled “Test Variable List:.” Then, use the lower arrow button to place your grouping variable, or independent variable (IV), in the box labeled “Grouping Variable.” As shown in Figure 4.6, we have placed the “Score” variable in the “Test Variable List” and the “Method” variable in the “Grouping Variable” box. Click on the “Define Groups . . .” button to assign a reference value to your IV (i.e., “Grouping Variable”).

As shown in Figure 4.7, type 1 into the box next to “Group 1:” and 2 in the box next to “Group 2:.” Then, click “Continue.” This step references the value labels you created when you defined your grouping variable in step 1. Now that the groups have been assigned, click “OK” to perform the analysis.

4.5.4 Interpret the Results from the SPSS Output Window

We first compare the samples with the Mann–Whitney U-test. SPSS Output 4.1 provides the sum of ranks and sample sizes for comparing the two groups. The second output table provides the Mann–Whitney U-test statistic ($U = 7.0$). As described in Figure 4.2, it also returns a similar nonparametric statistic called the Wilcoxon W-test statistic ($W = 35.0$). Notice that the Wilcoxon W is the smaller of the two rank sums in the table earlier.


Two-Independent-Samples Tests

Test Variable List: Score

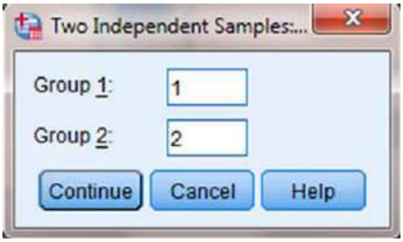
Grouping Variable: Method(? ?)

Test Type

☑ Mann-Whitney U ☑ Kolmogorov-Smirnov Z

□ Moses extreme reactions □ Wald-Wolfowitz runs

FIGURE 4.6


Two Independent Samples:...

Group 1: 1

Group 2: 2

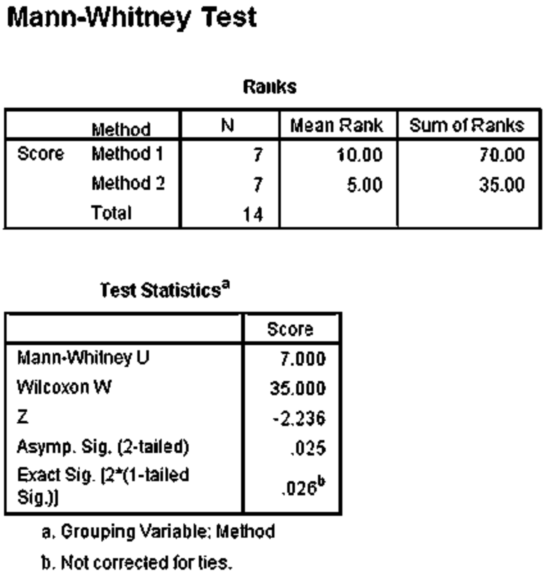
FIGURE 4.7

SPSS returns the critical *z*-score for large samples. In addition, SPSS calculates the two-tailed significance using two methods. The asymptotic significance is more appropriate with large samples. However, the exact significance is more appropriate with small samples or data that do not resemble a normal distribution.

Based on the results from SPSS, the ranked reading comprehension test scores of the two methods were significantly different ($U = 7$, $n_1 = 7$, $n_2 = 7$, $p < 0.05$). The sum of ranks for method 1 ($\Sigma R_1 = 70$) was larger than the sum of ranks for method 2 ($\Sigma R_2 = 35$).

Next, we analyzed the data with the Kolmogorov–Smirnov two-sample test. SPSS Output 4.2 provides the most extreme differences, $D_{\max} = 0.857$. The second output table provides the Kolmogorov–Smirnov two-sample test statistic, $Z = 1.604$, and the two-tailed significance, $p = 0.012$.

The results from the Kolmogorov–Smirnov two-sample test ($D = 0.857$, $p < 0.05$) indicate a significant difference between the two methods. Therefore, we



Mann-Whitney Test

Ranks

	Method	N	Mean Rank	Sum of Ranks
Score	Method 1	7	10.00	70.00
	Method 2	7	5.00	35.00
	Total	14		

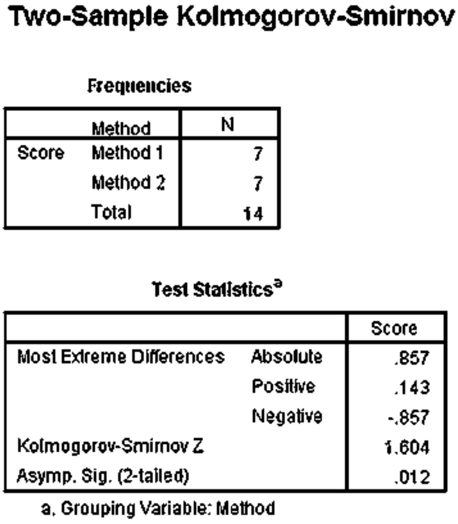
Test Statistics[a]

	Score
Mann-Whitney U	7.000
Wilcoxon W	35.000
Z	-2.236
Asymp. Sig. (2-tailed)	.025
Exact Sig. [2*(1-tailed Sig.)]	.026[b]

a. Grouping Variable: Method

b. Not corrected for ties.

SPSS OUTPUT 4.1



Two-Sample Kolmogorov-Smirnov Test

Frequencies

	Method	N
Score	Method 1	7
	Method 2	7
	Total	14

Test Statistics[a]

		Score
Most Extreme Differences	Absolute	.857
	Positive	.143
	Negative	-.857
Kolmogorov-Smirnov Z		1.604
Asymp. Sig. (2-tailed)		.012

a. Grouping Variable: Method

SPSS OUTPUT 4.2

can state that the data support the pull-out program as a more effective reading program for teaching comprehension to 4th-grade children at this school.

4.6 EXAMPLES FROM THE LITERATURE

Listed are varied examples of the nonparametric procedures described in this chapter. We have summarized each study's research problem and researchers' rationale(s) for choosing a nonparametric approach. We encourage you to obtain these studies if you are interested in their results.

Odaci (2007) investigated depression, submissive social behaviors, and frequency of automatic negative thoughts in Turkish adolescents. Obese participants were compared with participants of normal weight. After the Shapiro–Wilk statistic revealed that the data were not normally distributed, Odaci applied a Mann–Whitney *U*-test to compare the groups.

Bryant and Trockel (1976) investigated the impact of stressful life events on undergraduate females' locus of control. The authors compared accrued life changing units for participants with internal control against external using the Mann–Whitney *U*-test. This nonparametric procedure was selected since the data pertaining to stressful life events were ordinal in nature.

Re et al. (2007) investigated the expressive writing of children with attention-deficit/hyperactivity disorder (ADHD). The authors used a Mann–Whitney *U*-test to compare students showing symptoms of ADHD behaviors with a control group of students not displaying such behaviors. After examining their data with a Kolmogorov–Smirnov test, the researchers chose the nonparametric procedure due to significant deviations in the data distributions.

In an effort to understand the factors that have motivated minority students to enter the social worker profession, Limb and Organista (2003) studied data from nearly 7000 students in California entering a social worker program. The authors used a Wilcoxon rank sum test to compare sums of student group ranks. They chose this nonparametric test due to a concern that statistical assumptions were violated regarding sample normality and homogeneity of variances.

Schulze and Tomal (2006) examined classroom climate perceptions among undergraduate students. Since the student questionnaires used an interval scale, they analyzed their findings with a Mann–Whitney *U*-test.

Hegedus (1999) performed a pilot study to evaluate a scale designed to examine the caring behaviors of nurses. Care providers were compared with the consumers. She used a Wilcoxon rank sum test in her analysis because study participants were directed to rank the items on the scale.

The nature of expertise in astronomy was investigated across a broad spectrum of ages and experience (Bryce and Blown 2012). For each age and experience level, the researchers compared groups in New Zealand with respective groups in China using several Kolmogorov–Smirnov two-sample tests. In other words, each set of the two independent samples were from New Zealand versus China. The researchers chose a nonparametric procedure since their data were categorized with an ordinal scale.

4.7 SUMMARY

Two samples that are not related may be compared using a nonparametric procedure. Examples include the Mann–Whitney *U*-test (or the Wilcoxon rank sum test) and the Kolmogorov–Smirnov two-sample test. The parametric equivalent to these tests is known as the *t*-test for independent samples.

In this chapter, we described how to perform and interpret the Mann–Whitney *U*-test and the Kolmogorov–Smirnov two-sample test. We demonstrated both small samples and large samples for each test. We also explained how to perform the procedures using SPSS. Finally, we offered varied examples of these nonparametric statistics from the literature. The next chapter will involve comparing more than two samples that are related.

4.8 PRACTICE QUESTIONS

1. The data in Table 4.8 were obtained from a reading-level test for 1st-grade children. Compare the performance gains of the two different methods for teaching reading.

TABLE 4.8

Method	Gain score	Method	Gain score
One on one	16	Small group	11
One on one	13	Small group	2
One on one	16	Small group	10
One on one	16	Small group	4
One on one	13	Small group	9
One on one	9	Small group	8
One on one	12	Small group	5
One on one	12	Small group	6
One on one	20	Small group	4
One on one	17	Small group	16

Use two-tailed Mann–Whitney *U* and Kolmogorov–Smirnov two-sample tests to determine which method was better for teaching reading. Set $\alpha = 0.05$. Report your findings.

2. A research study was conducted to see if an active involvement in a hobby had a positive effect on the health of a person who retires after age 65. The data in Table 4.9 describe the health (number of doctor visits in 1 year) for participants who are involved in a hobby almost daily and those who are not.

TABLE 4.9

No hobby group	Hobby group
12	9
15	5
8	10
11	3
9	4
17	2

Use one-tailed Mann–Whitney U and Kolmogorov–Smirnov two-sample tests to determine whether the hobby tends to reduce the need for doctor visits. Set $\alpha = 0.05$. Report your findings.

3. Table 4.10 shows assessment scores of two different classes who are being taught computer skills using two different methods.

TABLE 4.10

Method 1	Method 2
53	91
41	18
17	14
45	21
44	23
12	99
49	16
50	10

Use two-tailed Mann–Whitney U and Kolmogorov–Smirnov two-sample tests to determine which method was better for teaching computer skills. Set $\alpha = 0.05$. Report your findings.

4. Two methods of teaching reading were compared. Method 1 used the computer to interact with the student, and diagnose and remediate the student based on misconceptions. Method 2 was taught using workbooks in classroom groups. Table 4.11 shows the data obtained on an assessment after 6 weeks of instruction. Calculate the ES using the z-score from the analysis.

TABLE 4.11

Method 1	Method 2
27	9
38	42
15	21
85	83
36	110
95	19
93	29
57	40
63	30
81	23
65	18
77	32
59	101

(*Continued*)

TABLE 4.11 (*Continued*)

Method 1	Method 2
89	7
41	50
26	37
102	22
55	71
46	16
82	45
24	35
87	28
66	91
12	86
90	20

5. Two methods were used to provide instruction in science for 7th grade. Method 1 included a laboratory each week and method 2 had only classroom work with lecture and worksheets. Table 4.12 shows end-of-course test performance for the two methods. Construct a 95% median confidence interval based on the difference between two independent samples to compare the two methods.

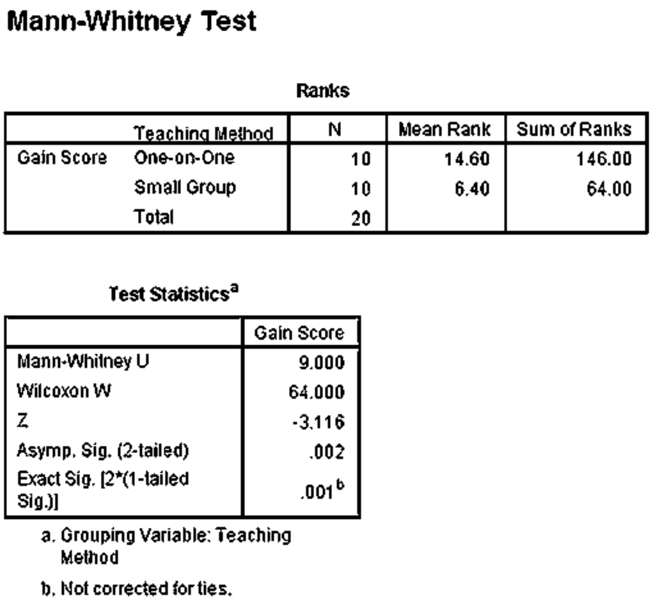
TABLE 4.12

Method 1	Method 2
15	8
23	15
9	10
12	13
18	17
22	5
17	18
20	7

4.9 SOLUTIONS TO PRACTICE QUESTIONS

1. The results from the analysis are displayed in SPSS Outputs 4.3 and 4.4.
The results from the Mann–Whitney U-test ($U = 9$, $n_1 = 10$, $n_2 = 10$, $p < 0.05$) indicated that the two methods were significantly different. Moreover, the one-on-one method produced a higher sum of ranks ($\Sigma R_1 = 146$) than the small group method ($\Sigma R_2 = 64$).

The results from the Kolmogorov–Smirnov two-sample test ($D = 1.789$, $p < 0.05$) also suggested that the two methods were significantly different.



Mann-Whitney Test

Ranks

	Teaching Method	N	Mean Rank	Sum of Ranks
Gain Score	One-on-One	10	14.60	146.00
	Small Group	10	6.40	64.00
	Total	20		

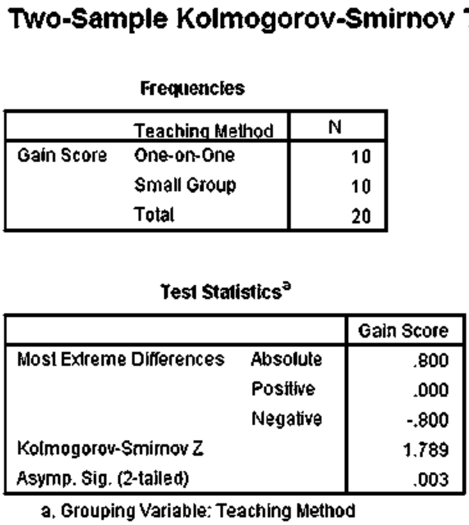
Test Statistics[a]

	Gain Score
Mann-Whitney U	9.000
Wilcoxon W	64.000
Z	-3.116
Asymp. Sig. (2-tailed)	.002
Exact Sig. [2*(1-tailed Sig.)]	.001[b]

a. Grouping Variable: Teaching Method

b. Not corrected for ties.

SPSS OUTPUT 4.3



Two-Sample Kolmogorov-Smirnov Test

Frequencies

	Teaching Method	N
Gain Score	One-on-One	10
	Small Group	10
	Total	20

Test Statistics[a]

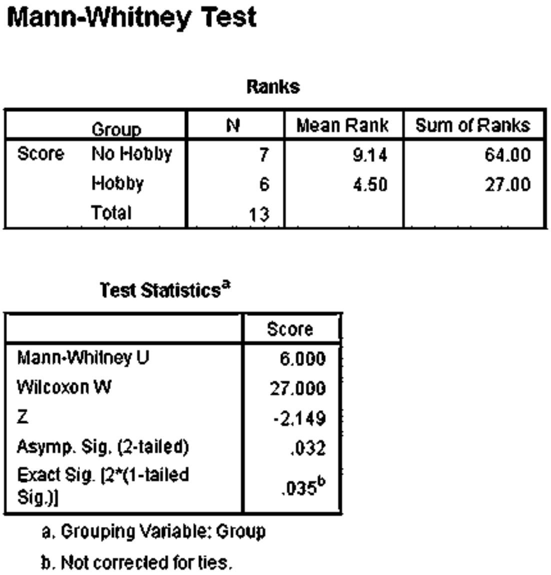
		Gain Score
Most Extreme Differences	Absolute	.800
	Positive	.000
	Negative	-.800
Kolmogorov-Smirnov Z		1.789
Asymp. Sig. (2-tailed)		.003

a. Grouping Variable: Teaching Method

SPSS OUTPUT 4.4

Therefore, based on both statistical tests, 1st-grade children displayed significantly higher reading levels when taught with a one-on-one method.

2. The results from the analysis are displayed in SPSS Outputs 4.5 and 4.6.
The results from the Mann–Whitney U-test ($U = 6$, $n_1 = 7$, $n_2 = 6$, $p < 0.05$) indicated that the two samples were significantly different. Moreover, the sample



Mann-Whitney Test

Ranks

	Group	N	Mean Rank	Sum of Ranks
Score	No Hobby	7	9.14	64.00
	Hobby	6	4.50	27.00
	Total	13		

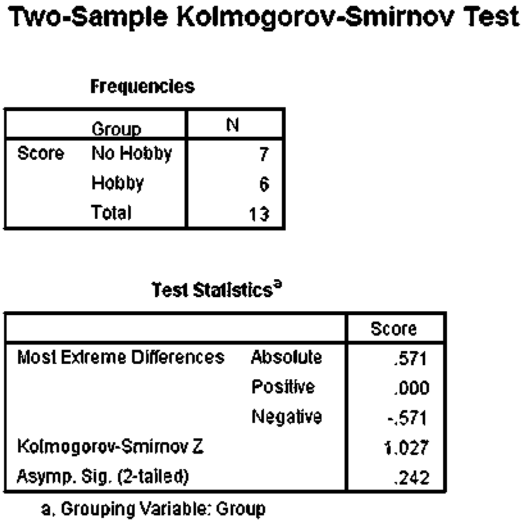
Test Statistics[a]

	Score
Mann-Whitney U	6.000
Wilcoxon W	27.000
Z	-2.149
Asymp. Sig. (2-tailed)	.032
Exact Sig. [2*(1-tailed Sig.)]	.035[b]

a. Grouping Variable: Group

b. Not corrected for ties.

SPSS OUTPUT 4.5



Two-Sample Kolmogorov-Smirnov Test

Frequencies

	Group	N
Score	No Hobby	7
	Hobby	6
	Total	13

Test Statistics[a]

		Score
Most Extreme Differences	Absolute	.571
	Positive	.000
	Negative	-.571
Kolmogorov-Smirnov Z		1.027
Asymp. Sig. (2-tailed)		.242

a. Grouping Variable: Group

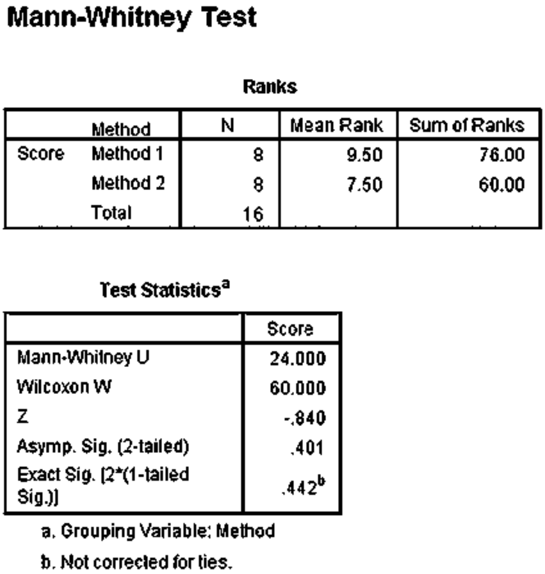
SPSS OUTPUT 4.6

with no hobby produced a higher sum of ranks ($\Sigma R_1 = 64$) than the sample with a hobby ($\Sigma R_2 = 27$).

The results from the Kolmogorov–Smirnov two-sample test ($D = 1.027$, $p > 0.05$) suggested, however, that the two methods were not significantly different.

The conflicting results from the two statistical tests prevent us from making a conclusive statement about this study. Study replication with larger sample sizes is recommended.

3. The results from the analysis are displayed in SPSS Outputs 4.7 and 4.8.



Mann-Whitney Test

Ranks

	Method	N	Mean Rank	Sum of Ranks
Score	Method 1	8	9.50	76.00
	Method 2	8	7.50	60.00
	Total	16		

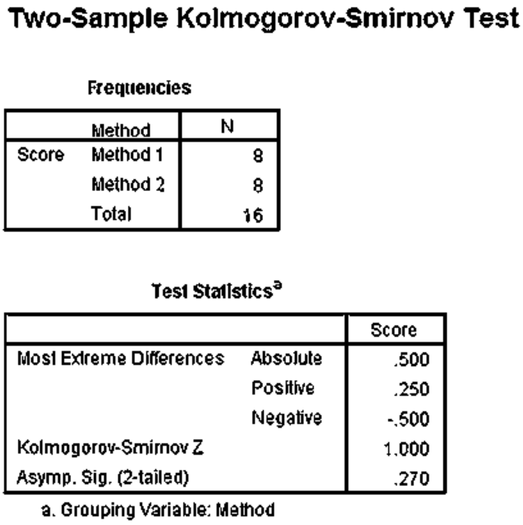
Test Statistics[a]

	Score
Mann-Whitney U	24.000
Wilcoxon W	60.000
Z	-.840
Asymp. Sig. (2-tailed)	.401
Exact Sig. [2*(1-tailed Sig.)]	.442[b]

a. Grouping Variable: Method

b. Not corrected for ties.

SPSS OUTPUT 4.7



Two-Sample Kolmogorov-Smirnov Test

Frequencies

	Method	N
Score	Method 1	8
	Method 2	8
	Total	16

Test Statistics[a]

		Score
Most Extreme Differences	Absolute	.500
	Positive	.250
	Negative	-.500
Kolmogorov-Smirnov Z		1.000
Asymp. Sig. (2-tailed)		.270

a. Grouping Variable: Method

SPSS OUTPUT 4.8

The results from the Mann–Whitney U-test ($U = 24$, $n_1 = 8$, $n_2 = 8$, $p > 0.05$) and the results from the Kolmogorov–Smirnov two-sample test ($D = 1.000$, $p > 0.05$) indicated that the two samples were not significantly different. Therefore, based on this study, neither method resulted in significantly different assessment scores for computer skills.

4. The results from the analysis are as follows:

$$U_1 = 199 \text{ and } U_2 = 426$$
$$x_u = 312.5$$
$$s_u = 51.54$$
$$z^* = -2.20$$
$$ES = 0.31$$

The ES is moderate.

5. For our example, $n_1 = 8$ and $n_2 = 8$. For 0.05/2 = 0.025, $w_{\alpha/2} = 14$. Based on these results, we are 95% certain that the median difference between the two methods is between 0 and 11.

CHAPTER 5

COMPARING MORE THAN TWO RELATED SAMPLES: THE FRIEDMAN TEST

5.1 OBJECTIVES

In this chapter, you will learn the following items:

- How to compute the Friedman test.
- How to perform contrasts to compare samples.
- How to perform the Friedman test and associated sample contrasts using SPSS®.

5.2 INTRODUCTION

Most public school divisions take pride in the percentage of their graduates admitted to college. A large school division might want to determine if these college admission rates are changing or stagnant. The division could compare the percentages of graduates admitted to college from each of its 10 high schools over the past 5 years. Each year would constitute a group, or sample, of percentages from each school. In other words, the study would include five groups, and each group would include 10 values.

The samples in the example are dependent, or related, since each school has a percentage for each year. The Friedman test is a nonparametric statistical procedure for comparing more than two samples that are related. The parametric equivalent to this test is the repeated measures analysis of variance (ANOVA).

When the Friedman test leads to significant results, then at least one of the samples is different from the other samples. However, the Friedman test does not identify where the difference(s) occur. Moreover, it does not identify how many differences occur. In order to identify the particular differences between sample pairs, a researcher might use sample contrasts, or *post hoc* tests, to analyze the

Nonparametric Statistics: A Step-by-Step Approach, Second Edition. Gregory W. Corder and Dale I. Foreman.

specific sample pairs for significant difference(s). The Wilcoxon signed rank test (see Chapter 3) is a useful method for performing sample contrasts between related sample sets.

In this chapter, we will describe how to perform and interpret a Friedman test followed with sample contrasts. We will also explain how to perform the procedures using SPSS. Finally, we offer varied examples of these nonparametric statistics from the literature.

5.3 COMPUTING THE FRIEDMAN TEST STATISTIC

The Friedman test is used to compare more than two dependent samples. When stating our hypotheses, we state them in terms of the population. Moreover, we examine the population medians, θ_i, when performing the Friedman test.

To compute the Friedman test statistic F_r, we begin by creating a table of our data. List the research subjects to create the rows. Place the values for each condition in columns next to the appropriate subjects. Then, rank the values for each subject across each condition. If there are no ties from the ranks, use Formula 5.1 to determine the Friedman test statistic F_r:

$$F_r = \left[\frac{12}{nk(k+1)}\sum_{i=1}^{k} R_i^2\right] - 3n(k+1) \tag{5.1}$$

where n is the number of rows, or subjects, k is the number of columns, or conditions, and R_i is the sum of the ranks from column, or condition, i.

If ranking of values results in any ties, use Formula 5.2 to determine the Friedman test statistic F_r:

$$F_r = \frac{n(k-1)\left[\sum_{i=1}^{k}\frac{R_i^2}{n} - C_F\right]}{\sum r_{ij}^2 - C_F} \tag{5.2}$$

where n is the number of rows, or subjects, k is the number of columns, or conditions, R_i is the sum of the ranks from column, or condition, i, C_F is the ties correction, $\frac{1}{4}nk(k+1)^2$, and r_{ij} is the rank corresponding to subject j in column i.

The degrees of freedom for the Friedman test is determined by using Formula 5.3:

$$df = k - 1 \tag{5.3}$$

Where df is the degrees of freedom and k is the number of groups.

Once the test statistic F_r is computed, it can be compared with a table of critical values (see Table B.5 in Appendix B) to examine the groups for significant differences. However, if the number of groups, k, or the number of values in a group, n, exceeds those available from the table, then a large sample approximation may be performed. Use a table with the χ^2 distribution (see Table B.2 in Appendix B) to obtain a critical value when performing a large sample approximation.

If the F_r statistic is not significant, then no differences exist between any of the related conditions. However, if the F_r statistic is significant, then a difference exists between at least two of the conditions. Therefore, a researcher might use sample contrasts between individual pairs of conditions, or *post hoc* tests, to determine which of the condition pairs are significantly different.

When performing multiple sample contrasts, the type I error rate tends to become inflated. Therefore, the initial level of risk, or α, must be adjusted. We demonstrate the Bonferroni procedure, shown in Formula 5.4, to adjust α:

$$\alpha_B = \frac{\alpha}{k} \tag{5.4}$$

where α_B is the adjusted level of risk, α is the original level of risk, and k is the number of comparisons.

5.3.1 Sample Friedman's Test (Small Data Samples without Ties)

A manager is struggling with the chronic tardiness of her seven employees. She tries two strategies to improve employee timeliness. First, over the course of a month, she punishes employees with a $10 paycheck deduction for each day that they arrive late. Second, the following month, she punishes employees by docking their pay $20 for each day that they do not arrive on time.

Table 5.1 shows the number of times each employee was late in a given month. The baseline shows the employees' monthly tardiness before the strategies. Month 1 shows the employees' monthly tardiness after a month of the $10 paycheck deductions. Month 2 shows the employees' monthly tardiness after a month of the $20 paycheck deductions.

TABLE 5.1

	Monthly tardiness		
Employee	Baseline	Month 1	Month 2
1	16	13	12
2	10	5	2
3	7	8	9
4	13	11	5
5	17	2	6
6	10	7	9
7	11	6	7

We want to determine if either of the paycheck deduction strategies reduced employee tardiness. Since the sample sizes are small ($n < 20$), we require a nonparametric test. The Friedman test is the best statistic to analyze the data and test the hypothesis.

5.3.1.1 State the Null and Research Hypotheses The null hypothesis states that neither of the manager's strategies will change employee tardiness. The research hypothesis states that one or both of the manager's strategies will reduce employee tardiness.

The null hypothesis is

$$H_O\colon \theta_B = \theta_{M1} = \theta_{M2}$$

The research hypothesis is

H_A: One or both of the manager's strategies will reduce employee tardiness.

5.3.1.2 Set the Level of Risk (or the Level of Significance) Associated with the Null Hypothesis The level of risk, also called an alpha (α), is frequently set at 0.05. We will use $\alpha = 0.05$ in our example. In other words, there is a 95% chance that any observed statistical difference will be real and not due to chance.

5.3.1.3 Choose the Appropriate Test Statistic The data are obtained from three dependent, or related, conditions that report employees' number of monthly tardiness. The three samples are small with some violations of our assumptions of normality. Since we are comparing three dependent conditions, we will use the Friedman test.

5.3.1.4 Compute the Test Statistic First, rank the values from each employee, or subject (see Table 5.2).

TABLE 5.2

	Ranks of monthly tardiness		
Employee	Baseline	Month 1	Month 2
1	3	2	1
2	3	2	1
3	1	2	3
4	3	2	1
5	3	1	2
6	3	1	2
7	3	1	2

Next, compute the sum of ranks for each condition. The ranks in each group are added to obtain a total R-value for the group.

For the baseline condition,

$$R_B = 3+3+1+3+3+3+3 = 19$$

For month 1,

$$R_{M1} = 2+2+2+2+1+1+1 = 11$$

For month 2,

$$R_{M2} = 1+1+3+1+2+2+2 = 12$$

These R-values are used to compute the F_r test statistic. Use Formula 5.1 since there were no ties involved in the ranking:

$$
\begin{aligned}
F_r &= \left[\frac{12}{nk(k+1)}\sum_{i=1}^{k} R_i^2\right] - 3n(k+1) \\
&= \left(\frac{12}{7(3)(3+1)}\right)\left(19^2+11^2+12^2\right) - 3(7)(3+1) \\
&= \left(\frac{12}{84}\right)(361+121+144) - 84 = (0.1429)(626) - 84 = 89.4286 - 84 \\
F_r &= 5.429
\end{aligned}
$$

5.3.1.5 Determine the Value Needed for Rejection of the Null Hypothesis Using the Appropriate Table of Critical Values for the Particular Statistic

We will use the critical value table for the Friedman test (see Table B.5 in Appendix B) since it includes the number of groups, k, and the number of samples, n, for our data. In this case, we look for the critical value for $k = 3$ and $n = 7$ with $\alpha = 0.05$. Table B.5 returns a critical value for the Friedman test of 7.14.

5.3.1.6 Compare the Obtained Value with the Critical Value

The critical value for rejecting the null hypothesis is 7.14 and the obtained value is $F_r = 5.429$. If the critical value is less than or equal to the obtained value, we must reject the null hypothesis. If instead, the critical value exceeds the obtained value, we do not reject the null hypothesis. Since the critical value exceeds the obtained value, we do not reject the null hypothesis.

5.3.1.7 Interpret the Results

We did not reject the null hypothesis, suggesting that no significant difference exists between any of the three conditions. Therefore, no further comparisons are necessary with these data.

5.3.1.8 Reporting the Results

The reporting of results for the Friedman test should include such information as the number of subjects, the F_r statistic, degrees of freedom, and p-value's relation to α.

For this example, the frequencies of employees' ($n = 7$) tardiness were compared over three conditions. The Friedman test was not significant ($F_{r(2)} = 5.429$, $p > 0.05$). Therefore, we can state that the data do not support punishing tardy employees with \$10 or \$20 paycheck deductions.

5.3.2 Sample Friedman's Test (Small Data Samples with Ties)

After the manager's failure to reduce employee tardiness with paycheck deductions, she decided to try a different approach. This time, she rewarded employees when they arrived to work on-time. Again, she tries two strategies to improve employee timeliness. First, over the course of a month, she rewards employees with a \$10 bonus for each day that they arrive on-time. Second, the following month, she rewards employees with a \$20 bonus for each day that they arrive on-time.

TABLE 5.3

	Monthly tardiness		
Employee	Baseline	Month 1	Month 2
1	16	17	11
2	10	5	2
3	7	8	0
4	13	9	5
5	17	2	2
6	10	10	9
7	11	6	5

Table 5.3 shows the number of times each employee was late in a given month. The baseline shows the employees' monthly tardiness before any of the strategies in either example. Month 1 shows the employees' monthly tardiness after a month of the \$10 bonuses. Month 2 shows the employees' monthly tardiness after a month of the \$20 bonuses.

We want to determine if either of the strategies reduced employee tardiness. Again, since the sample sizes are small ($n < 20$), we use a nonparametric test. The Friedman test is a good statistic to analyze the data and test the hypothesis.

5.3.2.1 State the Null and Research Hypotheses

The null hypothesis states that neither of the manager's strategies will change employee tardiness. The research hypothesis states that one or both of the manager's strategies will reduce employee tardiness.

The null hypothesis is

$$H_O\colon \theta_B = \theta_{M1} = \theta_{M2}$$

The research hypothesis is

H_A: One or both of the manager's strategies will reduce employee tardiness.

5.3.2.2 Set the Level of Risk (or the Level of Significance) Associated with the Null Hypothesis

The level of risk, also called an alpha (α), is frequently set at 0.05. We will use $\alpha = 0.05$ in our example. In other words, there is a 95% chance that any observed statistical difference will be real and not due to chance.

5.3.2.3 Choose the Appropriate Test Statistic

The data are obtained from three dependent, or related, conditions that report employees' number of monthly tardiness. The three samples are small with some violations of our assumptions of normality. Since we are comparing three dependent conditions, we will use the Friedman test.

5.3.2.4 Compute the Test Statistic

First, rank the values from each employee, or subject (see Table 5.4).

TABLE 5.4

	Ranks of monthly tardiness		
Employee	Baseline	Month 1	Month 2
1	2	3	1
2	3	2	1
3	2	3	1
4	3	2	1
5	3	1.5	1.5
6	2.5	2.5	1
7	3	2	1

Next, compute the sum of ranks for each condition. The ranks in each group are added to obtain a total R-value for the group.

For the baseline condition,

$$R_B = 2+3+2+3+3+2.5+3 = 18.5$$

For month 1,

$$R_{M1} = 3+2+3+2+1.5+2.5+2 = 16$$

For month 2,

$$R_{M2} = 1+1+1+1+1.5+1+1 = 7.5$$

These R-values are used to compute the F_r test statistic. Since there were ties involved in the rankings, we must use Formula 5.2. Finding the values for C_F and Σr_{ij}^2 first will simplify the calculation:

$$C_F = \frac{1}{4}nk(k+1)^2 = \left(\frac{1}{4}\right)(7)(3)(3+1)^2$$

$$C_F = 84$$

To find Σr_{ij}^2, square all of the ranks. Then, add all of the squared ranks together (see Table 5.5):

$$\Sigma r_{ij}^2 = 50.25+38.50+8.25$$

$$\Sigma r_{ij}^2 = 97.0$$

Now that we have C_F and Σr_{ij}^2, we are ready for Formula 5.2:

$$F_r = \frac{n(k-1)\left[\sum_{i=1}^{k}\frac{R_i^2}{n} - C_F\right]}{\sum r_{ij}^2 - C_F} = \frac{7(3-1)\left[\frac{18.5}{7}+\frac{16.0}{7}+\frac{7.5}{7}-84\right]}{97-84}$$

$$= \frac{7(2)[48.89+36.57+8.04-84]}{13} = \frac{7(2)9.5}{13}$$

$$F_r = 10.23$$

TABLE 5.5

	Ranks of monthly tardiness		
Employee	Baseline	Month 1	Month 2
1	4	9	1
2	9	4	1
3	4	9	1
4	9	4	1
5	9	2.25	2.25
6	6.25	6.25	1
7	9	4	1
Σr_i^2	50.25	38.50	8.25

5.3.2.5 Determine the Value Needed for Rejection of the Null Hypothesis Using the Appropriate Table of Critical Values for the Particular Statistic We will use the critical value table for the Friedman test (see Table B.5 in Appendix B) since it includes the number of groups, k, and the number of samples, n, for our data. In this case, we look for the critical value for $k = 3$ and $n = 7$ with $\alpha = 0.05$. Table B.5 returns a critical value for the Friedman test of 7.14.

5.3.2.6 Compare the Obtained Value with the Critical Value The critical value for rejecting the null hypothesis is 7.14 and the obtained value is $F_r = 10.23$. If the critical value is less than or equal to the obtained value, we must reject the null hypothesis. If instead, the critical value exceeds the obtained value, we do not reject the null hypothesis. Since the obtained value exceeds the critical value, we reject the null hypothesis.

5.3.2.7 Interpret the Results We rejected the null hypothesis, suggesting that a significant difference exists between one or more of the three conditions. In particular, both strategies seemed to result in less tardiness among employees. However, describing specific differences in this manner is speculative. Therefore, we need a technique for statistically identifying difference between conditions, or contrasts.

Sample Contrasts, or Post Hoc Tests The Friedman test identifies if a statistical difference exists; however, it does not identify how many differences exist and which conditions are different. To identify which conditions are different and which are not, we use a procedure called contrasts or *post hoc* tests. An appropriate test to use when comparing two related samples at a time is the Wilcoxon signed rank test described in Chapter 3.

It is important to note that performing several two-sample tests has a tendency to inflate the type I error rate. In our example, we would compare three groups, $k = 3$. At an $\alpha = 0.05$, the type I error rate would equal $1 - (1 - 0.05)^3 = 0.14$.

To compensate for this error inflation, we demonstrate the Bonferroni procedure (see Formula 5.4). With this technique, we use a corrected α with the Wilcoxon signed rank tests to determine significant differences between conditions. For our example, we are only comparing month 1 and month 2 against the baseline. We are not comparing month 1 against month 2. Therefore, we are only making two comparisons and $k = 2$:

$$\alpha_B = \frac{\alpha}{k} = \frac{0.05}{2}$$
$$\alpha_B = 0.025$$

When we compare the three samples with the Wilcoxon signed rank tests using α_B, we obtain the results presented in Table 5.6. Notice that the significance is one-tailed, or directional, since we were seeking a decline in tardiness.

TABLE 5.6

Condition comparison	Wilcoxon T statistic	Rank sum difference	One-tailed significance
Baseline–Month 1	3.0	18.0 − 3.0 = 15.0	0.057
Baseline–Month 2	0.0	28.0 − 0.0 = 28.0	0.009

Using $\alpha_B = 0.025$, we notice that the baseline–month 1 comparison does not demonstrate a significant difference ($p > 0.025$). However, the baseline–month 2 comparison does demonstrate a significant difference ($p < 0.025$). Therefore, the data indicate that the \$20 bonus reduces tardiness while the \$10 bonus does not.

Note that if you are not comparing all of the samples for the Friedman test, then k is only the number of comparisons you are making with the Wilcoxon signed rank tests. Therefore, comparing fewer samples will increase the chances of finding a significant difference.

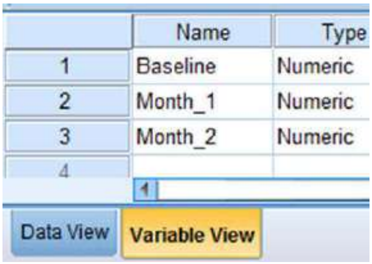
5.3.2.8 Reporting the Results The reporting of results for the Friedman test should include such information as the number of subjects, the F_r statistic, degrees of freedom, and p-value's relation to α.

For this example, the frequencies of employees' ($n = 7$) tardiness were compared over three conditions. The Friedman test was significant ($F_{r(2)} = 10.23$, $p < 0.05$). In addition, follow-up contrasts using Wilcoxon signed rank tests revealed that \$20 bonus reduces tardiness, while the \$10 bonus does not.

5.3.3 Performing the Friedman Test Using SPSS

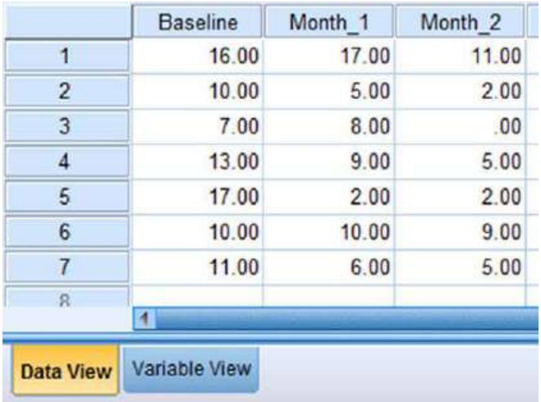
We will analyze the data from the example earlier using SPSS.

5.3.3.1 Define Your Variables First, click the "Variable View" tab at the bottom of your screen. Then, type the names of your variables in the "Name"



	Name	Type
1	Baseline	Numeric
2	Month_1	Numeric
3	Month_2	Numeric
4		

FIGURE 5.1



	Baseline	Month_1	Month_2
1	16.00	17.00	11.00
2	10.00	5.00	2.00
3	7.00	8.00	.00
4	13.00	9.00	5.00
5	17.00	2.00	2.00
6	10.00	10.00	9.00
7	11.00	6.00	5.00
8			

FIGURE 5.2

column. As shown in Figure 5.1, we have named our variables “Baseline,” “Month_1,” and “Month_2.”

5.3.3.2 Type in Your Values Click the “Data View” tab at the bottom of your screen and type your data under the variable names. As shown in Figure 5.2, we are comparing “Baseline,” “Month_1,” and “Month_2.”

5.3.3.3 Analyze Your Data As shown in Figure 5.3, use the pull-down menus to choose “Analyze,” “Nonparametric Tests,” “Legacy Dialogs,” and “K Related Samples. . . .”

Select each of the variables that you want to compare and click the button in the middle to move it to the “Test Variables:” box as shown in Figure 5.4. Notice that the “Friedman” box is checked by default. After the variables are in the “Test Variables:” box, click “OK” to perform the analysis.

5.3.3.4 Interpret the Results from the SPSS Output Window The first output table in SPSS Output 5.1 provides the mean ranks of each condition. The second output table provides the Friedman test statistic, 10.231. Since this test uses

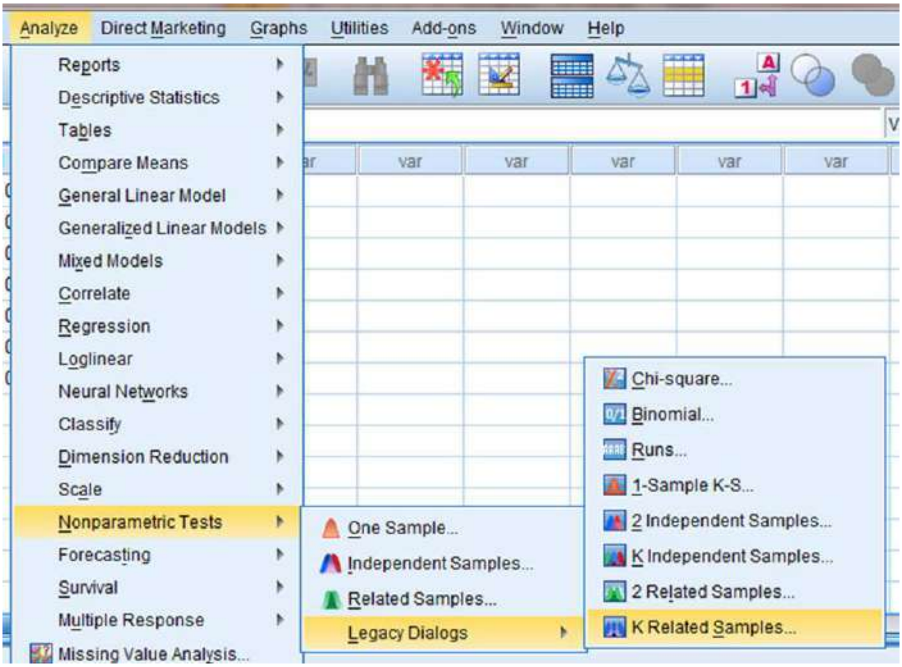
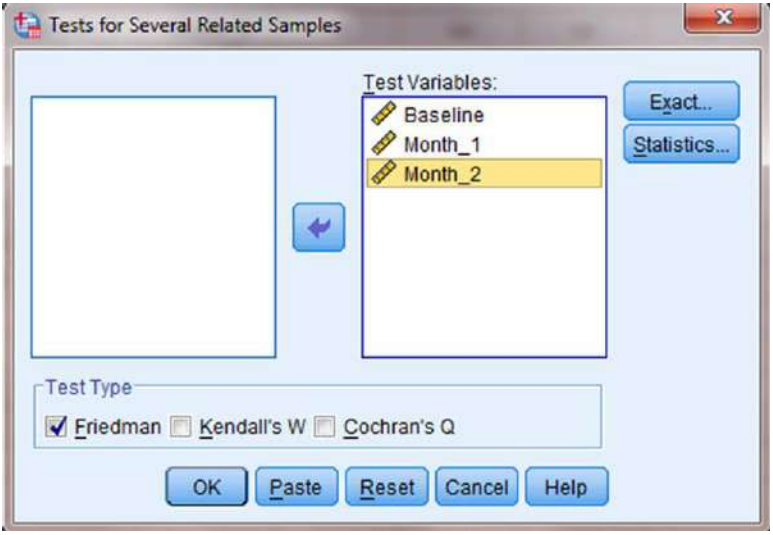


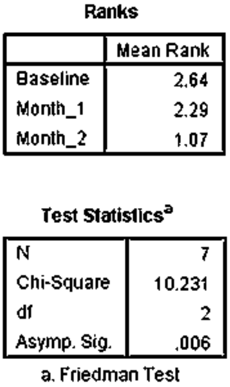
FIGURE 5.3



Test Variables:
Baseline
Month_1
Month_2

Test Type
☑ Friedman □ Kendall's W □ Cochran's Q

FIGURE 5.4



Ranks

	Mean Rank
Baseline	2.64
Month_1	2.29
Month_2	1.07

Test Statistics[a]

N	7
Chi-Square	10.231
df	2
Asymp. Sig.	.006

a. Friedman Test

SPSS OUTPUT 5.1

a χ^2 distribution, SPSS calls the F_r statistic "Chi-Square." This table also returns the number of subjects ($n = 7$) degrees of freedom ($df = 2$) and the significance ($p = 0.006$).

Based on the results from SPSS, three conditions were compared among employees ($n = 7$). The Friedman test was significant ($F_{r(2)} = 10.23$, $p < 0.05$). In order to compare individual pairs of conditions, contrasts may be used.

Note that to perform Wilcoxon signed rank tests for sample contrasts, remember to use your corrected level of risk, α_B, when examining your significance.

5.3.4 Sample Friedman's Test (Large Data Samples without Ties)

After hearing of the manager's success, the head office transferred her to a larger branch office. The transfer was strategic because this larger branch is dealing with tardiness issues among employees. The manager suggests that she use the same successful incentives for employee timeliness. Due to financial limitations, however, she is limited to offering employees smaller bonuses. First, over the course of a month, she rewards employees with a $2 bonus for each day that they arrive on-time. Second, the following month, she rewards employees with a $5 bonus for each day that they arrive on-time.

Table 5.7 shows the number of times each employee was late in a given month. The baseline shows the employees' monthly tardiness before any of the strategies in either example. Month 1 shows the employees' monthly tardiness after a month with $2 bonuses. Month 2 shows the employees' monthly tardiness after a month with $5 bonuses.

We want to determine if either of the paycheck bonus strategies reduced employee tardiness. Since the sample sizes are large ($n > 20$), we will use χ^2 for the critical value. The Friedman test is a good statistic to analyze the data and test the hypothesis.

5.3.4.1 State the Null and Research Hypotheses The null hypothesis states that neither of the manager's strategies will change employee tardiness. The research

TABLE 5.7

	Monthly tardiness		
Employee	Baseline	Month 1	Month 2
1	16	13	12
2	10	5	12
3	7	8	9
4	13	11	5
5	17	2	6
6	10	17	9
7	11	6	7
8	9	8	10
9	12	13	7
10	10	7	8
11	5	8	4
12	11	6	12
13	13	7	6
14	4	6	10
15	10	5	7
16	8	9	6
17	8	3	12
18	15	10	12
19	2	3	11
20	2	4	5
21	10	3	1
22	12	5	6
23	8	12	3
24	11	6	1
25	4	14	5

hypothesis states that one or both of the manager's strategies will reduce employee tardiness.

The null hypothesis is

$$H_O\text{: } \theta_B = \theta_{M1} = \theta_{M2}$$

The research hypothesis is

H_A: One or both of the manager's strategies will reduce employee tardiness.

5.3.4.2 Set the Level of Risk (or the Level of Significance) Associated with the Null Hypothesis

The level of risk, also called an alpha (α), is frequently set at 0.05. We will use $\alpha = 0.05$ in our example. In other words, there is a 95% chance that any observed statistical difference will be real and not due to chance.

5.3.4.3 Choose the Appropriate Test Statistic

The data are obtained from three dependent, or related, conditions that report employees' number of monthly tardiness. Since we are comparing three dependent conditions, we will use the Friedman test.

5.3.4.4 Compute the Test Statistic

First, rank the values from each employee or subject (see Table 5.8).

Next, compute the sum of ranks for each condition. The ranks in each group are added to obtain a total *R*-value for the group.

TABLE 5.8

	Ranks of monthly tardiness		
Employee	Baseline	Month 1	Month 2
1	3	2	1
2	2	1	3
3	1	2	3
4	3	2	1
5	3	1	2
6	2	3	1
7	3	1	2
8	2	1	3
9	2	3	1
10	3	1	2
11	2	3	1
12	2	1	3
13	3	2	1
14	1	2	3
15	3	1	2
16	2	3	1
17	2	1	3
18	3	1	2
19	1	2	3
20	1	2	3
21	3	2	1
22	3	1	2
23	2	3	1
24	3	2	1
25	1	3	2

For the baseline condition,

$$R_B = 56$$

For month 1,

$$R_{M1} = 46$$

For month 2,

$$R_{M2} = 48$$

These R-values are used to compute the F_r test statistic. Use Formula 5.1 since there were no ties involved in the ranking:

$$F_r = \left[\frac{12}{nk(k+1)}\sum_{i=1}^{k} R_i^2\right] - 3n(k+1)$$

$$= \left(\frac{12}{25(3)(3+1)}\right)\left(56^2 + 46^2 + 48^2\right) - 3(25)(3+1)$$

$$= \left(\frac{12}{300}\right)(3136 + 2116 + 2304) - 300 = (0.04)(7556) - 300 = 302.24 - 300$$

$$F_r = 2.24$$

5.3.4.5 Determine the Value Needed for Rejection of the Null Hypothesis Using the Appropriate Table of Critical Values for the Particular Statistic Since the data are a large sample, we will use the χ^2 distribution (see Table B.2 found in Appendix B) to find the critical value for the Friedman test. In this case, we look for the critical value for $df = 2$ and $\alpha = 0.05$. Using the table, the critical value for rejecting the null hypothesis is 5.99.

5.3.4.6 Compare the Obtained Value with the Critical Value The critical value for rejecting the null hypothesis is 5.99 and the obtained value is $F_r = 2.24$. If the critical value is less than or equal to the obtained value, we must reject the null hypothesis. If instead, the critical value exceeds the obtained value, we do not reject the null hypothesis. Since the critical value exceeds the obtained value, we do not reject the null hypothesis.

5.3.4.7 Interpret the Results We did not reject the null hypothesis, suggesting that no significant difference exists between one or more of the three conditions. In particular, the data suggest that neither strategy seemed to result in less tardiness among employees.

5.3.4.8 Reporting the Results The reporting of results for the Friedman test should include such information as the number of subjects, the F_r statistic, degrees of freedom, and p-value's relation to α. For this example, the frequencies of employees' ($n = 25$) tardiness were compared over three conditions. The Friedman test was not significant ($F_{r(2)} = 2.24$, $p > 0.05$). Therefore, we can state that the data do not support providing employees with the $2 or $5 paycheck incentive.

5.4 EXAMPLES FROM THE LITERATURE

Varied examples of the nonparametric procedures described in this chapter are to be shown later. We have summarized each study's research problem and researchers' rationale(s) for choosing a nonparametric approach. We encourage you to obtain these studies if you are interested in their results.

Marston (1996) examined teachers' attitudes toward three models for servicing elementary students with mild disabilities. He compared special education resource teachers' ratings of the three models (inclusion only, combined services, and pull-out only) using a Friedman test. He chose this nonparametric test because the teachers' attitude responses were based on rankings. When the Friedman test produced significant results, he modified the α with the Bonferroni procedure in order to avoid a ballooned type I error rate with follow-up comparisons.

From a Russian high school's English as a foreign language program, Savignon and Sysoyev (2002) examined 30 students' responses to explicit training in coping strategies for particular social and cultural situations. Since the researchers considered each student a block in a randomized block study, they used a Friedman test to compare the 30 students, or groups. A nonparametric test was chosen because there were only two possible responses for each strategy ($1 =$ *strategy was difficult*; $0 =$ *strategy was not difficult*). When the Friedman test produced significant results, they used a follow-up sign test to examine each pair for differences in response to find out which of seven strategies were more difficult than others.

Cady et al. (2006) examined math teachers beliefs about the teaching and learning of mathematics over time. Since their sample size was small ($n = 12$), they used a Friedman test to compare scores of participants' survey responses. When participants' scores on the surveys differed significantly, the researchers performed follow-up pairwise analyses with the Wilcoxon signed rank test.

Hardré et al. (2007) sought to determine if computer-based, paper-based, and web-based test administrations produce the same results. They compared university students' performance on each of the three test styles. Since normality violations were observed, the researchers used a Friedman test to compare correlations of the three methods. Follow-up contrast tests were not performed since no significant differences were observed.

5.5 SUMMARY

More than two samples that are related may be compared using the Friedman test. The parametric equivalent to this test is known as the repeated measures ANOVA. When the Friedman test produces significant results, it does not identify which nor how many pairs of conditions are significantly different. The Wilcoxon signed rank test, with a Bonferroni procedure to avoid type I error rate inflation, is a useful method for comparing individual condition pairs.

In this chapter, we described how to perform and interpret a Friedman test followed with sample contrasts. We also explained how to perform the procedures using SPSS. Finally, we offered varied examples of these nonparametric statistics from the literature. The next chapter will involve a nonparametric procedure for comparing more than two unrelated samples.

5.6 PRACTICE QUESTIONS

1. A graduate student performed a pilot study for his dissertation. He wanted to examine the effects of animal companionship on elderly males. He selected 10 male participants from a nursing home. Then he used an ABAB research design, where A represented a week with the absence of a cat and B represented a week with the presence of a cat. At the end of each week, he administered a 20-point survey to measure quality of life satisfaction. The survey results are presented in Table 5.9.

TABLE 5.9

Participants	Week 1	Week 2	Week 3	Week 4
1	7	6	8	9
2	9	8	10	7
3	15	18	16	17
4	7	6	8	9
5	7	8	10	11
6	10	14	13	11
7	12	19	11	13
8	7	4	2	5
9	8	7	9	5
10	12	16	14	15

Use a Friedman test to determine if one or more of the groups are significantly different. Since this is pilot study, use $\alpha = 0.10$. If a significant difference exists, use Wilcoxon signed rank tests to identify which groups are significantly different. Use the Bonferroni procedure to limit the type I error rate. Report your findings.

2. A physical education teacher conducted an action research project to examine a strength and conditioning program. Using 12 male participants, she measures the number of curl ups they could do in 1 min. She measured their performance before the programs. Then, she measured their performance at 1 month intervals. Table 5.10 presents the performance results.

TABLE 5.10

	Number of curl ups in one minute		
Participants	Baseline	Month 1	Month 2
1	66	67	69
2	49	50	56
3	51	52	49

(*Continued*)

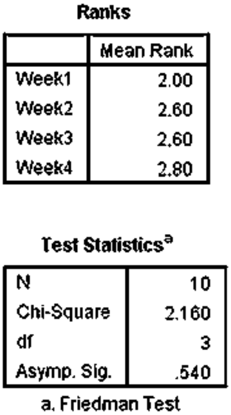
TABLE 5.10 (*Continued*)

	Number of curl ups in one minute		
Participants	Baseline	Month 1	Month 2
4	65	65	69
5	42	43	46
6	38	39	40
7	33	31	39
8	41	41	44
9	46	47	48
10	45	46	46
11	36	33	34
12	51	55	67

Use a Friedman test with $\alpha = 0.05$ to determine if one or more of the groups are significantly different. The teacher is expecting performance gains, so if a significant difference exists, use one-tailed Wilcoxon signed rank tests to identify which groups are significantly different. Use the Bonferroni procedure to limit the type I error rate. Report your findings.

5.7 SOLUTIONS TO PRACTICE QUESTIONS

1. The results from the Friedman test are displayed in SPSS Output 5.2.



Ranks

	Mean Rank
Week1	2.00
Week2	2.60
Week3	2.60
Week4	2.80

Test Statistics[a]

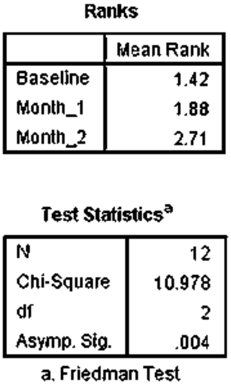
N	10
Chi-Square	2.160
df	3
Asymp. Sig.	.540

a. Friedman Test

SPSS OUTPUT 5.2

According to the data, the results from the Friedman test indicated that the four conditions were not significantly different ($F_{r(3)} = 2.160$, $p > 0.10$). Therefore, no follow-up contrasts are needed.

2. The results from the Friedman test are displayed in SPSS Output 5.3a.



Ranks

	Mean Rank
Baseline	1.42
Month_1	1.88
Month_2	2.71

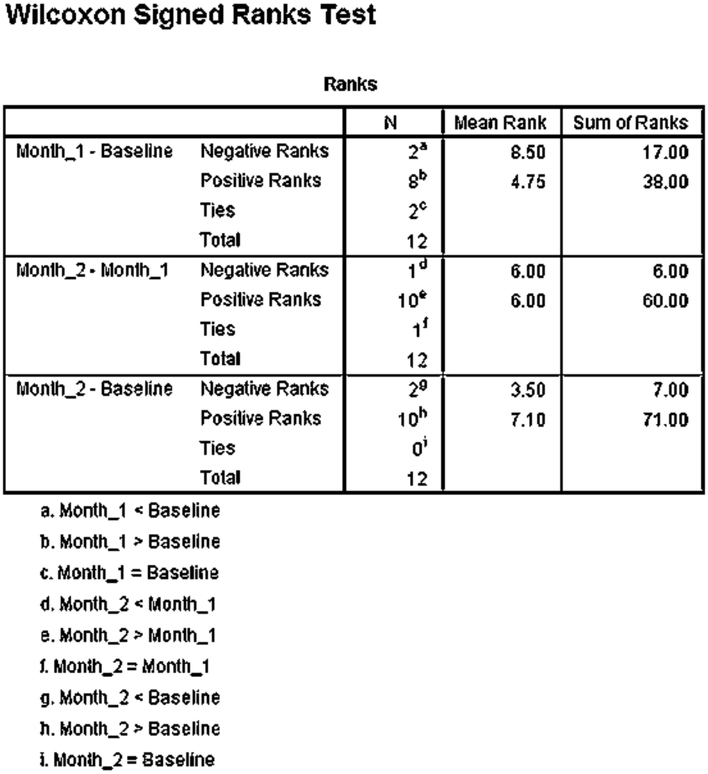
Test Statistics[a]

N	12
Chi-Square	10.978
df	2
Asymp. Sig.	.004

a. Friedman Test

SPSS OUTPUT 5.3

According to the data, the results from the Friedman test indicated that one or more of the three groups are significantly different ($F_{r(2)} = 10.978$, $p < 0.05$). Therefore, we must examine each set of samples with follow-up contrasts to find the differences between groups. We compare the samples with Wilcoxon signed rank tests. Since there are $k = 3$ groups, use $\alpha_B = 0.0167$ to avoid type I error rate inflation. The results from the Wilcoxon signed rank tests are displayed in SPSS Outputs 5.4 and 5.5.



Wilcoxon Signed Ranks Test

Ranks

		N	Mean Rank	Sum of Ranks
Month_1 - Baseline	Negative Ranks	2[a]	8.50	17.00
	Positive Ranks	8[b]	4.75	38.00
	Ties	2[c]		
	Total	12		
Month_2 - Month_1	Negative Ranks	1[d]	6.00	6.00
	Positive Ranks	10[e]	6.00	60.00
	Ties	1[f]		
	Total	12		
Month_2 - Baseline	Negative Ranks	2[g]	3.50	7.00
	Positive Ranks	10[h]	7.10	71.00
	Ties	0[i]		
	Total	12		

a. Month_1 < Baseline

b. Month_1 > Baseline

c. Month_1 = Baseline

d. Month_2 < Month_1

e. Month_2 > Month_1

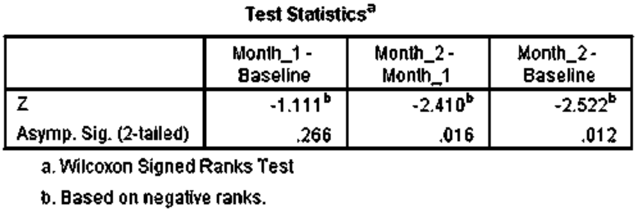
f. Month_2 = Month_1

g. Month_2 < Baseline

h. Month_2 > Baseline

i. Month_2 = Baseline

SPSS OUTPUT 5.4



Test Statistics[a]

	Month_1 - Baseline	Month_2 - Month_1	Month_2 - Baseline
Z	-1.111[b]	-2.410[b]	-2.522[b]
Asymp. Sig. (2-tailed)	.266	.016	.012

a. Wilcoxon Signed Ranks Test

b. Based on negative ranks.

SPSS OUTPUT 5.5

a. *Baseline–Month 1 Comparison.* The results from the Wilcoxon signed rank test ($T = 17.0$, $n = 12$, $p > 0.0167$) indicated that the two samples were not significantly different.

b. *Month 1–Month 2 Comparison.* The results from the Wilcoxon signed rank test ($T = 6.0$, $n = 12$, $p < 0.0167$) indicated that the two samples were significantly different.

c. *Baseline–Month 2 Comparison.* The results from the Wilcoxon signed rank test ($T = 7.0$, $n = 12$, $p < 0.0167$) indicated that the two samples were significantly different.

CHAPTER 6

COMPARING MORE THAN TWO UNRELATED SAMPLES: THE KRUSKAL–WALLIS H-*TEST*

6.1 OBJECTIVES

In this chapter, you will learn the following items.

- How to compute the Kruskal–Wallis *H*-test.
- How to perform contrasts to compare samples.
- How to perform the Kruskal–Wallis *H*-test and associated sample contrasts using SPSS®.

6.2 INTRODUCTION

A professor asked her students to complete end-of-course evaluations for her Psychology 101 class. She taught four sections of the course and wants to compare the evaluation results from each section. Since the evaluations were based on a five-point rating scale, she decides to use a nonparametric procedure. Moreover, she recognizes that the four sets of evaluation results are independent or unrelated. In other words, no single score in any single class is dependent on any other score in any other class. This professor could compare her sections using the Kruskal–Wallis *H*-test.

The Kruskal–Wallis *H*-test is a nonparametric statistical procedure for comparing more than two samples that are independent or not related. The parametric equivalent to this test is the one-way analysis of variance (ANOVA).

When the Kruskal–Wallis *H*-test leads to significant results, then at least one of the samples is different from the other samples. However, the test does not identify where the difference(s) occurs. Moreover, it does not identify how many differences occur. In order to identify the particular differences between sample pairs, a researcher might use sample contrasts, or *post hoc* tests, to analyze the specific sample pairs

Nonparametric Statistics: A Step-by-Step Approach, Second Edition. Gregory W. Corder and
Dale I. Foreman.

for significant difference(s). The Mann–Whitney *U*-test is a useful method for performing sample contrasts between individual sample sets.

In this chapter, we will describe how to perform and interpret a Kruskal–Wallis *H*-test followed with sample contrasts. We will also explain how to perform the procedures using SPSS. Finally, we offer varied examples of these nonparametric statistics from the literature.

6.3 COMPUTING THE KRUSKAL–WALLIS *H*-TEST STATISTIC

The Kruskal–Wallis *H*-test is used to compare more than two independent samples. When stating our hypotheses, we state them in terms of the population. Moreover, we examine the population medians, θ_i, when performing the Kruskal–Wallis *H*-test.

To compute the Kruskal–Wallis *H*-test statistic, we begin by combining all of the samples and rank ordering the values together. Use Formula 6.1 to determine an *H* statistic:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1) \tag{6.1}$$

where N is the number of values from all combined samples, R_i is the sum of the ranks from a particular sample, and n_i is the number of values from the corresponding rank sum.

The degrees of freedom, *df*, for the Kruskal–Wallis *H*-test are determined by using Formula 6.2:

$$df = k - 1 \tag{6.2}$$

where *df* is the degrees of freedom and k is the number of groups.

Once the test statistic H is computed, it can be compared with a table of critical values (see Table B.6 in Appendix B) to examine the groups for significant differences. However, if the number of groups, k, or the numbers of values in each sample, n_i, exceed those available from the table, then a large sample approximation may be performed. Use a table with the χ^2 distribution (see Table B.2 in Appendix B) to obtain a critical value when performing a large sample approximation.

If ranking of values results in any ties, a tie correction is required. In that case, find a new H statistic by dividing the original H statistic by the tie correction. Use Formula 6.3 to determine the tie correction value;

$$C_H = 1 - \frac{\sum (T^3 - T)}{N^3 - N} \tag{6.3}$$

where C_H is the ties correction, T is the number of values from a set of ties, and N is the number of values from all combined samples.

If the H statistic is not significant, then no differences exist between any of the samples. However, if the H statistic is significant, then a difference exists between at least two of the samples. Therefore, a researcher might use sample contrasts between individual sample pairs, or *post hoc* tests, to determine which of the sample pairs are significantly different.

When performing multiple sample contrasts, the type I error rate tends to become inflated. Therefore, the initial level of risk, or α, must be adjusted. We demonstrate the Bonferroni procedure, shown in Formula 6.4, to adjust α:

$$\alpha_B = \frac{\alpha}{k} \tag{6.4}$$

where α_B is the adjusted level of risk, α is the original level of risk, and k is the number of comparisons.

6.3.1 Sample Kruskal–Wallis *H*-Test (Small Data Samples)

Researchers were interested in studying the social interaction of different adults. They sought to determine if social interaction can be tied to self-confidence. The researchers classified 17 participants into three groups based on the social interaction exhibited. The participant groups were labeled as follows:

High = constant interaction; talks with many different people; initiates discussion

Medium = interacts with a variety of people; some periods of isolation; tends to focus on fewer people

Low = remains mostly isolated from others; speaks if spoken to, but leaves interaction quickly

After the participants had been classified into the three social interaction groups, they were directed to complete a self-assessment of self-confidence on a 25-point scale. Table 6.1 shows the scores obtained by each of the participants, with 25 points being an indication of high self-confidence.

TABLE 6.1

Original ordinal self-confidence scores placed within social interaction groups

High	Medium	Low
21	19	7
23	5	8
18	10	15
12	11	3
19	9	6
20		4

The original survey scores obtained were converted to an ordinal scale prior to the data analysis. Table 6.1 shows the ordinal values placed in the social interaction groups.

We want to determine if there is a difference between any of the three groups in Table 6.1. Since the data belong to an ordinal scale and the sample sizes are small ($n < 20$), we will use a nonparametric test. The Kruskal–Wallis H-test is a good choice to analyze the data and test the hypothesis.

6.3.1.1 State the Null and Research Hypotheses

The null hypothesis states that there is no tendency for self-confidence to rank systematically higher or lower for any of the levels of social interaction. The research hypothesis states that there is a tendency for self-confidence to rank systematically higher or lower for at least one level of social interaction than at least one of the other levels. We generally use the concept of "systematic differences" in the hypotheses.

The null hypothesis is

$$H_O\text{: } \theta_L = \theta_M = \theta_H$$

The research hypothesis is

H_A: There is a tendency for self-confidence to rank systematically higher or lower for at least one level of social interaction when compared with the other levels.

6.3.1.2 Set the Level of Risk (or the Level of Significance) Associated with the Null Hypothesis

The level of risk, also called an alpha (α), is frequently set at 0.05. We will use $\alpha = 0.05$ in our example. In other words, there is a 95% chance that any observed statistical difference will be real and not due to chance.

6.3.1.3 Choose the Appropriate Test Statistic

The data are obtained from three independent, or unrelated, samples of adults who are being assigned to three different social interaction groups by observation. They are then being assessed using a self-confidence scale with a total of 25 points. The three samples are small with some violations of our assumptions of normality. Since we are comparing three independent samples, we will use the Kruskal–Wallis H-test.

6.3.1.4 Compute the Test Statistic

First, combine and rank the three samples together (see Table 6.2).

Place the participant ranks in their social interaction groups to compute the sum of ranks R_i for each group (see Table 6.3).

Next, compute the sum of ranks for each social interaction group. The ranks in each group are added to obtain a total R-value for the group.

For the high group,

$$R_H = 10 + 12 + 13.5 + 15 + 16 + 17 = 83.5$$

$$n_H = 6$$

TABLE 6.2

Original ordinal score	Participant rank	Social interaction group
3	1	Low
4	2	Low
5	3	Medium
6	4	Low
7	5	Low
8	6	Low
9	7	Medium
10	8	Medium
11	9	Medium
12	10	High
15	11	Low
18	12	High
19	13.5	Medium
19	13.5	High
20	15	High
21	16	High
23	17	High

TABLE 6.3

Ordinal data ranks			
High	Medium	Low	
10	3	1	$N=17$
12	7	2	
13.5	8	4	
15	9	5	
16	13.5	6	
17		11	

For the medium group,

$$R_M = 3+7+8+9+13.5 = 40.5$$
$$n_M = 5$$

For the low group,

$$R_L = 1+2+4+5+6+11 = 29$$
$$n_L = 6$$

These R-values are used to compute the Kruskal–Wallis H-test statistic (see Formula 6.1). The number of participants in each group is identified by a lowercase n. The total group size in the study is identified by the uppercase N.

Now, using the data from Table 6.3, compute the H-test statistic using Formula 6.1:

$$H=\frac{12}{N(N+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i}-3(N+1)$$

$$H=\frac{12}{17(17+1)}\left(\frac{83.5^2}{6}+\frac{40.5^2}{5}+\frac{29^2}{6}\right)-3(17+1)$$

$$=0.0392(1162.04+328.05+140.17)-54=0.0392(1630.26)-54=63.93-54$$

$$H=9.93$$

Since there was a tie involved in the ranking, correct the value of H. First, compute the tie correction (see Formula 6.2). Then, divide the original H statistic by the ties correction C_H:

$$C_H=1-\frac{\sum(T^3-T)}{N^3-N}=1-\frac{(2^3-2)}{17^3-17}=1-\frac{(8-2)}{(4913-17)}=1-0.0001$$

$$C_H=0.9988$$

Next, we divide to find the corrected H statistic:

$$\text{corrected } H=\text{original } H\div C_H=9.93\div 0.9988=9.94$$

For this set of data, notice that the corrected H does not differ greatly from the original H. With the correction, $H = 9.94$.

6.3.1.5 Determine the Value Needed for Rejection of the Null Hypothesis Using the Appropriate Table of Critical Values for the Particular Statistic

We will use the critical value table for the Kruskal–Wallis H-test (see Table B.6 in Appendix B) since it includes the number of groups, k, and the numbers of samples, n, for our data. In this case, we look for the critical value for $k = 3$ and $n_1 = 6$, $n_2 = 6$, and $n_3 = 5$ with $\alpha = 0.05$. Table B.5 returns a critical value for the Kruskal–Wallis H-test of 5.76.

6.3.1.6 Compare the Obtained Value with the Critical Value

The critical value for rejecting the null hypothesis is 5.76 and the obtained value is $H = 9.94$. If the critical value is less than or equal to the obtained value, we must reject the null hypothesis. If instead, the critical value exceeds the obtained value, we do not reject the null hypothesis. Since critical value is less than the obtained value, we must reject the null hypothesis.

At this point, it is worth mentioning that larger samples often result in more ties. While comparatively small, as observed in step 4, corrections for ties can make a difference in the decision regarding the null hypothesis. If the H were near the critical value of 5.99 for $df = 2$ (e.g., $H = 5.80$), and the tie correction calculated to be 0.965, the decision would be to reject the null hypothesis with the correction ($H = 6.01$), but to not reject the null hypothesis without the correction. Therefore, it is important to perform tie corrections.

6.3.1.7 Interpret the Results We rejected the null hypothesis, suggesting that a real difference in self-confidence exists between one or more of the three social interaction types. In particular, the data show that those who were classified as fitting the definition of the "low" group were mostly people who reported poor self-confidence, and those who were in the "high" group were mostly people who reported good self-confidence. However, describing specific differences in this manner is speculative. Therefore, we need a technique for statistically identifying difference between groups, or contrasts.

Sample Contrasts, or Post Hoc Tests The Kruskal–Wallis *H*-test identifies if a statistical difference exists; however, it does not identify how many differences exist and which samples are different. To identify which samples are different and which are not, we can use a procedure called contrasts or *post hoc* tests. Methods for comparing two samples at a time are described in Chapters 3 and 4. The examples in this chapter compare unrelated samples so we will use the Mann–Whitney *U*-test.

It is important to note that performing several two-sample tests has a tendency to inflate the type I error rate. In our example, we would compare three groups, $k = 3$. At $\alpha = 0.05$, the type I error rate would be $1 - (1 - 0.05)^3 = 0.14$.

To compensate for this error inflation, we demonstrate the Bonferroni procedure (see Formula 6.4). With this technique, we use a corrected α with the Mann–Whitney *U*-tests to determine significant differences between samples. For our example,

$$\alpha_B = \frac{\alpha}{k} = \frac{0.05}{3}$$

$$\alpha_B = 0.0167$$

When we compare each set of samples with the Mann–Whitney *U*-tests and use α_B, we obtain the following results presented in Table 6.4.

TABLE 6.4

Group comparison	Mann–Whitney *U* statistic	Rank sum difference	Significance
High–medium	2.5	48.5 − 17.5 = 31.0	0.017
Medium–low	7.0	38.0 − 28.0 = 10.0	0.177
High–low	1.0	56.0 − 22.0 = 34.0	0.004

Since $\alpha_B = 0.0167$, we notice that the high–low group comparison is indeed significantly different. The medium–low group comparison is not significant. The high–medium group comparison requires some judgment since it is difficult to tell if the difference is significant or not; the way the value is rounded off could change the result.

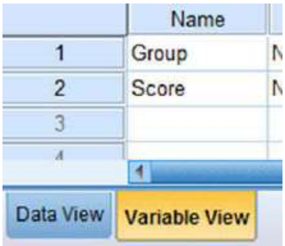
Note that if you are not comparing all of the samples for the Kruskal–Wallis *H*-test, then *k* is only the number of comparisons you are making with the Mann–Whitney *U*-tests. Therefore, comparing fewer samples will increase the chances of finding a significant difference.

6.3.1.8 Reporting the Results The reporting of results for the Kruskal–Wallis *H*-test should include such information as sample size for all of the groups, the *H* statistic, degrees of freedom, and *p*-value's relation to α. For this example, three social interaction groups were compared: high ($n_H = 6$), medium ($n_M = 5$), and low ($n_L = 6$). The Kruskal–Wallis *H*-test was significant ($H_{(2)} = 9.94$, $p < 0.05$). In order to compare each set of samples, contrasts may be used as described earlier in this chapter.

6.3.2 Performing the Kruskal–Wallis *H*-Test Using SPSS

We will analyze the data from the example earlier using SPSS.

6.3.2.1 Define Your Variables First, click the "Variable View" tab at the bottom of your screen. Then, type the names of your variables in the "Name" column. Unlike the Friedman's ANOVA described in Chapter 5, you cannot simply enter each sample into a separate column to execute the Kruskal–Wallis *H*-test. You must use a grouping variable. In Figure 6.1, the first variable is the grouping variable that we called "Group." The second variable that we called "Score" will have our actual values.



	Name
1	Group
2	Score
3	
4	

FIGURE 6.1

When establishing a grouping variable, it is often easiest to assign each group a whole number value. In our example, our groups are "High," "Medium," and "Low." Therefore, we must set our grouping variables for the variable "Group." First, we selected the "Values" column and clicked the gray square as shown in Figure 6.2. Then, we set a value of 1 to equal "High," a value of 2 to equal "Medium," and a value of 3 equal to "Low." Each value label is established and moved to the list when we click the "Add" button. Once we click the "OK" button, we are returned to the SPSS Variable View.

6.3.2.2 Type in Your Values Click the "Data View" tab at the bottom of your screen as shown in Figure 6.3. Type in the values for all three samples in the "Score" column. As you do so, type in the corresponding grouping variable in the "Group" column. For example, all of the values for "High" are signified by a value of 1 in the grouping variable column that we called "Group."