CHAPTER 5

Tests for Two Related Samples: Pretest-Posttest Measures for a Single Sample

- McNemar test
- Sign test
- Wilcoxon signed-ranks test

In Chapter 4, we examined nonparametric tests that could be used to determine the extent to which a single sample is similar to a hypothesized theoretical sample. In health care research, we frequently try to assess whether a particular intervention is effective with a certain population. In our hypothetical example from Chapters 1 and 3, we were interested in evaluating the effects of a staff-initiated intervention to reduce the number of sleep environment interruptions for hospitalized pediatric cancer patients on the frequency of their nocturnal awakenings, levels of fatigue, and distress compared to a usual-care group. In this study, there may be certain characteristics of the sample, such as type of cancer, age, or gender of the child, that are known from prior research to confound and potentially misrepresent the outcomes of the intervention.

Two approaches that can be used to address this problem prior to the intervention are the following: (1) matching the subjects with regard to these extraneous confounding variables and then randomly assigning one of the pairs to the control and the other to the intervention group and (2) using each subject as his or her own control. Data being analyzed then become paired, either through the use of related samples or through repeated observations on a single sample. A third occurrence of matched pairs can occur when the researcher has sampled observations in pairs (e.g., husband and wife). Although each member of the pair may have separate scores on a dependent variable (e.g., marital satisfaction), there is reason to believe that knowing the scores of one member of the pair (e.g., wife's marital satisfaction) will give information about the scores of the other member (husband's marital satisfaction).

There are many types of naturally occurring situations in which respondents are paired, such as mothers and fathers, husbands and wives, twins, and parents and children. A research design could also be longitudinal, with multiple observations of a single sample but lacking a control group. In these situations, a statistical test for independent groups is inappropriate because the data are paired. We are restricted instead to paired tests that acknowledge the dependence of the observations in the samples.

If we were using parametric statistics, the paired *t* test typically would be used to analyze paired samples. This type of *t* test is very useful and robust to violations of its assumptions. Sometimes, however, it is not possible to use the paired *t* test because the data do not sufficiently meet the test's assumptions. Perhaps the sample size is too small, the continuous dependent variable is severely skewed, or the data are categorical. In these situations, the statistical test of choice would be nonparametric.

Several nonparametric statistical procedures are suitable for use with paired observations or repeated measures using a single sample across two time periods. Three tests will be examined in this chapter: the McNemar test, the sign test, and the Wilcoxon signed-ranks test. Two additional tests for repeated measures across more than two time periods (Cochran's *Q* and the Friedman test) will be examined in Chapter 6. The McNemar test and Cochran's *Q* test are used when the measurement of the dependent variable is dichotomous, whereas the remaining tests are used when the dependent variable is at least ordinal in its level of measurement.

THE MCNEMAR TEST

The McNemar test is especially useful when the researcher has a pretest-posttest design in which the subjects serve as their own controls and the dependent variable is dichotomous (Bennett & Underwood, 1970; Feuer & Kessler, 1989; Siegel & Castellan, 1988). In health care research, for example, we might be interested in comparing the performance of two procedures, comparing the opinions of two experts, or determining whether an educational program altered people's preferences for a particular type of health provider (e.g., nurse practitioner vs. physician). The McNemar test examines the extent of change in the dichotomous variable from pretest to posttest. If the proportion of changed responses in one direction is sufficiently greater than what would be expected by chance, the null hypothesis of no disproportionate change is rejected.

An Appropriate Research Question for the McNemar Test

Numerous examples from the research literature illustrate the versatility of the McNemar test. Demir and Erdil (2013) used both the McNemar and other nonparametric tests to evaluate the effectiveness of home monitoring in hip replacement surgery patients. Price (2013) used the same test to assess creatinine normalization of workplace urine drug tests, and Elizondo-Montemayor et al. (2013) used the test to evaluate a school-based individualized lifestyle intervention to decrease obesity and the metabolic syndrome in Mexican children. Using this same test, Snoey and colleagues (1994) compared error rates of physicians and cardiologists regarding their interpretations of emergency department electrocardiograms (ECGs). In all these examples, the data that were analyzed using the McNemar test were dichotomous variables that were paired either through matched samples or through a design in which the subjects served as their own controls.

In our hypothetical intervention study, the children in both the staff-initiated intervention and usual-care groups were asked before and after the intervention whether they were distressed about their current hospitalization. Because the pretest-posttest measure is dichotomous (the children answered "yes" or "no" to the question concerning distress) and the data are paired, the use of the McNemar test is appropriate. An example of a research question that could be answered using this test would be as follows:

> Is the staff-initiated intervention effective in reducing children's distress concerning their current hospitalization?

Note that we are examining the pretest and posttest responses of one sample of children who have served as their own controls. We are not comparing the two groups.

Null and Alternative Hypotheses

Table 5.1 presents the null and alternative hypothesis for the McNemar test that would follow from the research question. Although nondirectional tests are possible, our alternative hypothesis predicts a direction for the probability statement; therefore, the test is directional, and we will be using a one-tailed alpha level (e.g., $\alpha = .05$).

Table 5.1 Example of Null and Alternative Hypotheses Appropriate for the McNemar Test

Null Hypothesis	
H_0:	Among those children who took part in the intervention and who change their reported distress concerning their current hospitalization, the probability that a child's distress is reduced is the same as the probability that the child's distress is increased. $P_{reduced} = P_{increased}$
Alternative Hypothesis	
H_a:	Among those children who took part in the intervention and who change their reported distress concerning their current hospitalization, the probability that a child's distress is reduced is greater than the probability that the child's distress is increased. $P_{reduced} > P_{increased.}$

Notice that the McNemar test focuses only on those children in the sample who change their opinions regarding their distress and does not include those children whose reported distress does not change. The test also does not compare intervention and usual-care groups, because this is a test of change in a single group.

Overview of the Procedure

To undertake a McNemar test, the data first need to be cast into a 2 × 2 table that represents the change in an individual's response from before to after the intervention. If the original response data are not nominal, they need to be reduced to a form such that the coding scheme (e.g., 0s and 1s) represents identical values for the paired variables being examined. Table 5.2 presents the form that such a table typically takes.

Table 5.2 2 × 2 Tables for the McNemar Test for Children's Expressed Distress Concerning Hospitalization

			After Intervention	
Expressed Distress			*Yes* *(1)*[a]	*No* *(0)*
Before Intervention				
	No	(0)[a]	A ①	B
	Yes	(1)	C	D ②

[a] The inconsistent ordering of the categories was deliberate, as a means of duplicating the pattern presented on the computer printout.

In our hypothetical example, "A" represents the number of children in the intervention group who reported not being distressed regarding their hospitalization prior to the intervention but reported being distressed following the intervention ①. "D" represents the number of children in the experimental group whose distress regarding their hospitalization decreased following the intervention ②. "A + D" therefore is the total number of subjects who changed their responses from before to after the intervention.

When sample sizes are reasonably large (i.e., when the total frequency of changes is greater than 10 (Siegel & Castellan, 1988)), the McNemar test uses a chi-square test to compare the number of subjects who changed in a specific direction (i.e., A – D) with the frequency of change that would be expected under the null hypothesis of a change being equally likely in both directions ([A + D]/2). This chi-square statistic is similar to the goodness-of-fit statistic that was presented in Chapter 4:

$$\chi^2 = \sum_{1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where

O_i = the observed number of cases in cell *i,*

E_i = the expected number of cases in cell *i,*

k = the number of cells that represent change.

For the McNemar test, we are interested only in the *observed* and *expected* values of two cells, A and D, because these are the only cells that represent change. The *expected* values for each of these cells are (A + D)/2. The chi-square statistic for the McNemar test is given by

$$\chi^2 = \sum_{1}^{k} \frac{(O_i - E_i)^2}{E_i} = \frac{[\mathrm{A} - (\mathrm{A} + \mathrm{D})/2]^2}{(\mathrm{A} + \mathrm{D})/2} + \frac{[\mathrm{D} - (\mathrm{A} + \mathrm{D})/2]^2}{(\mathrm{A} + \mathrm{D})/2}$$

If the sum of the differences between the observed and expected values across the change cells is sufficiently large, the resulting chi-square value with 1 degree of freedom ($df = k - 1 = 2 - 1$) will also be large, increasing the likelihood that the null hypothesis will be rejected. When the sample size is smaller (e.g., when the expected change frequency, (A+D/2) is less than 10), a binomial test is used instead (see Chapter 4) (Siegel & Castellan, 1988). In our case, (A + D)/2 = (0 + 6)/2 = 3, which is smaller than the criterion. The printout in SPSS for Windows, therefore, represents the binomial test instead.

Critical Assumptions of the McNemar Test

There are four major assumptions of the McNemar test.

1. *The dichotomous variable being assessed has assigned values for each level (e.g., 0 and 1), the meaning of which is similar across both time periods.* This assumption implies that the dichotomous variable for the pretest and posttest is coded similarly. In our hypothetical study, for example, children who expressed no distress concerning their hospitalization were assigned the value of 0, whereas children who did express distress were assigned the value of 1 for both the pretest and the postintervention measure.

2. *The data being examined represent frequencies, not scores.* As with all chi-square statistics, the assumption is made that the data being examined are frequency data, not scores. In our hypothetical study, the children answered "yes" (1) or "no" (0) to the question of whether they were distressed about their hospitalization. These data represent frequencies or counts and, therefore, meet this assumption.

3. *The dichotomous measures are paired observations of the same randomly selected subjects or matched pairs.* It is expected that the data consist of paired responses from a set of randomly selected subjects or matched samples of subjects. In our hypothetical study, the children were not randomly selected because this was a sample of convenience. The observations were paired, however, because the children served as their own controls by responding to the question concerning their distress at two different points in time.

4. *The levels of the dichotomous variable are mutually exclusive; that is, a subject can be assigned to only one level of the dichotomous variable that is being examined over time.* This assumption implies that a subject cannot be assigned to both a 0 and a 1 on the dichotomous variable at pretest and posttest. In our hypothetical intervention study, a child could not report *both* a *"yes" and* a "no" on the pre- or postintervention distress variables.

Computer Commands

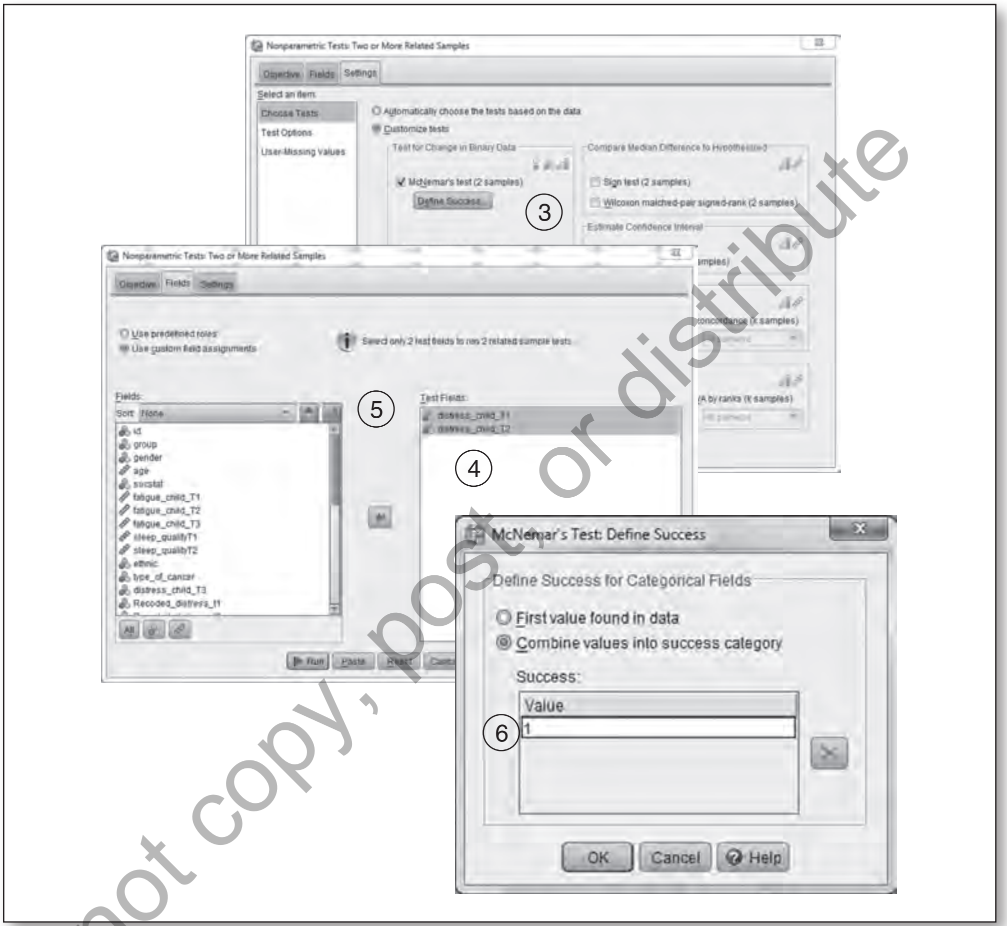
The data set that we will be using for this example is "hospitalized children with cancer-20 cases.sav." This file can be found on the SAGE website (study.sagepub.com/pett2e). Figure 5.1 presents the SPSS for Windows dialog boxes that are used to generate the McNemar test. This dialog box was opened by choosing the following items from the menu: *Analyze . . . Nonparametric Tests . . . Related Samples.* Because we were interested in restricting the analysis to only those children who received the intervention (*n* = 10), the sample was restricted by first using the *Select Cases* command (*Data . . . Select Cases . . . If the Condition is Satisfied . . .*) and instructing the computer to select only those cases that met the condition that they had been assigned to the intervention group (i.e., Group = 1).

Selecting *Customize Analysis . . .* will allow us to choose the McNemar test ③. Next, the pair of dichotomous variables to be examined is selected from the menu. These variables can be either repeated measures from subjects who are being used as their own controls (e.g., *Distress_Child_T1* = preintervention distress and Distress_Child_T2 = postintervention distress) ④ or dichotomous variables obtained from matched samples (e.g., husbands' and wives' marital satisfaction). Be sure that both variables have been labeled "nominal" in the data set ⑤. We will also instruct the computer that the value of "1" for pre- and postintervention distress is the "success" value. If desired, values could also be combined into the success category ⑥.

Computer-Generated Output

Figure 5.2 presents the SPSS for Windows syntax commands and computer-generated output for the McNemar test. The syntax commands can be saved for future analyses. They indicate that the sample selection has been restricted to only those children from the staff-initiated intervention group (Group = 1) because a filter has been placed on the data set (filter_$) ⑦ and that a McNemar test has been run on the restricted sample ⑧. The subcommand, */missing scope = analysis,* indicates that we

Figure 5.1 SPSS for Windows commands for generating the McNemar test (study.sagepub.com/pett2e).



Nonparametric Tests: Two or More Related Samples

Settings — Select an item: Choose Tests

○ Automatically choose the tests based on the data
● Customize tests

Test for Change in Binary Data
☑ McNemar's test (2 samples)

Compare Median Difference to Hypothesized
☐ Sign test (2 samples)
☐ Wilcoxon matched-pair signed-rank (2 samples)

Nonparametric Tests: Two or More Related Samples

Fields

○ Use predefined roles
● Use custom field assignments

Select only 2 test fields to run 2 related sample tests

Test Fields:
distress_child_T1
distress_child_T2

McNemar's Test: Define Success

Define Success for Categorical Fields
○ First value found in data
● Combine values into success category

Success:

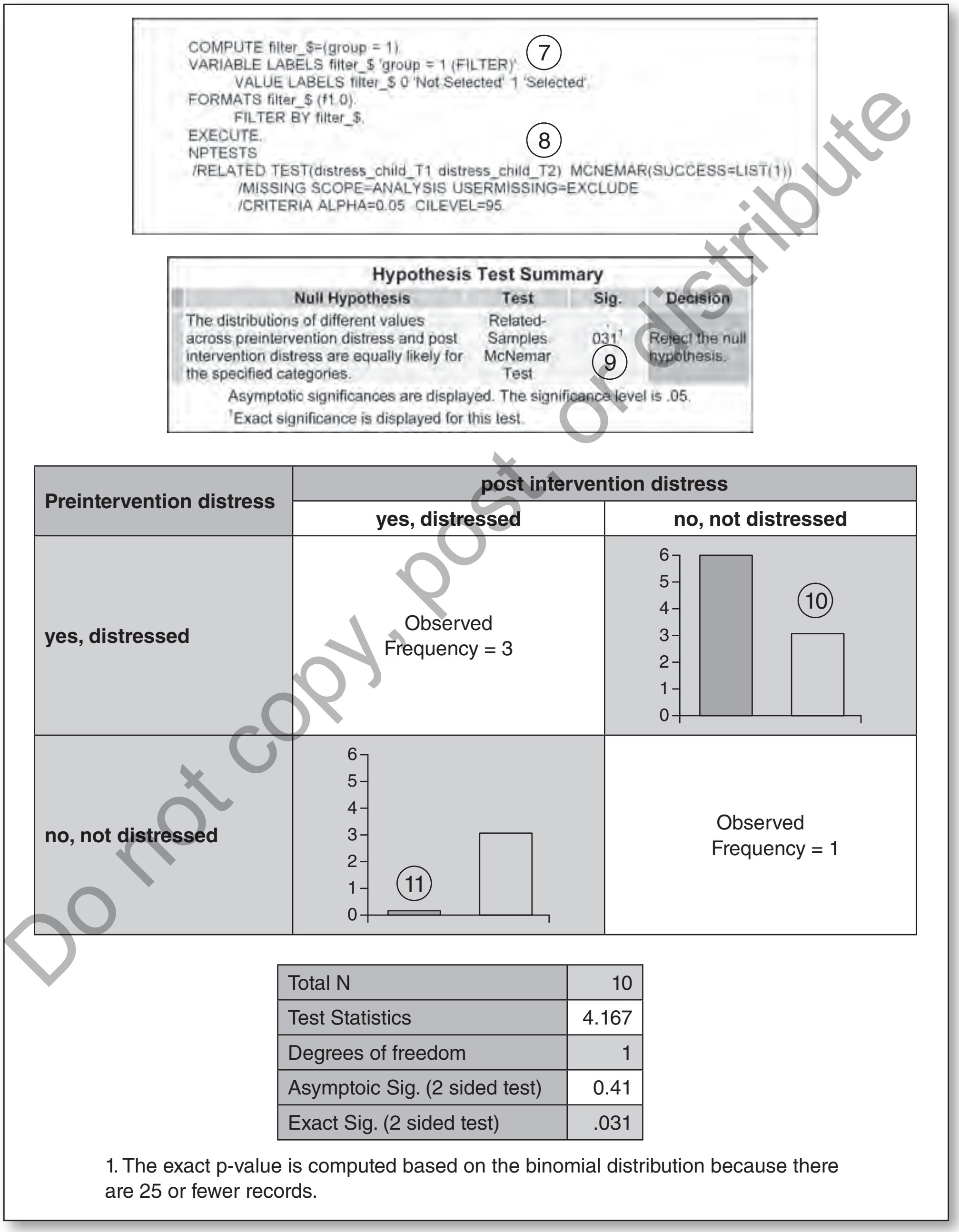
Value
1

Reprints Courtesy of International Business Machines Corporation, © International Business Machines Corporation

have selected the default option for the handling of missing values: Cases will be omitted on a test-by-test basis.

In the printout, we are presented with the Hypothesis Test Summary. The significance level is .031, and the "decision" is to reject the null hypothesis (9). Before we accept what the output suggests, we should first reexamine our null and alternative hypotheses. Were they directional or nondirectional? Since our alternative hypothesis was directional, we have a directional test. Since the *p* value (.031) in the SPSS output is two-tailed, we need to divide that value in half (.031/2 = .016) and compare the resulting value to α = .05. The null hypothesis will be rejected if our generated *p* value (.016) is less than our alpha (.05), which indeed it is.

Figure 5.2 Syntax and computer-generated output obtained for the McNemar test in SPSS for Windows (v. 22–23). Data file: hospitalized children with cancer-20 cases.sav (study.sagepub.com/pett2e).



```
COMPUTE filter_$=(group = 1).
VARIABLE LABELS filter_$ 'group = 1 (FILTER)'.
        VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
        FILTER BY filter_$.
EXECUTE.
NPTESTS
 /RELATED TEST(distress_child_T1 distress_child_T2)  MCNEMAR(SUCCESS=LIST(1))
         /MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE
         /CRITERIA ALPHA=0.05  CILEVEL=95.
```

Hypothesis Test Summary

Null Hypothesis	Test	Sig.	Decision
The distributions of different values across preintervention distress and post intervention distress are equally likely for the specified categories.	Related-Samples McNemar Test	.031[1]	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

[1]Exact significance is displayed for this test.

Preintervention distress	post intervention distress	
	yes, distressed	no, not distressed
yes, distressed	Observed Frequency = 3	
no, not distressed		Observed Frequency = 1

Total N	10
Test Statistics	4.167
Degrees of freedom	1
Asymptoic Sig. (2 sided test)	0.41
Exact Sig. (2 sided test)	.031

1. The exact p-value is computed based on the binomial distribution because there are 25 or fewer records.

Reprints Courtesy of International Business Machines Corporation, © International Business Machines Corporation

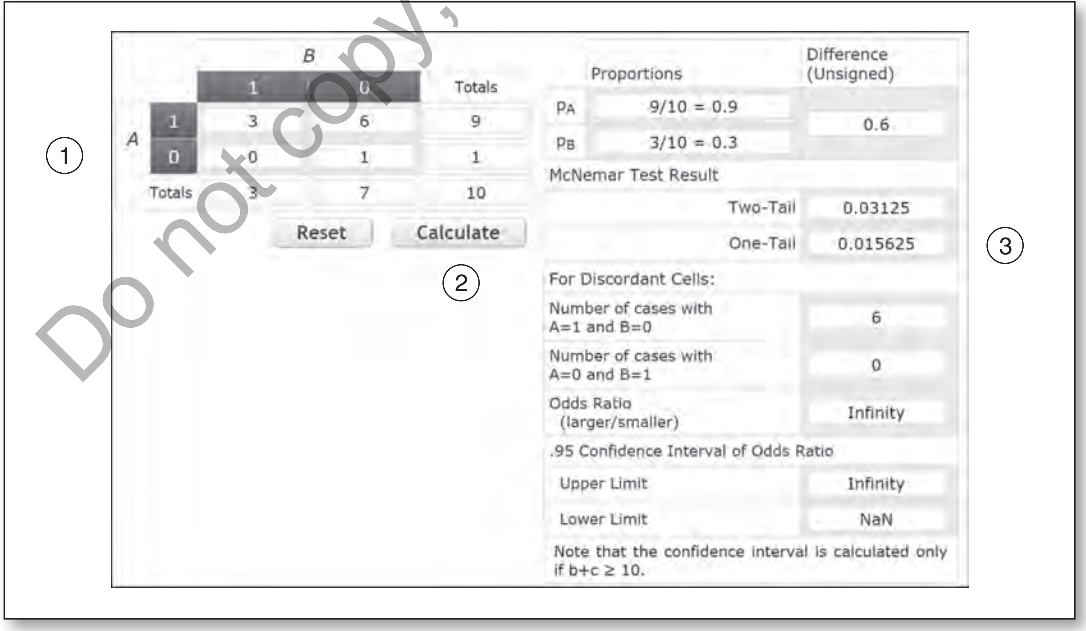
How will we interpret these results? To examine whether our decision to reject the null hypothesis is in our hypothesized direction, we need to examine the results presented for the change test. The results from this printout indicate that six children from our staff-initiated intervention group who reported preintervention distress no longer reported distress following the intervention ⑩. No children reported increased distress following the intervention ⑪. Our conclusion, therefore, is that, among the 10 children in the staff-initiated intervention group, the change in distress levels was in the direction of lowering distress.

Determining the Outcome of a McNemar Test Using a Website

Several useful websites provide interactive calculators that enable the researcher to calculate a McNemar test without the need for a statistical computer package. Because some of these websites have publishing restrictions, we will illustrate how to use this type of calculator using www.vassarstats.net, a website that is in the public domain.

In the vassarstats website, click on *Proportions . . . McNemar's test for correlated proportions.* Once there, we can input our summary data for the distress levels of the children in the staff-initiated intervention group ①. By clicking on the *Calculate* button ②, we are given the one- and two-tailed McNemar test result ③. Because our research

Figure 5.3 Internet-generated output for the McNemar test.



		B		
		1	0	Totals
A	1	3	6	9
	0	0	1	1
	Totals	3	7	10

	Proportions	Difference (Unsigned)
PA	9/10 = 0.9	0.6
PB	3/10 = 0.3	

McNemar Test Result

Two-Tail	0.03125
One-Tail	0.015625

For Discordant Cells:

Number of cases with A=1 and B=0	6
Number of cases with A=0 and B=1	0
Odds Ratio (larger/smaller)	Infinity

.95 Confidence Interval of Odds Ratio

Upper Limit	Infinity
Lower Limit	NaN

Note that the confidence interval is calculated only if b+c ≥ 10.

SOURCE: www.vassarstats.net.

hypothesis was directional, we will focus on the one-tailed result (p = .0156). This is the same result we obtained from the output generated in SPSS for Windows. Again, we can conclude that the staff-initiated intervention was effective in lowering the children's distress levels from pre- to postintervention.

Presentation of Results

Table 5.3 presents a suggested approach to presenting the results of the McNemar test. Percentages are omitted from the table to prevent confusion resulting from such a small sample. An alternative method would be to present the results in written format:

> The results of the McNemar test indicate that the staff-initiated intervention was effective in changing distress levels in the direction of reduced distress from pre- to postintervention (one-tailed p = .016).

Table 5.3 Suggested Presentation of McNemar Test Results

	Distress Following Intervention?			
	Yes	*No*	*Total*	*p (1-tailed)*
Distress prior to intervention?				
Yes	3	6	9	.016
No	0	1	1	
Total	3	7	10	

Advantages, Limitations, and Alternatives to the McNemar Test

The McNemar test is especially useful because it can be used to examine change in nominal-level data. A disadvantage of this test is that it does not examine extent of change, only whether change has occurred. In the hypothetical example, it is not possible to ascertain the extent of change in the children's distress, only whether they changed and, if so, in what direction.

The McNemar test also does not allow for a comparison group (e.g., the control group) because this is a test using dependent observations. Feuer and Kessler (1989) have presented a two-sample situation for the McNemar test in which the marginal changes in a nominal-level variable for two independent cohorts (a control and an intervention cohort) are examined across two time periods. If the researcher wanted to compare two groups, two additional nonparametric approaches are possible. McNemar tests could be run on each of the independent groups independently and their results compared. Alternatively, a 2 × 2 contingency table representing group membership (intervention, control) by change classification (change in a positive direction, change in a negative direction) could be created. The Fisher's exact test (Chapter 7) could then be used to examine group-by-time interaction between the two dichotomous variables, group and direction of change.

Because the McNemar test is used primarily with nominal-level data, there is no parametric counterpart to this test. If there are more than two periods of data collection (e.g., pretest, posttest, and follow-up), Cochran's *Q* test is recommended (see Chapter 6). If the data are continuous and meaningfully ranked, it would be advisable to use more sensitive nonparametric tests that use paired data, such as the sign test and the Wilcoxon test.

Examples From Published Research

Demir, S. G., & Erdil, F. (2013). Effectiveness of home monitoring according to the Model of Living in hip replacement surgery patients. *Journal of Clinical Nursing, 22*(9/10), 1226–1241.

Elizondo-Montemayor, L., Gutierrez, N. G., Moreno, D. M., Martínez, U., Tamargo, D., & Treviño, M. (2013). School-based individualised lifestyle intervention decreases obesity and the metabolic syndrome in Mexican children. *Journal of Human Nutrition & Dietetics, 26,* 82–89.

Price, J. W. (2013). Creatinine normalization of workplace urine drug tests: Does it make a difference? *Journal of Addiction Medicine,* 7(2), 129–132.

Snoey, E., Housset, B., Guyon, P., ElHaddad, S., Valty, J., & Hericord, P. (1994). Analysis of emergency department interpretation of electrocardiograms. *Journal of Accident & Emergency Medicine, 11*(3), 149–153.

THE SIGN TEST

The sign test is one of the oldest of all nonparametric tests, dating back to 1710 (Conover, 1999). It is called the sign test because the statistic is generated from data that have been reduced to +'s or −'s. The test can be used with paired data that are either dichotomous, ordinal, or continuous and that have been collected across a single sample or matched pairs. If the data are dichotomous, the two categories making up the variable need to have some rank order to their measurement (e.g., "success" vs. "failure" or "yes" vs. "no"), and the test reduces to the McNemar situation.

The sign test can be used in any situation in which the researcher can determine whether one of two paired or matched observations is "greater" or "less" than the other with regard to some identified attribute. The exact quantitative amount of the difference does not need to be determined for this test because the focus of the analysis is on the signs of the differences between each pair of variables.

An Appropriate Research Question for the Sign Test

The sign test has not been very popular in the research literature, possibly because it has been overshadowed by the more powerful Wilcoxon signed-ranks test. Several examples from the literature, however, demonstrate the potential usefulness of the sign test. Graves, Carter, Anderson, and Winett (2003) used the sign test to evaluate an 8-week pilot intervention based on social cognitive theory to improve quality of life for women with breast cancer. A similar sign test was used by Miletic, Sekulic, and Ostojic (2007) to examine body physique and prior training experience as determinants of Self-Estimated Functional Inability Because of Pain (SEFIP) scores for university dancers, and Whellan et al. (2012) used the same test to determine time from admission to 50% improvement in impedance and to when a heart patient was medically ready for discharge.

In our hypothetical health care intervention study, suppose we were interested in examining the reduction in self-reported levels of fatigue from pretest to posttest within the group of 10 children who received the staff-initiated intervention. An example of a research question for which a sign test would be appropriate would be as follows:

> Do the children in the staff-initiated intervention group reduce their self-reported fatigue from pretest to posttest?

Note that because the pretest and posttest fatigue measures that we are using are scales of ordinal level of measurement that range from 1 to 7, we cannot use the McNemar test.

Null and Alternative Hypotheses

Table 5.4 presents the null and alternative hypotheses for which a sign test would be appropriate given the research question outlined above. Note that the alternative hypothesis is directional; therefore, the test will be one-tailed. Because the fatigue variable consists of a 7-point Likert scale (range: 1 = *not at all fatigued* to 7 = *very fatigued*), we are predicting that the number of negative differences formed by Fatigue_T2 – Fatigue_T1 will be greater than the number of positive differences. That is, self-reported fatigue was greater at pretest than at posttest.

Table 5.4 Example of Null and Alternative Hypotheses Appropriate for the Sign Test

Null Hypothesis
H_0: The self-reported level of fatigue for the children who took part in the staff-initiated intervention will not change from pretest to posttest; that is, the number of children whose fatigue is reduced from pretest to posttest is the same as the number of children whose fatigue is increased.
Alternative Hypothesis
H_a: The self-reported level of fatigue for the children who took part in the staff-initiated intervention will decrease from pretest to posttest; that is, the number of children whose fatigue is reduced from pretest to posttest is greater than the number of children whose fatigue is increased.

Overview of the Procedure

To compute the sign test, the differences in the paired data are obtained and the direction of the differences ("+" or "–") recorded and summed. The test then takes the form of a binomial test (see Chapter 4 for calculations) in which the sum of the negative signs is compared to the sum of the positive signs, ignoring the instances of no differences. Depending on the direction stated in the alternative hypothesis, the null hypothesis of no difference between the number of positive and negative signs will be rejected if the probability of obtaining as extreme an occurrence of the obtained values is less than the prestated alpha level.

Either a one- or a two-tailed test may be used, depending on the wording of the alternative hypothesis. In a one-tailed test, the alternative hypothesis states which sign, positive or negative, will occur more frequently. A two-tailed alternative test merely states that there will be a difference in the number of positive and negative signs.

When the sample size is relatively large ($N > 25$ in SPSS for Windows), the normal approximation to the binomial distribution is used for the sign test. This distribution has a mean, μ_x, that is equal to np and a variance, σ_x^2, equal to $n$pq. The value of the z statistic with a continuity correction for categorical data and p = q = .5 is given as follows:

$$\text{z} = \frac{x - \mu_x}{\sigma_x} = \frac{(x \pm 0.5) - n\text{p}}{\sqrt{n\text{pq}}} = \frac{(x \pm 0.5) - 0.5n}{0.5\sqrt{n}}$$

where

x = the number of +'s or −'s, depending on the stated direction of the alternative hypothesis, and

n = the number of paired observations that have been assigned a "+" or "−" value.

When calculating the z statistic, $(x + .5)$ is used when $x < .5n$ and $(x - .5)$ is used when $x > .5n$. The calculated value of this z statistic is then compared to the critical value of the standard normal distribution at the prestated one- or two-tailed alpha level.

Hand-Calculating the Value of the Sign Test

We could hand-calculate the value of this sign test by using the formula given above. The difference in fatigue scores (postintervention fatigue minus preintervention fatigue) is presented in Table 5.5. The data set we are using is hospitalized children with cancer-20 cases.sav (study.sagepub.com/pett2e). These difference scores were obtained in SPSS for Windows (v. 22–23) by using the SPSS commands *Transform . . . Compute Variable . . .* and creating a difference variable by subtracting the preintervention fatigue variable (*Fatigue_T1*) from the postintervention fatigue variable (*Fatigue_T2*) for the staff-initiated intervention group (Group = 1).

Table 5.5 indicates that there were one positive and eight negative changes in the children's fatigue levels. Since our research hypothesis indicated that there would be a reduction in the children's fatigue levels from pre- to postintervention, we are interested in the negative values (see discussion below). Our "x," therefore, is 8, and $n = 9$ since there was one tie. We will also use (x – .5) since 8 > (.5) (9) > 4.5. We will reject the null hypothesis of no change in fatigue levels if and only if our generated z value is less than our one-tailed critical value ($z = 1.64$) at $\alpha = .05$.

Using the formula above, we obtain the following actual value of z:

$$z = \frac{(x - 0.5) - (0.5)n}{0.5\sqrt{n}} = \frac{(9 - 0.5) - (0.5)9}{0.5\sqrt{9}} = \frac{(8.5) - 4.5}{1.5} = 2.67$$

Table 5.5 Difference in Fatigue Scores (Postintervention Minus Preintervention Generated in SPSS for Windows [v. 22–23])

Difference scores: Postintervention Minus Preintervention

		Frequency	*Percent*	*Valid Percent*	*Cumulative Percent*
Valid	–4.00	1	10.0	10.0	10.0
	–3.00	2	20.0	20.0	30.0
	–2.00	2	20.0	20.0	50.0
	–1.00	3	30.0	30.0	80.0
	.00	1	10.0	10.0	90.0
	3.00	1	10.0	10.0	100.0
	Total	10	100.0	100.0	

Reprints Courtesy of International Business Machines Corporation, © International Business Machines Corporation

Since our actual *z* (2.67) is greater than our critical value of *z* (1.64), we will reject the null hypothesis and conclude that the children in the staff-initiated intervention group reduced their self-reported fatigue levels from pre- to postintervention.

Critical Assumptions of the Sign Test

One of the advantages of the sign test is that there are not many assumptions attached to it. Unlike the paired *t* test, the sign test makes no assumptions regarding the form of the distribution of differences between the two variables being examined. The assumptions for this test are as follows.

1. *The data to be analyzed may be dichotomous or continuous. For dichotomous data, there must be some order implied in the coding system (e.g., "0" and "1").* The data that we are examining, pre- and postintervention fatigue, have been measured on a 7-point Likert-type scale and are, therefore, at the ordinal level of measurement.

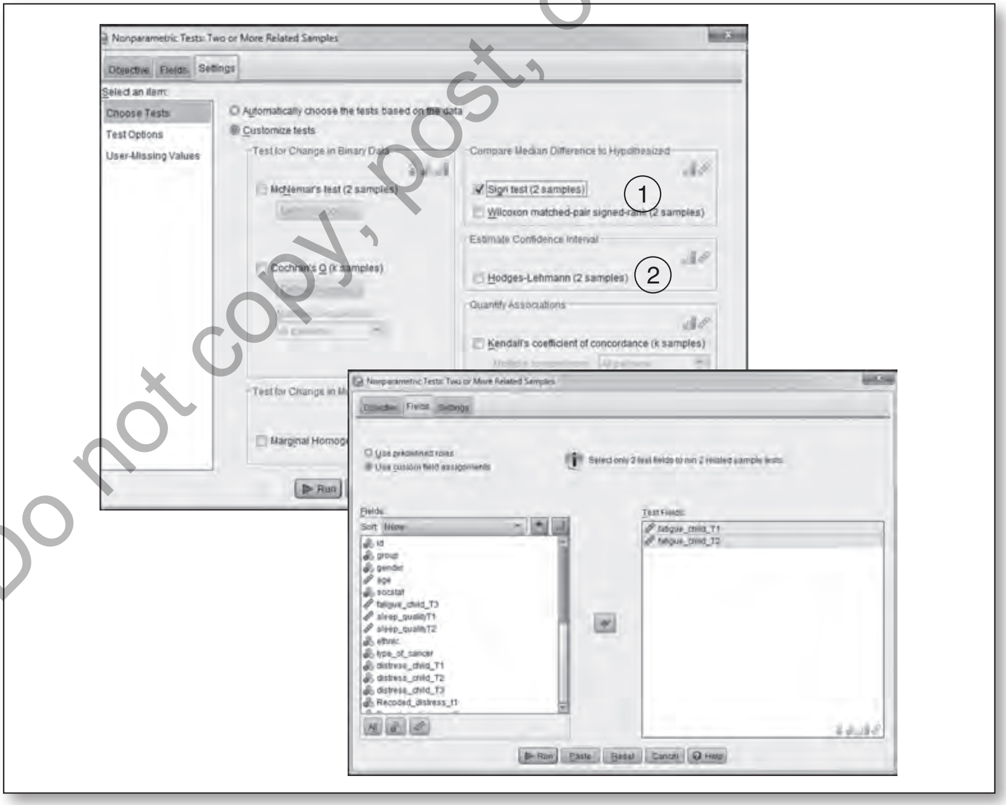
2. *The randomly selected data are paired observations from a single sample, constructed either through matched pairs or through using subjects as their own controls.* The data from our hypothetical intervention study consist of a pre- and postintervention measure that has been conducted on the same sample of children and are, therefore, paired observations. The data are not, however, randomly selected.

Computer Commands

Figure 5.4 presents the SPSS for Windows dialog boxes used to generate the sign test for pre- and postintervention fatigue levels for the 10 children who received the staff-initiated intervention. The data set that we are using is hospitalized children with cancer-20 cases.sav that is found on the SAGE website (study.sagepub.com/pett2e).

As with the McNemar test (Figure 5.2), only those children who were in the staff-initiated intervention were selected (*Data . . . Select Cases . . . If the Condition is Satisfied . . . If Group = 1*). Next, the following items were selected from the drop-down menu: *Analyze . . . Nonparametric tests . . . Related samples*. Note that the Wilcoxon signed-ranks test also can be generated for the fatigue data from the same dialog box by clicking on the appropriate test box (1). A confidence interval for the difference in the pre- and postintervention fatigue medians using the Hodges-Lehmann confidence interval estimation procedure (Hodges & Lehmann, 1963) could also be selected (2). Since the sign test is not concerned with medians but rather with the number of positive versus negative changes, estimating a confidence interval for the medians will be addressed when discussing the Wilcoxon signed-ranks test.

Figure 5.4 SPSS for Windows (v. 22–23) dialog boxes used to generate the sign test. Data set: hospitalized children with cancer-20 cases (study.sagepub.com/pett2e).


Nonparametric Tests: Two or More Related Samples

Objective | Fields | Settings

Select an item:
Choose Tests
Test Options
User-Missing Values

○ Automatically choose the tests based on the data
● Customize tests

Test for Change in Binary Data
□ McNemar's test (2 samples)
□ Cochran's Q (k samples)

Test for Change in M...
□ Marginal Homoge...

Compare Median Difference to Hypothesized
☑ Sign test (2 samples) (1)
□ Wilcoxon matched-pair signed-rank (2 samples)

Estimate Confidence Interval
□ Hodges-Lehmann (2 samples) (2)

Quantify Associations
□ Kendall's coefficient of concordance (k samples)

Nonparametric Tests: Two or More Related Samples

Objective | Fields | Settings

○ Use predefined roles
● Use custom field assignments

Select only 2 test fields to run 2 related sample tests.

Fields:
Sort: None
id
group
gender
age
socstat
fatigue_child_T3
sleep_qualityT1
sleep_qualityT2
ethnic
type_of_cancer
distress_child_T1
distress_child_T2
distress_child_T3
Recoded_distress_t1

Test Fields:
fatigue_child_T1
fatigue_child_T2

Reprints Courtesy of International Business Machines Corporation, © International Business Machines Corporation

Computer-Generated Output

Figure 5.5 presents the syntax commands and computer-generated output for the sign test obtained in SPSS for Windows (v. 22–23). As indicated, the sample has been filtered to include only the 10 children in the staff-initiated intervention group ③. The syntax commands for the sign test are then presented ④.

Recall from Table 5.4 that the alternative hypothesis is directional. We are hypothesizing that the children in the staff-initiated intervention group will report lower levels of fatigue from pre- to postintervention. The test, therefore, is one-tailed. We will reject the null hypothesis if and only if our one-tailed p value is less than the stated alpha (e.g., .05) and if the results are in the predicted direction.

According to the output presented in Figure 5.5, the two-tailed p value is .0391 ⑤, which means that the one-tailed p value is .0391/2 or .0195, which is less than $\alpha = .05$. Note that the output suggests that the null hypothesis should be rejected ⑥. Should we then reject the null? Before that decision is made, it is imperative that we first determine whether the results are in the predicted direction. To do that, those "positive" ($n = 1$) and "negative" ($n = 8$) differences that are presented in the output need to be interpreted ⑦.

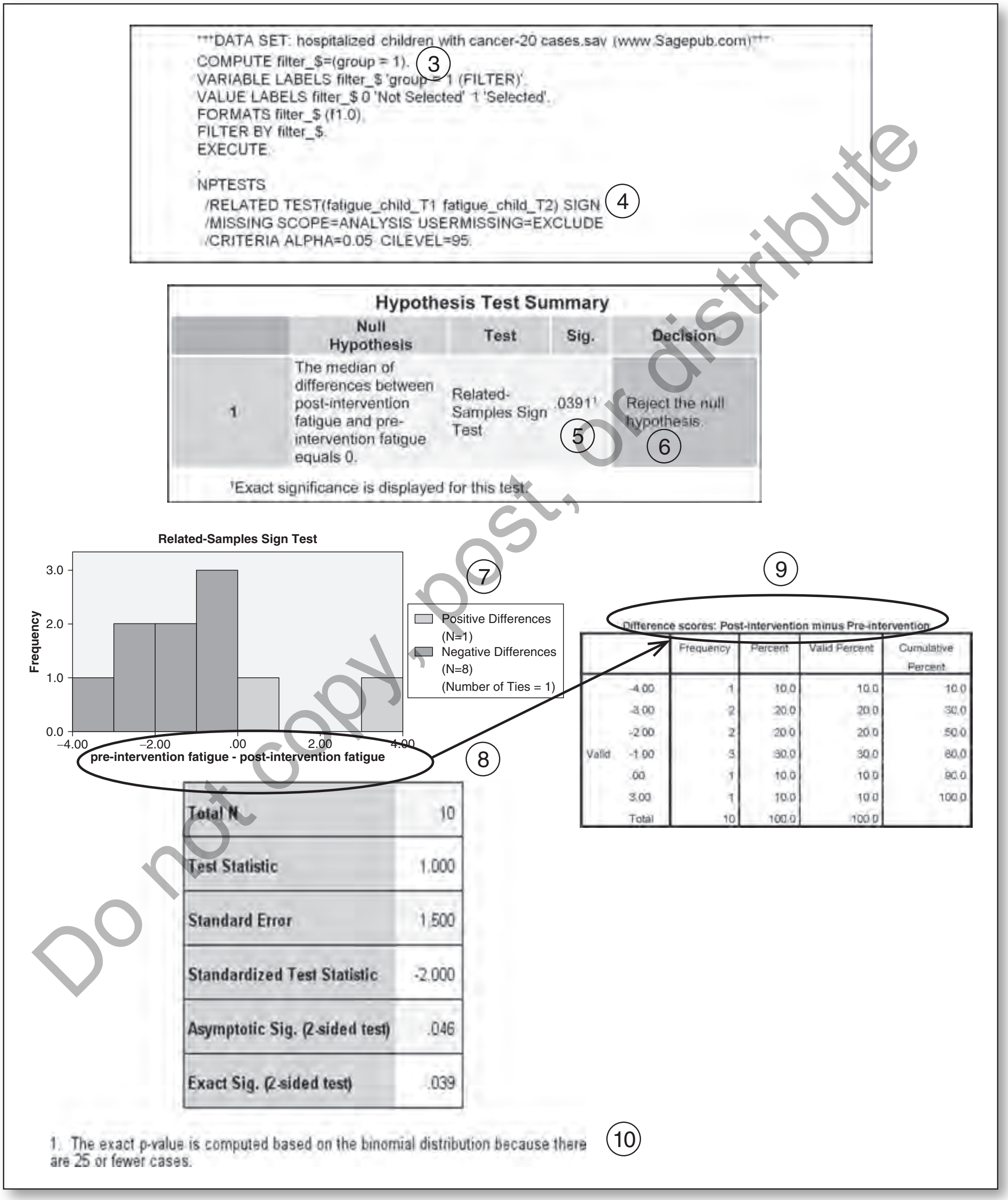
It is important to review the scoring of the fatigue variables. The children indicated on a scale of 1 to 7 (1 = *not at all fatigued* to 7 = *very fatigued*) the extent to which they felt "fatigued." Lower scores, therefore, indicated lower fatigue. Please note that the table element for the results table presented to us on the X-axis (*preintervention fatigue – postintervention fatigue*) is confusing ⑧. If indeed the computer subtracted postintervention scores from preintervention scores, then negative values would mean that postintervention fatigue scores were *higher* than the preintervention scores, a disaster for our staff-initiated intervention. However, the stated null hypothesis suggests the opposite, that the preintervention fatigue scores were subtracted from the postintervention scores ⑨. If so, the negative differences ($n = 8$) represent the number of children whose fatigue at postintervention was *lower* than at preintervention, that the number of positive differences ($n = 1$) represents those children whose fatigue at postintervention was *higher* than at preintervention, and that the number of ties ($n = 1$) represent those children for whom there was no change in fatigue scores. Note, too, that because the sample size was small ($n < 25$), a binomial test was undertaken ⑩.

To determine the meaning of the positive and negative differences, a new variable (*Post_Pre_Fatigue*) was created by using the commands *Transform . . . Compute . . . . Post_Pre_Fatigue = fatigueT2 - fatigueT1*. This new variable represented the differences when the preintervention fatigue scores were subtracted from the postintervention fatigue scores. These results (Figure 5.5 ⑨) indicate that indeed the negative values represent those children who reported higher fatigue scores at preintervention. Given that these results are in the predicted direction, we will reject the null hypothesis and conclude that there was a significant reduction in self-reported levels of fatigue from pretest to posttest among the children who took part in the intervention.

Using Internet Resources to Determine the Outcome of a Sign Test

As with the McNemar test, several useful websites provide interactive statistical resources that enable the researcher to generate findings from a sign test without the

Figure 5.5 SPSS for Windows (v. 22–23) syntax and computer-generated output for the sign test.



```
***DATA SET: hospitalized children with cancer-20 cases.sav (www.Sagepub.com)***
COMPUTE filter_$=(group = 1).
VARIABLE LABELS filter_$ 'group = 1 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.
.
NPTESTS
  /RELATED TEST(fatigue_child_T1 fatigue_child_T2) SIGN
  /MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE
  /CRITERIA ALPHA=0.05 CILEVEL=95.
```

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The median of differences between post-intervention fatigue and pre-intervention fatigue equals 0.	Related-Samples Sign Test	.039[1]	Reject the null hypothesis.

[1]Exact significance is displayed for this test.

Difference scores: Post-intervention minus Pre-intervention

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	-4.00	1	10.0	10.0	10.0
	-3.00	2	20.0	20.0	30.0
	-2.00	2	20.0	20.0	50.0
	-1.00	3	30.0	30.0	80.0
	.00	1	10.0	10.0	90.0
	3.00	1	10.0	10.0	100.0
	Total	10	100.0	100.0	

Total N	10
Test Statistic	1.000
Standard Error	1.500
Standardized Test Statistic	-2.000
Asymptotic Sig. (2-sided test)	.046
Exact Sig. (2-sided test)	.039

1. The exact p-value is computed based on the binomial distribution because there are 25 or fewer cases.

Reprints Courtesy of International Business Machines Corporation, © International Business Machines Corporation

need for a stand-alone statistical computer package. One such website that is in the public domain is http://www.socr.ucla.edu. While this is not the only free site available, it should illustrate how one might use such a website to generate the sign test. We will also use the same site to generate other nonparametric statistics such as the Wilcoxon signed-ranks test.

Given that the data for pre- and postintervention fatigue are ordinal level of measurement, we can no longer simply enter the numbers into a 2 × 2 table as we did for the McNemar test. Instead, the data need to be downloaded into the site via a spreadsheet program such as Excel. This example is available to you in Excel format (hospitalized children with cancer-20 cases.xlsx) at study.sagepub.com/pett2e. You may also need to upgrade your Java script.

After accessing the website and the spreadsheet (Figure 5.6), you will need to first click on the *Analyses* button ⑪ and then indicate which test you wish to undertake (e.g., *two paired samples sign test*) ⑫. Next, return to your own spreadsheet containing the data of interest, copy, and, using the *Paste* button, ⑬ paste the values that will be used for the sign test into the SOCR spreadsheet presented on the website (Figure 5.6 ⑭). Note that the data of interest in this example are those children who were in the staff-initiated intervention. If desired, you can also change the variable names from C1 and C2 to Fatigue_t1 and Fatigue_t2 ⑮.

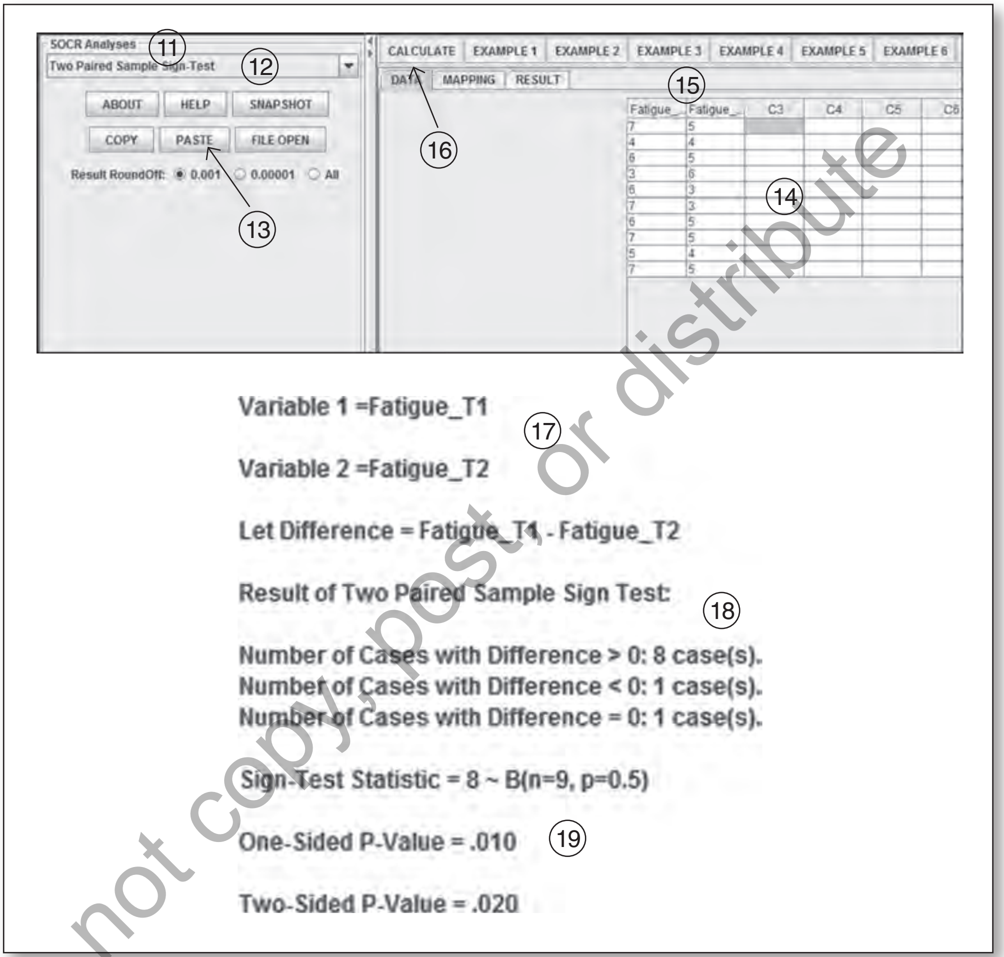
By clicking on the *Calculate* button ⑯, the output for the sign test is generated (Figure 5.6). Notice that the difference between the two variables is *Fatigue_T1 – Fatigue_T2* ⑰. That means that the number of cases with differences >0 are those whose fatigue scores were higher at preintervention than postintervention ⑱. Our one-tailed p value is .01 ⑲, which is lower than the SPSS-generated one-tailed p value (.0195) but still small enough to reject the null hypothesis. Why this discrepancy? Because the sample size was less than 25, both programs estimated the p values based on the binomial distribution, so one would expect the resulting p values to be similar. In fact, Conover (1999) gives a similar example (pp. 161–162) in which the resulting one-tailed p value is .0195 when $n = 9$ (ties are not counted), the number of positive (as opposed to negative) differences = 8, and $p = .50$. It is apparent, therefore, that the discrepancy seems to be with the p value generated from the SOCR website.

Presentation of Results

Table 5.6 presents an example of how the results for the sign test might be reported. The information for this type of table can be obtained in SPSS for Windows (v. 22–23) by clicking on *Analyze . . . Tables . . . Custom Tables . . .* bringing the variables of interest (e.g., *Fatigue_T1* and *Fatigue_T2*) into the drawing frame and choosing the summary statistics that are desired. Be sure that the appropriate cases have been selected for presentation (e.g., *Data . . . Select Cases . . . If the condition is satisfied (Group = 1)*).

Notice that the median is presented along with the mean and standard deviation. Given that the sign test is nonparametric, some authors might prefer to limit the presentation to only the medians. Because the sample size was small, the binomial distribution was used to evaluate the sign test. For that test, only the p value can be presented in the

Figure 5.6 Spreadsheet format and output generated for the sign test from the website http://www.socr.ucla.edu.



Fatigue_	Fatigue_	C3	C4	C5	C6
7	5				
4	4				
6	5				
3	6				
6	3				
7	3				
6	5				
7	5				
5	4				
7	5				

```
Variable 1 =Fatigue_T1

Variable 2 =Fatigue_T2

Let Difference = Fatigue_T1 - Fatigue_T2

Result of Two Paired Sample Sign Test:

Number of Cases with Difference > 0: 8 case(s).
Number of Cases with Difference < 0: 1 case(s).
Number of Cases with Difference = 0: 1 case(s).

Sign-Test Statistic = 8 ~ B(n=9, p=0.5)

One-Sided P-Value = .010

Two-Sided P-Value = .020
```

Table 5.6 Suggested Presentation of Sign Test Results

Fatigue Scores	N	*Mean*	*Median*	*Standard Deviation*	p[a]
Preintervention	10	5.8	6.0	1.4	.019
Postintervention	10	4.4	4.5	1.0	

[a]The calculated one-tailed *p* value is for the sign test.

table. For larger sample sizes, the normal approximation to the binomial distribution is used. For that reason, the generated *z* statistic also could be presented.

The results from statistical analysis using the sign test could also be more easily presented in the text as follows:

> The results of the sign test analysis indicated that the 10 children who took part in the staff-initiated intervention significantly reduced their median fatigue levels from preintervention (*Md* = 6.0) to postintervention (*Md* = 4.5) (one-tailed *p* = .019).

Advantages, Limitations, and Alternatives to the Sign Test

The sign test is a versatile, simple, and easy-to-apply statistical test that can be used to determine whether one variable tends to be larger than another. It also can be used to test for trends in a series of ordinal measurements (Conover, 1999) or as a quick assessment of direction in an exploratory study. The disadvantage of this test is that it does not take into account the order of magnitude of the differences between two paired variables. When data are at least ordinal in level of measurement, the Wilcoxon signed-ranks test is preferred.

The parametric alternative to the sign test is the paired *t* test. Both Siegel and Castellan (1988) and Walsh (1946) report that the sign test is about 95% as efficient as the paired *t* test. Recall from Chapter 3 that *power efficiency* refers to the sample size that is required for one test (e.g., the sign test) to be as powerful as its rival (e.g., the paired *t* test) given the same alpha level and that the assumptions of both tests have been met. A 95% efficiency rating implies that, for small samples, only 20 cases are needed for the sign test to achieve the same power as the paired *t* test with 19 cases (i.e., N_2/N_1 [100%] = 19/20 [100%] = 95%). This suggests that the sign test is especially useful for small sample sizes and in situations in which meeting the assumptions of the robust paired *t* test either is not possible (e.g., the data are nominal) or is questionable (e.g., a severely skewed distribution with small sample sizes). A more powerful nonparametric alternative to the sign test when the data are at least ordinal in level of measurement is the Wilcoxon signed-ranks test, which makes better use of the quantitative differences between the paired observations.

Examples From Published Research

Graves, K. D., Carter, C. L., Anderson, E. S., & Winett, R. A. (2003). Quality of life pilot intervention for breast cancer patients: Use of social cognitive theory. *Palliative & Supportive Care, 1*(2), 121–134.

Miletic, D., Sekulic, D., & Ostojic, L. (2007). Body physique and prior training experience as determinants of SEFIP score for university dancers. *Medical Problems of Performing Artists, 22*(3), 110–115.

Whellan, D. J., Droogan, C. J., Fitzpatrick, J., Adams, S., McCarey, M. M., Andrel, J., . . . Keith, S. (2012). Change in intrathoracic impedance measures during acute decompensated heart failure admission: Results from the Diagnostic Data for Discharge in Heart Failure Patients (3D-HF) pilot study. *Journal of Cardiac Failure, 18*(2), 107–112.

THE WILCOXON SIGNED-RANKS TEST

The reduction of data in the sign test to +'s or –'s results in the loss of potentially important quantitative information: the *size* of the differences between two paired variables. In our fatigue data, for example, no use is made by the sign test of the information that 5 of the 10 children reduced their fatigue by more than 2 points and that one child increased his fatigue by 3 points (Table 5.5). By taking into account the magnitude and the direction of changes, the Wilcoxon signed-ranks test, which was developed by Wilcoxon (1945), produces a more sensitive statistical test. It is used with paired data that are measured on at least the ordinal scale and is especially effective when the sample size is small and the distribution of the data to be examined does not meet the assumptions of normality, as is required in the paired *t* test.

An Appropriate Research Question for the Wilcoxon Signed-Ranks Test

The Wilcoxon signed-ranks test has been used widely in the health care research literature. It is a very flexible test that can be used in a variety of situations with different sample sizes and few restrictions. The only requirements are that the data be at least ordinal level of measurement and be paired observations; that is, there are either pretest-posttest measures for a single sample, or subjects who have been matched on certain criteria.

This test has been used frequently in the research literature to evaluate changes in attitudes on a variety of topics, such as changes over time in satisfaction with health care and medical mistrust among Native American cancer patients (Guadagnolo, Cina, Koop, Brunette, & Petereit, 2011) and nursing students' attitudes toward Australian Aborigines (Hayes, Quine, & Bush, 1994). It has been particularly useful in evaluating the effectiveness of interventions, such as the effects of a cardiac rehabilitation paradigm for nonacute ischemic stroke patients (Lennon, Carey, Gaffney, Stephenson, & Blake, 2008), a pilot walking program for Mexican American women living in colonias at the border (Mier et al., 2011), magnetic resonance imaging in patients with low-tension glaucoma (Stroman, Stewart, Golnik, Curé, & Olinger, 1995), and the effects of a mindfulness stress reduction program on distress in a community-based sample (Evans, Ferrando, Carr, & Haglin, 2011).

Numerous assessments have also been made between two alternative approaches to data collection methods. For example, Vereecken, Covents, and Maes (2010) compared a food frequency questionnaire with an online dietary assessment tool for assessing preschool children's dietary intake. Waninge and colleagues (Waninge, Evenhuis, van Wijck, & van der Schans, 2011; Waninge, van der Weide, Evenhuis, van Wijck, & van der Schans, 2009) used the Wilcoxon to evaluate the feasibility and reliability of body composition measurements and two different walking tests in adults with severe intellectual and sensory disabilities. The Wilcoxon signed-ranks test was also used by Bowring et al. (2012) to measure the accuracy of self-reported height and weight in a community-based sample of young people. Bottom line, there are limitless possibilities for the application of the Wilcoxon signed-ranks test.

In our hypothetical intervention study, we will continue to use the fatigue data that were collected on the children at preintervention and then immediately following the staff-initiated intervention. This will enable us to compare the results that we obtain from the Wilcoxon signed-ranks test with those from the sign test. A research question similar to that of the sign test could, therefore, be asked:

> Do the children in the staff-initiated intervention group reduce their fatigue from pretest to posttest?

As with the sign test, the Wilcoxon signed-ranks test can only examine changes in one group over time.

Null and Alternative Hypotheses

Table 5.7 presents an example of null and alternative hypotheses that would be appropriate for the Wilcoxon signed-ranks test. Note that this nonparametric test examines the differences between medians, not means. Because our alternative hypothesis is directional (i.e., we are predicting a drop in fatigue level following our intervention), the test that will be undertaken is one-tailed. Our level of alpha for this test will remain the same as before ($\alpha = .05$).

Table 5.7 Example of Null and Alternative Hypotheses Appropriate for the Wilcoxon Signed-Ranks Test

Null Hypothesis
H_0: The median fatigue scores of the children who took part in the staff-initiated intervention will not change from pretest to posttest (i.e., $Md_{pretest} = Md_{posttest}$).
Alternative Hypothesis
H_a: The median posttest fatigue scores of the children who took part in the staff-initiated intervention will be lower than at pretest (i.e., $Md_{pretest} > Md_{posttest}$).

Overview of the Procedure

To conduct the Wilcoxon signed-ranks test, the differences between the paired data are calculated and the absolute values of these differences are recorded. Next, the absolute values of the differences between the two variables are ranked from lowest to highest. Finally, each rank is given a positive or negative sign depending on the sign of the original difference. The positive and negative ranks are then summed and averaged. Pairs that indicate no change are dropped from the analysis.

A z statistic is used to test the null hypothesis of no differences in the matched pairs. This z statistic takes the following form:

$$z = \frac{x - \mu}{\sigma} = \frac{T - [n\,(n+1)/4]}{\sqrt{n(n+1)\,(2n+1)/24}}$$

where

T = the absolute value of the sum of the positive or negative ranks, depending on the proposed alternative hypothesis, and

n = the number of positive and negative ranks, excluding ties.

If the null hypothesis is true, the absolute value of the sum of the positive ranks should be nearly equal to the absolute value of the sum of the negative ranks. If the differences in positive and negative ranks are sufficiently large, the null hypothesis is rejected. Either a one- or a two-tailed test is undertaken, depending on the wording of the alternative hypothesis.

Hand-Calculating the Value of the Wilcoxon Signed-Ranks Test

We could arrive at a hand-calculated value for the Wilcoxon signed-ranks test using the test statistic outlined above. We have eight negative ranks whose sum (T), according to Table 5.5, is the absolute value of $(-4)1 + (-3)2 + (-2)2 + (-1)3 = |-38| = 38$. Since $n = 9$ (1 positive + 8 negative ranks = 9), the actual value of our z, therefore, would be as follows:

$$z = \frac{T - \left[n\left(n+1\right)/4\right]}{\sqrt{\frac{n\left(n+1\right)\left(2n+1\right)}{24}}} = \frac{38 - \left[9\left(9+1\right)/4\right]}{\sqrt{\frac{9\left(9+1\right)\left(\left(2\right)9+1\right)}{24}}} = \frac{38 - 22.5}{8.44} = 1.84$$

Since the research hypothesis in Table 5.7 is directional and our one-tailed $\alpha = .05$, the critical value of our z statistic will be +1.64. We will reject the null hypothesis if and only if the actual value of our z statistic is greater than our critical value. Since 1.84 is greater than 1.64, we will reject the null hypothesis and conclude that, according to the Wilcoxon signed-ranks test, the children in the staff-initiated intervention reported statistically significantly lower levels of fatigue following the intervention.

Critical Assumptions of the Wilcoxon Signed-Ranks Test

The assumptions of the Wilcoxon signed-ranks test are fairly liberal.

1. *The data are paired observations from a single randomly selected sample, constructed either through matched pairs or through using subjects as their own controls.* It is assumed either that the data being analyzed are test-retest measures of the same

group of randomly selected subjects or that the data have been collected from subjects who have been paired on one or more variables. The data for our hypothetical study only partially meet this assumption. Although the fatigue data consist of Time 1 and Time 2 measures for the same sample of 10 children who took part in the staff-initiated intervention, our sample is a nonrandom sample of convenience.

2. *The data to be analyzed must be at least ordinal in level of measurement, both within and between pairs of observations.* This assumption means that not only must the variables themselves be at least ordinal in level of measurement, but the generated values of the difference scores must also be at least ordinal level of measurement. In fact, Daniel (2000) indicates that these differences should be measured on at least an interval scale.

The fatigue data from our hypothetical intervention study consist of two pretest and posttest Likert-type scale measurements (1 = *not at all fatigued* to 7 = *extremely fatigued*). Both of these scales and their difference scores are at least ordinal in level of measurement.

3. *There is symmetry of the difference scores about the true median for the population.* This assumption implies that, if it were possible to view the distribution of the difference scores in the population, the distribution of these difference scores would be symmetric (though not necessarily normal) about the population median (Daniel, 2000). One approach to assessing this third assumption might be to plot the difference scores for the sample to assess their symmetry. Figure 5.7 presents the plot that was generated for the DIFF12 variable that was created by subtracting the children's Fatigue_T1 scores from their Fatigue_T2 scores using the *Transform . . . Compute* commands. This histogram was obtained by opening the *Statistics . . . Summarize . . . Frequencies* dialog box and selecting histograms from the *Charts* option. The histogram indicates that although the data for the 20 children are not completely symmetric, they are not badly skewed. We could conclude, therefore, that we approach meeting this assumption.

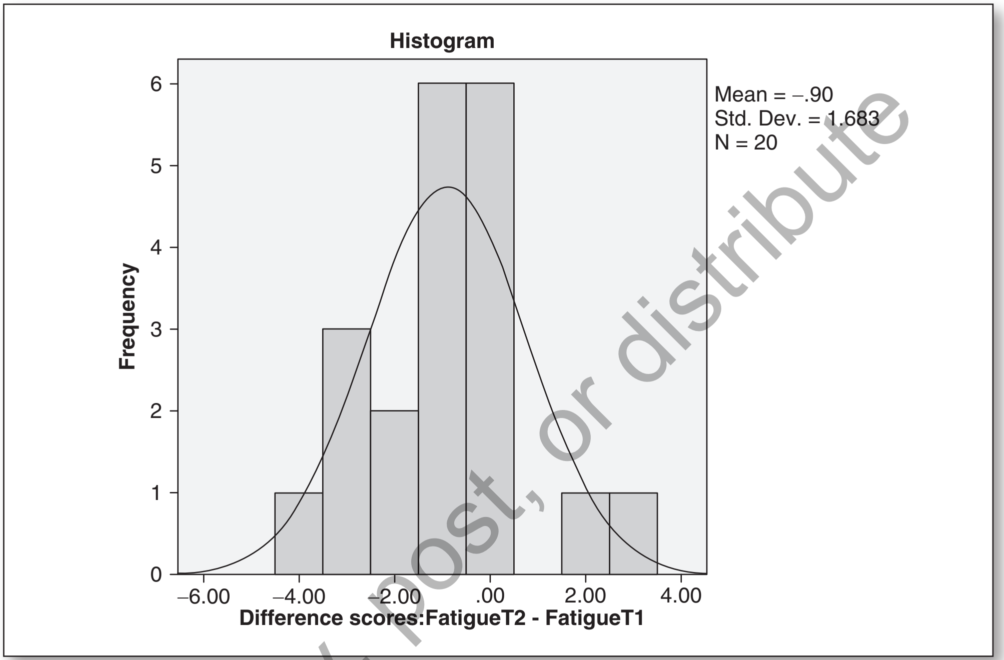
Computer Commands

The SPSS for Windows (v. 22–23) dialog box that was used to generate the sign test (Figure 5.4) also produces the Wilcoxon signed-ranks test. The data set used was hospitalized children with cancer-20 cases.sav (study.sagepub.com/pett2e). The Wilcoxon signed-ranks test was obtained by clicking on the *Wilcoxon* box under *Compare median differences to hypothesized.* Note, too, that we are also asking for the Hodges-Lehmann 95% confidence interval for the difference in the medians as well.

Computer-Generated Output

Figure 5.8 presents the syntax commands and computer-generated output from SPSS for Windows (v. 22–23) for both the Wilcoxon signed-ranks test and the Hodges-Lehmann 95% confidence interval. As with the sign test, we are interested only in the results for the 10 children in the intervention group since the Wilcoxon can examine

Figure 5.7 Histogram of difference scores: Fatigue2 – Fatigue1 generated in SPSS for Windows (v. 22-23). Data set: Hospitalized children with cancer-20 cases.sav (study.sagepub.com/pett2e).



Reprints Courtesy of International Business Machines Corporation, © International Business Machines Corporation

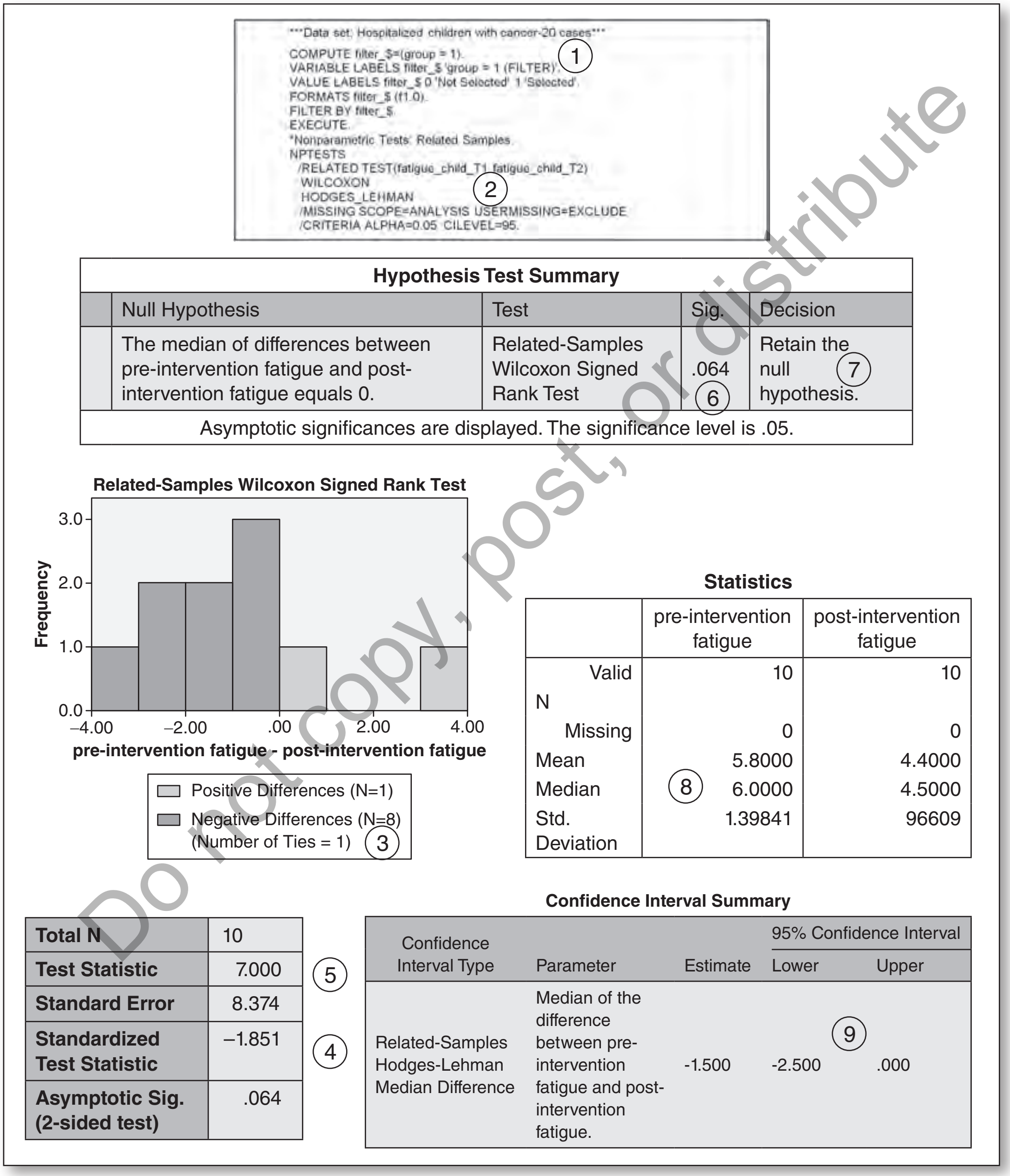
only one group at a time. Therefore, the *Select Cases . . .* command obtained from the *Data* menu is operative (1). The SPSS for Windows syntax commands for the Wilcoxon signed-ranks test are then presented along with the request for the Hodges-Lehmann confidence interval (2).

The computer-generated printout for the Wilcoxon signed-ranks test indicates that eight children had a negative rank (their scores postintervention were lower than preintervention), one child's fatigue level increased from pre- to postintervention, and one child did not alter his or her fatigue level (3). This is the same information that we obtained for the sign test.

Notice that the *z* statistic generated in SPSS for Windows (–1.851) (4) is a negative value. The *T* value used by SPSS for Windows was the *sum* of the positive differences ([7][1] = 7) instead of the sum of the negative values (38) (5). It is also not exactly clear why the absolute value of the *z* statistic (1.851) should be slightly larger than our hand-calculated value (1.84).

The two-tailed *p* value for this *z* statistic is .064 (6). If our alternative hypothesis test had been nondirectional, we would not have been able to reject the null hypothesis

Figure 5.8 Syntax and SPSS output for the Wilcoxon signed-ranks test, the Hodges-Lehmann 95% confidence interval, and the descriptive statistics for pre- and postintervention fatigue.



```
***Data set: Hospitalized children with cancer-20 cases***
COMPUTE filter_$=(group = 1).
VARIABLE LABELS filter_$ 'group = 1 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.
*Nonparametric Tests: Related Samples.
NPTESTS
 /RELATED TEST(fatigue_child_T1 fatigue_child_T2)
  WILCOXON
  HODGES_LEHMAN
 /MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE
 /CRITERIA ALPHA=0.05 CILEVEL=95.
```

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
	The median of differences between pre-intervention fatigue and post-intervention fatigue equals 0.	Related-Samples Wilcoxon Signed Rank Test	.064	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Related-Samples Wilcoxon Signed Rank Test

Frequency
3.0
2.0
1.0
0.0
–4.00
–2.00
.00
2.00
4.00
pre-intervention fatigue - post-intervention fatigue
Positive Differences (N=1)
Negative Differences (N=8)
(Number of Ties = 1)
1
2
3
4
5
6
7
8
9

Statistics

		pre-intervention fatigue	post-intervention fatigue
N	Valid	10	10
	Missing	0	0
Mean		5.8000	4.4000
Median		6.0000	4.5000
Std. Deviation		1.39841	96609

Total N	10
Test Statistic	7.000
Standard Error	8.374
Standardized Test Statistic	–1.851
Asymptotic Sig. (2-sided test)	.064

Confidence Interval Summary

Confidence Interval Type	Parameter	Estimate	95% Confidence Interval Lower	Upper
Related-Samples Hodges-Lehman Median Difference	Median of the difference between pre-intervention fatigue and post-intervention fatigue.	-1.500	-2.500	.000

Reprints Courtesy of International Business Machines Corporation, © International Business Machines Corporation

because this two-tailed *p* value (.064) is greater than a two tailed α = .05. Note that the decision suggested by the output is to retain the null hypothesis ⑦. Our alternative hypothesis, however, was directional, in that we stated that the children would have lower median fatigue scores at postintervention than at preintervention (Table 5.7). The output that we have obtained from running the frequency statistics for the two variables (*Analyze . . . Frequencies . . .*) (Figure 5.8) supports our predicted direction in that the median postintervention fatigue score (4.5) *is* lower than the median pretest value (6.0) ⑧.

To determine whether this difference in medians is large enough to reject the null hypothesis, we need to compare our one-tailed α = .05 to half the presented two-tailed *p* value (.064/2 = .032). Because this one-tailed *p* value, .032, is less than our α = .05 and our medians are in the direction predicted, we can reject the null hypothesis of equal medians. Our conclusion, therefore, is that the intervention group significantly reduced its self-reported fatigue from pre- to postintervention.

Note that this generated one-tailed *p* value for the Wilcoxon test, .032, is greater than the *p* value that was generated for the sign test, .019 (Table 5.6). The reason for this discrepancy is that the Wilcoxon signed-ranks test is picking up the *size* of the negative and positive ranks, not just noting the direction of differences. The significance of the Wilcoxon statistic has been influenced by the fact that the one child who did increase his or her fatigue from pretest to posttest did so substantially.

Hodges-Lehmann Confidence Interval for the Wilcoxon Signed-Ranks Test

The (1 – α)100% confidence interval for median differences for both paired and independent samples was developed by Hodges and Lehmann (1963) and is described in detail in Lehmann's (2006) textbook on nonparametric statistics. Using the normal approximation to the Wilcoxon signed-ranks distribution, an asymptotic confidence interval can be obtained as illustrated in Han (2009) and Newson (2006, 2007).

For our data set, the interpretation for such a confidence interval would be as follows: *We can be (1 – α)100% confident that the true median difference in the children's pre- and postintervention fatigue scores for the population lies between ____ and ____.* If "0" falls in that confidence interval, we cannot reject the null hypothesis because it could conceivably be the true median difference.

The Hodges-Lehmann 95% confidence interval (CI) for the median differences is presented in Figure 5.8 ⑨. This 95% CI has a lower limit of –2.500 and an upper limit of .000. Note that because this CI contains "0," we should not reject the null hypothesis. Yet our one-tailed *p* value (.032) indicated that we should reject the null since .032 < .05. What is going on here?

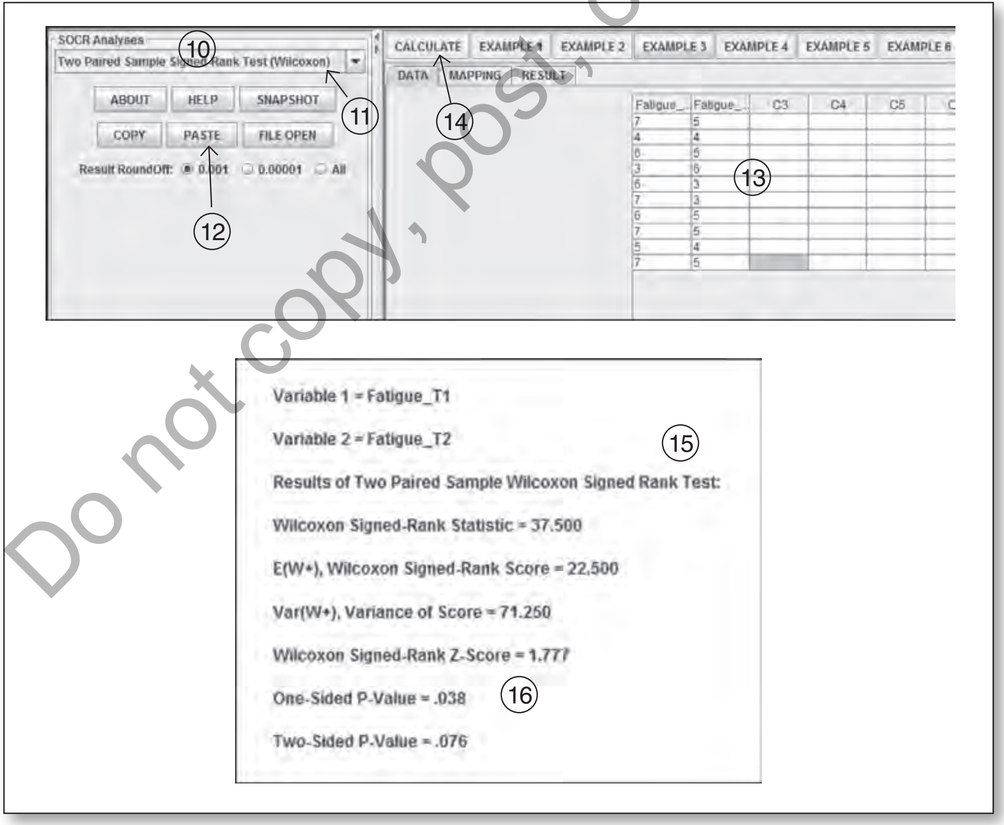
The answer is rather straightforward. The 95% CI is two-tailed. Therefore, using a 95% CI, we could not reject the null as indicated in the printout in Figure 5.8 ⑦. To build a more appropriate CI when we have a one-tailed test (and α = .05), we should use a 90% CI instead. This 90% CI would result in a range of values between –2.500 and –.500. Since this range does not contain "0," we can reject the null hypothesis of no difference in medians.

Using Internet Resources to Determine the Outcome of the Wilcoxon Signed-Ranks Test

As we did with the sign test, we will generate the results of the Wilcoxon signed-ranks test using the website http://www.socr.ucla.edu (Figure 5.9). Again, this is not the only site that is available to calculate the Wilcoxon but is illustrative of what is possible using free websites available in the public domain.

After accessing the website and the spreadsheet (Figure 5.9), we will first click on the *Analyses* button ⑩ and then indicate that we would like to undertake the Wilcoxon signed-ranks test (e.g., *two paired sample signed rank test[Wilcoxon]*) ⑪. Next, using the Excel workbook (Hospitalized Children with Cancer-20 cases.xlsx) that is available on the website, study.sagepub.com/pett2e, highlight and copy the Fatigue_T1 and Fatigue_T2 data for the group that we are interested in evaluating (e.g., the staff-initiated

Figure 5.9 Internet-generated output for the Wilcoxon signed-ranks test (http://www.socr.ucla.edu).



Fatigue_...	Fatigue_...	C3	C4	C5
7	5			
4	4			
6	5			
3	6			
6	3			
7	3			
6	5			
7	5			
5	4			
7	5			

Variable 1 = Fatigue_T1

Variable 2 = Fatigue_T2

Results of Two Paired Sample Wilcoxon Signed Rank Test:

Wilcoxon Signed-Rank Statistic = 37.500

E(W+), Wilcoxon Signed-Rank Score = 22.500

Var(W+), Variance of Score = 71.250

Wilcoxon Signed-Rank Z-Score = 1.777

One-Sided P-Value = .038

Two-Sided P-Value = .076

CC-BY License: http://socr.ucla.edu/htmls/SOCR_CitingLicense.html intervention or Group = 1). Bring the data over into the www.socr.ucla.edu spreadsheet and, using the *Paste* button ⑫, paste the values that will be used for the Wilcoxon signed-ranks test into the SOCR spreadsheet presented on the website (Figure 5.9). Note that the data of interest in this example are those children who were in the staff-initiated intervention. If desired, you can also change the variable names from C1 and C2 to Fatigue_t1 and Fatigue_t2 ⑬.

By clicking on the *Calculate* button ⑭, the output for the Wilcoxon signed-ranks test is generated (Figure 5.9) ⑮. While the resulting one-tailed p value (.033) ⑯ is negligibly higher than the SPSS-generated one-tailed p value (.032) (Figure 5.8), it is still low enough to reject the null hypothesis. It does not appear that this website provides us with the Hodges-Lehmann confidence interval for the median difference.

Presentation of Results

The Wilcoxon signed-ranks test can be presented in tabular form in much the same way as the sign test (Table 5.6). The only change that would be required would be the size of the p value (.03) and the type of test reported (Wilcoxon signed-ranks test). The results could also be presented in the text as follows:

> The results of the Wilcoxon signed-ranks test indicated that the 10 children who took part in the staff-initiated intervention significantly reduced their self-reported fatigue levels from preintervention (*Md* = 6.0) to postintervention (*Md* = 4.5) (p = .03). The Hodges-Lehmann 90% confidence interval for the median difference ranged from –2.500 to –.500.

Advantages, Limitations, and Alternatives to the Wilcoxon Signed-Ranks Test

The Wilcoxon signed-ranks test is very easy to apply and has an advantage over the sign test in that it uses more of the information provided by the data. There are no specific limitations that have been identified for this test.

The parametric alternative to the Wilcoxon signed-ranks test is the t test for related samples or paired t test. This is the statistical test of choice if the data are found to be normally distributed. The paired t test, however, has been found to be unsatisfactory when the distributions of the variables being considered have heavy tails (Wilcox, 1992). Blair and Higgins (1985) used Monte Carlo methods to assess the relative power of the paired t test and the Wilcoxon signed-ranks test under 10 different distributional shapes. They report that, when the data were not normally distributed, the Wilcoxon was more often the more powerful test and that the magnitude of the Wilcoxon's power advantage over the paired t test often increased with the sample size. Lehmann (2006) also compared the Wilcoxon to the t test and reported that, when the distribution is normal, the t test is more powerful but that the efficiency loss of the Wilcoxon is slight (about 5%). When the shape of the distribution is unknown, however, the Wilcoxon may be considerably more efficient than the t test.

The nonparametric alternatives to this test are the sign and binomial tests. As indicated above, these tests are not as sensitive as the Wilcoxon signed-ranks test but could be used if that test's assumptions are not sufficiently met.

Examples From Published Research

Chaplin, D., Deitz, J., & Jaffe, K. M. (1993). Motor performance in children after traumatic brain injury. *Archives of Physical Medicine and Rehabilitation, 74,* 161–164.

Guadagnolo, B. A., Cina, K., Koop, D., Brunette, D., & Petereit, D. G. (2011). A pre-post survey analysis of satisfaction with health care and medical mistrust after patient navigation for American Indian cancer patients. *Journal of Health Care for the Poor and Underserved, 22*(4), 1331–1343. doi: 10.1353/hpu.2011.0115

Hayes, L., Quine, S., & Bush, J. (1994). Attitude change amongst nursing students towards Australian Aborigines. *International Journal of Nursing Studies, 31*(1), 67–76.

Lennon, O., Carey, A., Gaffney, N., Stephenson, J., & Blake, C. (2008). A pilot randomized controlled trial to evaluate the benefit of the cardiac rehabilitation paradigm for the non-acute ischaemic stroke population. *Clinical Rehabilitation, 22*(2), 125–133. doi: 10.1177/02692 15507081580

Mier, N., Tanguma, J., Millard, A. V., Villarreal, E. K., Alen, M., & Ory, M. G. (2011). A pilot walking program for Mexican-American women living in colonias at the border. *American Journal of Health Promotion, 25*(3), 172–175. doi: 10.4278/ajhp.090325-ARB-115

Novack, C. M., Waffarn, F, Sills, J. H., Pousti, T. J., Warden, M. J., & Cunningham, M. D. (1994). Focal intestinal perforation in the extremely-low-birth-weight infant. *Journal of Perinatology, 14,* 450–453.

Summary

In this chapter, we have examined three nonparametric statistical techniques that could be used when data collected from a single sample are paired through using subjects either as their own controls (e.g., pretest and posttest measures) or as matched pairs (e.g., husband-wife pairs). The McNemar test is used when the pretest-posttest data being examined are dichotomous. The sign and Wilcoxon tests require that the distribution of the variable being considered be continuous. Although the sign test can be used with ordered dichotomous data, the Wilcoxon signed-ranks test assumes at least an ordinal level of measurement.

In Chapter 6, several nonparametric tests that can be used when data have been collected over more than two time periods will be presented. These tests include Cochran's *Q* test, which is used with dichotomous data, and the Friedman test, which is used with continuous data.

TEST YOUR KNOWLEDGE

Here is a "test" of your knowledge on the main points regarding the various nonparametric statistics that have been discussed in this chapter. You will want to reread the chapter should you find that you cannot recall their content.

1. What are the main differences between the McNemar, sign, and Wilcoxon signed-rank tests, and when would you consider using each of these nonparametric tests?
2. What are the critical assumptions of each of these three nonparametric tests? Did the distress and fatigue data that were used in this chapter meet those assumptions?
3. Why was it necessary to separate out the staff-initiated intervention group from the usual-care group when running the three nonparametric tests discussed in this chapter?
4. Why was it preferable to change from a 95% to a 90% confidence interval when reporting the median difference for our one-tailed test at $\alpha = .05$?

COMPUTER EXERCISES

The following computer exercises should enable you to build on your skills in using SPSS for Windows, Excel, and/or Internet-based programs.

1. Using either the SPSS for Windows data set (hospitalized children with cancer -20 cases.sav) or Excel spreadsheet (hospitalized children with cancer -20 cases.xlsx) made available to you at study.sagepub.com/pett2e, undertake the sign and Wilcoxon signed-ranks test for the pre- and postintervention fatigue data (*fatigue_T1, fatigue_t2*) for the usual-care group (Hint: be sure to select only those cases for whom *Group = 2*). Set a two-tailed alpha at .05 and request a Hodges-Lehmann 95% confidence interval (CI) for the median difference.

   a. Looking at your results in Question 1, what was your decision with regard to the null hypotheses for these two tests? Does the Hodges-Lehmann 95% CI support your decision? Why or why not?
   b. Compare the results that you have obtained for the usual-care group with that of the staff-initiated intervention as detailed in this chapter. Were the results for the two groups different or similar? What conclusions would you draw from your results?

2. Eighty participants, ages 48 to 49 years, who were enrolled in a workplace health maintenance plan attended a 3-hour educational workshop addressing the need for routine colorectal cancer screening after age 50. Prior to the workshop, the participants were asked whether they would be likely to obtain a routine colonoscopy once they reached age 50. This same question was asked again immediately following the workshop. The research hypothesis for this study was that the educational workshop would increase the likelihood that the participants would undergo a colonoscopy screening once they reached age 50.

(Continued)

(Continued)

a. What nonparametric statistical test would you use to analyze these data?

b. State the null and alternative hypotheses for this analysis.

c. Is this a one- or a two-tailed test? Please justify your answer.

d. Using the SPSS for Windows data set provided to you at study.sagepub.com/pett2e (colorectal cancer screening data.sav), undertake your analysis of these data using $\alpha = .05$. Alternatively, use an available website (e.g., www.vassarstats.net) to undertake your analyses using the data outlined in the table below.

e. Based on the results you have obtained, were the participants more likely to seek an immediate colonoscopy once they reached age 50? Justify your answer.

f. Create a table and written paragraph that would reflect your results.

		Following Workshop		
Recommend Colonoscopy Screening?		*Likely*	*Not Likely*	*Total*
Prior to workshop	Likely	25	10	35
	Not likely	22	23	45
	Total	47	33	80

Visit **study.sagepub.com/pett2e** to access SAS output, SPSS datasets, SAS datasets, and SAS examples.