COMPARING TWO RELATEDSAMPLES:

WILCOXONSIGNED RANK AND SIGN TEST

OBJECTIVES

In this lecture, you will learn the following items:

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

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.

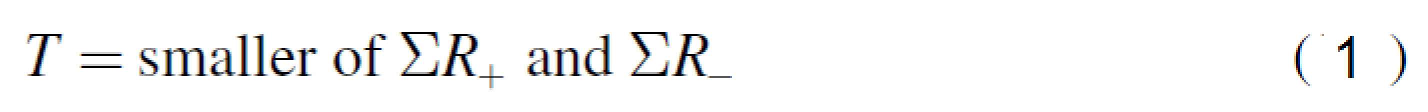
The two sets of test scores in the previous scenario are related or paired. This is because each person was tested twice. In other words, each test score in one group of scores has another test score counterpart. The Wilcoxon signed rank test and the sign test are nonparametric statistical procedures 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 lecture:

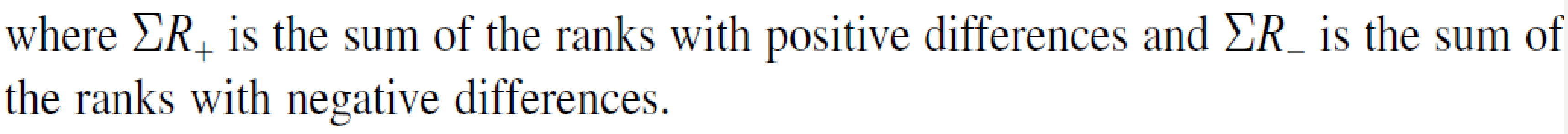
- we will describe how to perform and interpret a Wilcoxon signed rank test and a sign test, using both small samples and large samples.
- Finally, we offer varied examples of these non-parametric statistics from the literature.

COMPUTING THE WILCOXON SIGNED RANK TEST STATISTIC

The formula for computing the Wilcoxon *T* for small samples n < 20is shown in Formula 1. The signed ranks are the values that are used to compute the positive and negative differences values in the formula:



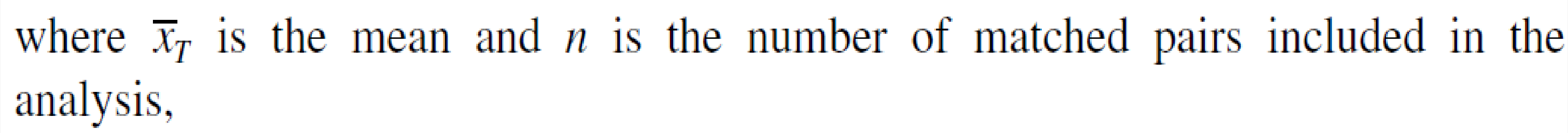
$$T = \text{smaller of } \Sigma R_+ \text{ and } \Sigma R_- \qquad (1)$$



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 30 those available from the table, then a large sample approximation may be performed. For large samples 20<= n <=30, 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 2, Formula 3, and Formula 4 are used to find the *z*-score of a Wilcoxon signed rank test for large samples:

$$\bar{x}_T = \frac{n(n+1)}{4} \qquad (2)$$



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

$$s_T = \sqrt{\frac{n(n+1)(2n+1)}{24}} \qquad (3)$$

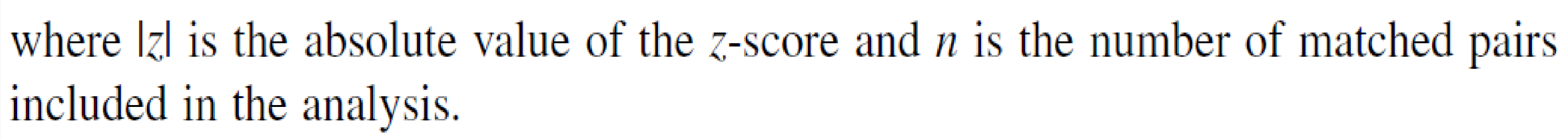
where s_T is the standard deviation,

$$z^* = \frac{T - \bar{x}_T}{s_T} \qquad (4)$$

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 5 to calculate the ES:

$$ES = \frac{|z|}{\sqrt{n}} \qquad (5)$$



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.

Example:

Wilcoxon Signed Rank Test

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 1, the 12 elementary school counselors, or participants, reported the percentage of successful interventions last year and the percentage this year.

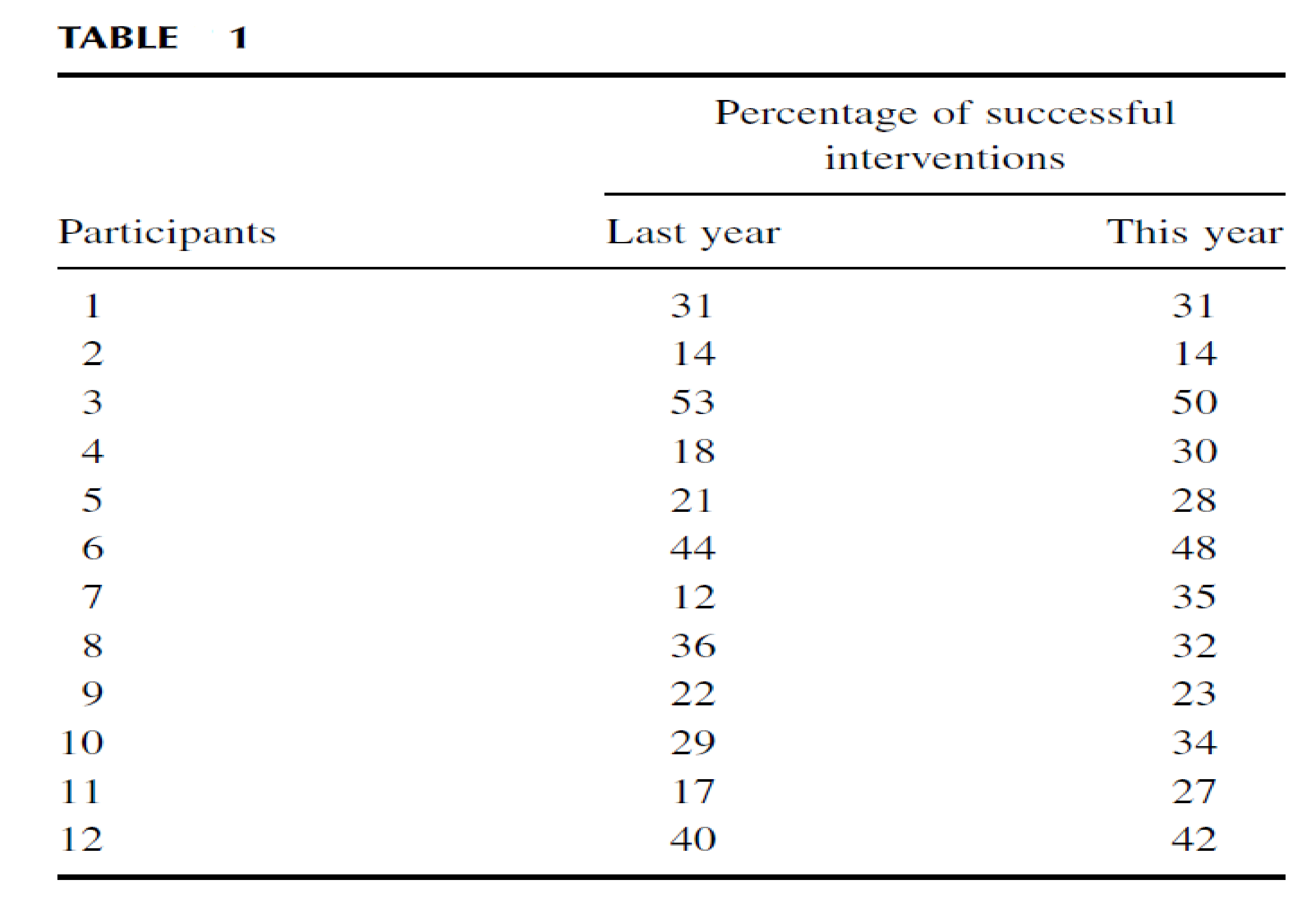


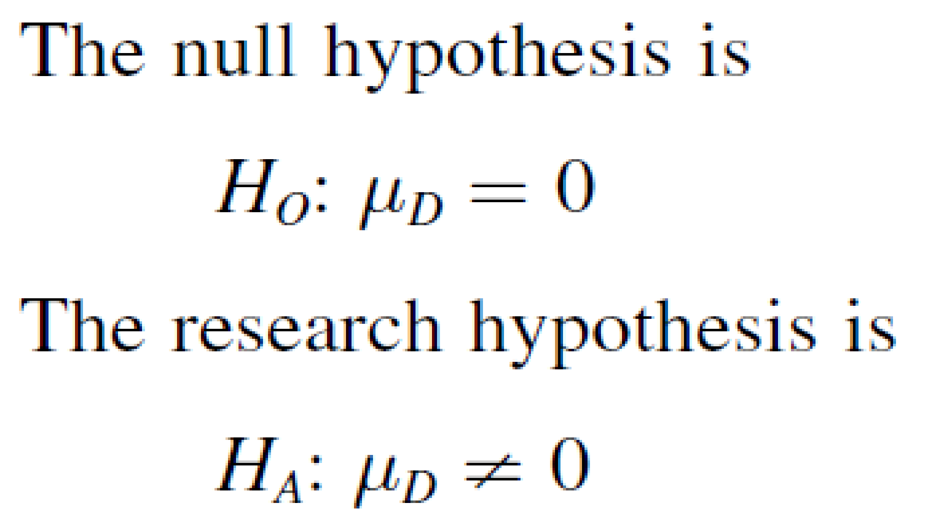
TABLE 1

	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.

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, no directional 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

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

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 α = 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. 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.

4. Compute the Test Statistic

(i). compute the difference between each sample pair.

(ii). Rank the absolute value of those computed differences.

(iii). Using this method, the differences of zero are ignored when ranking.

We have done this in Table 2.

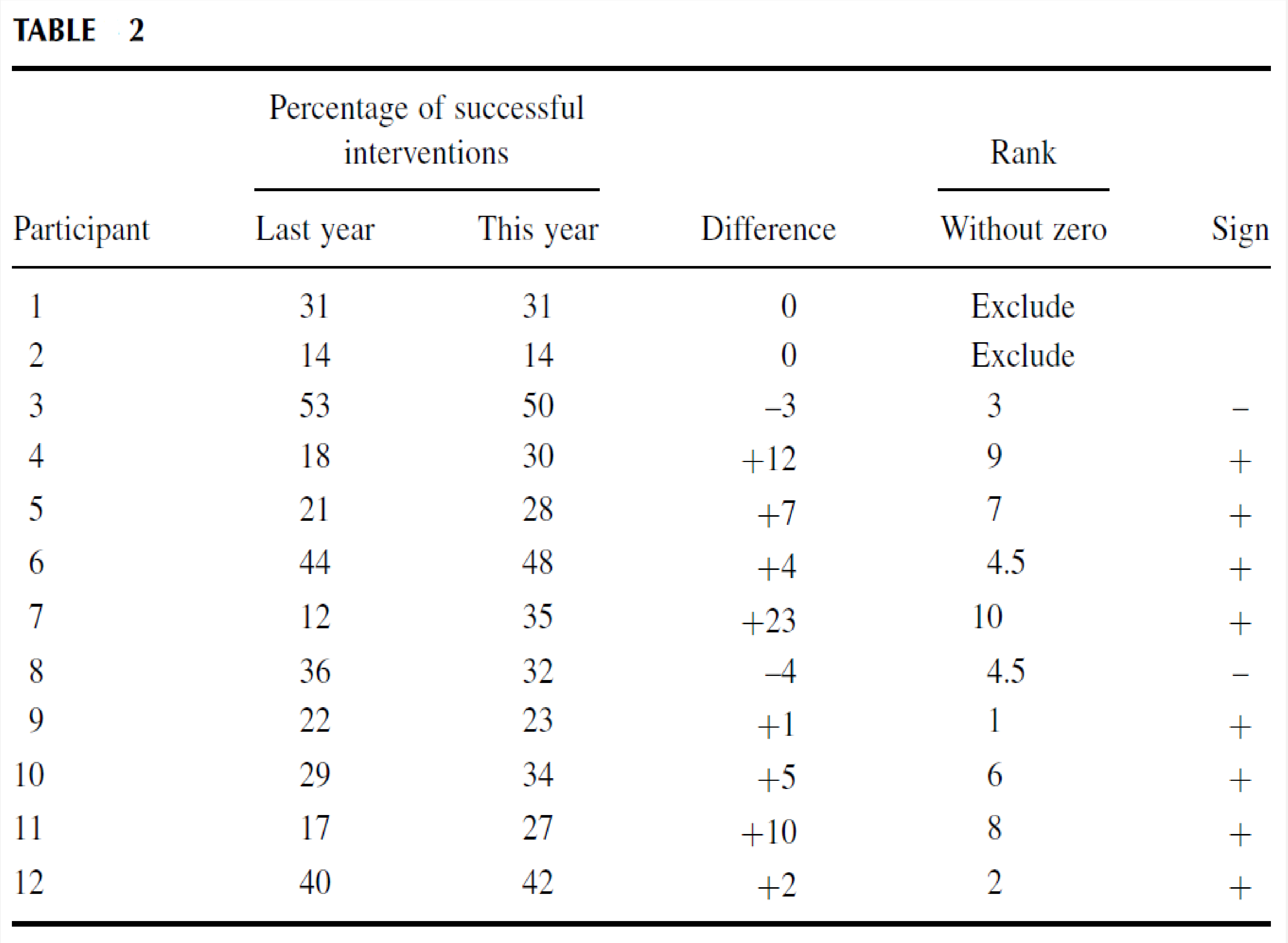


TABLE 2

	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 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 $\sum R+$ = 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 $\sum R-$ = 7.5.

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

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, which lists the critical values for the Wilcoxon T.

As noted earlier in Table 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 α = 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.

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.

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 ($\sum R+$) was **larger than** the negative difference ranks ($\sum R-$), 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.

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 ($\sum R +$) and sum of the negative difference ranks ($\sum R -$).

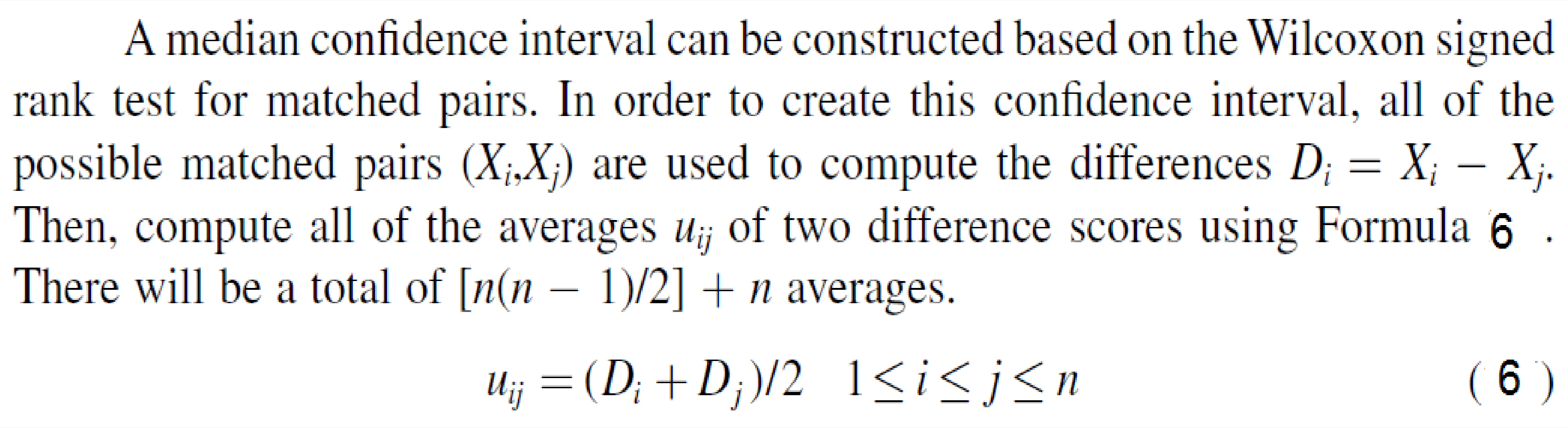
For this example, the Wilcoxon signed rank test (T = 7.5, n = 12, p < 0.05) indicated that the percentage of successful interventions was significantly different.

In addition, the sum of the positive difference ranks ($\sum R +$ = 47.5) was larger than the sum of the negative difference ranks ($\sum R -$ = 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.

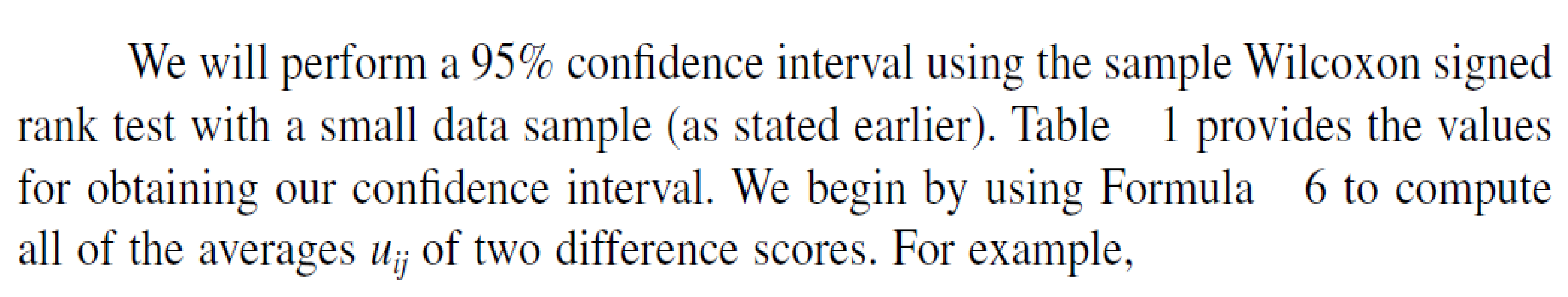
Confidence Interval for the Wilcoxon Signed Rank Test

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 $100(1 - \alpha)\%$.

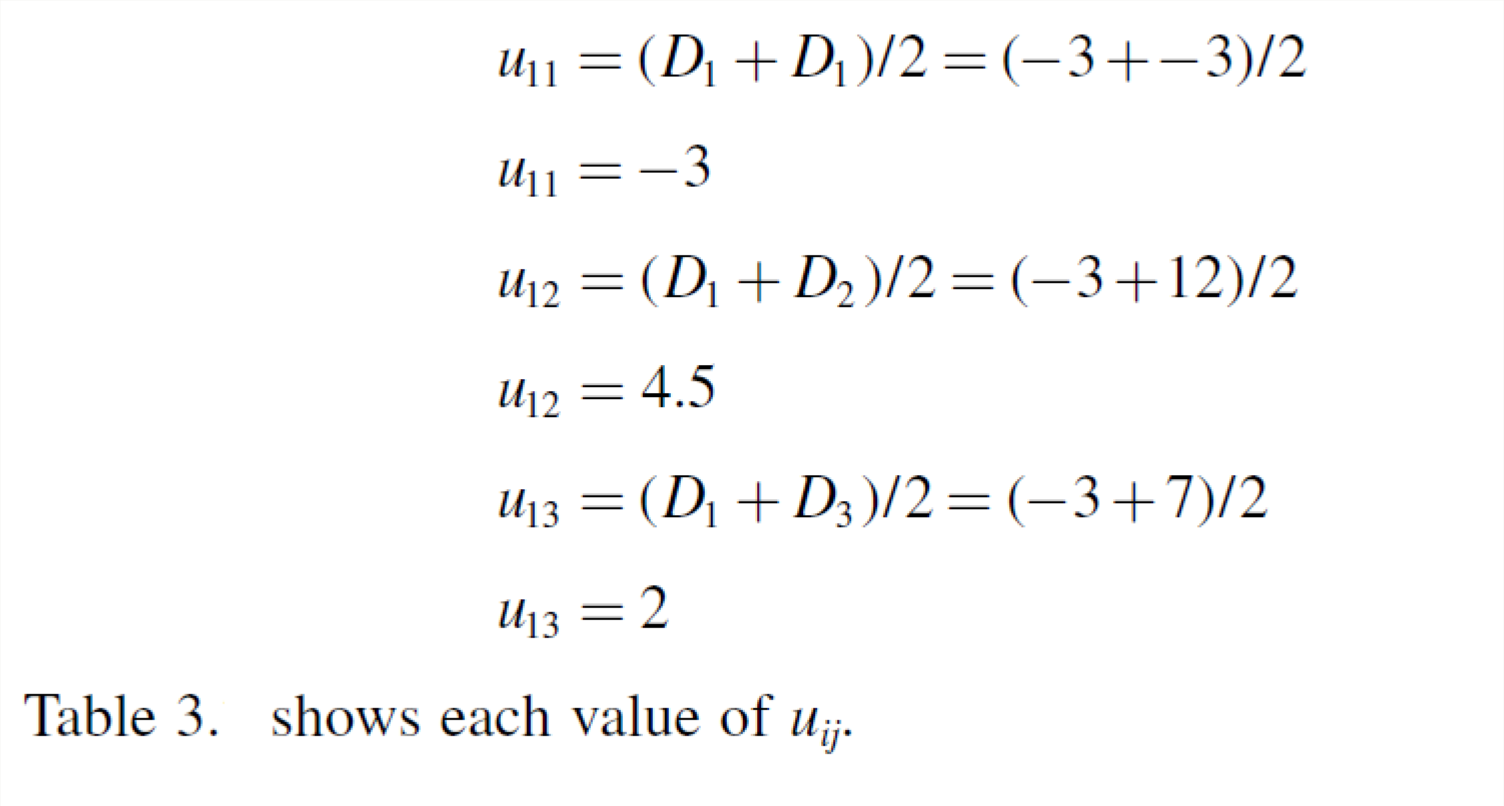


A median confidence interval can be constructed based on the Wilcoxon signed rank test for matched pairs. In order to create this confidence interval, all of the possible matched pairs (X_i,X_j) are used to compute the differences $D_i = X_i - X_j$. Then, compute all of the averages u_{ij} of two difference scores using Formula 6 . There will be a total of $[n(n-1)/2] + n$ averages.

$$u_{ij} = (D_i + D_j)/2 \quad 1 \le i \le j \le n \qquad (6)$$



We will perform a 95% confidence interval using the sample Wilcoxon signed rank test with a small data sample (as stated earlier). Table 1 provides the values for obtaining our confidence interval. We begin by using Formula 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. shows each value of u_{ij}.

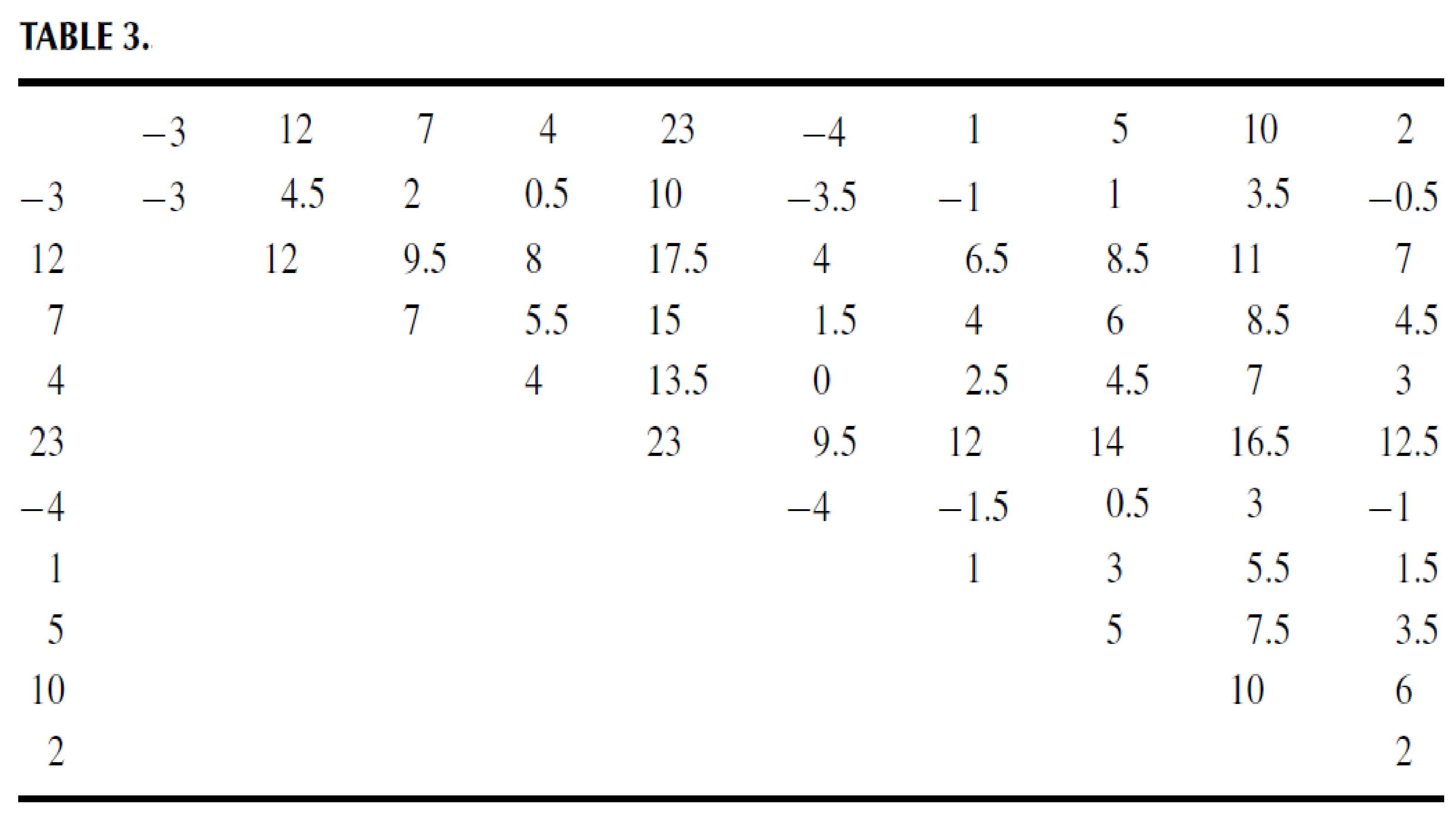
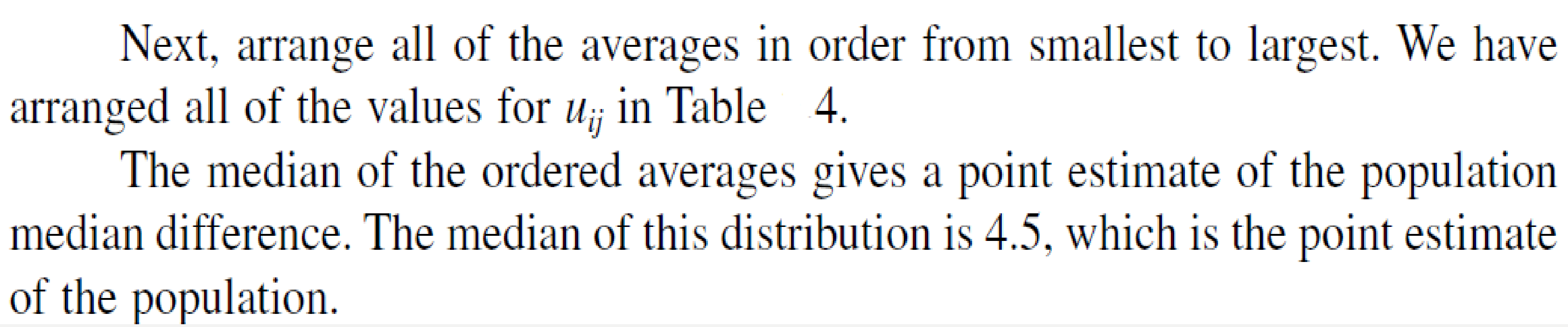


TABLE 3.

	−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 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 to find the endpoints of the confidence interval.

First, determine *T* from the table that corresponds with the sample size and desired

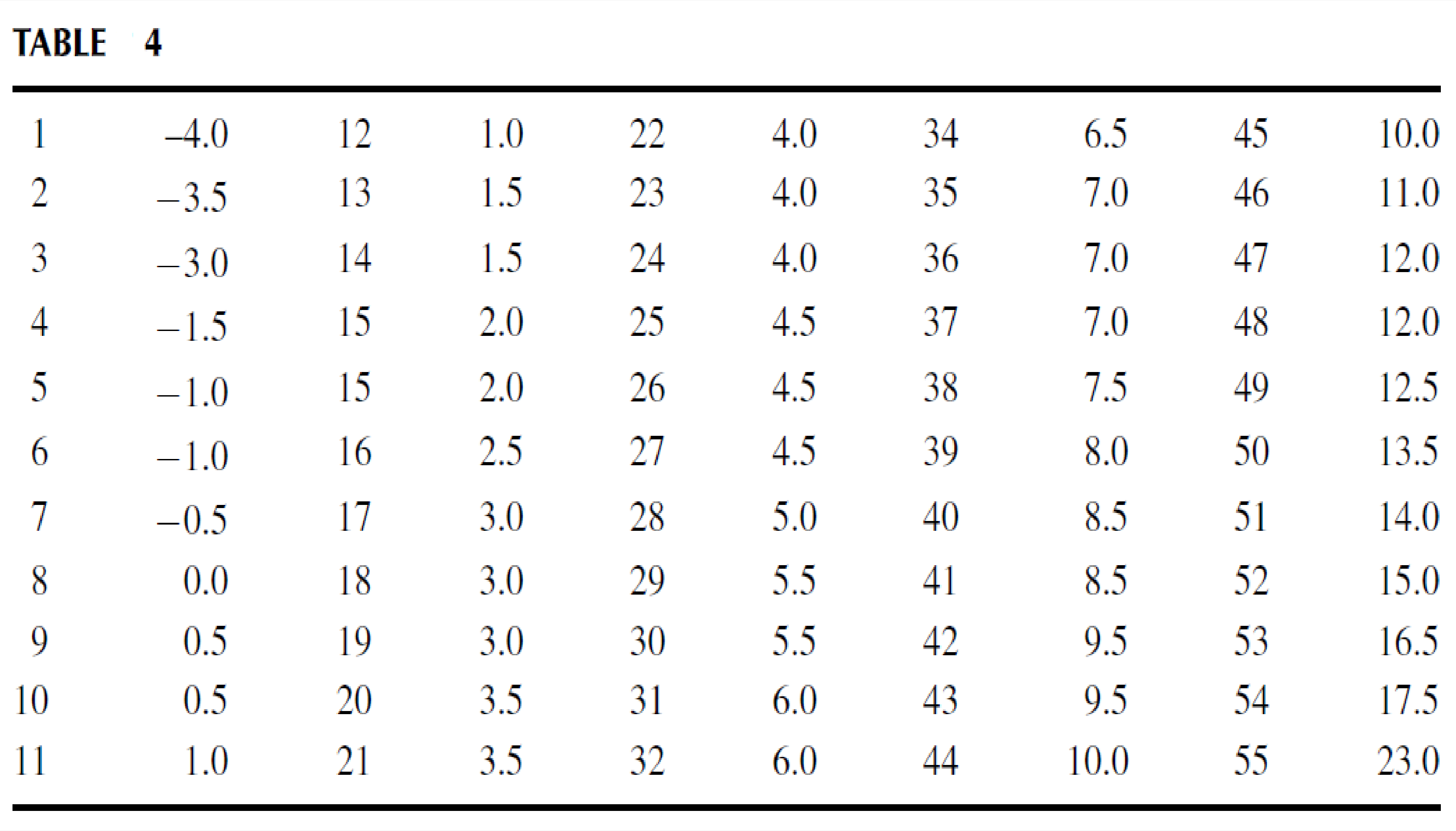
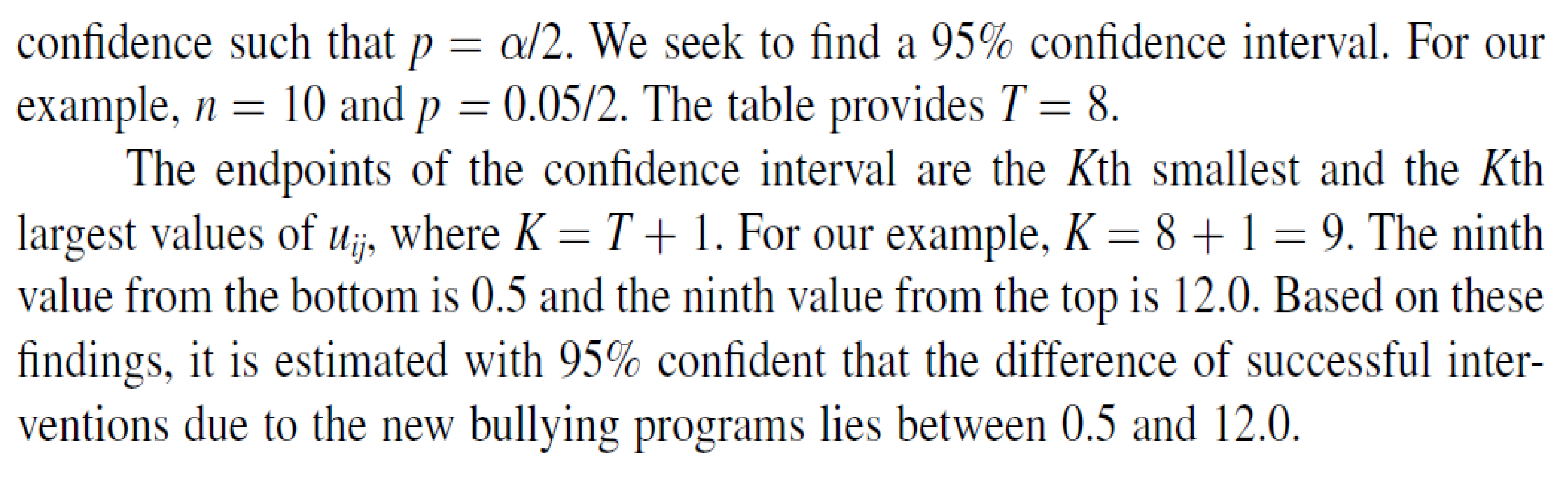


TABLE 4

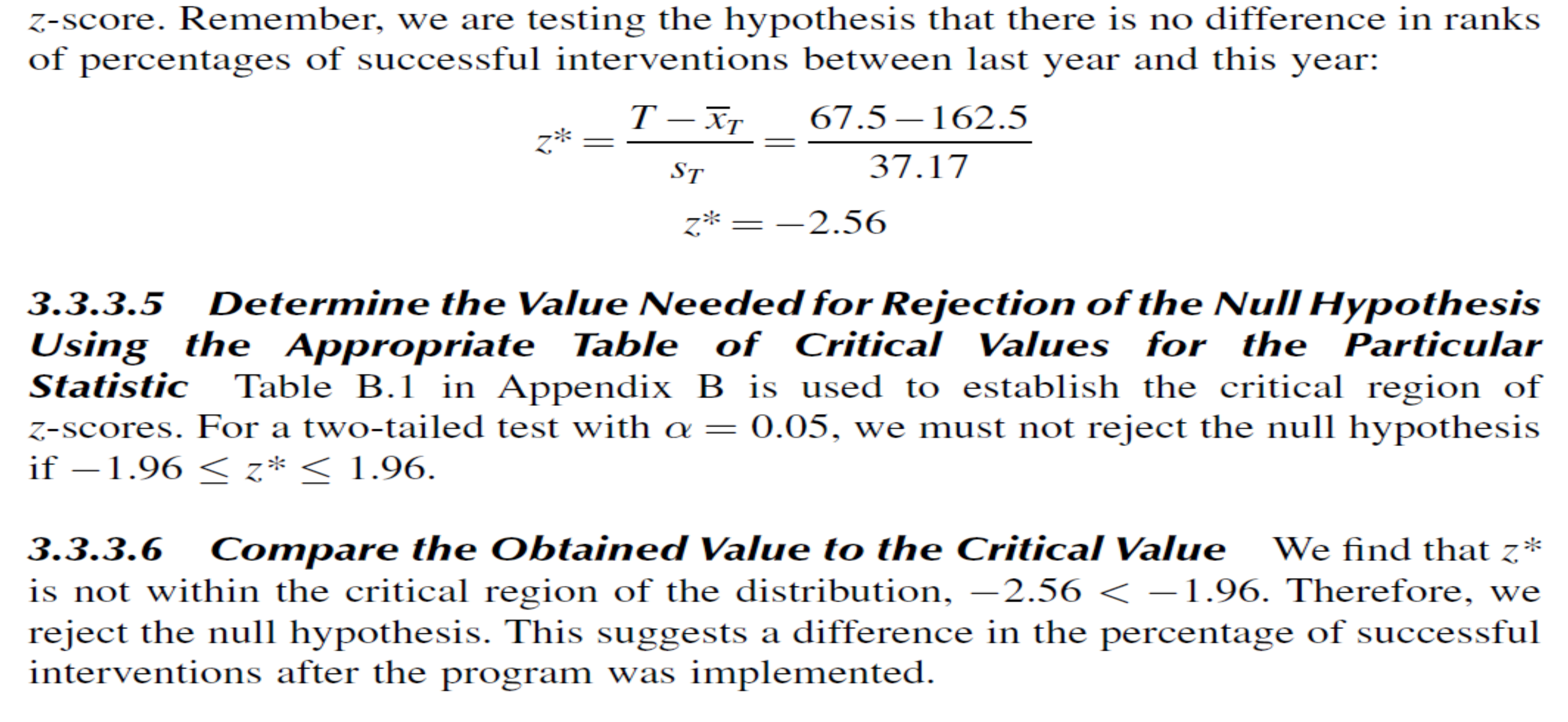
1	−4.0	12	1.0	22	4.0	34	6.5	45	10.0
2	−3.5	13	1.5	23	4.0	35	7.0	46	11.0
3	−3.0	14	1.5	24	4.0	36	7.0	47	12.0
4	−1.5	15	2.0	25	4.5	37	7.0	48	12.0
5	−1.0	15	2.0	26	4.5	38	7.5	49	12.5
6	−1.0	16	2.5	27	4.5	39	8.0	50	13.5
7	−0.5	17	3.0	28	5.0	40	8.5	51	14.0
8	0.0	18	3.0	29	5.5	41	8.5	52	15.0
9	0.5	19	3.0	30	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.

SAMPLE WILCOXON SIGNED RANK TEST (LARGE DATA SAMPLES)



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:

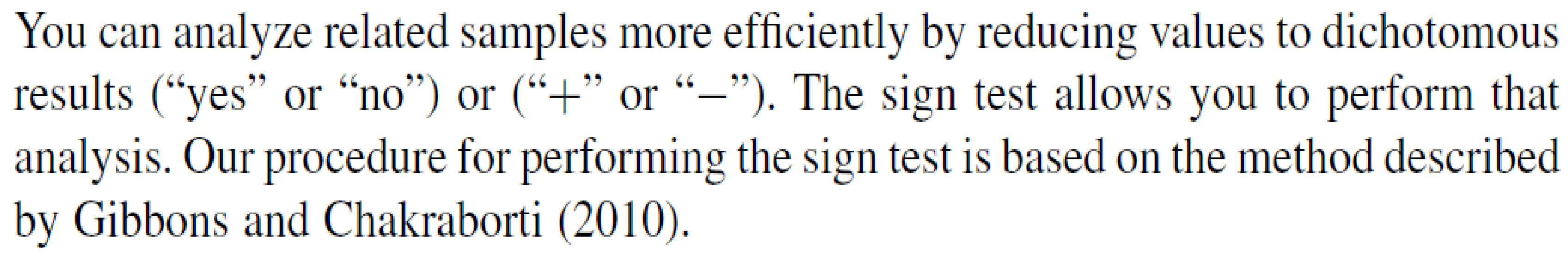
$$z^* = \frac{T - \bar{x}_T}{s_T} = \frac{67.5 - 162.5}{37.17}$$

$$z^* = -2.56$$

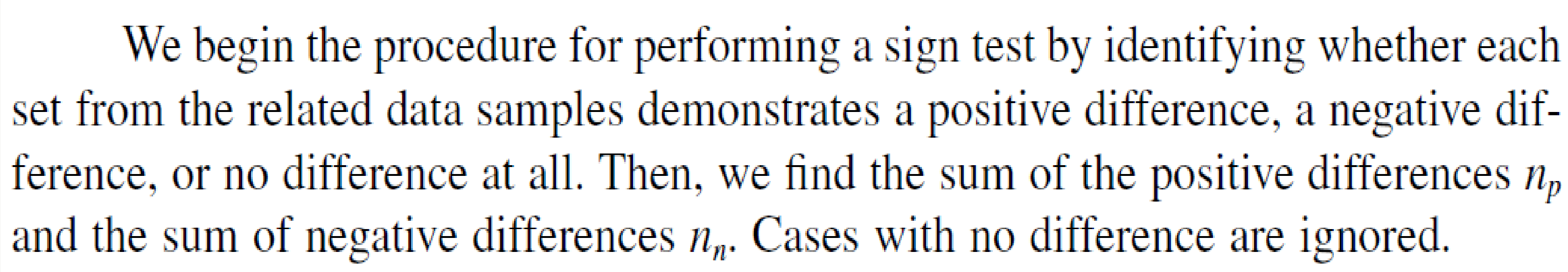
3.3.3.5 Determine the Value Needed for Rejection of the Null Hypothesis Using the Appropriate Table of Critical Values for the Particular Statistic Table B.1 in Appendix B is used to establish the critical region of z-scores. For a two-tailed test with $\alpha = 0.05$, we must not reject the null hypothesis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

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.

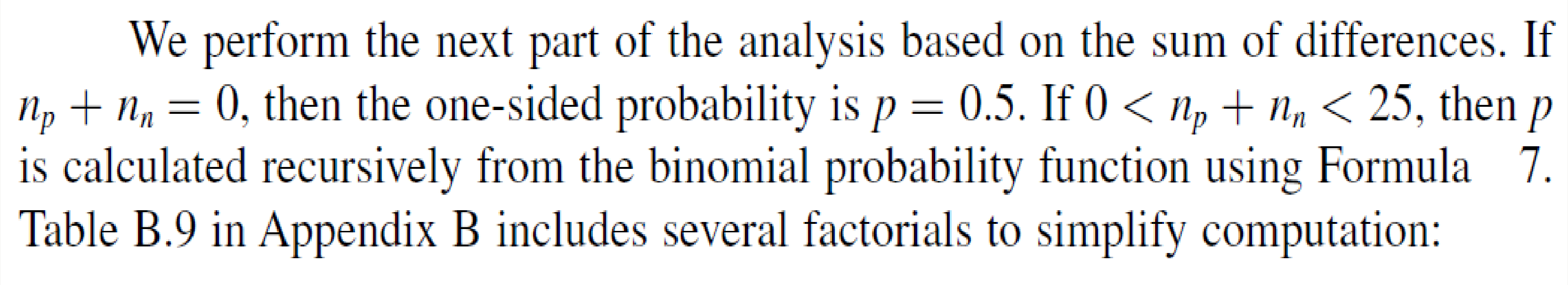
COMPUTING THE SIGN TEST



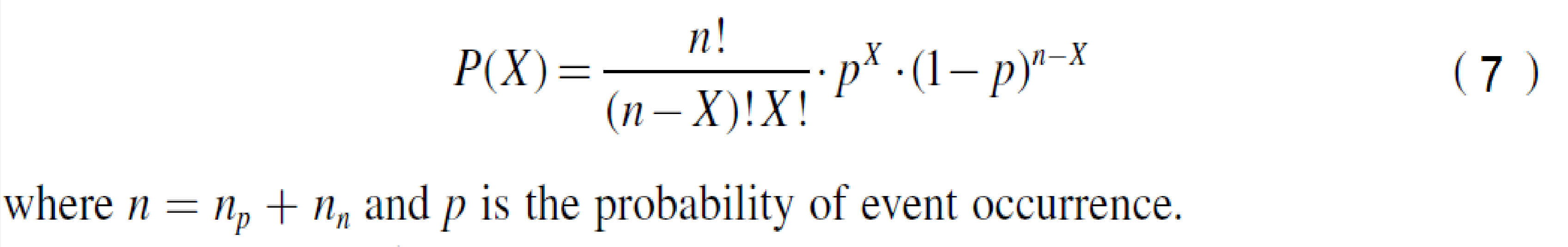
You can analyze related samples more efficiently by reducing values to dichotomous results (“yes” or “no”) or (“+” or “–”). The sign test allows you to perform that analysis. Our procedure for performing the sign test is based on the method described by Gibbons and Chakraborti (2010).



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.

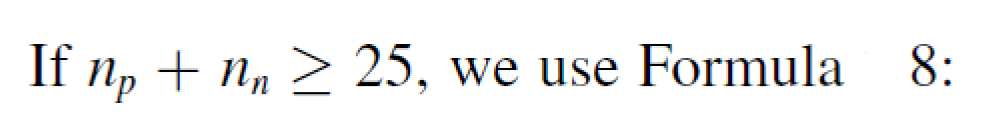


We perform the next part of the analysis based on the sum of differences. If $n_p + n_n = 0$, then the one-sided probability is $p = 0.5$. If $0 < n_p + n_n < 25$, then p is calculated recursively from the binomial probability function using Formula 7. Table B.9 in Appendix B includes several factorials to simplify computation:

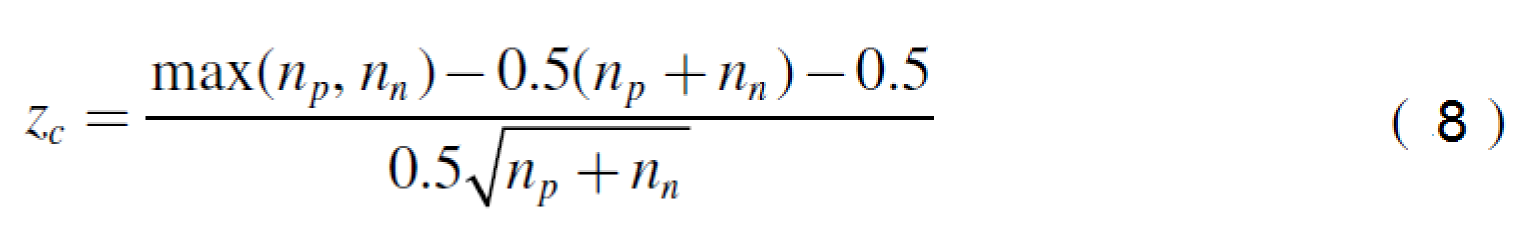


$$P(X) = \frac{n!}{(n-X)!X!} \cdot p^X \cdot (1-p)^{n-X} \qquad (7)$$

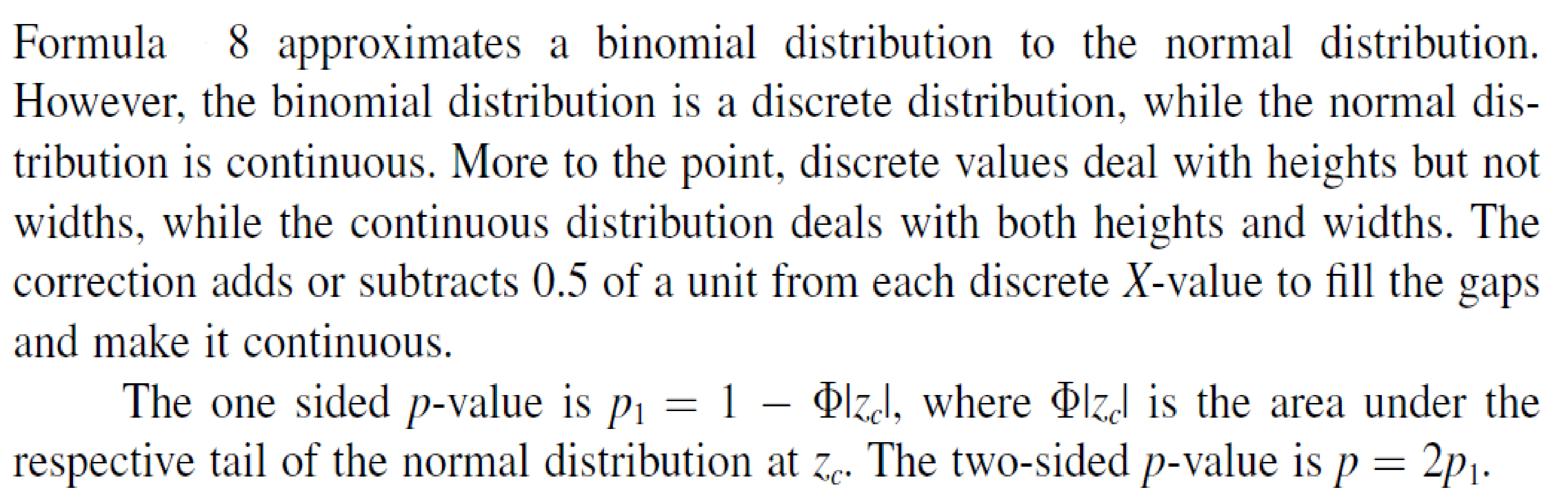
where $n = n_p + n_n$ and p is the probability of event occurrence.



If $n_p + n_n \geq 25$, we use Formula 8:



$$z_c = \frac{\max(n_p, n_n) - 0.5(n_p + n_n) - 0.5}{0.5\sqrt{n_p + n_n}} \qquad (8)$$



Formula 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 $p = 2p_1$.

Example:

Sign Test

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 1 are being reduced to a binomial distribution for use with the sign test. The relatively small sample size warrants a nonparametric procedure.

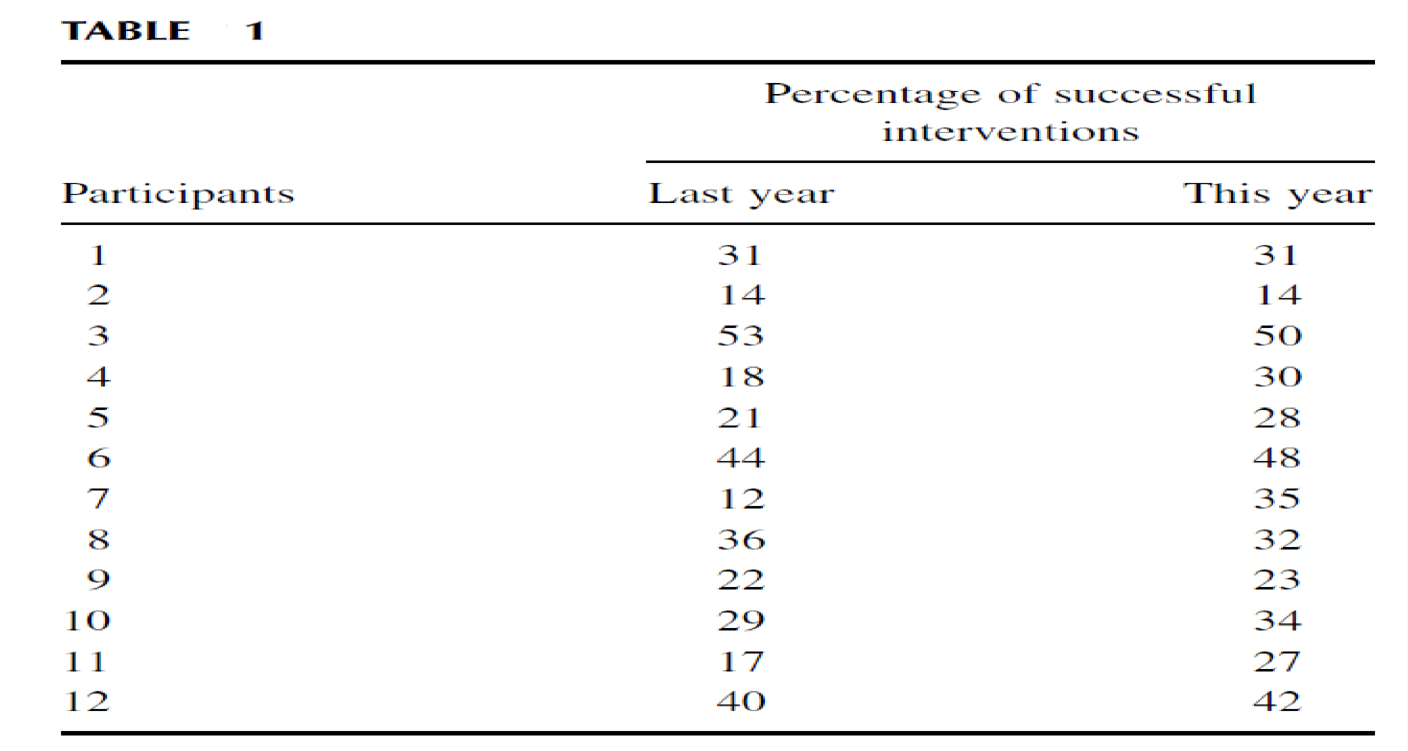
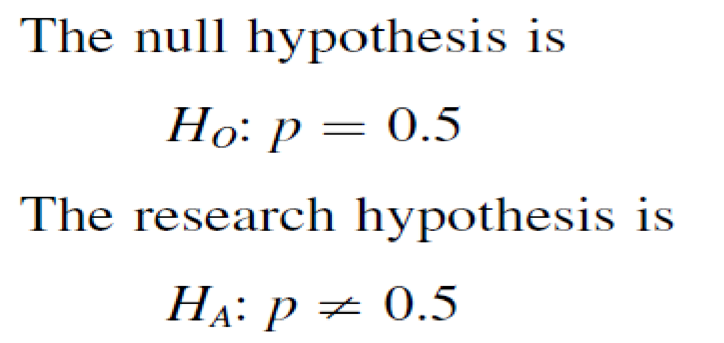


TABLE 1

	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

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, no-directional hypothesis because it indicates a difference, but in no particular direction.



The null hypothesis is

$$H_O\colon p = 0.5$$

The research hypothesis is

$$H_A\colon p \neq 0.5$$

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 α= 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. 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 sign test.

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 7, we see that eight participants showed positive differences, $n_p = 8$. Count the number of “−” or negative differences. When we look at Table 7, we see only two negative differences, $n_n = 2$.

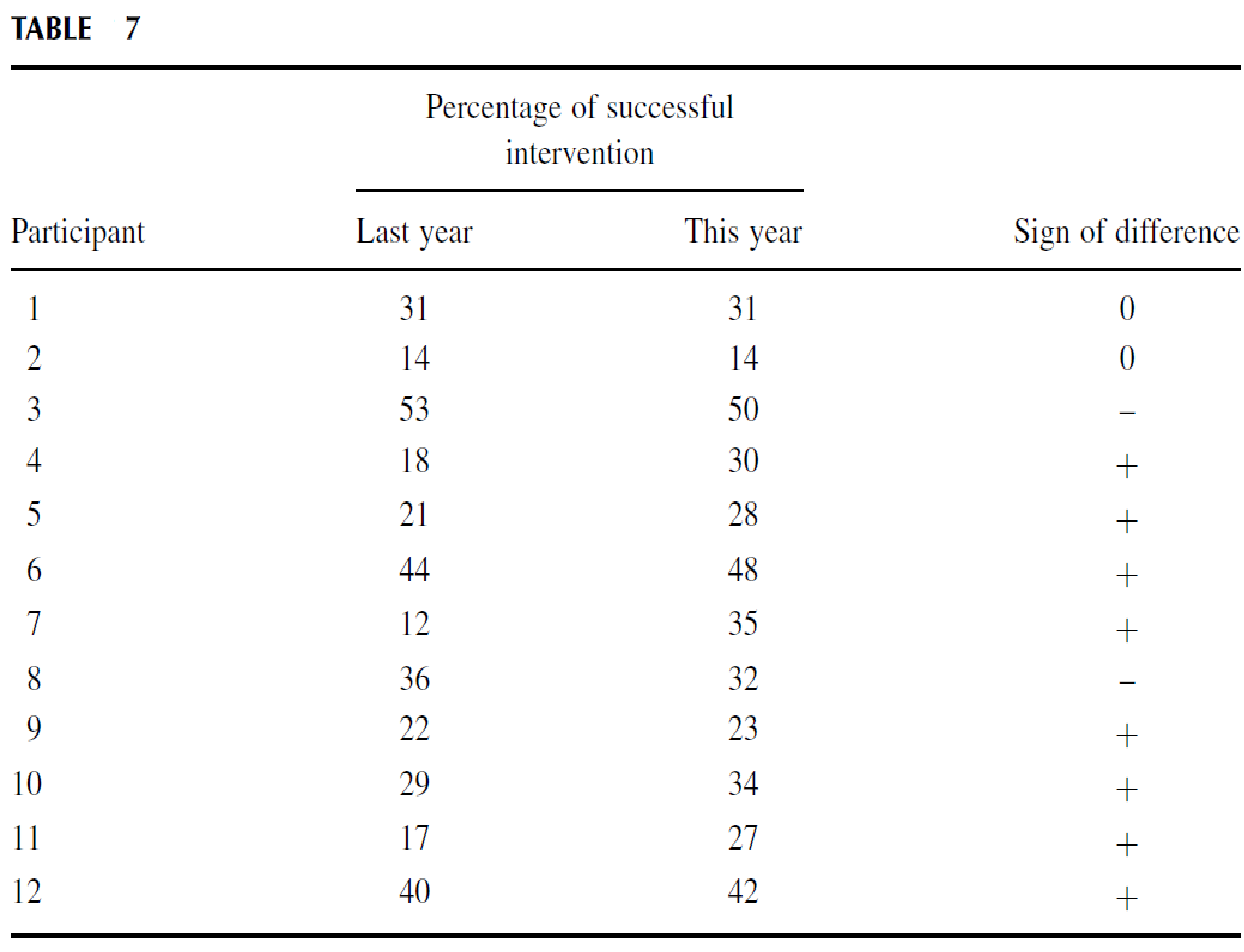


TABLE 7

	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 = 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 1.

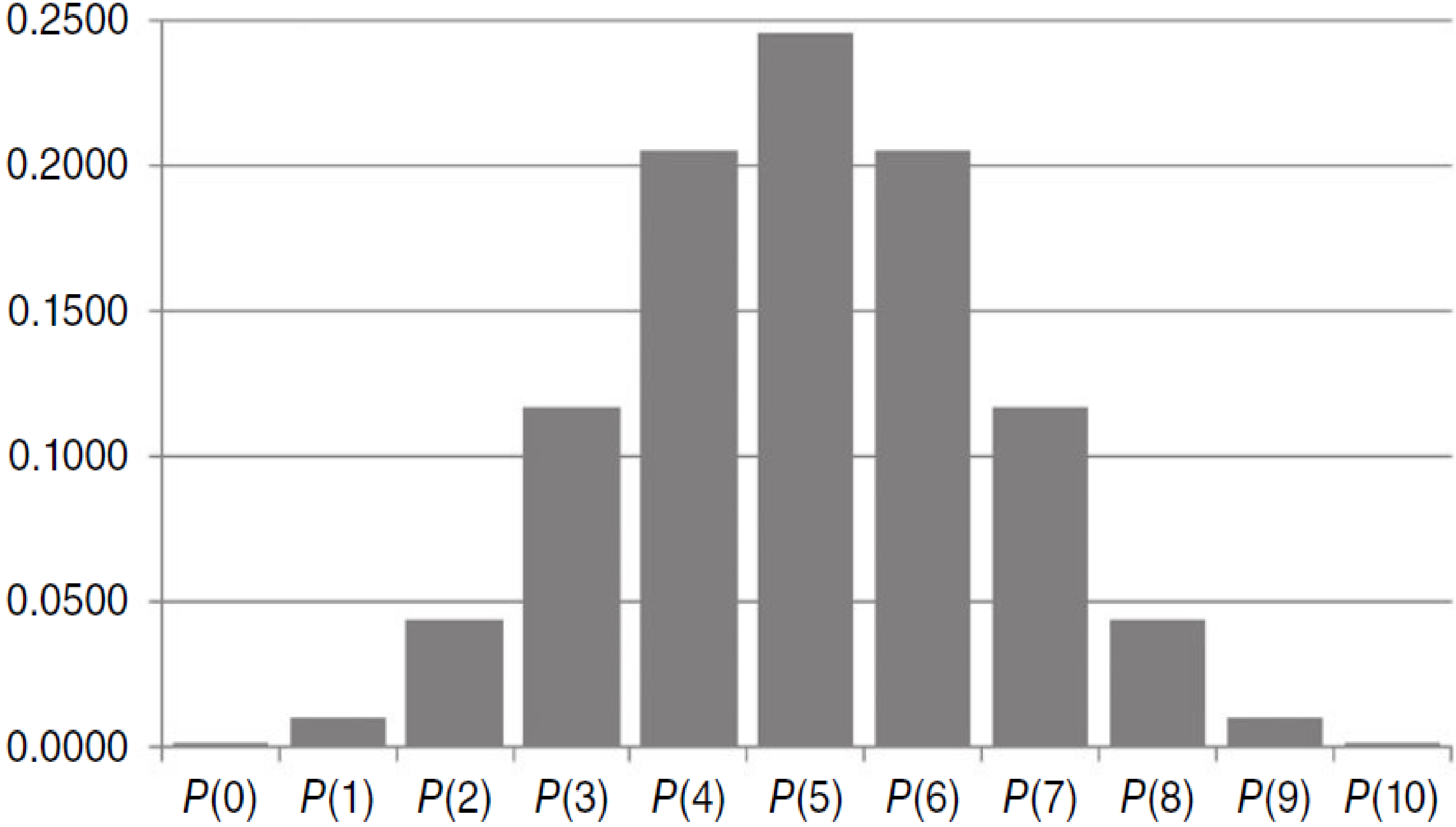
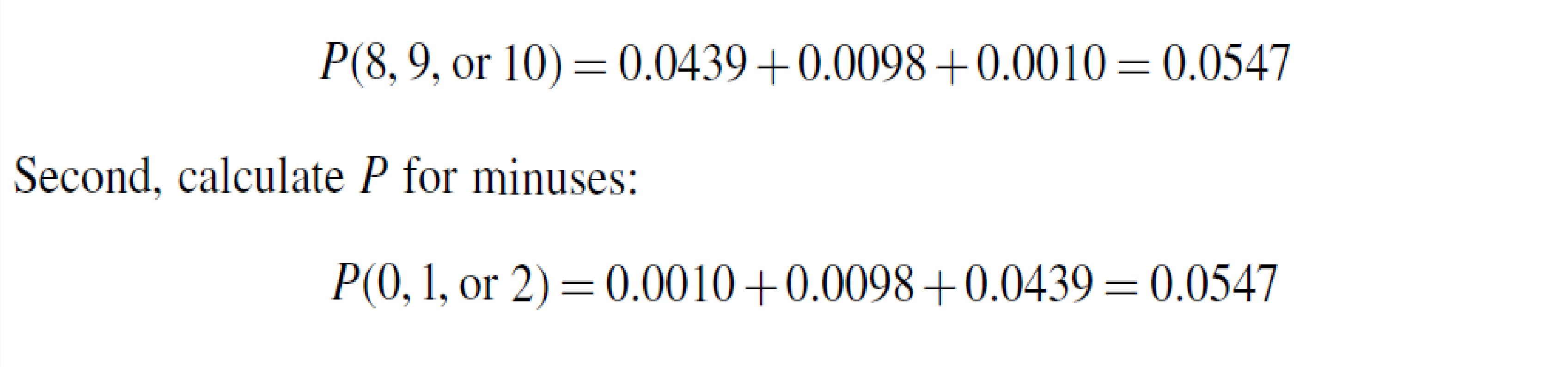


FIGURE 1

Using this distribution, we find the p-values for each tail.

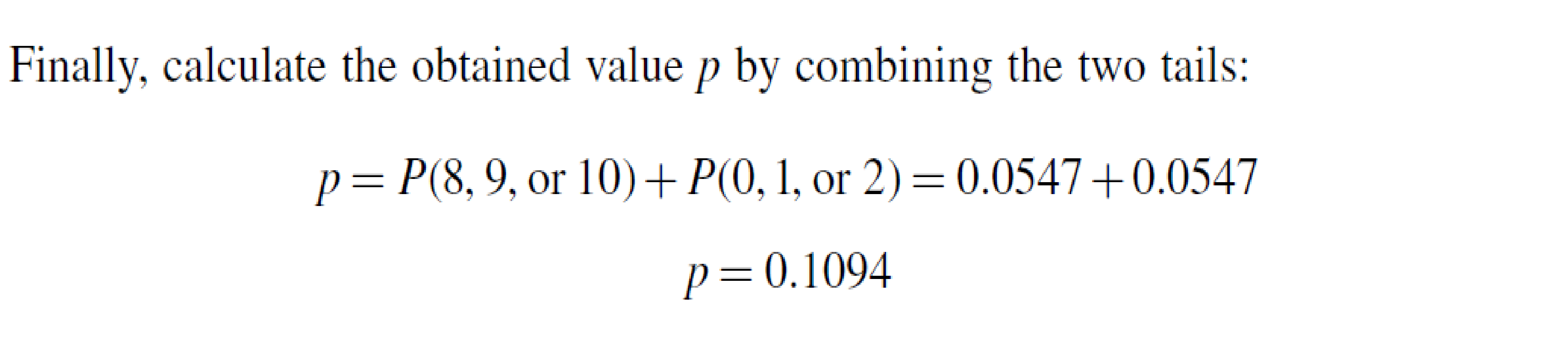
To do that, we sum the probabilities for each tail until we find a probability equal to or greater than $\alpha/2$ = 0.025. First, calculate P for pluses:



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

Second, calculate *P* for minuses:

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



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

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

$$p = 0.1094$$

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, $\alpha = 0.05$.

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.

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 lecture.

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, α = 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.

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.

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.

Both sections analyzed the same data; however, one part 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.

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 lecture, we described how to perform and interpret a Wilcoxon signed rank test and a sign test, using small samples.