COMPARING MORE THAN TWO RELATED SAMPLES:

THE FRIEDMAN TEST

OBJECTIVE

In this lecture, you will learn the following items:

- How to compute the Friedman test.
- How to perform contrasts to compare samples.

INTRODUCTION

Most public school divisions take pride in the percentage of their graduates admitted to college.

A large school division might want to determine if these college admission rates are changing or stagnant.

The division could compare the percentages of graduates admitted to college from each of its 10 high schools over the past 5 years.

Each year would constitute a group, or sample, of percentages from each school. In other words, the study would include five groups, and each group would include 10 values.

The samples in the example are dependent, or related, since each school has percentage for each year.

The Friedman test is a nonparametric statistical procedure for comparing more than two samples that are related.

The parametric equivalent to this test is the repeated measures analysis of variance (ANOVA).

When the Friedman test leads to significant results, then at least one of the samples is different from the other samples.

However, the Friedman test does not identify where the difference(s) occur. Moreover, it does not identify how many differences occur.

In order to identify the particular differences between sample pairs, a researcher might use sample contrasts, or post hoc tests, to analyze the specific sample pairs for significant difference(s).

The Wilcoxon signed rank test is a useful method for performing sample contrasts between related sample sets.

In this lecture, we will describe how to perform and interpret a Friedman test followed with sample contrasts.

COMPUTING THE FRIEDMAN TEST STATISTIC

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

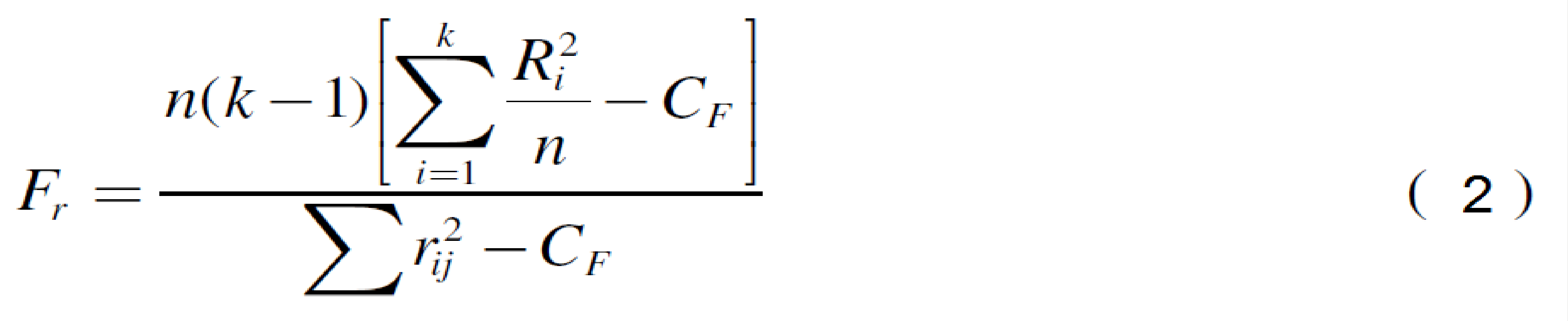
To compute the Friedman test statistic F_r , we begin by creating a table of our data

List the research subjects to create the rows. Place the values for each condition in columns next to the appropriate subjects. Then, rank the values for each subject across each condition. If there are no ties from the ranks, use Formula 1 to deter- mine the Friedman test statistic F_r:

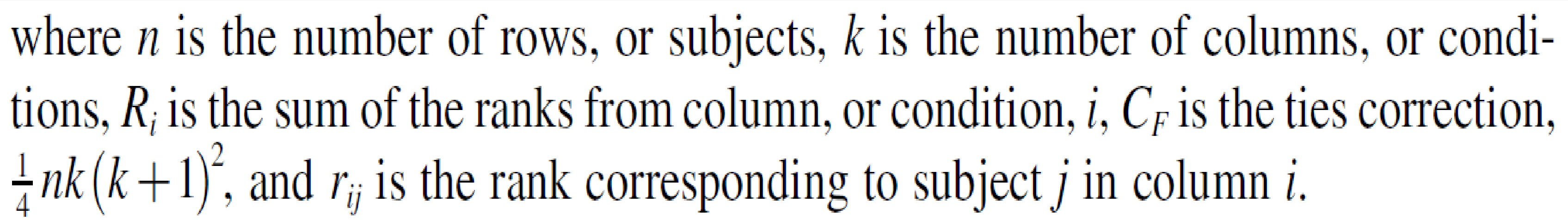
$$F_r = \left[\frac{12}{nk(k+1)} \sum_{i=1}^{k} R_i^2 \right] - 3n(k+1) \qquad (1)$$

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

If ranking of values results in any **ties**, use Formula 2 to determine the Fried- man test statistic F_r:



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



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

The degrees of freedom for the Friedman test is determined by using Formula 3:

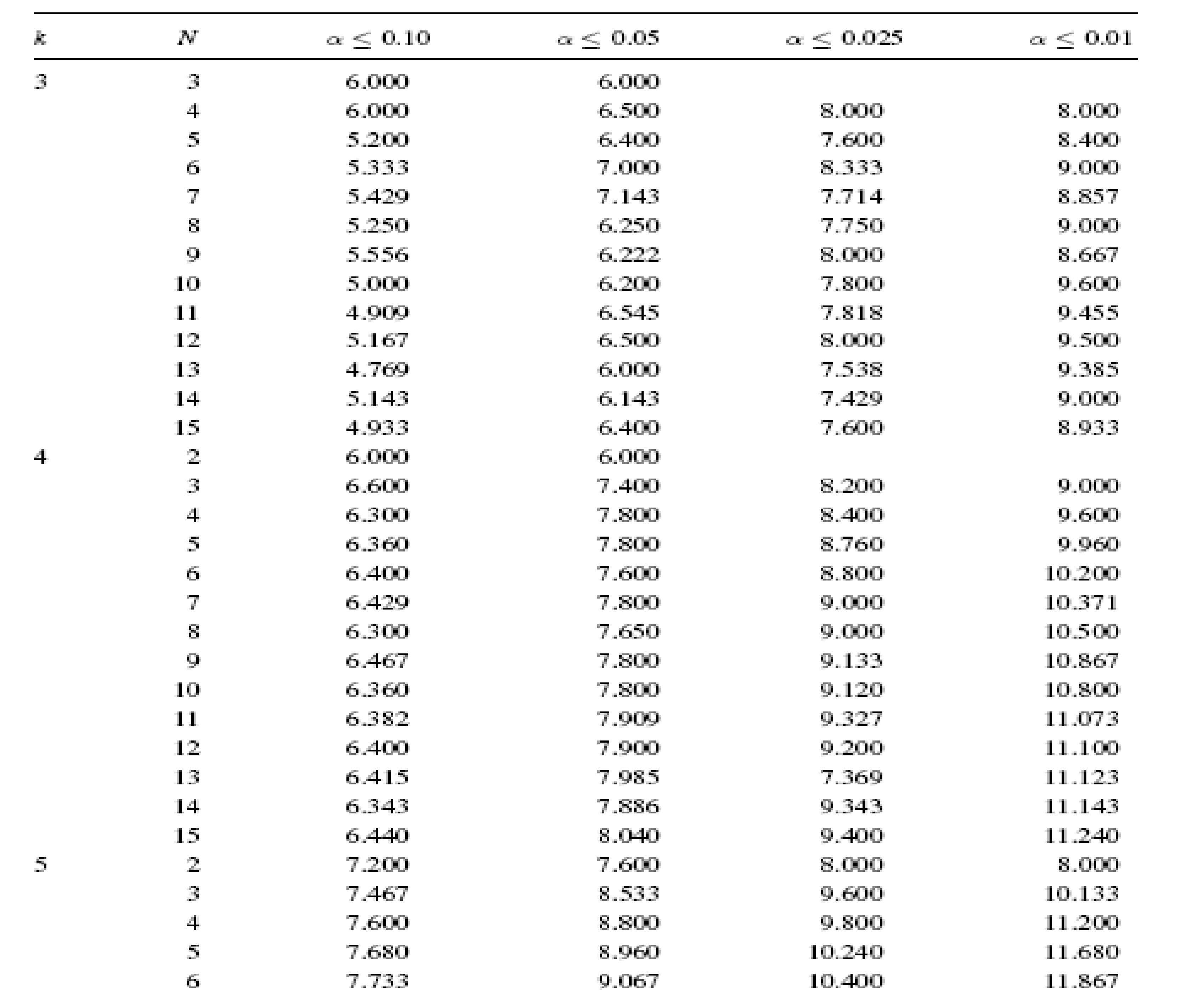
$$\mathbf{df = k - 1} \qquad (3)$$

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

Once the test statistic F_r is computed, it can be compared with a table of critical values (Table B.5) to examine the groups for significant differences.

However, if the number of groups, **k**, or the number of values in a group, **n**, exceeds those available from the table, then a large sample approximation may be performed.

Use a table with the χ2 distribution (Table B.2) to obtain a critical value when performing a large sample approximation.



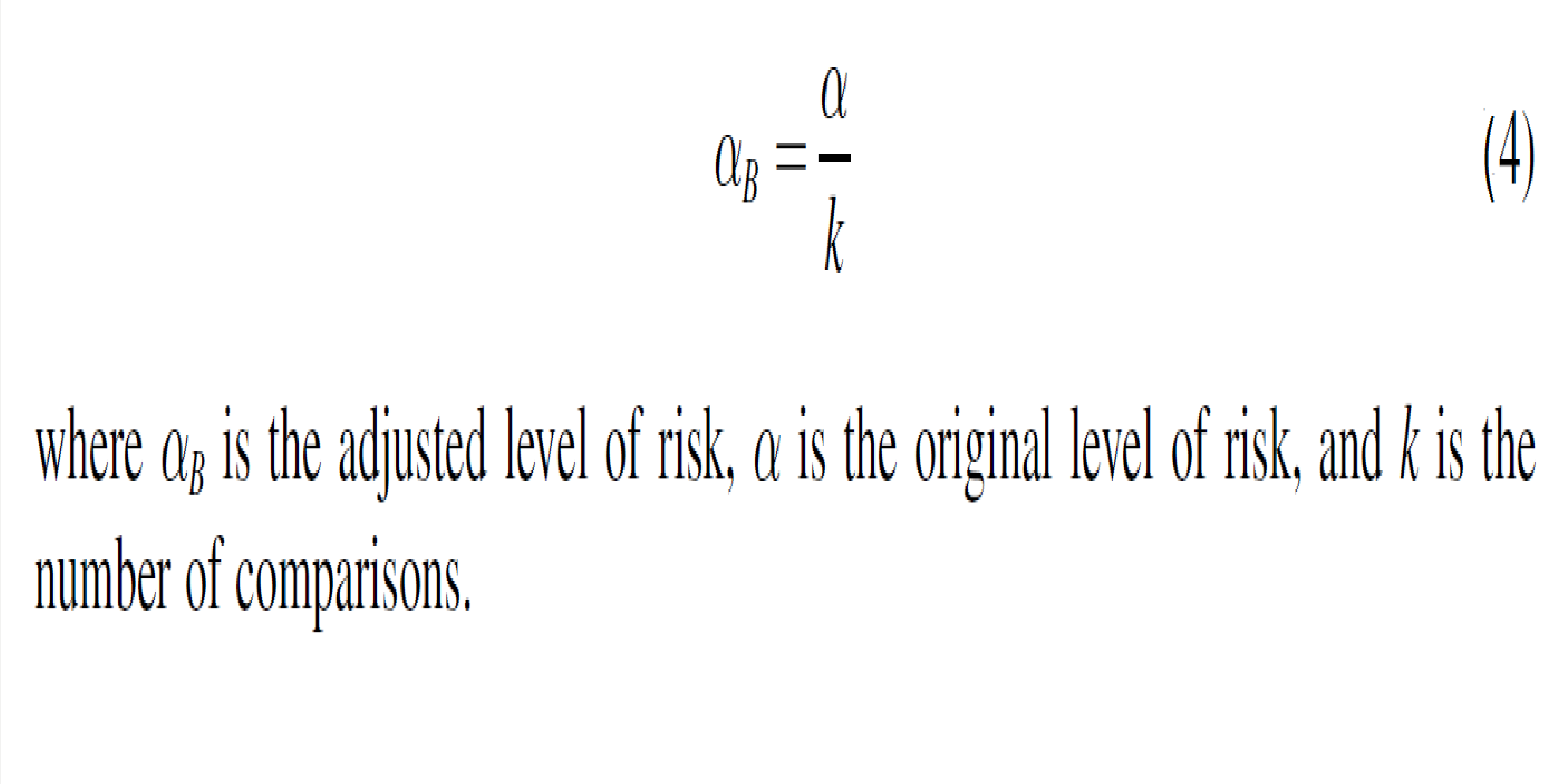
k	N	$\alpha \le 0.10$	$\alpha \le 0.05$	$\alpha \le 0.025$	$\alpha \le 0.01$
3	3	6.000	6.000		
	4	6.000	6.500	8.000	8.000
	5	5.200	6.400	7.600	8.400
	6	5.333	7.000	8.333	9.000
	7	5.429	7.143	7.714	8.857
	8	5.250	6.250	7.750	9.000
	9	5.556	6.222	8.000	8.667
	10	5.000	6.200	7.800	9.600
	11	4.909	6.545	7.818	9.455
	12	5.167	6.500	8.000	9.500
	13	4.769	6.000	7.538	9.385
	14	5.143	6.143	7.429	9.000
	15	4.933	6.400	7.600	8.933
4	2	6.000	6.000		
	3	6.600	7.400	8.200	9.000
	4	6.300	7.800	8.400	9.600
	5	6.360	7.800	8.760	9.960
	6	6.400	7.600	8.800	10.200
	7	6.429	7.800	9.000	10.371
	8	6.300	7.650	9.000	10.500
	9	6.467	7.800	9.133	10.867
	10	6.360	7.800	9.120	10.800
	11	6.382	7.909	9.327	11.073
	12	6.400	7.900	9.200	11.100
	13	6.415	7.985	7.369	11.123
	14	6.343	7.886	9.343	11.143
	15	6.440	8.040	9.400	11.240
5	2	7.200	7.600	8.000	8.000
	3	7.467	8.533	9.600	10.133
	4	7.600	8.800	9.800	11.200
	5	7.680	8.960	10.240	11.680
	6	7.733	9.067	10.400	11.867

If the F_r statistic is not significant, then no differences exist between any of the related conditions.

However, if the F_r statistic is significant, then a difference exists between at least two of the conditions.

Therefore, a researcher might use sample contrasts between individual pairs of conditions, or post hoc tests, to determine which of the condition pairs are significantly different.

When performing multiple sample contrasts, the type I error rate tends to become inflated. Therefore, the initial level of risk, or , must be adjusted. We demonstrate the Bonferroni procedure, shown in Formula 4, to adjust :



$$\alpha_B = \frac{\alpha}{k} \tag{4}$$

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

Example

Friedman’s Test
(Small Data Samples without Ties)

A manager is struggling with the chronic tardiness of her seven employees. She tries two strategies to improve employee timeliness.

First, over the course of a month, she punishes employees with a $10 paycheck deduction for each day that they arrive late.

Second, the following month, she punishes employees by docking their pay $20 for each day that they do not arrive on time.

Table 5.1 shows the number of times each employee was late in a given month.

The baseline shows the employees’ monthly tardiness before the strategies.

Month 1 shows the employees’ monthly tardiness after a month of the $10 paycheck deductions.

Month 2 shows the employees’ monthly tardiness after a month of the $20 paycheck deductions.

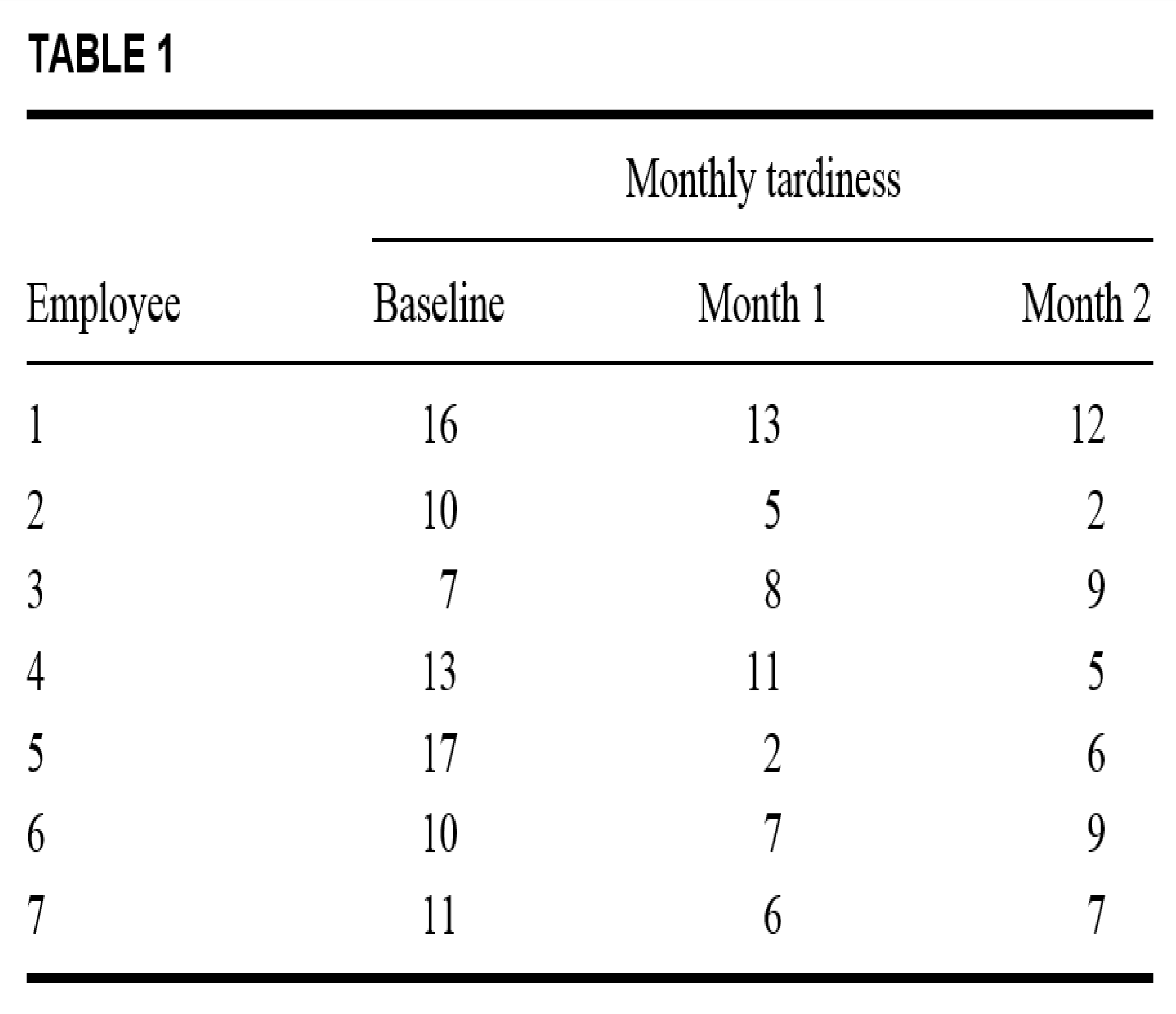


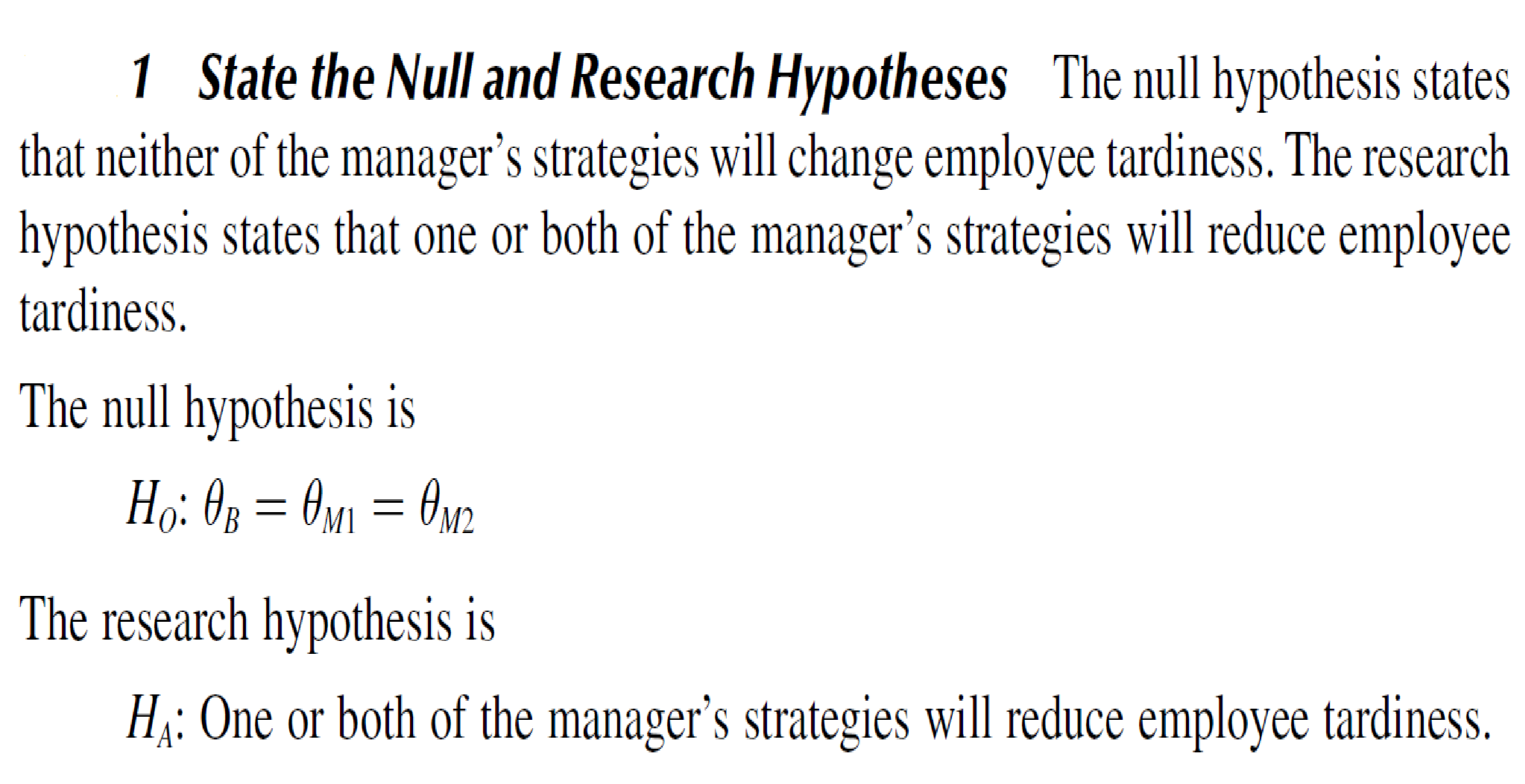
TABLE 1

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

We want to determine if either of the paycheck deduction strategies reduced employee tardiness.

Since the sample sizes are small (n < 20), we require a non-parametric test.

The Friedman test is the best statistic to analyze the data and test the hypothesis.



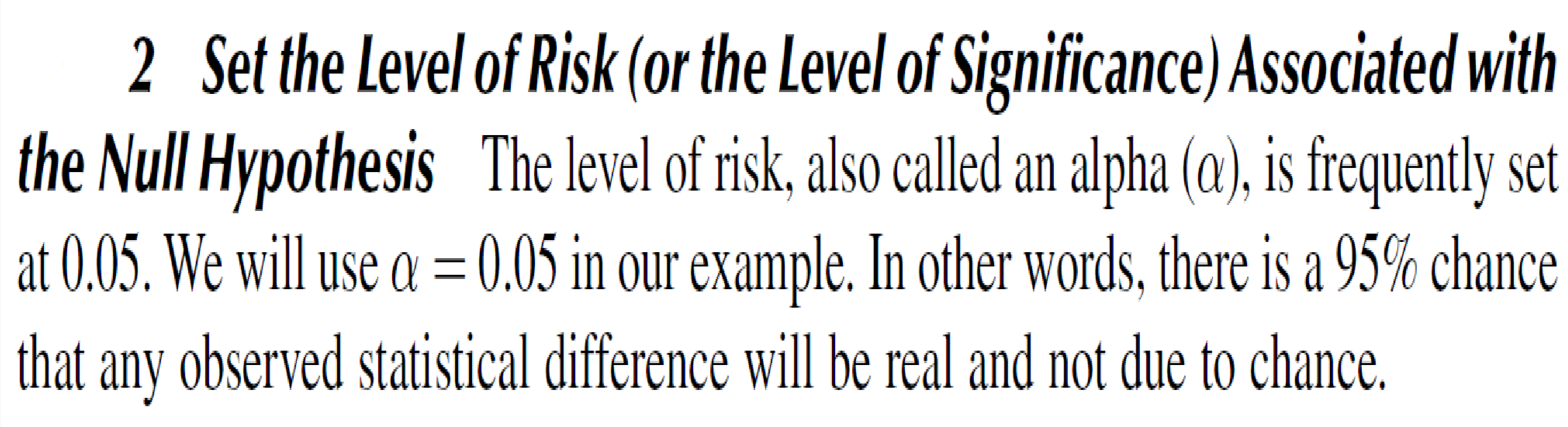
1 State the Null and Research Hypotheses The null hypothesis states that neither of the manager's strategies will change employee tardiness. The research hypothesis states that one or both of the manager's strategies will reduce employee tardiness.

The null hypothesis is

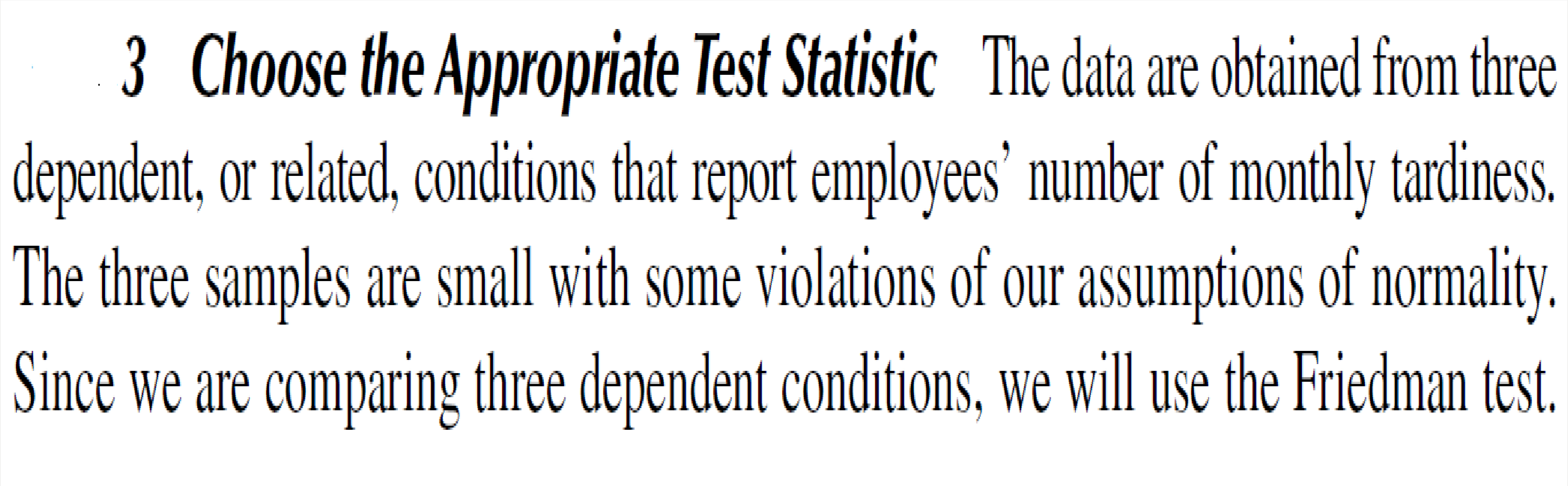
$$H_O: \theta_B = \theta_{M1} = \theta_{M2}$$

The research hypothesis is

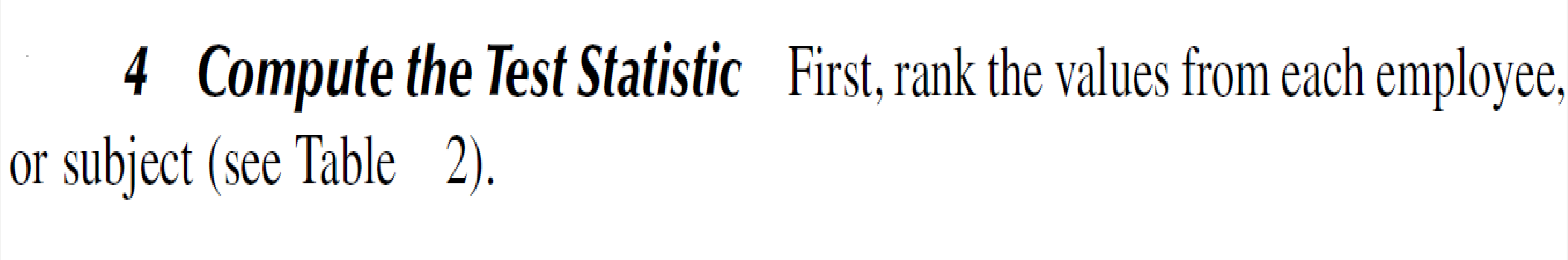
H_A: One or both of the manager's strategies will reduce employee tardiness.



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



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



4 *Compute the Test Statistic* First, rank the values from each employee, or subject (see Table 2).

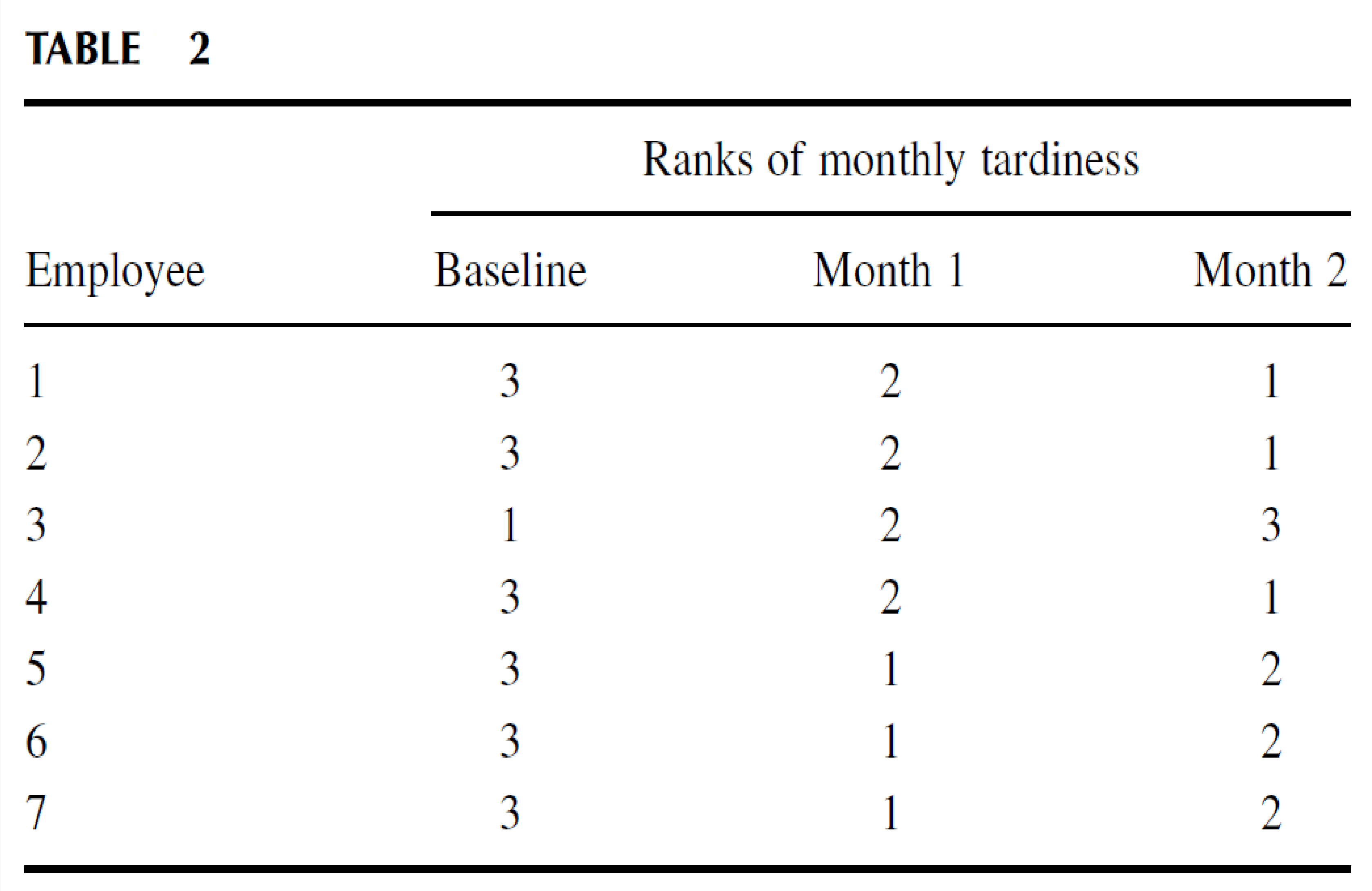
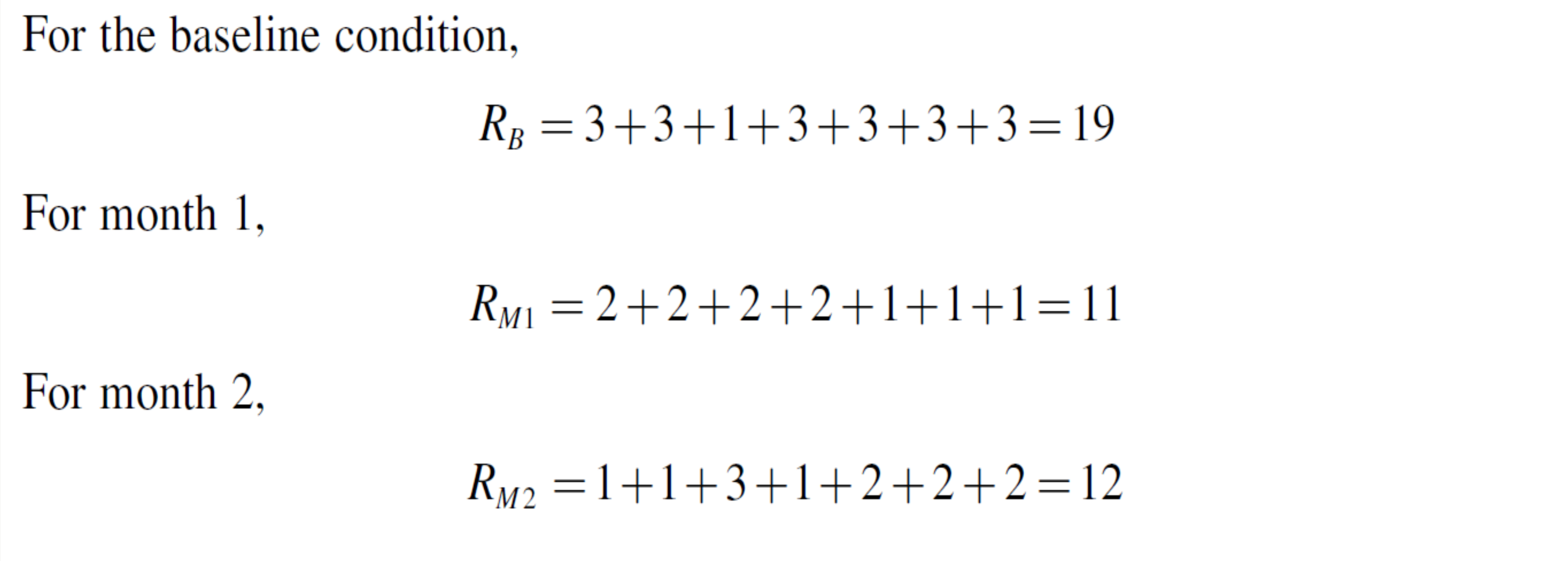


TABLE 2

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

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



For the baseline condition,

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

For month 1,

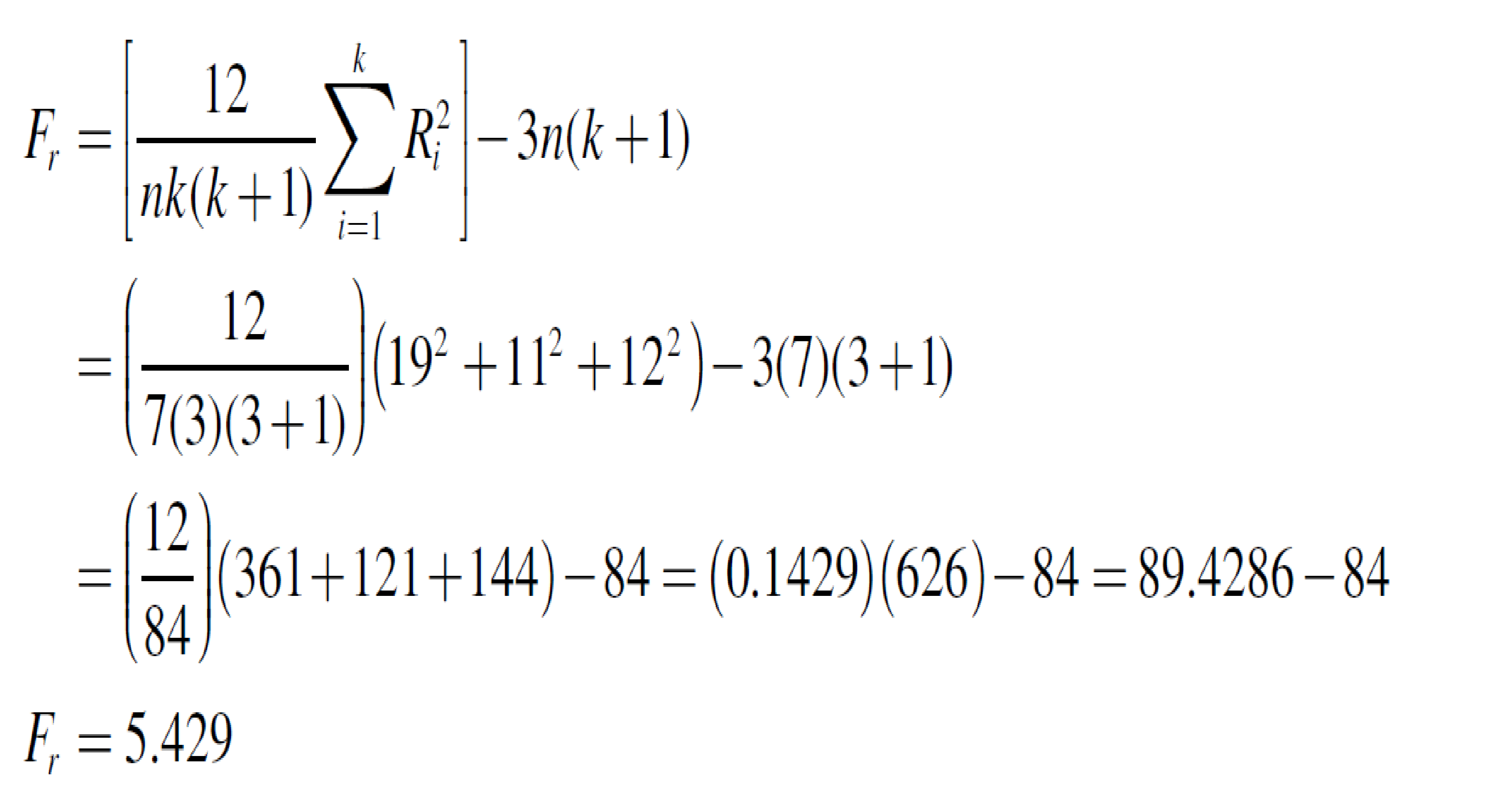
$$R_{M1} = 2+2+2+2+1+1+1 = 11$$

For month 2,

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

These R-values are used to compute the Fr test statistic.

Use Formula 1 since there were no ties involved in the ranking:



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

$$= \left(\frac{12}{7(3)(3+1)}\right)\left(19^2 + 11^2 + 12^2\right) - 3(7)(3+1)$$

$$= \left(\frac{12}{84}\right)(361 + 121 + 144) - 84 = (0.1429)(626) - 84 = 89.4286 - 84$$

$$F_r = 5.429$$

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

We will use the critical value table for the Friedman test (Table B.5) since it includes the number of groups, k, and the number of samples, n, for our data.

In this case, we look for the critical value for k = 3 and n = 7 with $\alpha = 0.05$.

Table B.5 returns a critical value for the Friedman test of 7.14.

6. Compare the Obtained Value with the Critical Value

The critical value for rejecting the null hypothesis is 7.14 and the obtained value is $F_r = 5.429$.

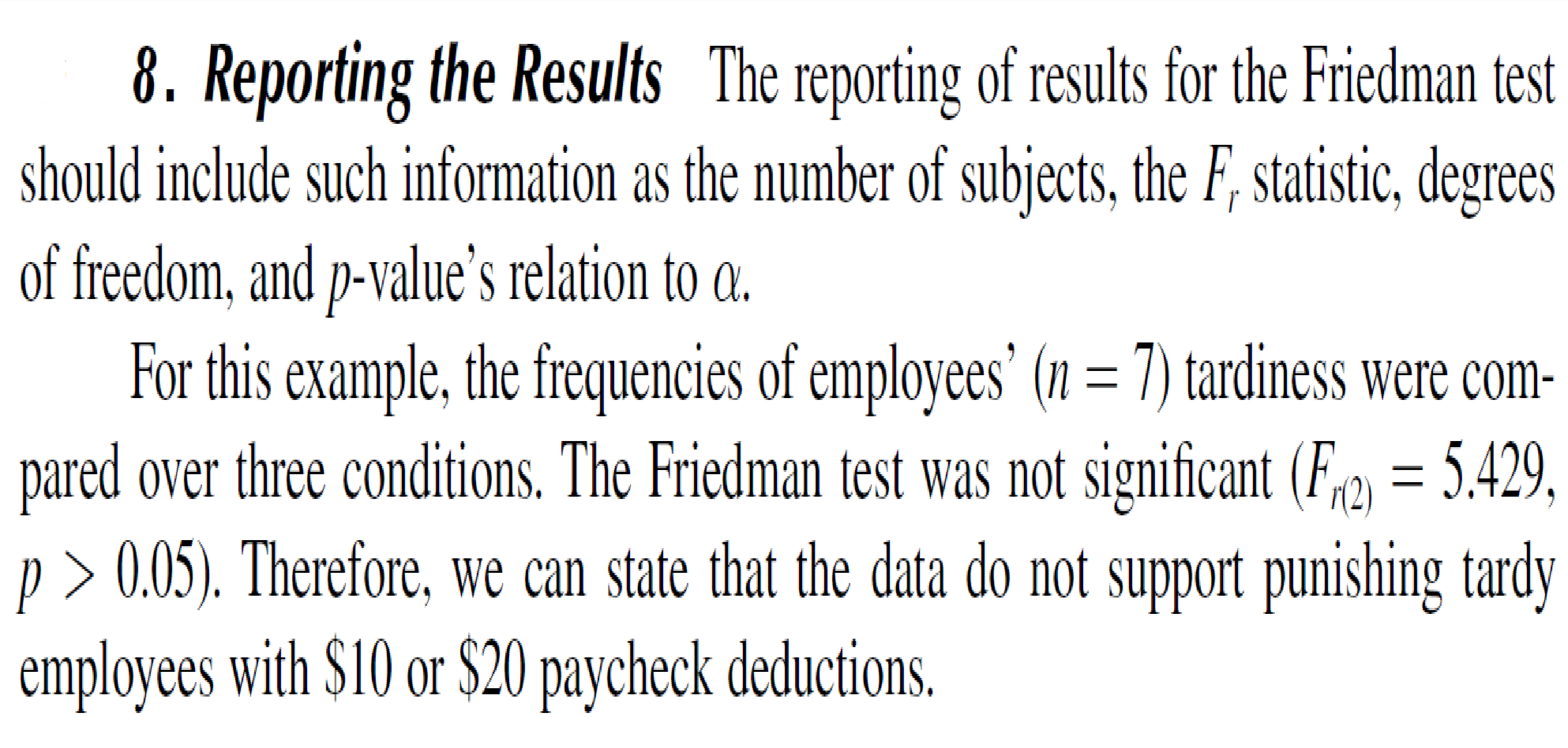
If the critical value is less than or equal to the obtained value, we must reject the null hypothesis.

If instead, the critical value exceeds the obtained value, we do not reject the null hypothesis.

Since the critical value exceeds the obtained value, we do not reject the null hypothesis.

7. Interpret the Results

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



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

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

Example

Friedman’s Test (Small Data Samples with Ties)

After the manager’s failure to reduce employee tardiness with paycheck deductions, she decided to try a different approach. This time, she rewarded employees when they arrived to work on-time. Again, she tries two strategies to improve employee timeliness.

First, over the course of a month, she rewards employees with a $10 bonus for each day that they arrive on-time.

Second, the following month, she rewards employees with a $20 bonus for each day that they arrive on-time.

Table 3 shows the number of times each employee was late in a given month.

The baseline shows the employees’ monthly tardiness before any of the strategies
in either example.

Month 1 shows the employees’ monthly tardiness after a month of the $10 bonuses.

Month 2 shows the employees’ monthly tardiness after a month of the $20 bonuses.

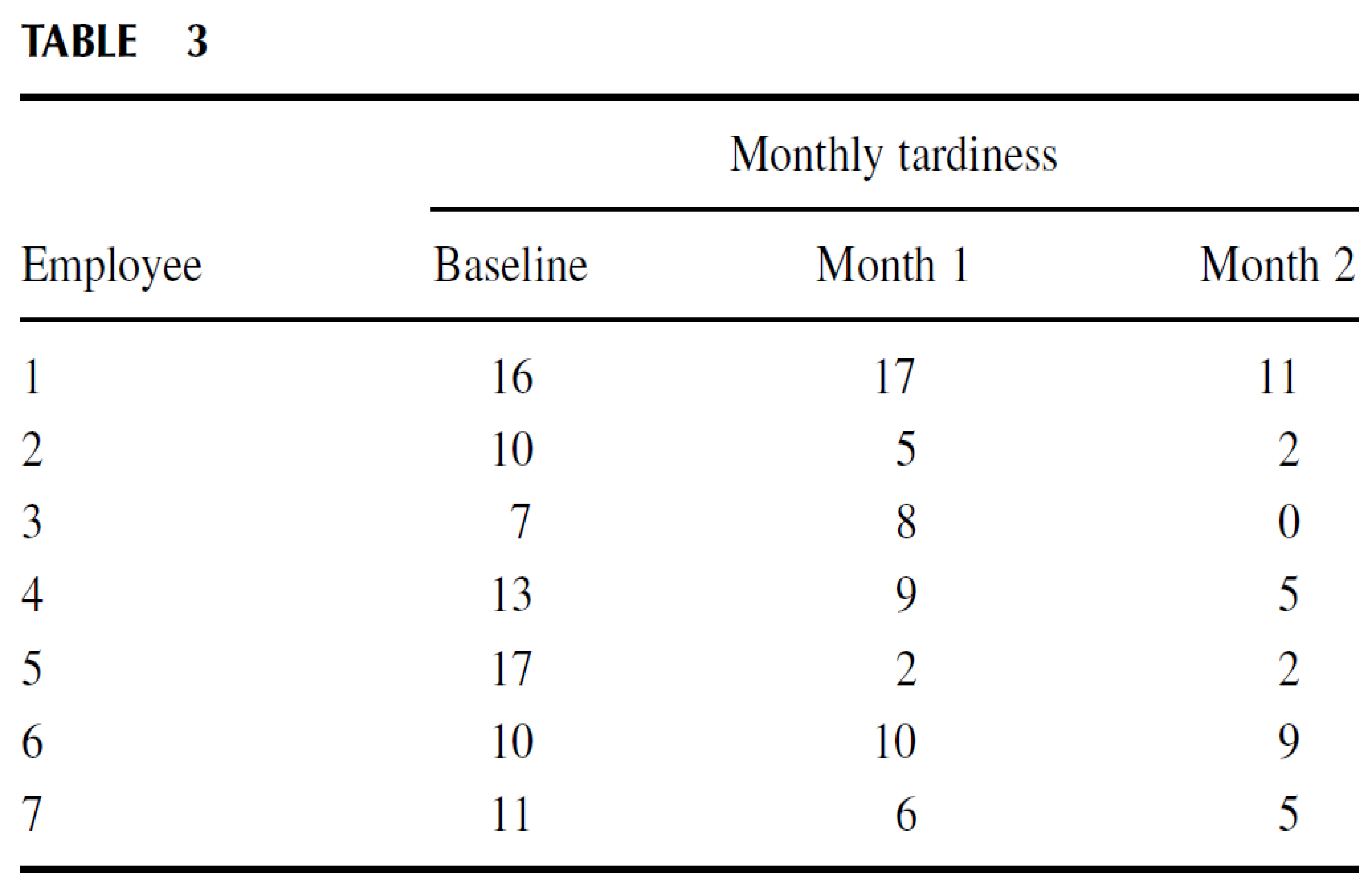


TABLE 3

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

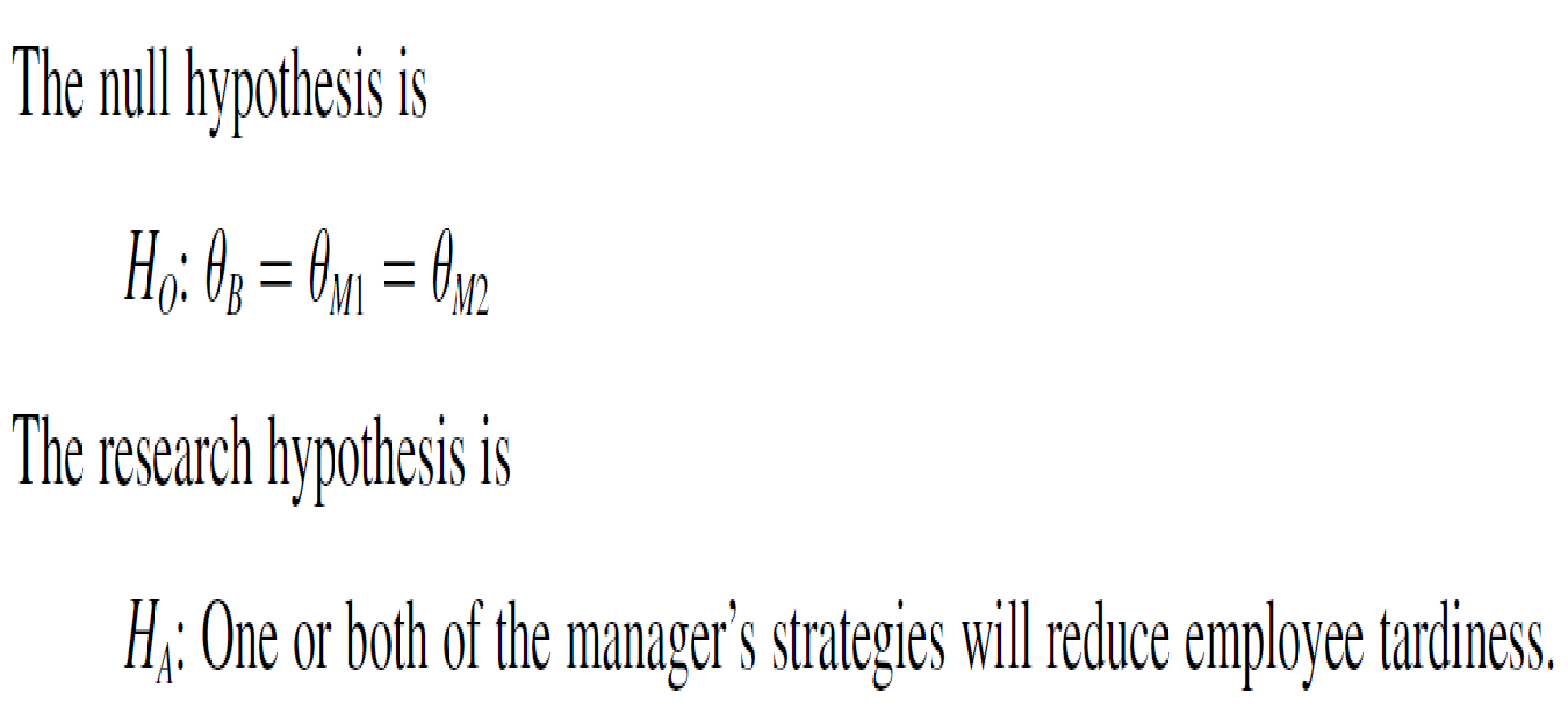
We want to determine if either of the strategies reduced employee tardiness.

Again, since the sample sizes are small (n < 20), we use a nonparametric test.

The Friedman test is a good statistic to analyze the data and test the hypothesis.

1. State the Null and Research Hypotheses

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

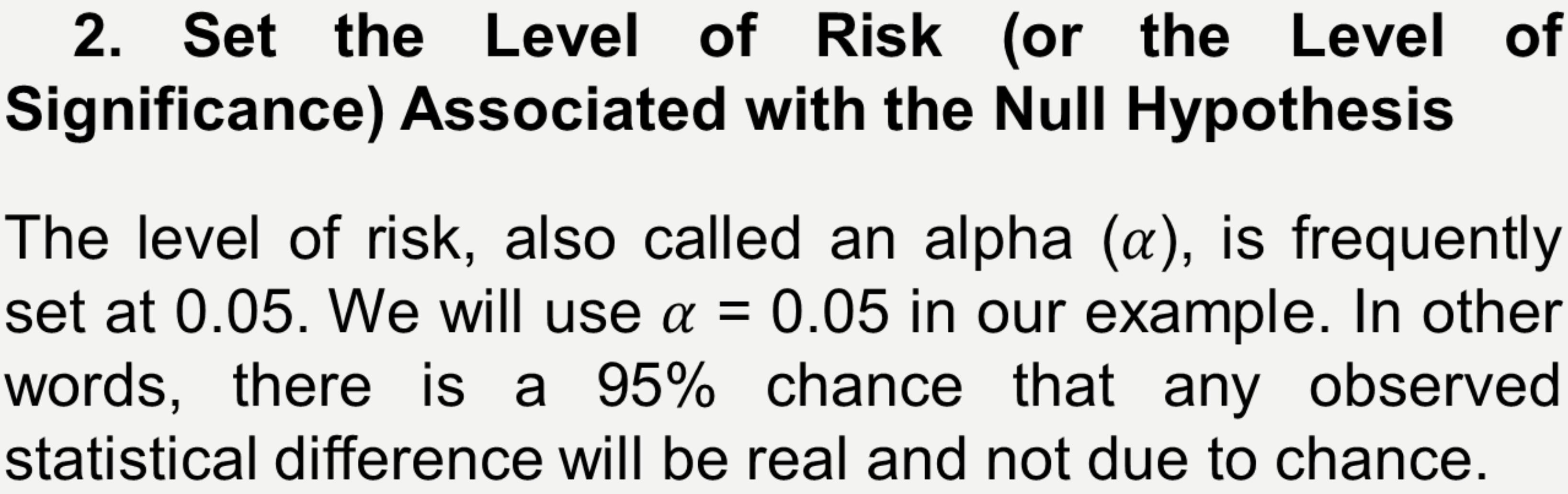


The null hypothesis is

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

The research hypothesis is

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



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

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

3. Choose the Appropriate Test Statistic

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

4. Compute the Test Statistic First, rank the values from each employee, or subject (Table 4).

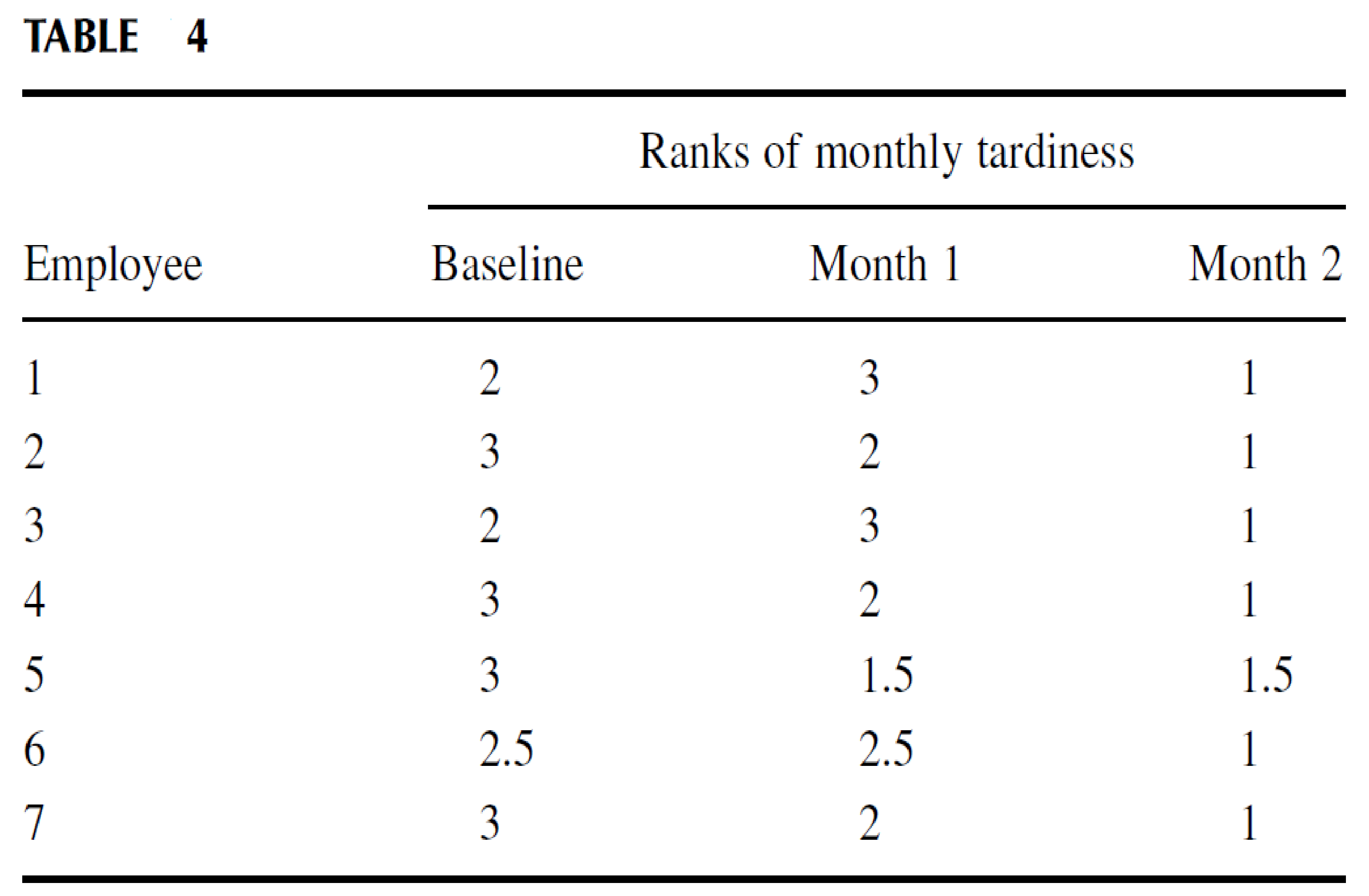
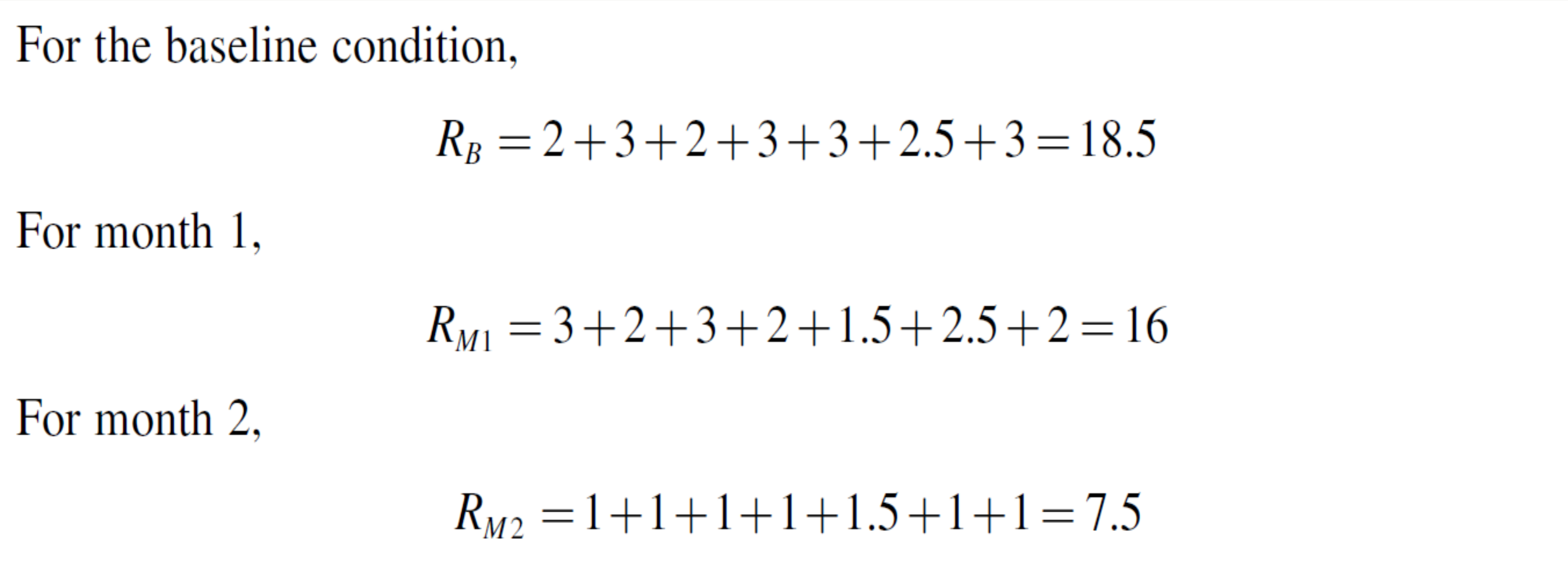


TABLE 4

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

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



For the baseline condition,

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

For month 1,

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

For month 2,

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

These R-values are used to compute the Fr test statistic. Since there were ties involved in the rankings, we must use Formula 2.

Finding the values for C_F and $\sum r^2_{ij}$ first will simplify the calculation:

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

$$C_F = 84$$

To find $\sum r^2_{ij}$, square all of the ranks. Then, add all of the squared ranks together (Table 5):

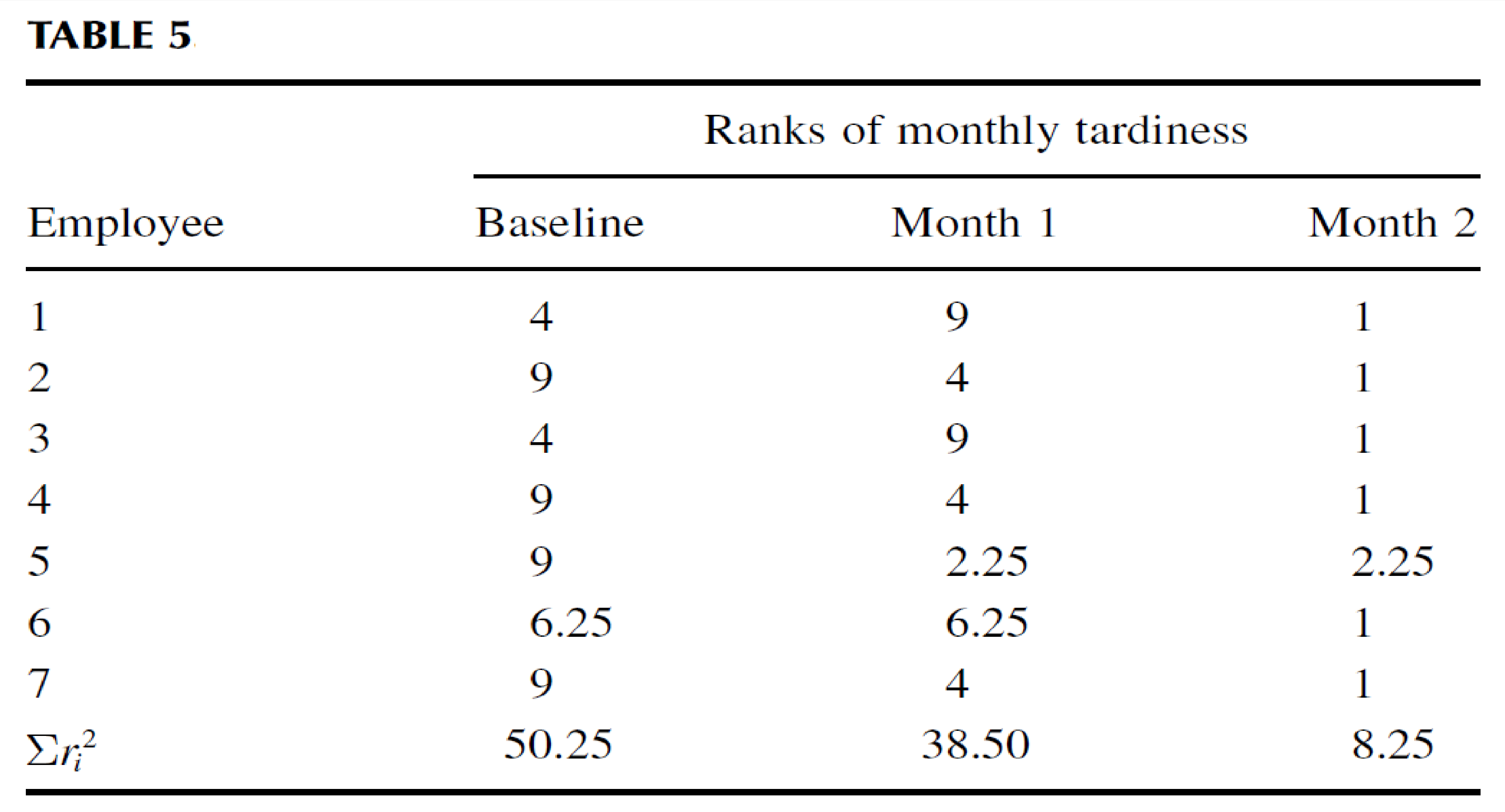
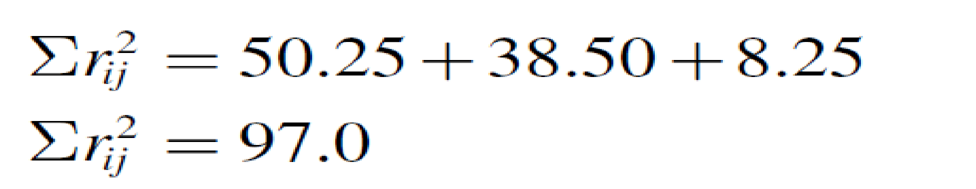


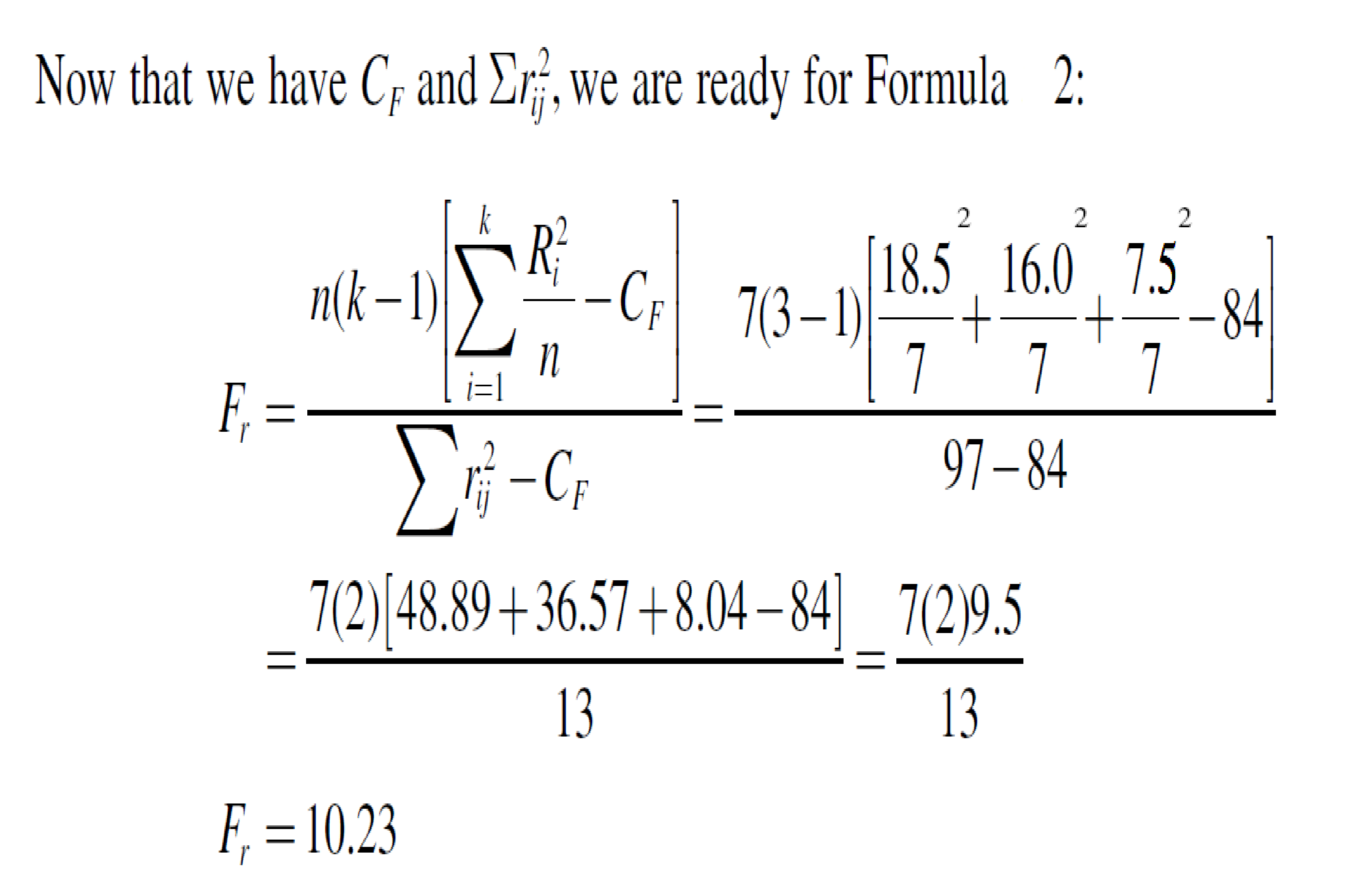
TABLE 5

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



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

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



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

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

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

$$F_r = 10.23$$

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

We will use the critical value table for the Friedman test (Table B.5) since it includes the number of groups, k, and the number of samples, n, for our data. In this case, we look for the critical value for k = 3 and n = 7 with α = 0.05.

Table B.5 returns a critical value for the Friedman test of 7.14.

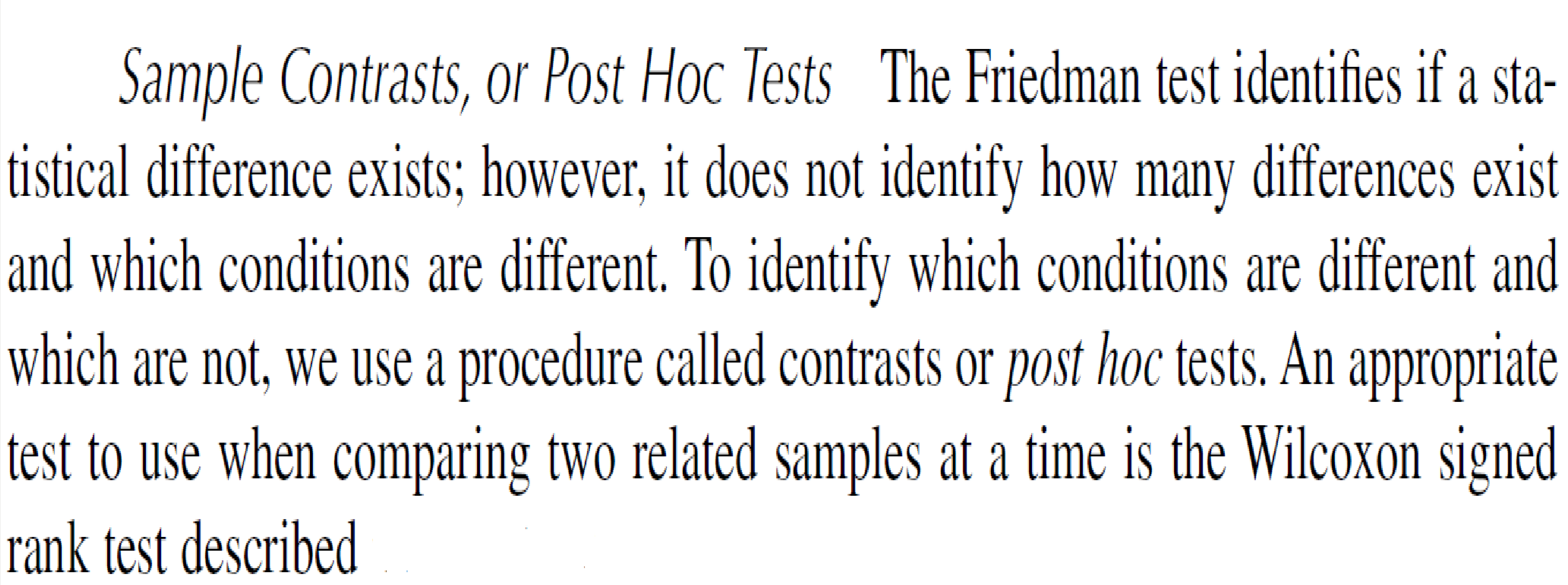
6. Compare the Obtained Value with the Critical Value

The critical value for rejecting the null hypothesis is 7.14 and the obtained value is F_r = 10.23.

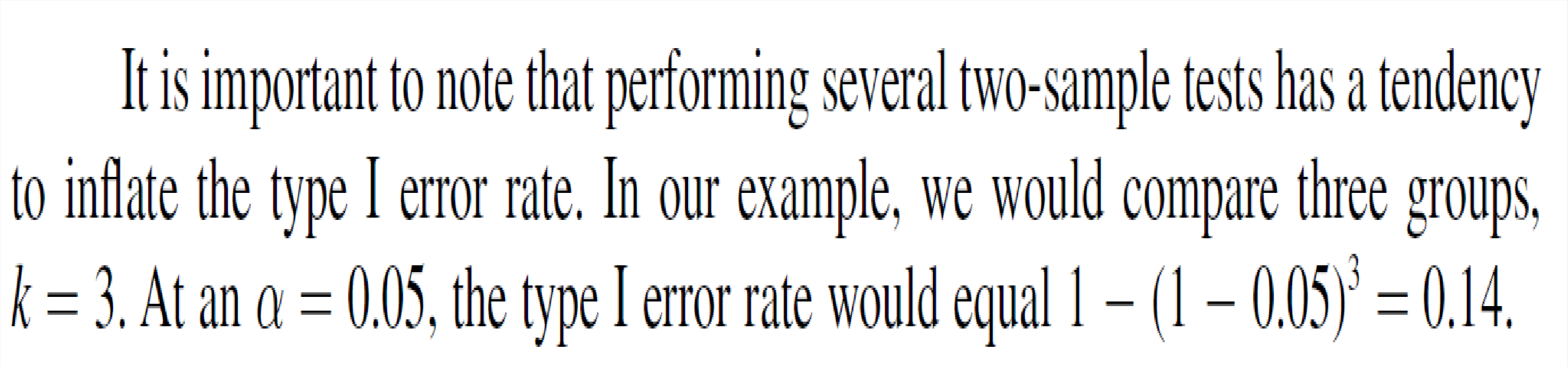
If the critical value is less than or equal to the obtained value, we must reject the null hypothesis. If instead, the critical value exceeds the obtained value, we do not reject the null hypothesis. Since the obtained value exceeds the critical value, we reject the null hypothesis.

7. Interpret the Results

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



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



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

To compensate for this error inflation, we demonstrate the Bonferroni procedure (Formula 4).

With this technique, we use a corrected with the Wilcoxon signed rank tests to determine significant differences between conditions.

For our example, we are only comparing Month 1 and Month 2 against the baseline.

We are not comparing Month 1 against Month 2. Therefore, we are only making two comparisons and k = 2:

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

$$\alpha_B = 0.025$$

When we compare the three samples with the Wilcoxon signed rank tests using α_B, we obtain the results presented in Table 6.

Notice that the significance is one-tailed, or directional, since we were seeking a decline in tardiness.

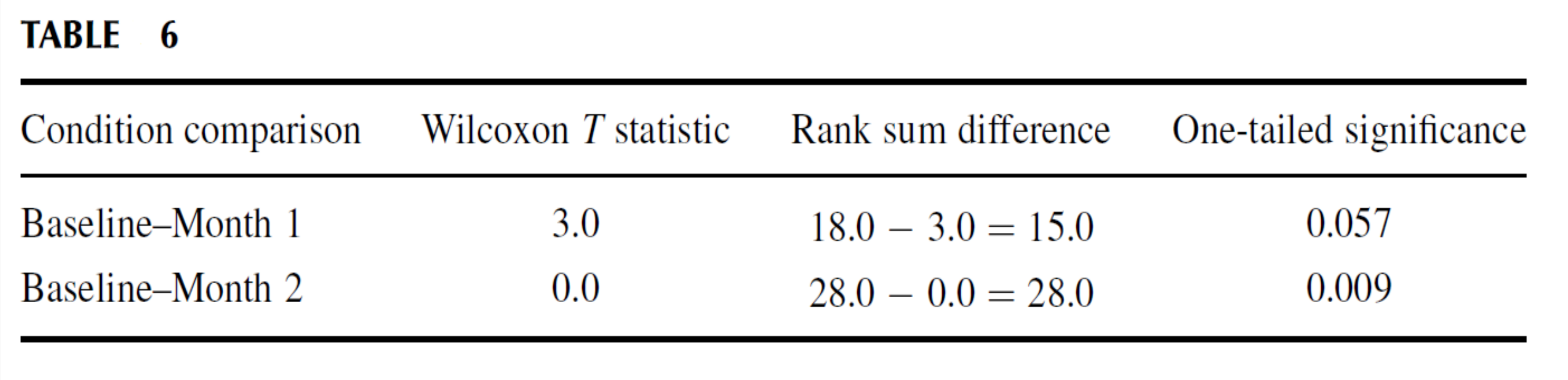
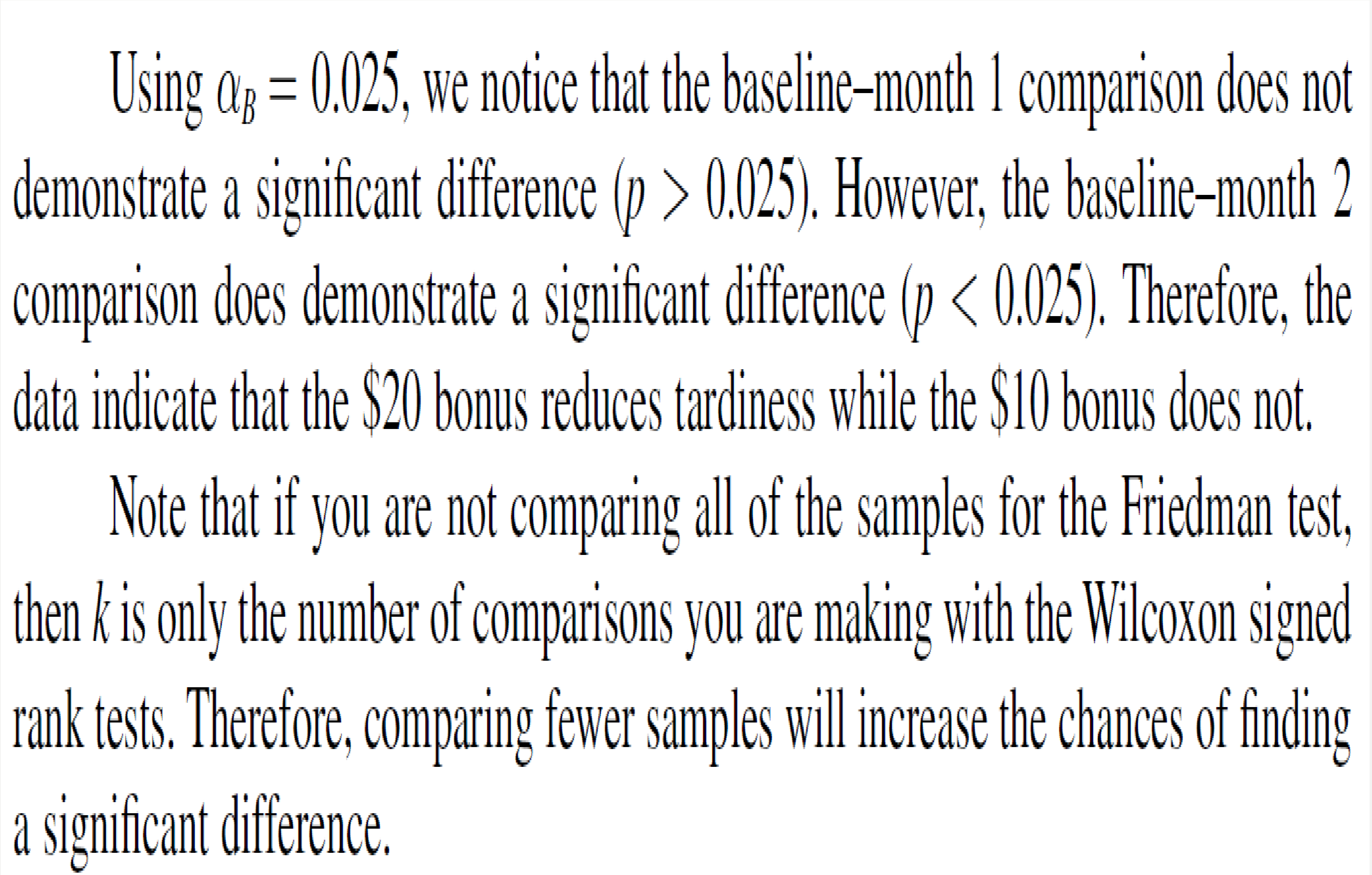


TABLE 6

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



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

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

Reporting the Results

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

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

SUMMARY

More than two samples that are related may be compared using the Friedman test.

The parametric equivalent to this test is known as the repeated measures ANOVA.

When the Friedman test produces significant results, it does not identify which nor how many pairs of conditions are significantly different.

The Wilcoxon signed rank test, with a Bonferroni procedure to avoid type I error rate inflation, is a useful method for comparing individual condition pairs.

In this lecture, we described how to perform and interpret a Friedman test followed with sample contrasts. The next lecture will explained how to perform the procedures using SPSS.