BAU
BAHÇEŞEHİR ÜNİVERSİTESİ

BME 3005
BIOSTATISTICS

Lecture 8: *paired t-test, McNemar’s Test, Mann-Whitney Test*

Burcu Tunç Çamlıbel

Chapter 6 Power of ANOVA

Power of ANOVA, Example

Suppose in the experimental study of rabbit sperm motility with three ($k = 3$) experimental conditions – ordinary control, stress control and cell phone exposure – we also wanted to measure the effect of cell phone exposure on sperm count. Normal sperm count in a rabbit is about 350 million sperm/mL with a standard deviation of about 20 million sperm/mL. What would be the power of the study with $n = 8$ rabbits per group we analyzed earlier (Box 3-1) to detect a change of 50 million sperm/mL at conventional statistical significance ($\alpha = .05$)?

Using this information, the noncentrality parameter is

$$\phi = \frac{\delta}{\sigma}\sqrt{\frac{n}{2k}} = \frac{50}{20}\sqrt{\frac{8}{2 \cdot 3}} = 2.88$$

There are $\nu_n = k - 1 = 3 - 1 = 2$ numerator and $\nu_d = k(n - 1) = 3\ (8 - 1) = 21$ denominator degrees of freedom. From the power chart in Figure 6-10, the power to detect a change of 50 million sperm/mL is .99, so we can be very confident of detecting this change.

Power of ANOVA, Example

This is an exceptionally high power. Given the cost of doing the experiments and a desire to minimize the number of animals used in the experiments, suppose that we would be happy with .80 power. ?????

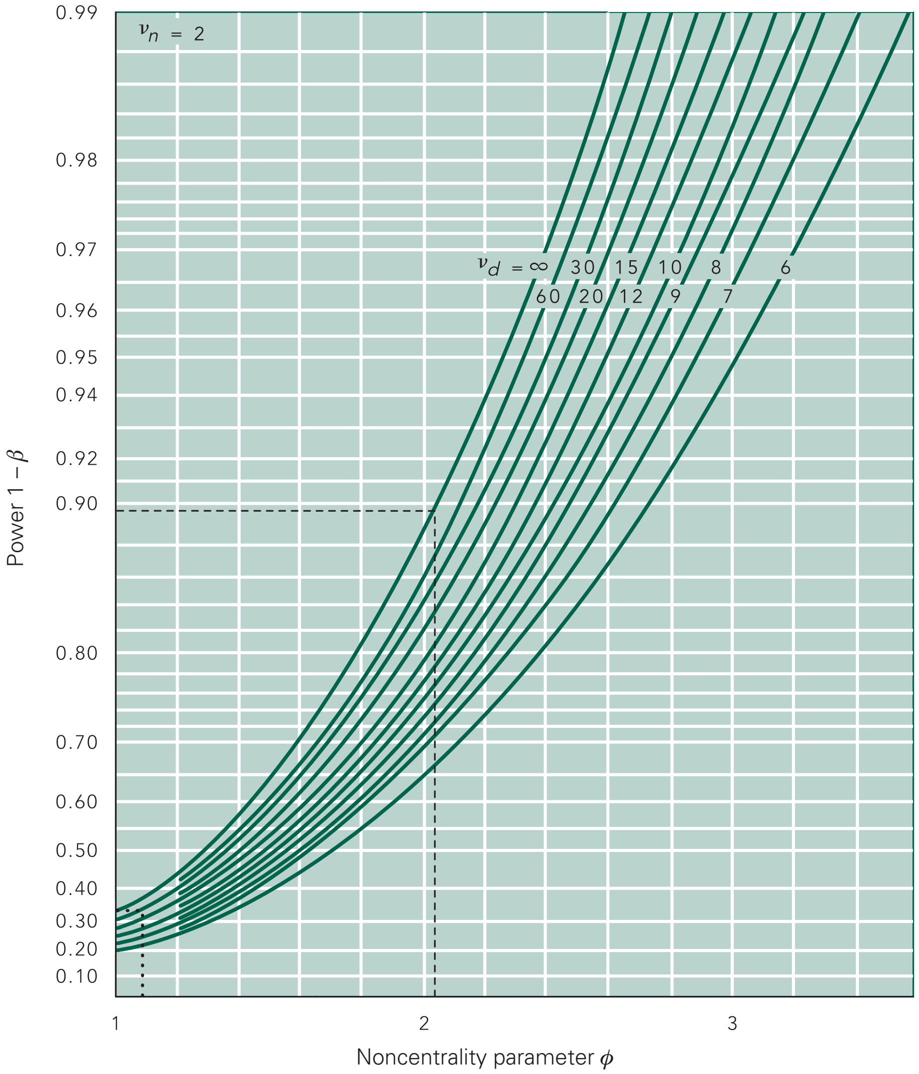
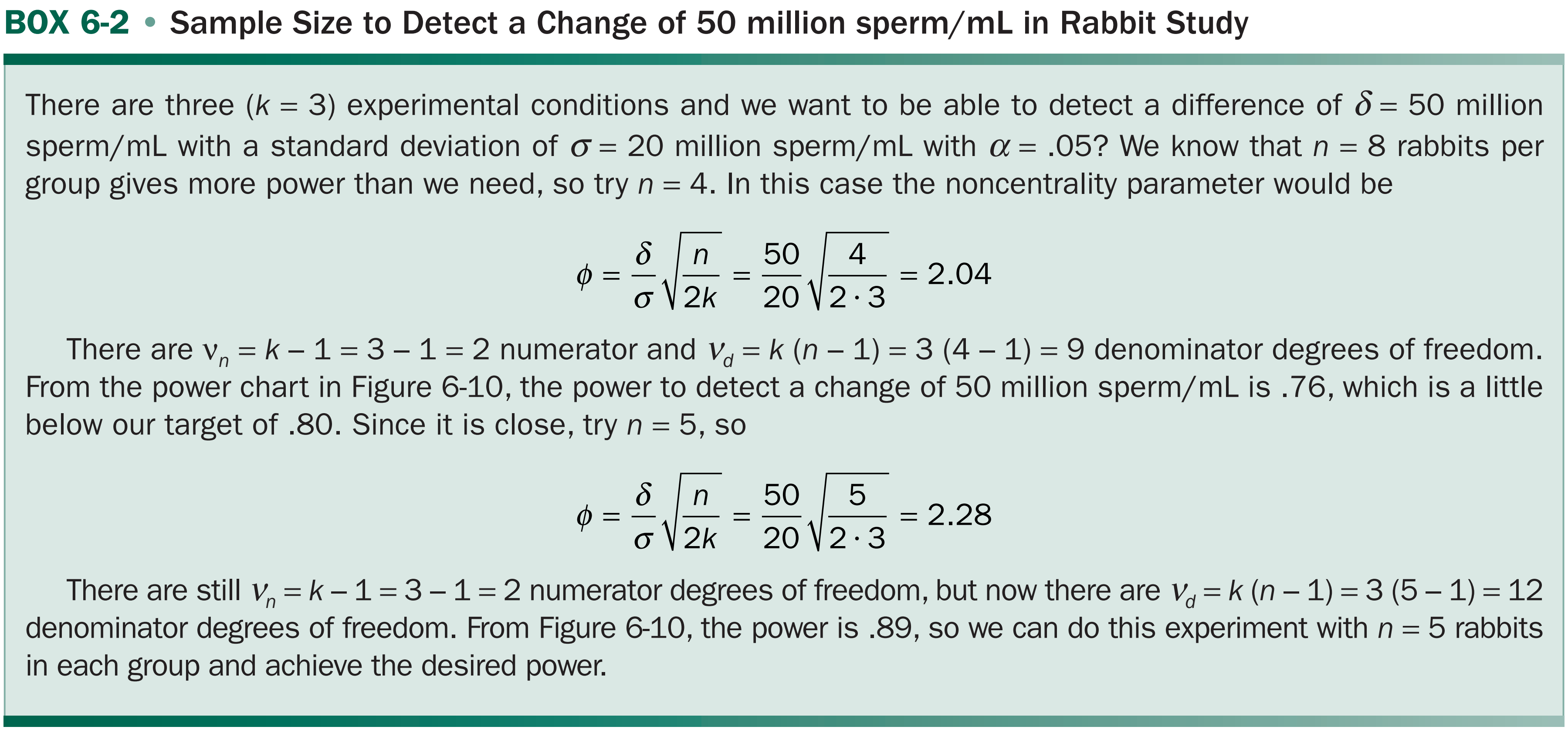


FIGURE 6-10. The power function for analysis of variance for $\nu_n = 2$ and $\alpha = 0.05$. Appendix B contains a complete set of power charts for a variety of values of ν_n and $\alpha = 0.05$ and .01. (*Source*: Adapted from Pearson ES, Hartley HO. Charts for the power function for analysis of variance tests, derived from the non-central f distribution. *Biometrika* 1951;38:112–130.)

Power of ANOVA, Example



BOX 6-2 • Sample Size to Detect a Change of 50 million sperm/mL in Rabbit Study

There are three ($k = 3$) experimental conditions and we want to be able to detect a difference of $\delta = 50$ million sperm/mL with a standard deviation of $\sigma = 20$ million sperm/mL with $\alpha = .05$? We know that $n = 8$ rabbits per group gives more power than we need, so try $n = 4$. In this case the noncentrality parameter would be

$$\phi = \frac{\delta}{\sigma}\sqrt{\frac{n}{2k}} = \frac{50}{20}\sqrt{\frac{4}{2 \cdot 3}} = 2.04$$

There are $\nu_n = k - 1 = 3 - 1 = 2$ numerator and $\nu_d = k\,(n - 1) = 3\,(4 - 1) = 9$ denominator degrees of freedom. From the power chart in Figure 6-10, the power to detect a change of 50 million sperm/mL is .76, which is a little below our target of .80. Since it is close, try $n = 5$, so

$$\phi = \frac{\delta}{\sigma}\sqrt{\frac{n}{2k}} = \frac{50}{20}\sqrt{\frac{5}{2 \cdot 3}} = 2.28$$

There are still $\nu_n = k - 1 = 3 - 1 = 2$ numerator degrees of freedom, but now there are $\nu_d = k\,(n - 1) = 3\,(5 - 1) = 12$ denominator degrees of freedom. From Figure 6-10, the power is .89, so we can do this experiment with $n = 5$ rabbits in each group and achieve the desired power.

Power of ANOVA, Example

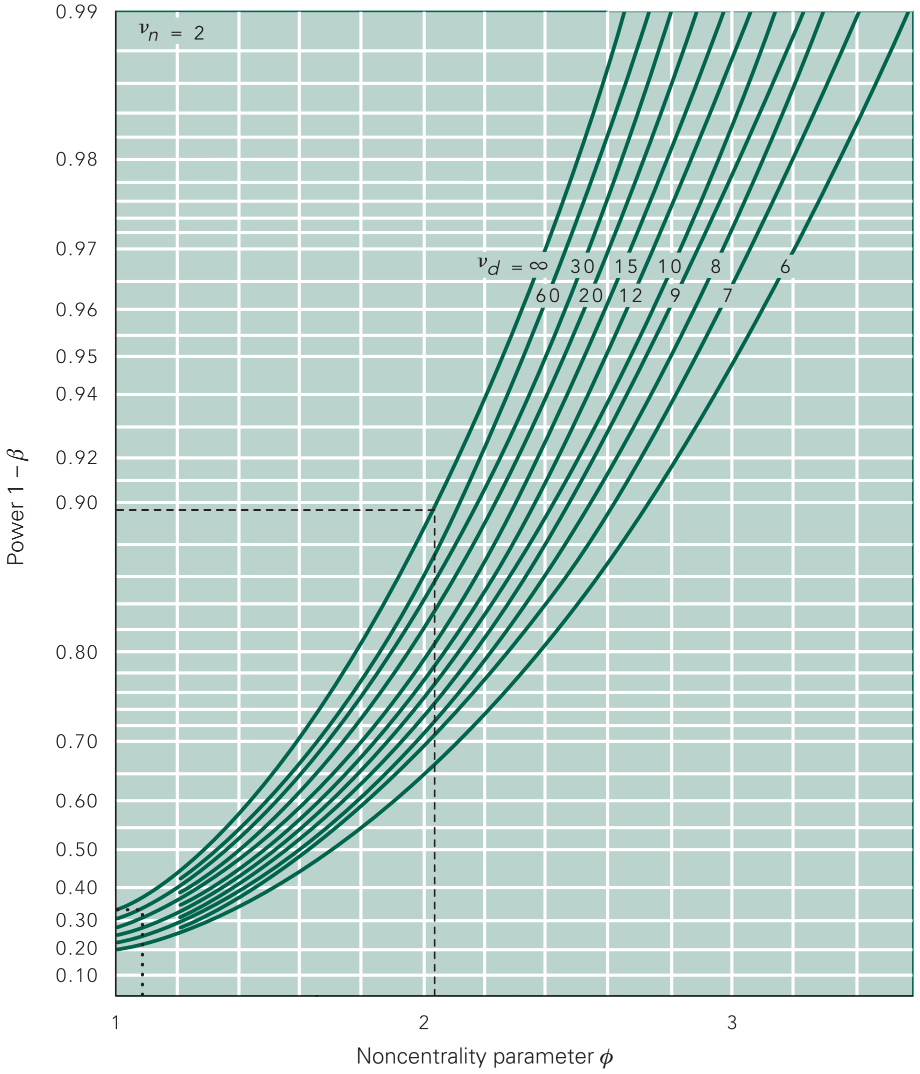
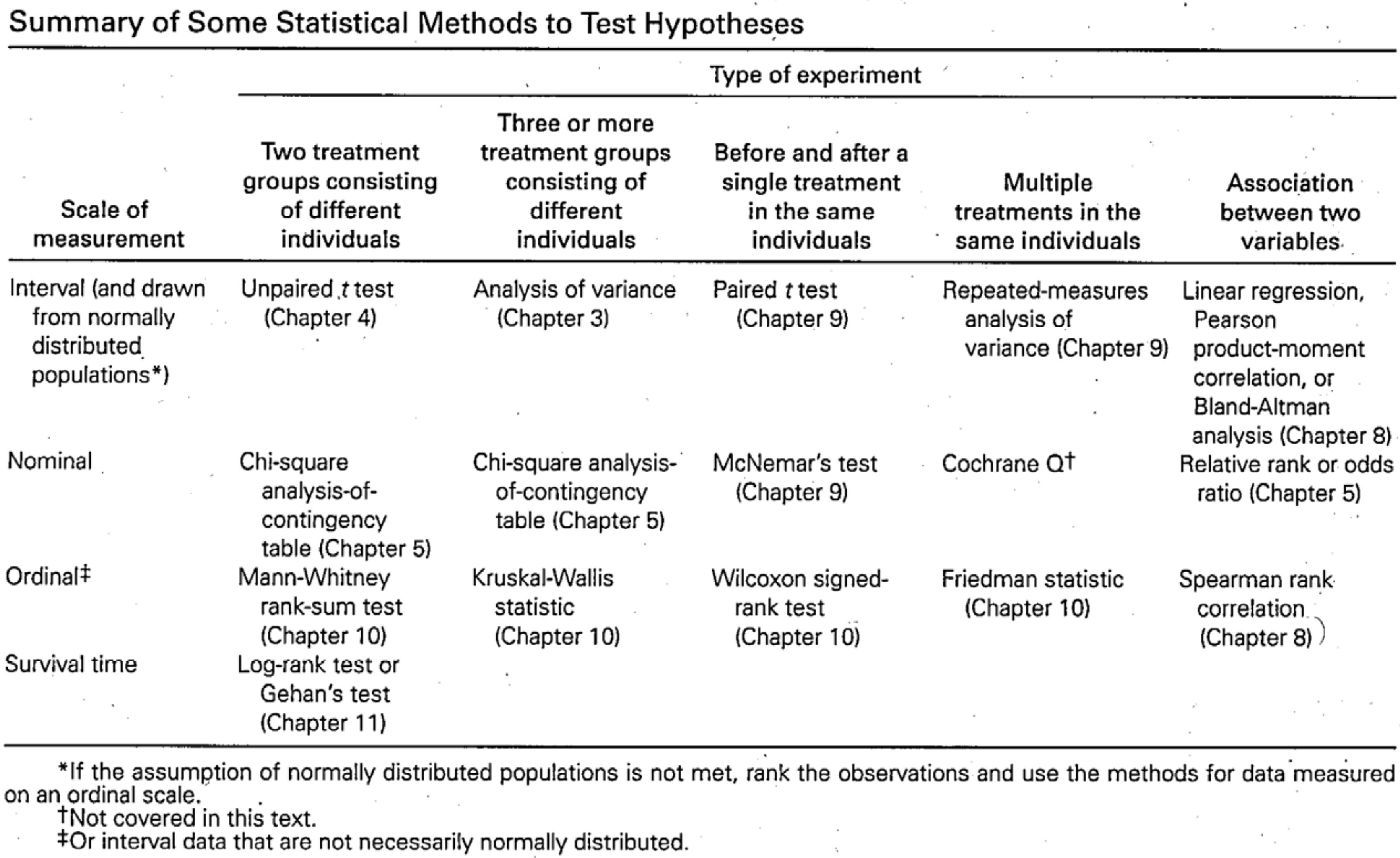


FIGURE 6-10. The power function for analysis of variance for $\nu_n = 2$ and $\alpha = 0.05$. Appendix B contains a complete set of power charts for a variety of values of ν_n and $\alpha = 0.05$ and .01. (*Source*: Adapted from Pearson ES, Hartley HO. Charts for the power function for analysis of variance tests, derived from the non-central f distribution. *Biometrika* 1951;38:112–130.)

Chapter 9
Experiments When Each Subject Receives More Than One Treatment



Summary of Some Statistical Methods to Test Hypotheses

	Type of experiment				
Scale of measurement	Two treatment groups consisting of different individuals	Three or more treatment groups consisting of different individuals	Before and after a single treatment in the same individuals	Multiple treatments in the same individuals	Association between two variables
Interval (and drawn from normally distributed populations*)	Unpaired *t* test (Chapter 4)	Analysis of variance (Chapter 3)	Paired *t* test (Chapter 9)	Repeated-measures analysis of variance (Chapter 9)	Linear regression, Pearson product-moment correlation, or Bland-Altman analysis (Chapter 8)
Nominal	Chi-square analysis-of-contingency table (Chapter 5)	Chi-square analysis-of-contingency table (Chapter 5)	McNemar's test (Chapter 9)	Cochrane Q†	Relative rank or odds ratio (Chapter 5)
Ordinal‡	Mann-Whitney rank-sum test (Chapter 10)	Kruskal-Wallis statistic (Chapter 10)	Wilcoxon signed-rank test (Chapter 10)	Friedman statistic (Chapter 10)	Spearman rank correlation (Chapter 8)
Survival time	Log-rank test or Gehan's test (Chapter 11)				

*If the assumption of normally distributed populations is not met, rank the observations and use the methods for data measured on an ordinal scale.

†Not covered in this text.

‡Or interval data that are not necessarily normally distributed.

EXPERIMENTS WHEN SUBJECTS ARE OBSERVED BEFORE AND AFTER A SINGLE TREATMENT: THE PAIRED t TEST

- In experiments in which it is possible to observe each experimental subject before and after administering a single treatment, we will test a hypothesis about the average change the treatment produces instead of the difference in average responses with and without the treatment.

- This approach reduces the variability in the observations due to differences between individuals and yields a more sensitive test.

The resulting value of *t* is compared with the critical value of ν = *n* − 1 degrees of freedom.

(

$$s_d = \sqrt{\frac{\sum (d - \bar{d})^2}{n-1}} \qquad s_{\bar{d}} = \frac{s_d}{\sqrt{n}} \qquad t = \frac{\bar{d}}{s_{\bar{d}}}$$

EXPERIMENTS WHEN SUBJECTS ARE OBSERVED BEFORE AND AFTER A SINGLE TREATMENT: THE PAIRED t TEST

(See page 187, cigarette smoking and platelet function example)

To recapitulate, when analyzing data from an experiment in which it is possible to observe each individual before and after applying a single treatment:

- Compute the change in response that accompanies the treatment in each individual *d*.
- Compute the mean change $\bar{d}$ and the standard error of the mean change $s_{\bar{d}}$.
- Use these numbers to compute t.
- Compare this t with the critical value of ν = *n* − 1 degrees of freedom.

$$s_d = \sqrt{\frac{\sum (d - \bar{d})^2}{n-1}}$$

$$s_{\bar{d}} = \frac{s_d}{\sqrt{n}}$$

$$t = \frac{\bar{d}}{s_{\bar{d}}}$$

EXPERIMENTS WHEN SUBJECTS ARE OBSERVED BEFORE AND AFTER A SINGLE TREATMENT: THE PAIRED t TEST

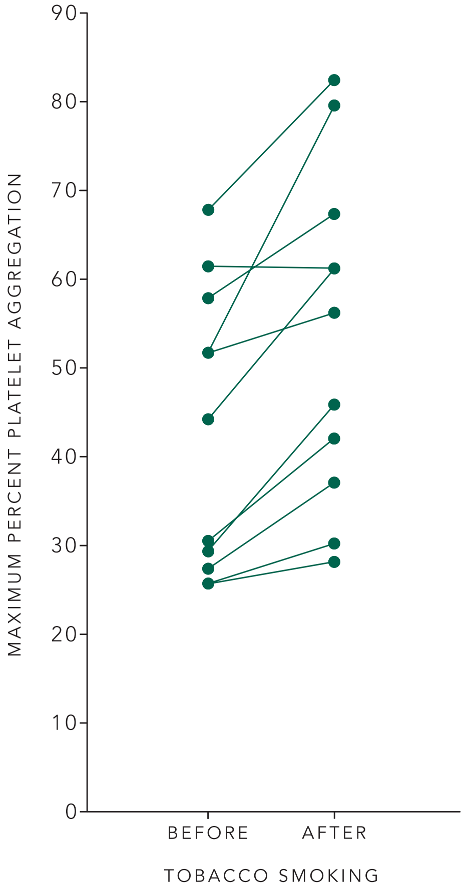
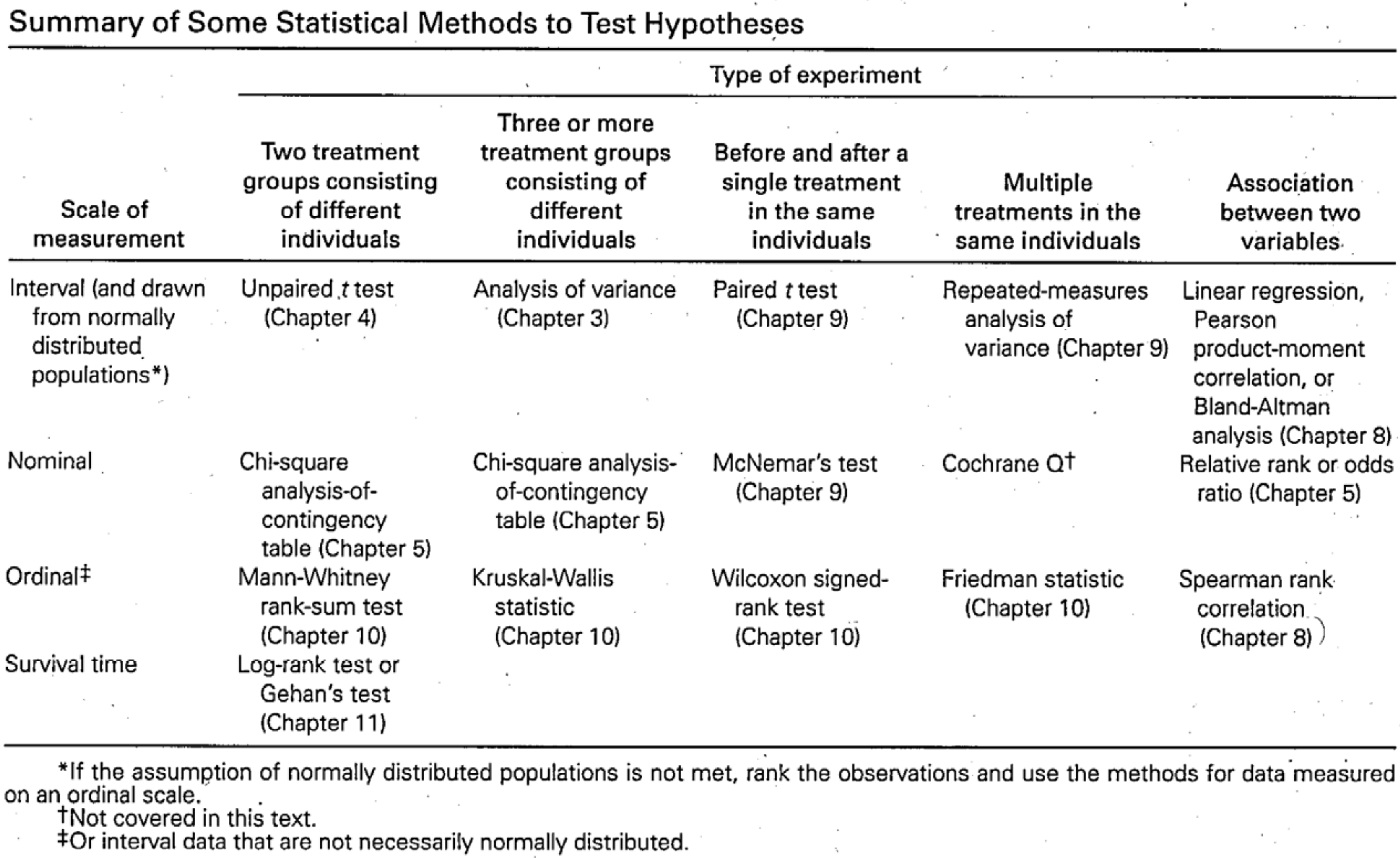


FIGURE 9-2. Maximum percentage platelet aggregation before and after smoking a tobacco cigarette in 11 people. (Adapted with permission of the American Heart Association, Inc. from Fig. 1 of Levine PH. An acute effect of cigarette smoking on platelet function: a possible link between smoking and arterial thrombosis. *Circulation*. 1973;48:619–623.)

The changes in maximum percent platelet aggregation that accompany smoking are (from Fig. 9-2) 2%, 4%, 10%, 12%, 16%, 15%, 4%, 27%, 9%, –1%, and 15%. Therefore, the mean change in percent platelet aggregation with smoking in these 11 people is $\bar{d} = 10.3\%$. The standard deviation of the change is 8.0%, so the standard error of the change is $s_{\bar{d}} = 8.0/\sqrt{11} = 2.41\%$. Finally, our test statistic is

$$t = \frac{\bar{d}}{s_{\bar{d}}} = \frac{10.3}{2.41} = 4.27$$

This value exceeds 3.169, the value that defines the most extreme 1% of the t distribution with $\nu = n - 1 = 11 - 1 = 10$ degrees of freedom (from Table 4-1). Therefore, we report that smoking increases platelet aggregation ($P < .01$).



Summary of Some Statistical Methods to Test Hypotheses

	Type of experiment				
Scale of measurement	Two treatment groups consisting of different individuals	Three or more treatment groups consisting of different individuals	Before and after a single treatment in the same individuals	Multiple treatments in the same individuals	Association between two variables
Interval (and drawn from normally distributed populations*)	Unpaired *t* test (Chapter 4)	Analysis of variance (Chapter 3)	Paired *t* test (Chapter 9)	Repeated-measures analysis of variance (Chapter 9)	Linear regression, Pearson product-moment correlation, or Bland-Altman analysis (Chapter 8)
Nominal	Chi-square analysis-of-contingency table (Chapter 5)	Chi-square analysis-of-contingency table (Chapter 5)	McNemar's test (Chapter 9)	Cochrane Q†	Relative rank or odds ratio (Chapter 5)
Ordinal‡	Mann-Whitney rank-sum test (Chapter 10)	Kruskal-Wallis statistic (Chapter 10)	Wilcoxon signed-rank test (Chapter 10)	Friedman statistic (Chapter 10)	Spearman rank correlation (Chapter 8)
Survival time	Log-rank test or Gehan's test (Chapter 11)				

*If the assumption of normally distributed populations is not met, rank the observations and use the methods for data measured on an ordinal scale.

†Not covered in this text.

‡Or interval data that are not necessarily normally distributed.

EXPERIMENTS WHEN OUTCOMES ARE MEASURED ON A NOMINAL SCALE: McNEMAR’S TEST

- The paired t test and repeated measures analysis of variance (you’re not responsible for problem solving part) can be used to analyze experiments in which the variable being studied can be measured on an interval scale (and satisfies the other assumptions required of parametric methods).

- What about experiments, analogous to the ones in Chapter 5, in which outcomes are measured on a nominal scale?

- This problem often arises when asking whether or not an individual responded to a treatment or when comparing the results of two different diagnostic tests that are classified as positive or negative in the same individuals.

- We will develop a procedure to analyze such experiments, McNemar’s test for changes, in the context of one such study.

EXPERIMENTS WHEN OUTCOMES ARE MEASURED ON A NOMINAL SCALE: McNEMAR’S TEST

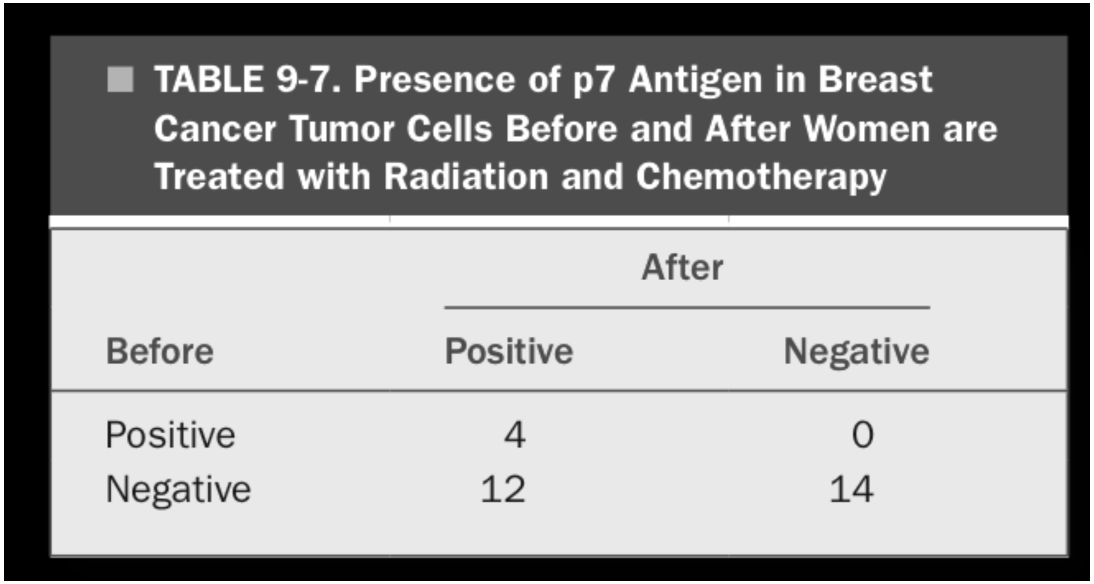
In sum, McNemar’s test for changes consists of the following procedure:

- Ignore individuals who responded the same way to both treatments.
- Compute the total number of individuals who responded differently to the two treatments.
- Compute the expected number of individuals who would have responded positively to each of the two treatments (but not the other) as half the total number of individuals who responded differently to the two treatments.
- Compare the observed and expected number of individuals that responded to one of the treatments by computing a χ 2 test statistic (including Yates correction for continuity).
- Compare this value of χ2 with the critical values of the χ2 distribution with 1 degree of freedom

- See example «p7 Antigen Expression in Human Breast Cancer» (page 200-201)

EXPERIMENTS WHEN OUTCOMES ARE MEASURED ON A NOMINAL SCALE: McNEMAR’S TEST

- See example «p7 Antigen Expression in Human Breast Cancer» (page 200-201)

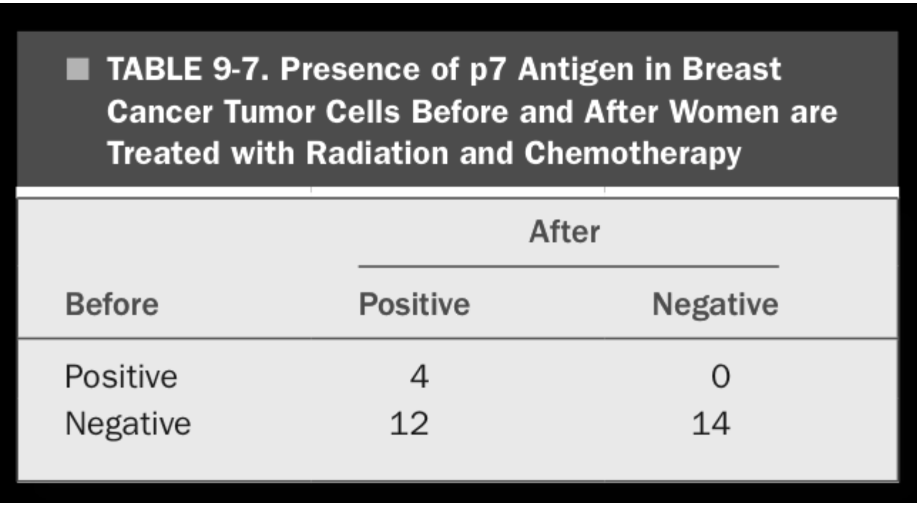


■ TABLE 9-7. Presence of p7 Antigen in Breast Cancer Tumor Cells Before and After Women are Treated with Radiation and Chemotherapy

	After	
Before	Positive	Negative
Positive	4	0
Negative	12	14

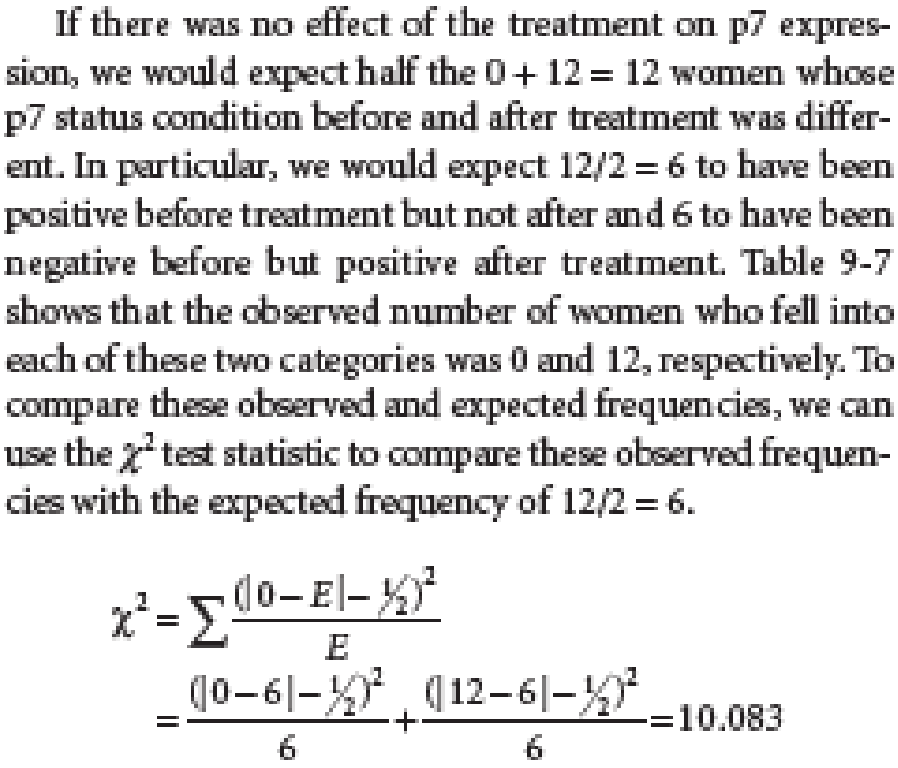
EXPERIMENTS WHEN OUTCOMES ARE MEASURED ON A NOMINAL SCALE: McNEMAR’S TEST

- See example «p7 Antigen Expression in Human Breast Cancer» (page 200-201)



■ TABLE 9-7. Presence of p7 Antigen in Breast Cancer Tumor Cells Before and After Women are Treated with Radiation and Chemotherapy

	After	
Before	Positive	Negative
Positive	4	0
Negative	12	14



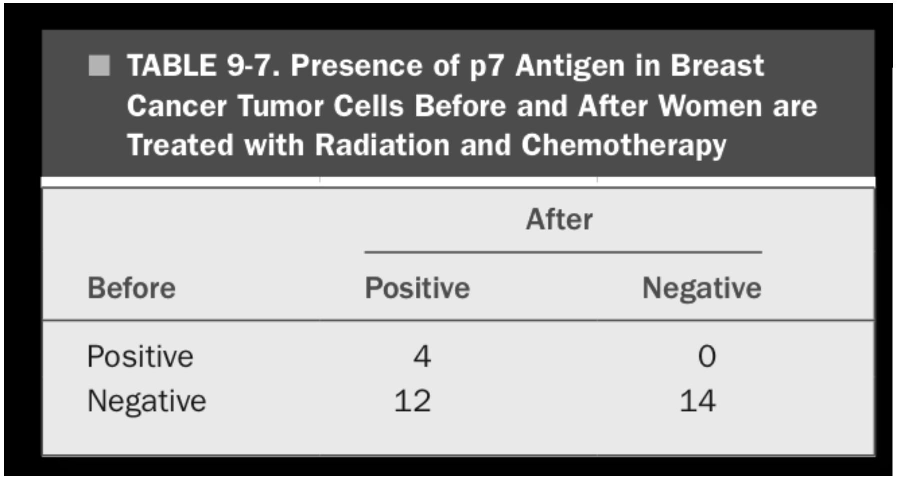
If there was no effect of the treatment on p7 expression, we would expect half the 0 + 12 = 12 women whose p7 status condition before and after treatment was different. In particular, we would expect 12/2 = 6 to have been positive before treatment but not after and 6 to have been negative before but positive after treatment. Table 9-7 shows that the observed number of women who fell into each of these two categories was 0 and 12, respectively. To compare these observed and expected frequencies, we can use the χ^2 test statistic to compare these observed frequencies with the expected frequency of 12/2 = 6.

$$\chi^2 = \sum \frac{(|O - E| - \frac{1}{2})^2}{E}$$

$$= \frac{(|0 - 6| - \frac{1}{2})^2}{6} + \frac{(|12 - 6| - \frac{1}{2})^2}{6} = 10.083$$

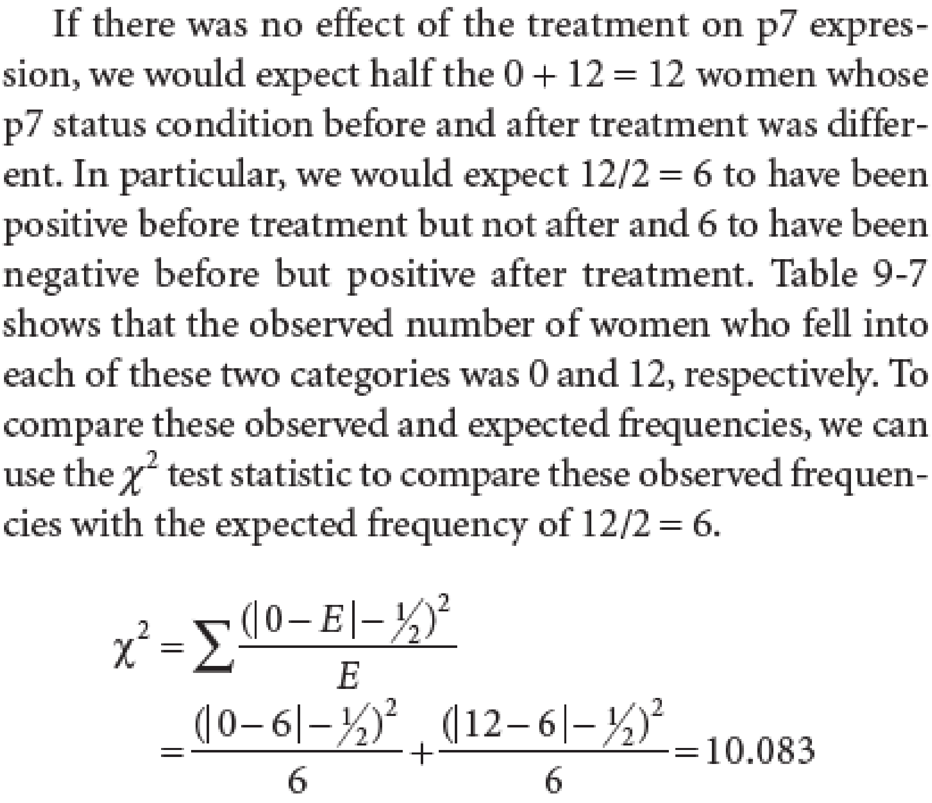
EXPERIMENTS WHEN OUTCOMES ARE MEASURED ON A NOMINAL SCALE: McNEMAR'S TEST

- See example «p7 Antigen Expression in Human Breast Cancer» (page 200-201)



■ **TABLE 9-7. Presence of p7 Antigen in Breast Cancer Tumor Cells Before and After Women are Treated with Radiation and Chemotherapy**

	After	
Before	Positive	Negative
Positive	4	0
Negative	12	14

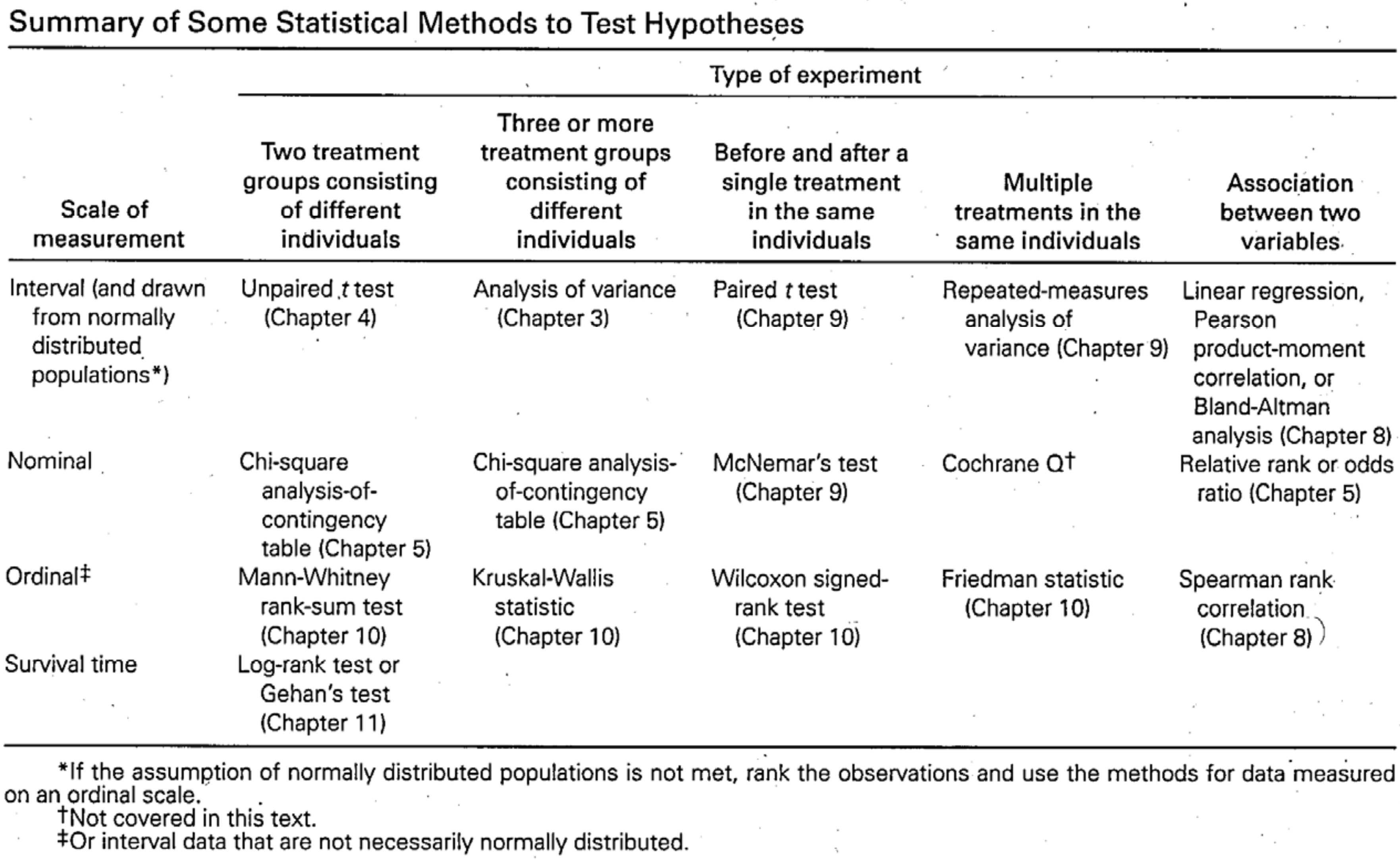


If there was no effect of the treatment on p7 expression, we would expect half the 0 + 12 = 12 women whose p7 status condition before and after treatment was different. In particular, we would expect 12/2 = 6 to have been positive before treatment but not after and 6 to have been negative before but positive after treatment. Table 9-7 shows that the observed number of women who fell into each of these two categories was 0 and 12, respectively. To compare these observed and expected frequencies, we can use the χ^2 test statistic to compare these observed frequencies with the expected frequency of 12/2 = 6.

$$\chi^2 = \sum \frac{(|O-E|-\frac{1}{2})^2}{E}$$

$$= \frac{(|0-6|-\frac{1}{2})^2}{6} + \frac{(|12-6|-\frac{1}{2})^2}{6} = 10.083$$

Chapter 10
Alternatives to Analysis of Variance and the t test Based on Ranks



Summary of Some Statistical Methods to Test Hypotheses

	Type of experiment				
Scale of measurement	Two treatment groups consisting of different individuals	Three or more treatment groups consisting of different individuals	Before and after a single treatment in the same individuals	Multiple treatments in the same individuals	Association between two variables
Interval (and drawn from normally distributed populations*)	Unpaired *t* test (Chapter 4)	Analysis of variance (Chapter 3)	Paired *t* test (Chapter 9)	Repeated-measures analysis of variance (Chapter 9)	Linear regression, Pearson product-moment correlation, or Bland-Altman analysis (Chapter 8)
Nominal	Chi-square analysis-of-contingency table (Chapter 5)	Chi-square analysis-of-contingency table (Chapter 5)	McNemar's test (Chapter 9)	Cochrane Q†	Relative rank or odds ratio (Chapter 5)
Ordinal‡	Mann-Whitney rank-sum test (Chapter 10)	Kruskal-Wallis statistic (Chapter 10)	Wilcoxon signed-rank test (Chapter 10)	Friedman statistic (Chapter 10)	Spearman rank correlation (Chapter 8)
Survival time	Log-rank test or Gehan's test (Chapter 11)				

*If the assumption of normally distributed populations is not met, rank the observations and use the methods for data measured on an ordinal scale.

†Not covered in this text.

‡Or interval data that are not necessarily normally distributed.

TWO DIFFERENT SAMPLES: THE MANN-WHITNEY RANK-SUM TEST

The procedure for testing the hypothesis that a treatment had no effect with this statistic is:

- *Rank all observations according to their magnitude, a rank of 1 being assigned to the smallest observation. Tied observations should be assigned the same rank, equal to the average of the ranks they would have been assigned had there been no tie (i.e., using the same procedure as in computing the Spearman rank correlation coefficient in Chapter 8).*
- *Compute T, the sum of the ranks in the smaller sample. (If both samples are the same size, you can compute T from either one.)*
- *Compare the resulting value of T with the distribution of all possible rank sums for experiments with samples of the same size to see whether the pattern of rankings is compatible with the hypothesis that the treatment had no effect. (use table 10.3 page 210 if sample size is smaller than 8) or calcute z_T (next slide) (use table 4.1)*

TWO DIFFERENT SAMPLES: THE MANN-WHITNEY RANK-SUM TEST

$$\mu_T = \frac{n_S(n_S + n_B + 1)}{2}$$

and standard deviation

$$\sigma_T = \sqrt{\frac{n_S n_B (n_S + n_B + 1)}{12}}$$

$$z_T = \frac{\left|T - \mu_T\right| - \frac{1}{2}}{\sigma_T}$$

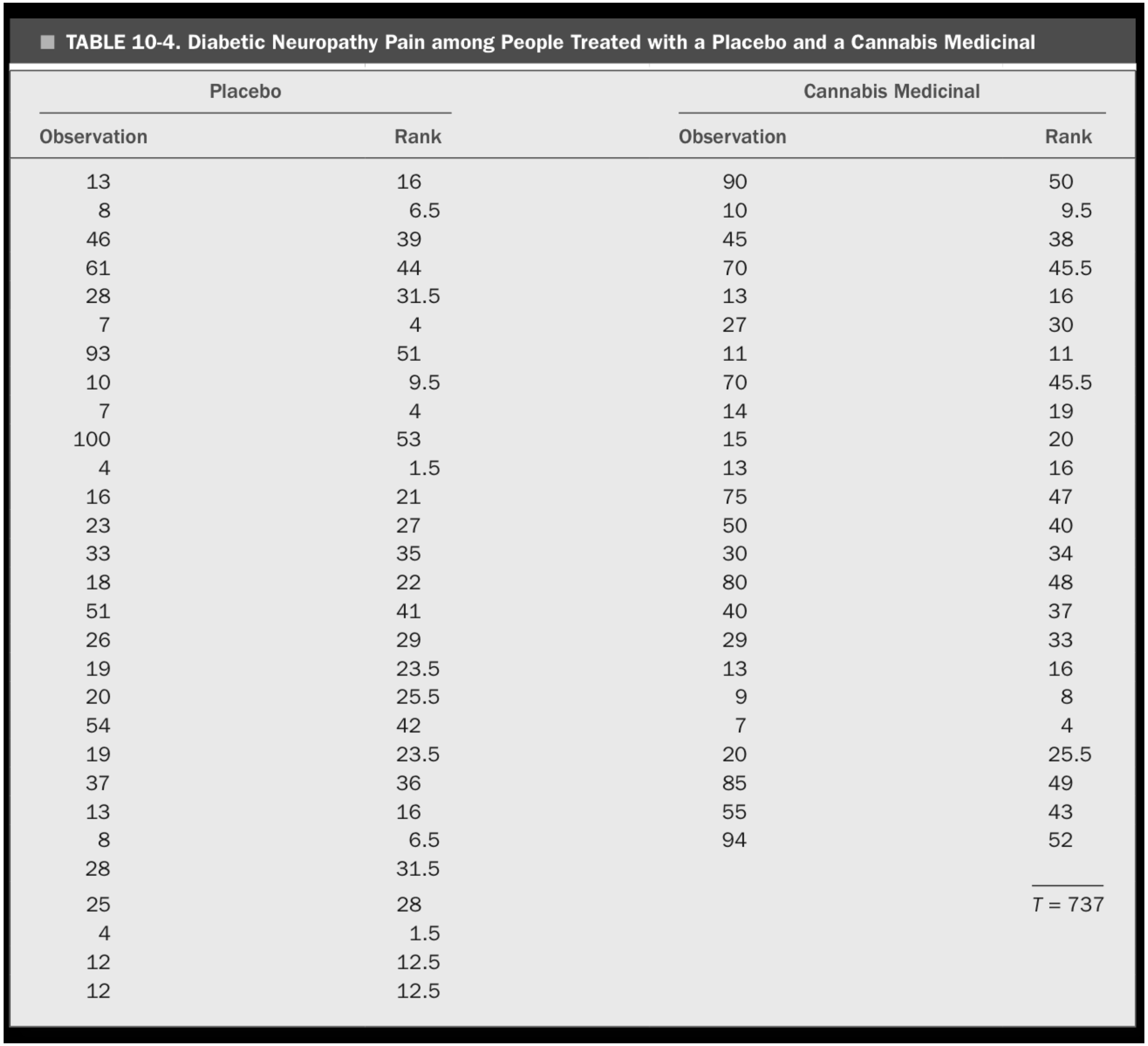
n_S is the size of the smaller sample.*

compare this statistic with the critical values of the normal distribution that define, say 5%, the most extreme possible values.

z_T can also be compared with the t distribution with an infinite number of degrees of freedom Table 4-1) because it is equals to the normal distribution.

Example page 211: USE OF A CANNABIS-BASED MEDICINE IN PAINFUL DIABETIC NEUROPATHY

USE OF A CANNABIS-BASED MEDICINE IN PAINFUL DIABETIC NEUROPATHY



■ TABLE 10-4. Diabetic Neuropathy Pain among People Treated with a Placebo and a Cannabis Medicinal

Placebo		Cannabis Medicinal	
Observation	Rank	Observation	Rank
13	16	90	50
8	6.5	10	9.5
46	39	45	38
61	44	70	45.5
28	31.5	13	16
7	4	27	30
93	51	11	11
10	9.5	70	45.5
7	4	14	19
100	53	15	20
4	1.5	13	16
16	21	75	47
23	27	50	40
33	35	30	34
18	22	80	48
51	41	40	37
26	29	29	33
19	23.5	13	16
20	25.5	9	8
54	42	7	4
19	23.5	20	25.5
37	36	85	49
13	16	55	43
8	6.5	94	52
28	31.5		
25	28		*T* = 737
4	1.5		
12	12.5		
12	12.5		

USE OF A CANNABIS-BASED MEDICINE IN PAINFUL DIABETIC NEUROPATHY

The cannabis medicinal group is the smaller sample, so we compute the test statistic T by summing all the ranks in that group, yielding $T = 737$. The cannabis group has $n_S =$ 24 people in it and the larger placebo group, $n_B = 29$, so the mean value of T for all studies of this size is

$$\mu_T = \frac{n_S(n_S + n_B + 1)}{2} = \frac{24(24 + 29 + 1)}{2} = 648$$

and the standard deviation is

$$\sigma_T = \sqrt{\frac{n_S n_B(n_S + n_B + 1)}{2}} = \sqrt{\frac{24 \cdot 29(24 + 29 + 1)}{12}} = 55.96$$

So

$$z_T = \frac{|T - \mu_T| - \frac{1}{2}}{\sigma_T} = \frac{|737 - 648| - \frac{1}{2}}{55.96} = 1.581$$

This value is smaller than 1.960, the value of z that defines the most extreme 5% of the normal distribution (from Table 4-1). Hence, this study does not provide substantial evidence that the cannabis medicinal was any more or less effective than placebo in controlling pain associated with diabetic neuropathy.