Chapter

8

Comparison of Paired Samples

Objectives

In this chapter we study comparisons of paired samples. We will

- demonstrate how to conduct a paired *t* test.
- demonstrate how to construct and interpret a confidence interval for the mean of a paired difference.
- discuss ways in which paired data arise and how pairing can be advantageous.
- consider the conditions under which a paired *t* test is valid.
- show how paired data may be analyzed using the sign test and the Wilcoxon signed-rank test.

8.1 Introduction

In Chapter 7 we considered the comparison of two independent samples when the response variable Y is a quantitative variable. In the present chapter we consider the comparison of two samples that are not independent but are paired. In a **paired design**, the observations (Y_1, Y_2) occur in pairs; the observational units in a pair are linked in some way, so that they have more in common with each other than with members of another pair. The following is an example of a paired design.

Example 8.1.1

Blood Flow Does drinking coffee affect blood flow, particularly during exercise? Doctors studying healthy subjects measured myocardial blood flow (MBF)* during bicycle exercise before and after giving the subjects a dose of caffeine that was equivalent to drinking two cups of coffee. Table 8.1.1 shows the MBF levels before (baseline) and after (caffeine) the subjects took a tablet containing 200 mg of caffeine.[1] Figure 8.1.1 shows parallel dotplots of these data, with line segments that connect the baseline and caffeine readings for each subject so that the change from "before" to "after" is evident for each subject. ■

In Example 8.1.1 the data arise in pairs; the data in a pair are linked by virtue of being measurements on the same person. A suitable analysis of the data should take advantage of this pairing. That is, we could imagine an experiment in which some subjects are studied after being given caffeine and others are studied without ever being given caffeine; such an experiment would provide two independent samples of data and could be analyzed using the methods of Chapter 7. But the current experiment used a paired design. Myocardial blood flow varies from person to person, with some subjects having high MBF levels both before and after consuming caffeine and others having low MBF levels. Knowing a subject's MBF level at baseline

*MBF was measured by taking positron emission tomography (PET) images after oxygen-15 labeled water was infused in the patients.

Table 8.1.1 Myocardial blood flow (ml/min/g) for eight subjects

Subject	MBF	
	Baseline y_1	Caffeine y_2
1	6.37	4.52
2	5.69	5.44
3	5.58	4.70
4	5.27	3.81
5	5.11	4.06
6	4.89	3.22
7	4.70	2.96
8	3.53	3.20
Mean	5.14	3.99
SD	0.83	0.86

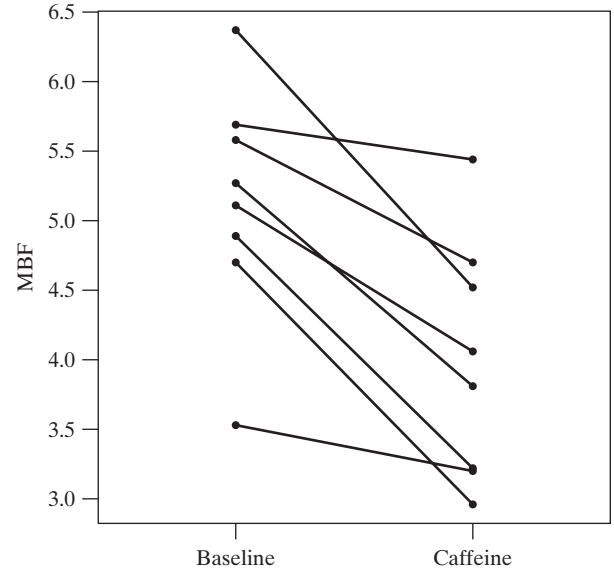

Figure 8.1.1 Dotplots of MBF readings before and after caffeine consumption, with line segments connecting readings on each subject

tells us something about how the subject did on caffeine, and vice versa. We want to use this information when we analyze the data.

In Section 8.2 we show how to analyze paired data using methods based on Student's *t* distribution. In Sections 8.4 and 8.5 we describe two nonparametric tests for paired data. Sections 8.3, 8.6, and 8.7 contain more examples and discussion of the paired design.

8.2 The Paired-Sample *t* Test and Confidence Interval

In this section we discuss the use of Student's *t* distribution to obtain tests and confidence intervals for paired data.

Analyzing Differences

In Chapter 7 we considered how to analyze data from two independent samples. When we have paired data, we make a simple shift of viewpoint: Instead of considering Y_1 and Y_2 separately, we consider the *difference* D, defined as

$$D = Y_1 - Y_2$$

Note that it is often natural to consider a difference as the response variable of interest in a study. For example, if we were studying the growth rates of plants, we might grow plants under control conditions for a while at the beginning of a study and then apply a treatment for one week. We would measure the growth that takes place during the week after the treatment is introduced as $D = Y_1 - Y_2$, where Y_1 = height one week after applying the treatment and Y_2 = height before the treatment is applied.* Sometimes data are paired in a way that is less obvious, but whenever we have paired data, it is the observed differences that we wish to analyze.

*Exercises 7.2.11 and 7.2.12 both involve such "before versus after" data.

Let us denote the mean of sample D's as $\overline{D}$. The quantity $\overline{D}$ is related to the individual sample means as follows:

$$\overline{D} = (\overline{Y}_1 - \overline{Y}_2)$$

The relationship between population means is analogous:

$$\mu_D = \mu_1 - \mu_2$$

Thus, we may say that *the mean of the difference is equal to the difference of the means.* Because of this simple relationship, a comparison of two paired means can be carried out by concentrating entirely on the D's.

The standard error for $\overline{D}$ is easy to calculate. Because $\overline{D}$ is just the mean of a single sample, we can apply the SE formula of Chapter 6 to obtain the following formula:

$$\text{SE}_{\overline{D}} = \frac{s_D}{\sqrt{n_D}}$$

where s_D is the standard deviation of the D's and n_D is the number of D's. The following example illustrates the calculation.

Example 8.2.1 Blood Flow Table 8.2.1 shows the blood flow data of Example 8.1.1 and the differences d.

Note that the mean of the difference is equal to the difference of the means:

$$\overline{d} = 1.15 = 5.14 - 3.99$$

Figure 8.2.1 shows the distribution of the 8 sample differences.

Table 8.2.1 Myocardial blood flow (ml/min/g) for eight subjects

Subject	MBF		
	Baseline y_1	Caffeine y_2	Difference $d = y_1 - y_2$
1	6.37	4.52	1.85
2	5.69	5.44	0.25
3	5.58	4.70	0.88
4	5.27	3.81	1.46
5	5.11	4.06	1.05
6	4.89	3.22	1.67
7	4.70	2.96	1.74
8	3.53	3.20	0.33
Mean	5.14	3.99	1.15
SD	0.83	0.86	0.63

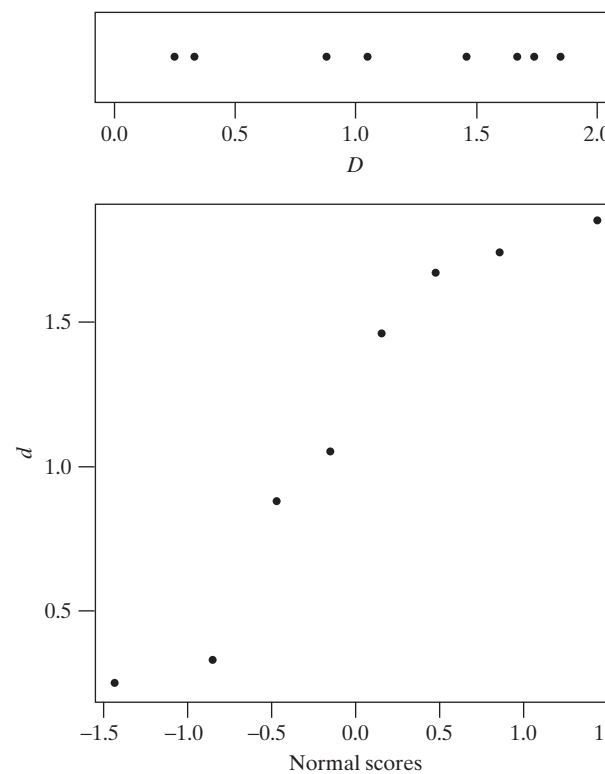

Figure 8.2.1 Dotplot of differences in MBF at baseline and after taking caffeine, along with a normal probability plot of the data

We calculate the standard error of the mean difference as follows:

$$s_D = 0.63$$

$$n_D = 8$$

$$\text{SE}_{\overline{D}} = \frac{0.63}{\sqrt{8}} = 0.22$$

While the mean of the difference is the same as the difference of the means, note that the standard error of the mean difference is *not* the difference of standard errors of the means. ■

Confidence Interval and Test of Hypothesis

The standard error described previously is the basis for the **paired-sample *t* method** of analysis, which can take the form of a confidence interval or a test of hypothesis.

A 95% confidence interval for μ_D is constructed as

$$\overline{d} \pm t_{n_D-1,\,0.025}\text{SE}_{\overline{D}}$$

where the constant $t_{n_D-1,\,0.025}$ is determined from Student's t distribution with

$$\text{df} = n_D - 1$$

Intervals with other confidence coefficients (such as 90%, 99%, etc.) are constructed analogously (using $t_{0.05}$, $t_{0.005}$, etc.). The following example illustrates the confidence interval.

Example 8.2.2

Blood Flow For the blood flow data, we have df $= 8 - 1 = 7$. From Table 4 we find that $t_{7,\,0.025} = 2.365$; thus, the 95% confidence interval for μ_D is

$$1.15 \pm (2.365)\left(\frac{0.63}{\sqrt{8}}\right)$$

or

$$1.15 \pm 0.53$$

or

$$(0.62, 1.68)$$ ■

We can also conduct a t test. To test the null hypothesis

$$H_0\colon \mu_D = 0$$

we use the test statistic

$$t_s = \frac{\overline{d} - 0}{\text{SE}_{\overline{D}}}$$

Critical values are obtained from Student's t distribution (Table 4) with df $= n_D - 1$. The following example illustrates the t test.

Example 8.2.3

Blood Flow For the blood flow data, let us formulate the null hypothesis and nondirectional alternative:

H_0: Mean myocardial blood flow is the same at baseline as it is after taking caffeine.

H_A: Mean myocardial blood flow is different after taking caffeine then at baseline.

or, in symbols,

$$H_0: \mu_D = 0$$
$$H_A: \mu_D \neq 0$$

Let us test H_0 against H_A at significance level $\alpha = 0.05$. The test statistic is

$$t_s = \frac{1.15 - 0}{0.63/\sqrt{8}} = 5.16$$

From Table 4, $t_{7,\,0.005} = 3.499$ and $t_{7,\,0.0005} = 5.408$. We reject H_0 and find that there is sufficient evidence ($0.001 < P < 0.01$) to conclude that mean myocardial blood flow is decreased after taking caffeine. (Using a computer gives the P-value as $P = 0.0013$.) (Note that even though there is significant evidence for a decrease in MBF after taking the caffeine, we cannot conclude that caffeine caused the decrease. For example, it may be that blood flow decreased due to the passage of time.) ■

Result of Ignoring Pairing

Suppose that a study is conducted using a paired design, but that the pairing is ignored in the analysis of the data. Such an analysis is not valid because it assumes that the samples are independent when in fact they are not. The incorrect analysis can be misleading, as the following example illustrates.

Example 8.2.4

Hunger Rating During a weight loss study each of nine subjects was given either the active drug *m*-chlorophenylpiperazine (mCPP) for two weeks and then a placebo for another two weeks, or else was given the placebo for the first two weeks and then mCPP for the second two weeks. As part of the study the subjects were asked to rate how hungry they were at the end of each two-week period. The hunger rating data are shown in Table 8.2.2.[2]

Table 8.2.2 Hunger Rating for Nine Women

	Hunger rating		
Subject	Drug (mCPP) y_1	Placebo y_2	Difference $d = y_1 - y_2$
1	79	78	1
2	48	54	−6
3	52	142	−90
4	15	25	−10
5	61	101	−40
6	107	99	8
7	77	94	−17
8	54	107	−53
9	5	64	−59
Mean	55	85	−30
SD	32	34	33

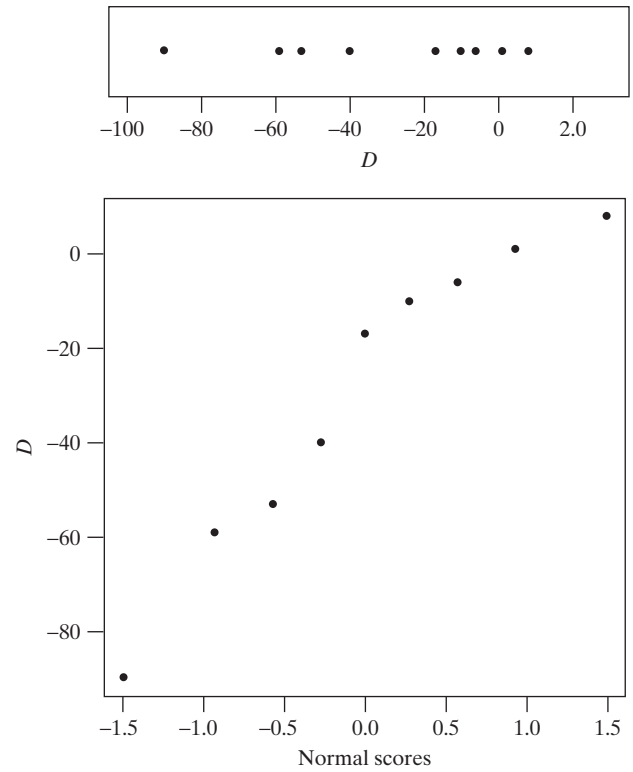

Figure 8.2.2 Dotplot of differences in hunger rating when on the drug and when on placebo, along with a normal probability plot of the data

For the hunger rating data, the SE for the mean difference is

$$\text{SE}_{\overline{D}} = \frac{33}{\sqrt{9}} = 11$$

Figure 8.2.2 shows the distribution of the nine sample differences.

A test of

$$H_0: \mu_D = 0$$

versus

$$H_A: \mu_D \neq 0$$

gives a test statistic of

$$t_s = \frac{-30 - 0}{11} = -2.72$$

This test statistic has 8 degrees of freedom. Using a computer gives the P-value as $P = 0.027$.

Figure 8.2.3 displays the drug and placebo data separately. There is considerable overlap in the two distributions. This plot does not show compelling evidence that the drug lowers hunger ratings (as determined from the paired analysis above) because this plot does not take into account the paired nature of these data.

Looking at the drug and placebo data separately, the two sample SDs are $s_1 = 32$ and $s_2 = 34$. If we proceed improperly as if the samples were independent and apply the SE formula of Chapter 7, we obtain

$$\text{SE}_{(\overline{Y}_1 - \overline{Y}_2)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$= \sqrt{\frac{32^2}{9} + \frac{34^2}{9}} = 15.6$$

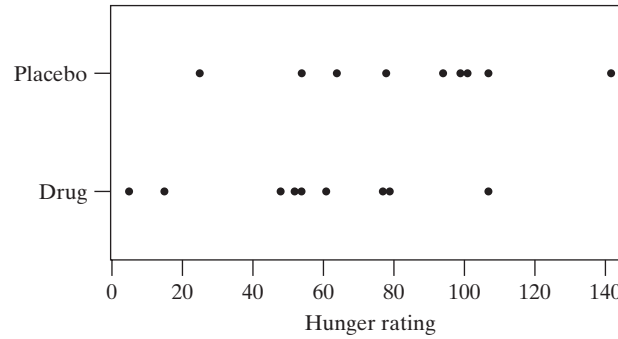

Figure 8.2.3 Parallel dotplots of hunger rating when on the drug and when on placebo

This SE is quite a bit larger than the value ($SE_{\overline{D}} = 11$) that we calculated using the pairing.

Continuing to (wrongly) proceed as if the samples were independent, the test statistic is

$$t_s = \frac{55 - 85}{15.6} = -1.92$$

The P-value for this test is 0.075, which is much greater than the P-value for the correct test, 0.027.

To further compare the paired and unpaired analyses, let us consider the 95% confidence interval for $(\mu_1 - \mu_2)$. For the unpaired analysis, formula (6.7.1) yields $15.9 \approx 16$ degrees of freedom; this gives a t multiplier of $t_{16,\,0.025} = 2.121$ and yields a confidence interval of

$$(55 - 85) \pm (2.121)(15.6)$$

or

$$-30 \pm 33.1$$

or

$$(-63.1, 3.1)$$

This erroneous confidence interval is wider than the correct confidence interval from a paired analysis. A paired analysis yields the narrower interval

$$-30 \pm (2.306)(11)$$

or

$$-30 \pm 25.4$$

or

$$(-55.4, -4.6)$$

The paired-sample interval is narrower because it uses a smaller SE; this effect is slightly offset by a larger value of $t_{0.025}$ (2.306 versus 2.121).

Why is the paired-sample SE smaller than the independent-samples SE calculated from the same data ($SE = 11$ versus $SE = 15.6$)? Table 8.2.2 reveals the reason. The data show that there is large variation from one subject to the next. For instance, subject 4 has low hunger ratings (both when on the drug and when on placebo) and subject 6 has high values. The independent-samples SE formula incorporates all this variation (expressed through s_1 and s_2); in the paired-sample approach, intersubject variation in hunger rating has no influence on the calculations because only the D's are used. By using each subject as her own control, the experimenter has increased the precision of the experiment. But if the pairing is ignored in the analysis, the extra precision is wasted. ■

The preceding example illustrates the gain in precision that can result from a paired design coupled with a paired analysis. The choice between a paired and an unpaired design will be discussed in Section 8.3.

Conditions for Validity of Student's *t* Analysis

The conditions for validity of the paired-sample *t* test and confidence interval are as follows:

1. It must be reasonable to regard the *differences* (the D's) as a random sample from some large population.
2. The population distribution of the D's must be normal. The methods are approximately valid if the population distribution is approximately normal or if the sample size (n_D) is large.

The preceding conditions are the same as those given in Chapter 6; in the present case, the conditions apply to the D's because the analysis is based on the D's. Verification of the conditions can proceed as described in Chapter 6. First, the design should be checked to assure that the D's are independent of each other, and especially that there is no hierarchical structure within the D's. (Note, however, that the Y_1's are not independent of the Y_2's because of the pairing.) Second, a histogram or dotplot of the D's can provide a rough check for approximate normality. A normal probability plot can also be used to assess normality.

Notice that normality of the Y_1's and Y_2's is not required, because the analysis depends only on the D's. The following example shows a case in which the Y_1's and Y_2's are not normally distributed, but the D's are.

Example 8.2.5

Squirrels If you walk toward a squirrel that is on the ground, it will eventually run to the nearest tree for safety. A researcher wondered whether he could get closer to the squirrel than the squirrel was to the nearest tree before the squirrel would start to run. He made 11 observations, which are given in Table 8.2.3. Figure 8.2.4 shows

Table 8.2.3 Distances (in inches) from person and from tree when squirrel started to run

Squirrel	From person y_1	From tree y_2	Difference $d = y_1 - y_2$
1	81	137	−56
2	178	34	144
3	202	51	151
4	325	50	275
5	238	54	184
6	134	236	−102
7	240	45	195
8	326	293	33
9	60	277	−217
10	119	83	36
11	189	41	148
Mean	190	118	72
SD	89	101	148

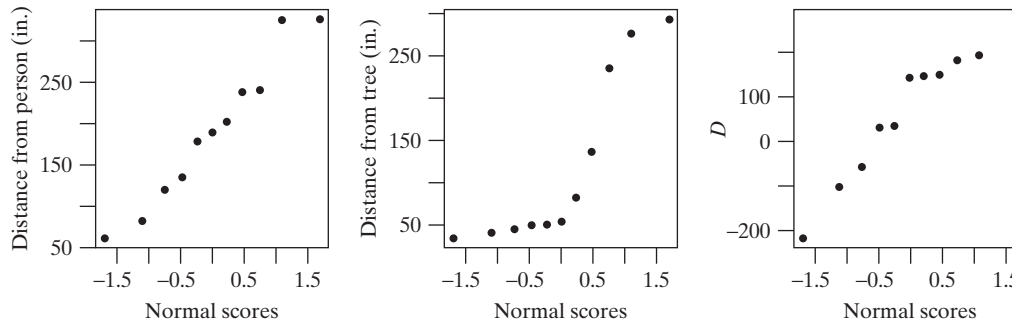

Figure 8.2.4 Normal probability plots of distance from squirrel to person and from squirrel to tree

that the distribution of distances from squirrel to person appear to be reasonably normal, but that the distances from squirrel to tree are far from being normally distributed. However, panel (c) of Figure 8.2.4 shows that the 11 differences do meet the normality condition. Since a paired t test analyzes the differences, a t test (or confidence interval) is valid here.[3] ■

Summary of Formulas

For convenient reference, we summarize the formulas for the paired-sample methods based on Student's t.

Standard Error of $\overline{D}$

$$\text{SE}_{\overline{D}} = \frac{s_D}{\sqrt{n_D}}$$

***t* Test**

$$H_0\colon \mu_D = 0$$

$$t_s = \frac{\overline{d} - 0}{\text{SE}_{\overline{D}}}$$

95% Confidence Interval for μ_d

$$\overline{d} \pm t_{0.025}\text{SE}_{\overline{D}}$$

Intervals with other confidence levels (e.g., 90%, 99%) are constructed analogously (e.g., using $t_{0.05}, t_{0.005}$).

Exercises 8.2.1–8.2.11

8.2.1 In an agronomic field experiment, blocks of land were subdivided into two plots of 346 square feet each. Each block provided two paired observations: one for each of the varieties of wheat. The plot yields (lb) of wheat are given in the table.[4]

	VARIETY		
BLOCK	1	2	DIFFERENCE
1	32.1	34.5	−2.4
2	30.6	32.6	−2.0
3	33.7	34.6	−0.9
4	29.7	31.0	−1.3
Mean	31.52	33.17	−1.65
SD	1.76	1.72	0.68

(a) Calculate the standard error of the mean difference between the varieties.

(b) Test for a difference between the varieties using a paired t test at $\alpha = 0.05$. Use a nondirectional alternative.

(c) Test for a difference between the varieties the wrong way, using an independent-samples test. Compare with the result of part (b).

8.2.2 In an experiment to compare two diets for fattening beef steers, nine pairs of animals were chosen from the herd; members of each pair were matched as closely as possible with respect to hereditary factors. The members of each pair were randomly allocated, one to each diet. The following table shows the weight gains (lb) of the animals over a 140-day test period on diet 1 (Y_1) and on diet 2 (Y_2).[5]

PAIR	DIET 1	DIET 2	DIFFERENCE
1	596	498	98
2	422	460	−38
3	524	468	56
4	454	458	−4
5	538	530	8
6	552	482	70
7	478	528	−50
8	564	598	−34
9	556	456	100
Mean	520.4	497.6	22.9
SD	57.1	47.3	59.3

(a) Calculate the standard error of the mean difference.

(b) Test for a difference between the diets using a paired t test at $\alpha = 0.10$. Use a nondirectional alternative.

(c) Construct a 90% confidence interval for μ_D.

(d) Interpret the confidence interval from part (c) in the context of this setting.

8.2.3 Cyclic adenosine monophosphate (cAMP) is a substance that can mediate cellular response to hormones. In a study of maturation of egg cells in the frog *Xenopus laevis,* oocytes from each of four females were divided into two batches; one batch was exposed to progesterone and the other was not. After two minutes, each batch was assayed for its cAMP content, with the results given in the table.[6] Use a t test to investigate the effect of progesterone on cAMP. Let H_A be nondirectional and let $\alpha = 0.10$.

	cAMP (pmol/oocyte)		
FROG	CONTROL	PROGESTERONE	d
1	6.01	5.23	0.78
2	2.28	1.21	1.07
3	1.51	1.40	0.11
4	2.12	1.38	0.74
Mean	2.98	2.31	0.68
SD	2.05	1.95	0.40

8.2.4 The following table shows the amount of weight loss (kg) for the nine subjects from Example 8.2.4 when taking the drug mCPP and when taking a placebo.[2] (Note that if a subject gained weight, then the recorded weight loss is negative, as is the case for subject 2 who gained 0.3 kg when on the placebo.) Use a t test to investigate the claim that mCPP affects weight loss. Let H_A be nondirectional and let $\alpha = 0.01$.

	WEIGHT CHANGE		
SUBJECT	MCPP	PLACEBO	DIFFERENCE
1	1.1	0.0	1.1
2	1.3	−0.3	1.6
3	1.0	0.6	0.4
4	1.7	0.3	1.4
5	1.4	−0.7	2.1
6	0.1	−0.2	0.3
7	0.5	0.6	−0.1
8	1.6	0.9	0.7
9	−0.5	−2.0	1.5
Mean	0.91	−0.09	1.00
SD	0.74	0.88	0.72

8.2.5 Refer to Exercise 8.2.4.

(a) Construct a 99% confidence interval for μ_D.

(b) Interpret the confidence interval from part (a) in the context of this setting.

8.2.6 Under certain conditions, electrical stimulation of a beef carcass will improve the tenderness of the meat. In one study of this effect, beef carcasses were split in half; one side (half) was subjected to a brief electrical current and the other side was an untreated control. For each side, a steak was cut and tested in various ways for tenderness. In one test, the experimenter obtained a specimen of connective tissue (collagen) from the steak and determined the temperature at which the tissue would shrink; a tender piece of meat tends to yield a low collagen shrinkage temperature. The data are given in the following table.[7]

(a) Construct a 95% confidence interval for the mean difference between the treated side and the control side.

(b) Construct a 95% confidence interval the wrong way, using the independent-samples method. How does this interval differ from the one you obtained in part (a)?

	COLLAGEN SHRINKAGE TEMPERATURE (°C)		
CARCASS	TREATED SIDE	CONTROL SIDE	DIFFERENCE
1	69.50	70.00	−0.50
2	67.00	69.00	−2.00
3	70.75	69.50	1.25
4	68.50	69.25	−0.75
5	66.75	67.75	−1.00
6	68.50	66.50	2.00
7	69.50	68.75	0.75
8	69.00	70.00	−1.00
9	66.75	66.75	0.00
10	69.00	68.50	0.50
11	69.50	69.00	0.50
12	69.00	69.75	−0.75
13	70.50	70.25	0.25
14	68.00	66.25	1.75
15	69.00	68.25	0.75
Mean	68.750	68.633	0.117
SD	1.217	1.302	1.118

8.2.7 Refer to Exercise 8.2.6. Use a *t* test to test the null hypothesis of no effect against the alternative hypothesis that the electrical treatment tends to reduce the collagen shrinkage temperature. Let $\alpha = 0.10$.

8.2.8 Trichotillomania is a psychiatric illness that causes its victims to have an irresistible compulsion to pull their own hair. Two drugs were compared as treatments for trichotillomania in a study involving 13 women. Each woman took clomipramine during one time period and desipramine during another time period in a double-blind experiment. Scores on a trichotillomania-impairment scale, in which high scores indicate greater impairment, were measured on each woman during each time period. The average of the 13 measurements for clomipramine was 6.2; the average of the 13 measurements for desipramine was 4.2.[8] A paired *t* test gave a value of $t_s = 2.47$ and a two-tailed *P*-value of 0.03. Interpret the result of the *t* test. That is, what does the test indicate about clomipramine, desipramine, and hair pulling?

8.2.9 A scientist conducted a study of how often her pet parakeet chirps. She recorded the number of distinct chirps the parakeet made in a 30-minute period, sometimes when the room was silent and sometimes when music was playing. The data are shown in the following table.[9] Construct a 95% confidence interval for the mean increase in chirps (per 30 minutes) when music is playing over when music is not playing.

	CHIRPS IN 30 MINUTES		
DAY	WITH MUSIC	WITHOUT MUSIC	DIFFERENCE
1	12	3	9
2	14	1	13
3	11	2	9
4	13	1	12
5	20	5	15
6	14	3	11
7	10	0	10
8	12	2	10
9	8	6	2
10	13	3	10
11	14	2	12
12	15	4	11
13	12	3	9
14	13	2	11
15	8	0	8
16	18	5	13
17	15	3	12
18	12	2	10
19	17	2	15
20	15	4	11
21	11	3	8
22	22	4	18
23	14	2	12
24	18	4	14
25	15	5	10
26	8	1	7
27	13	2	11
28	16	3	13
Mean	13.7	2.8	10.9
SD	3.4	1.5	3.0

8.2.10 Consider the data in Exercise 8.2.9. There are two outliers among the 28 differences: the smallest value, which is 2, and the largest value, which is 18. Delete these two observations and construct a 95% confidence interval for the mean increase, using the remaining 26 observations. Do the outliers have much of an effect on the confidence interval?

8.2.11 Invent a paired data set, consisting of five pairs of observations, for which $\bar{y}_1$ and $\bar{y}_2$ are not equal, and $\text{SE}_{\bar{Y}_1} > 0$ and $\text{SE}_{\bar{Y}_2} > 0$, but $\text{SE}_{\bar{D}} = 0$.

8.3 The Paired Design

Ideally, in a paired design the members of a pair are relatively similar to each other—that is, more similar to each other than to members of other pairs—with respect to extraneous variables. The advantage of this arrangement is that, when members of a pair are compared, the comparison is free of the extraneous variation that originates in between-pair differences. We will expand on this theme after giving some examples.

Examples of Paired Designs

Paired designs can arise in a variety of ways, including the following:

Experiments in which similar experimental units form pairs

Observational studies of identical twins

Repeated measurements on the same individual at two different times

Pairing by time

Experiments with Pairs of Units Often researchers who wish to compare two treatments will first form pairs of experimental units (pairs of animals, pairs of plots of land, etc.) that are similar (e.g., animals of the same age and sex or plots of land with the same type of soil and exposure to wind, rain, and sun). Then one member of a pair is randomly chosen to receive the first treatment and the other member is given the second treatment. The following is an example.

Example 8.3.1

Fertilizers for Eggplants In a greenhouse experiment to compare two fertilizer treatments for eggplants, individually potted plants are arranged on the greenhouse bench in pairs, such that two plants in the same pair are subject to the same amount of sunlight, the same temperature, and so on. Within each pair, one (randomly chosen) plant will receive treatment 1 and the other will receive treatment 2. ■

Observational Studies As noted in Section 7.4, randomized experiments are preferred over observational studies, due to the many confounding variables that can arise within an observational study. An observational study may tell us that X and Y are *associated*, but only an experiment can address the question of whether X *causes* Y. If no experiment is possible and an observational study must be carried out, then it is preferable (although rarely possible) to study identical twins as the observational units. For example, in a study of the effect of "secondhand smoke" it would be ideal to enroll several sets of nonsmoking twins for which, in each pair, one of the twins lived with a smoker and the other twin did not. Because sets of twins are rarely, if ever, available, **matched-pair designs**, in which two groups are matched with respect to various extraneous variables, are often used.[10] Here is an example.

Example 8.3.2

Smoking and Lung Cancer In a case-control study of lung cancer, 100 lung cancer patients were identified. For each case, a control was chosen who was individually matched to the case with respect to age, sex, and education level. The smoking habits of the cases and controls were compared. ■

Repeated Measurements Many biological investigations involve repeated measurements made on the same individual at different times. These include studies of growth and development, studies of biological processes, and studies in which measurements are made before and after application of a certain treatment. When only two times are involved, the measurements are paired, as in Example 8.1.1. The following is another example.

Example 8.3.3

Exercise and Serum Triglycerides Triglycerides are blood constituents that are thought to play a role in coronary artery disease. To see whether regular exercise could reduce triglyceride levels, researchers measured the concentration of triglycerides in the blood serum of seven male volunteers, before and after participation in a 10-week exercise program. The results are shown in Table 8.3.1.[11] Note that there is considerable variation from one participant to another. For instance, participant 1 had relatively low triglyceride levels both before and after, while participant 3 had relatively high levels. ■

Table 8.3.1 Serum triglycerides (mmol/L)

Participant	Before	After
1	0.87	0.57
2	1.13	1.03
3	3.14	1.47
4	2.14	1.43
5	2.98	1.20
6	1.18	1.09
7	1.60	1.51

Pairing by Time In some situations, pairs are formed implicitly when replicate measurements are made at different times. The following is an example.

Example 8.3.4

Growth of Viruses In a series of experiments on a certain virus (mengovirus), a microbiologist measured the growth of two strains of the virus—a mutant strain and a nonmutant strain—on mouse cells in petri dishes. Replicate experiments were run on 19 different days. The data are shown in Table 8.3.2. Each number represents the total growth in 24 hours of the viruses in a single dish.[12]

Note that there is considerable variation from one run to another. For instance, run 1 gave relatively large values (160 and 97), whereas run 2 gave relatively small values (36 and 55). This variation between runs arises from unavoidable small variations in the experimental conditions. For instance, both the growth of the viruses and the measurement technique are highly sensitive to environmental conditions such as the temperature and CO_2 concentration in the incubator. Slight fluctuations in the environmental conditions cannot be prevented, and these fluctuations cause the variation that is reflected in the data. In this kind of situation the advantage of running the two strains concurrently (that is, in pairs) is particularly striking. ■

Examples 8.3.3 and 8.3.4 both involve measurements at different times. But notice that the pairing structure in the two examples is entirely different. In Example 8.3.3 the members of a pair are measurements on the same individual at two times, whereas in Example 8.3.4 the members of a pair are measurements on

Table 8.3.2 Virus growth at twenty-four hours

Run	Nonmutant strain	Mutant strain	Run	Nonmutant strain	Mutant strain
1	160	97	11	61	15
2	36	55	12	14	10
3	82	31	13	140	150
4	100	95	14	68	44
5	140	80	15	110	31
6	73	110	16	37	14
7	110	100	17	95	57
8	180	100	18	64	70
9	62	6	19	58	45
10	43	7			

two petri dishes at the same time. Nevertheless, in both examples the principle of pairing is the same: Members of a pair are similar to each other with respect to extraneous variables. In Example 8.3.4 time is an extraneous variable, whereas in Example 8.3.3 the comparison between two times (before and after) is of primary interest and interperson variation is extraneous.

Purposes of Pairing

Pairing in an experimental design can serve to reduce bias, to increase precision, or both. Usually the primary purpose of pairing is to increase precision.

We noted in Section 7.4 that pairing or matching can reduce bias by controlling variation due to extraneous variables. The variables used in the matching are necessarily balanced in the two groups to be compared and therefore cannot distort the comparison. For instance, if two groups are composed of age-matched pairs of people, then a comparison between the two groups is free of any bias due to a difference in age distribution.

In randomized experiments, where bias can be controlled by randomized allocation, a major reason for pairing is to increase precision. Effective pairing increases precision by increasing the information available in an experiment. An appropriate analysis, which extracts this extra information, leads to more powerful tests and narrower confidence intervals. Thus, an effectively paired experiment is more efficient; it yields more information than an unpaired experiment with the same number of observations.

We saw an instance of effective pairing in the hunger rating data of Example 8.2.4. The pairing was effective because much of the variation in the measurements was due to variation between subjects, which did not enter the comparison between the treatments. As a result, the experiment yielded more precise information about the treatment difference than would a comparable unpaired experiment—that is, an experiment that would compare hunger ratings of nine women given mCPP to hunger ratings of nine different control women who were given the placebo.

The effectiveness of a given pairing can be displayed visually in a scatterplot of Y_2 against Y_1; each point in the scatterplot represents a single pair (Y_1, Y_2). Figure 8.3.1 shows a scatterplot for the virus growth data of Example 8.3.4, together with a boxplot of the differences; each point in the scatterplot represents a single run. Notice that the points in the scatterplot show a definite upward

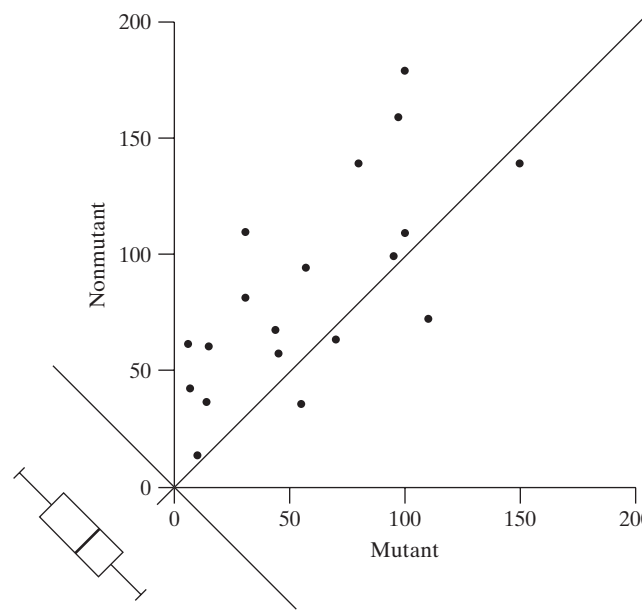

Figure 8.3.1 Scatterplot for the virus growth data, with boxplot of the differences

trend. This upward trend indicates the effectiveness of the pairing: Measurements on the same run (i.e., the same day) have more in common than measurements on different runs, so that a run with a relatively high value of Y_1 tends to have a relatively high value of Y_2, and similarly for low values.

Note that pairing is a strategy of *design,* not of analysis, and is therefore carried out *before* the Y's are observed. It is not correct to use the observations themselves to form pairs. Such a data manipulation could severely distort the experimental results and could be considered scientific fraud.

Randomized Pairs Design versus Completely Randomized Design

In planning a randomized experiment, the experimenter may need to decide between a paired design and a design that uses random assignment without any pairing, called a completely randomized design. We have said that effective pairing can greatly enhance the precision of an experiment. On the other hand, pairing in an experiment may not be effective, if the observed variable Y is not related to the factors used in the pairing. For instance, suppose pairs were matched on age only, but in fact Y turned out not to be age related. It can be shown that ineffective pairing can actually yield less precision than no pairing at all. For instance, in relation to a t test, ineffective pairing would not tend to reduce the SE, but it would reduce the degrees of freedom, and the net result would be a loss of power.

The choice of whether to use a paired design depends on practical considerations (pairing may be expensive or unwieldy) and on precision considerations. With respect to precision, the choice depends on how effective the pairing is expected to be. The following example illustrates this issue.

Example 8.3.5

Fertilizers for Eggplants A horticulturist is planning a greenhouse experiment with individually potted eggplants. Two fertilizer treatments are to be compared, and the observed variable is to be Y = yield of eggplants (pounds). The experimenter knows that Y is influenced by such factors as light and temperature, which vary somewhat from place to place on the greenhouse bench. The allocation of pots to

positions on the bench could be carried out according to a completely randomized design, or according to a paired design, as in Example 8.3.1. In deciding between these options, the experimenter must use her knowledge of how effective the pairing would be—that is, whether two pots sitting adjacent on the bench would be very much more similar in yield than pots farther apart. If she judges that the pairing would not be very effective, she may opt for the completely randomized design. ■

Note that effective pairing is *not* the same as simply holding experimental conditions constant. Pairing is a way of *organizing* the unavoidable variation that still remains after experimental conditions have been made as constant as possible. The ideal pairing organizes the variation in such a way that the variation within each pair is minimal and the variation between pairs is maximal.

Choice of Analysis

The analysis of data should fit the design of the study. If the design is paired, a paired-sample analysis should be used; if the design is unpaired, an independent-samples analysis (as in Chapter 7) should be used.

Note that the extra information made available by an effectively paired design is *entirely wasted* if an unpaired analysis is used. (We saw an illustration of this in Example 8.2.4.) Thus, the paired design does not increase efficiency unless it is accompanied by a paired analysis.

Exercises 8.3.1–8.3.4

8.3.1 (Sampling exercise) This exercise illustrates the application of a matched-pairs design to the population of 100 ellipses (shown with Exercise 3.1.1). The accompanying table shows a grouping of the 100 ellipses into 50 pairs.

PAIR	ELLIPSE ID NUMBERS		PAIR	ELLIPSE ID NUMBERS		PAIR	ELLIPSE ID NUMBERS	
01	20	45	18	11	46	35	16	66
02	03	49	19	09	29	36	18	58
03	07	27	20	19	39	37	30	50
04	42	82	21	00	10	38	76	86
05	81	91	22	40	55	39	17	83
06	38	72	23	21	56	40	04	52
07	60	70	24	08	62	41	12	64
08	31	61	25	24	78	42	23	57
09	77	89	26	67	93	43	98	99
10	01	41	27	35	80	44	36	96
11	14	48	28	74	88	45	44	84
12	59	87	29	94	97	46	06	51
13	22	68	30	02	28	47	85	90
14	47	79	31	26	71	48	37	63
15	05	95	32	25	65	49	43	69
16	53	73	33	15	75	50	34	54
17	13	33	34	32	92			

To better appreciate this exercise, imagine the following experimental setting. We want to investigate the effect of a certain treatment, T, on the organism *C. ellipticus*. We will observe the variable Y = length. We can measure each individual only once, and so we will compare n treated individuals with n untreated controls. We know that the individuals available for the experiment are of various ages, and we know that age is related to length, so we have formed 50 age-matched pairs, some of which will be used in the experiment. The purpose of the pairing is to increase the power of the experiment by eliminating the random variation due to age. (Of course, the ellipses do not actually have ages, but the pairing shown in the table has been constructed in a way that *simulates* age matching.)

(a) Use random digits (from Table 1 or your calculator) to choose a random sample of five pairs from the list.

(b) For each pair, use random digits (or toss a coin) to randomly allocate one member to treatment (T) and the other to control (C).

(c) Measure the lengths of all 10 ellipses. Then, to simulate a treatment effect, add 6 mm to each length in the T group.

(d) Apply a paired-sample t test to the data. Use a nondirectional alternative and let $\alpha = 0.05$.

(e) Did the analysis of part (d) lead you to a Type II error?

8.3.2 (Continuation of Exercise 8.3.1) Apply an independent-samples t test to your data. Use a nondirectional

alternative and let $\alpha = 0.05$. Does this analysis lead you to a Type II error?

8.3.3 (Sampling exercise) Refer to Exercise 8.3.1. Imagine that a matched-pairs experiment is not practical (perhaps because the ages of the individuals cannot be measured), so we decide to use a completely randomized design to evaluate the treatment T.

(a) Use random digits (from Table 1 or your calculator) to choose a random sample of 10 individuals from the ellipse population (shown with Exercise 3.1.1). From these 10, randomly allocate 5 to T and 5 to C. (Or, equivalently, just randomly select 5 from the population to receive T and 5 to receive C.)

(b) Measure the lengths of all 10 ellipses. Then, to simulate a treatment effect, add 6 mm to each length in the T group.

(c) Apply an independent-samples t test to the data. Use a nondirectional alternative and let $\alpha = 0.05$.

(d) Did the analysis of part (c) lead you to a Type II error?

8.3.4 Refer to each of the following exercises. Construct a scatterplot of the data. Does the appearance of the scatterplot indicate that the pairing was effective?

(a) Exercise 8.2.1

(b) Exercise 8.2.2

(c) Exercise 8.2.6

8.4 The Sign Test

The **sign test** is a nonparametric test that can be used to compare two paired samples. It is not particularly powerful, but it is very flexible in application and is especially simple to use and understand—a blunt but handy tool.

Method

Like the paired-sample t test, the sign test is based on the differences

$$D = Y_1 - Y_2$$

The only information used by the sign test is the *sign* (positive or negative) of each difference. If the differences are preponderantly of one sign, this is taken as evidence for the alternative hypothesis. The following examples illustrate the sign test.

Example 8.4.1

Skin Grafts Skin from cadavers can be used to provide temporary skin grafts for severely burned patients. The longer such a graft survives before its inevitable rejection by the immune system, the more the patient benefits. A medical team investigated the usefulness of matching graft to patient with respect to the HL-A (Human Leukocyte Antigen) antigen system. Each patient received two grafts, one with close HL-A compatibility and the other with poor compatibility. The survival times (in days) of the skin grafts are shown in the Table 8.4.1.[13]

Notice that a t test could not be applied here because two of the observations are incomplete; patient 3 died with a graft still surviving and the observation on patient 10 was incomplete for an unspecified reason. Nonetheless, we can proceed with a sign test, since the sign test depends only on the sign of the difference for each patient and we know that $Y_1 - Y_2$ is positive for both of these patients.

Let us carry out a sign test to compare the survival times of the two sets of skin grafts using $\alpha = 0.05$. A directional research (alternative) hypothesis is appropriate for this experiment:

H_A: Skin grafts tend to last longer when the HL-A compatibility is close.

The null hypothesis is

H_0: The survival time distribution is the same for close compatibility as it is for poor compatibility.

Table 8.4.1 Skin graft survival times

Patient	HL-A COMPATIBILITY: Close y_1	HL-A COMPATIBILITY: Poor y_2	Sign of $d = y_1 - y_2$
1	37	29	+
2	19	13	+
3	57+	15	+
4	93	26	+
5	16	11	+
6	23	18	+
7	20	26	−
8	63	43	+
9	29	18	+
10	60+	42	+
11	18	19	−

The first step is to determine the following counts:

$$N_+ = \text{Number of positive differences}$$

$$N_- = \text{Number of negative differences}$$

Because H_A is directional and it predicts that most of the differences will be positive, the test statistic B_s is

$$B_s = N_+$$

For the present data, we have

$$N_+ = 9$$

$$N_- = 2$$

$$B_s = 9$$

The next step is to find the P-value. We use the letter B in labeling the test statistic B_s because the distribution of B_s is based on the binomial distribution. Let p represent the probability that a difference will be positive. If the null hypothesis is true, then $p = 0.5$. Thus, the null distribution of B_s is a binomial with $n = 11$ and $p = 0.5$. That is, the null hypothesis implies that the sign of each difference is like the result of a coin toss, with heads corresponding to a positive difference and tails to a negative difference.

For the skin graft data, the P-value for the test is the probability of getting 9 or more positive differences in 11 patients if $p = 0.5$. This is the probability that a binomial random variable with $n = 11$ and $p = 0.5$ will be greater than or equal to 9. Using the binomial formula from Chapter 3, or a computer, we find that this probability is 0.03272.*

Because the P-value is less than α, we find significant evidence for H_A that skin grafts tend to last longer when the HL-A compatibility is close than when it is poor. ■

*Later in this section we shall learn how to use a table to compute these P-values; however, if you have covered the optional section on the binomial distribution, you can compute this probability using the binomial formula

$$_{11}C_9(0.5)^9(0.5)^2 + {}_{11}C_{10}(0.5)^{10}(0.5)^1 + {}_{11}C_{11}(0.5)^{11} = 0.02686 + 0.00537 + 0.00049 = 0.03272$$

Example 8.4.2

Growth of Viruses Table 8.4.2 shows the virus growth data of Example 8.3.4, together with the signs of the differences.

Table 8.4.2 Virus growth at twenty-four hours

Run	Nonmutant strain y_1	Mutant strain y_2	Sign of $d = y_1 - y_2$	Run	Nonmutant strain y_1	Mutant strain y_2	Sign of $d = y_1 - y_2$
1	160	97	+	11	61	15	+
2	36	55	−	12	14	10	+
3	82	31	+	13	140	150	−
4	100	95	+	14	68	44	+
5	140	80	+	15	110	31	+
6	73	110	−	16	37	14	+
7	110	100	+	17	95	57	+
8	180	100	+	18	64	70	−
9	62	6	+	19	58	45	+
10	43	7	+				

Let's carry out a sign test to compare the growth of the two strains, using $\alpha = 0.10$. The null hypothesis and nondirectional alternative are

H_0: The two strains of virus grow equally well.

H_A: One of the strains grows better than the other.

For these data,

$$N_+ = 15$$

$$N_- = 4$$

When the alternative is nondirectional, B_s is defined as

$$B_s = \text{Larger of } N_+ \text{ and } N_-$$

so for the virus growth data,

$$B_s = 15$$

The P-value for the test is the probability of getting 15 or more successes, plus the probability of getting 4 or fewer successes, in a binomial experiment with $n = 19$. We could use the binomial formula to calculate the P-value. As an alternative, critical values and P-values for the sign test are given in Table 7 (at the end of the book). Using Table 7 with $n_D = 19$, we obtain the critical values and corresponding P-values shown in Table 8.4.3:

Table 8.4.3 Critical values and P-values for the sign test when $n_D = 19$

n_D	0.20	0.10	0.05	0.02	0.01	0.002	0.001
19	**13** *0.167*	**14** *0.064*	**15** *0.019*	**15** *0.019*	**16** *0.004*	**17** *0.0007*	**17** *0.0007*

From the table we see that for $B_s = 15$ the P-value is 0.019, so there is significant evidence for H_A. That is, we reject H_0 and conclude that the data provide significant evidence that the nonmutant strain grows better (at 24 hours) than the mutant strain of the virus. ■

Bracketing the *P*-Value Like the Wilcoxon-Mann-Whitney test, the sign test has a discrete null distribution. Certain critical value entries in Table 7 are blank, for in some cases the most extreme data possible do not lead to a small *P*-value. Table 7 has another peculiarity that is not shared by the Wilcoxon-Mann-Whitney test: Some critical values appear more than once in the same row due to the discreteness of the null distribution.

Directional Alternative To use Table 7 if the alternative hypothesis is directional, we proceed with the familiar two-step procedure:

Step 1. Check directionality (see if the data deviate from H_0 in the direction specified by H_A).

(a) If not, the *P*-value is greater than 0.50.

(b) If so, proceed to step 2.

Step 2. The *P*-value is half what it would be if H_A were nondirectional.

Caution Note that Table 7, for the sign test, and Table 4, for the *t* test, are organized differently: Table 7 is entered with n_D, while Table 4 is entered with (df $= n_D - 1$).

Treatment of Zeros It may happen that some of the differences $(Y_1 - Y_2)$ are equal to zero. Should these be counted as positive or negative in determining B_s? A recommended procedure is to drop the corresponding pairs from the analysis and reduce the sample size n_D accordingly. In other words, each pair whose difference is zero is ignored entirely; such pairs are regarded as providing no evidence against H_0 in either direction. Notice that this procedure has no parallel in the *t* test; the *t* test treats differences of zero the same as any other value.

Example 8.4.3

Null Distribution Consider an experiment with 10 pairs, so that $n_D = 10$. If H_0 is true, then the probability distribution of N_+ is a binomial distribution with $n = 10$ and $p = 0.5$. Figure 8.4.1(a) shows this binomial distribution, together with the associated values of N_+ and N_-. Figure 8.4.1(b) shows the null distribution of B_s, which is a "folded" version of Figure 8.4.1(a). (We saw a similar relationship between parts (a) and (b) of Figure 7.10.4.)

If N_+ is 7 and H_A is directional (and predicts that positive differences are more likely than negative differences), then the *P*-value is the probability of 7 or more (+) signs in 10 trials. Using the binomial formula from Chapter 3, or a computer, we find

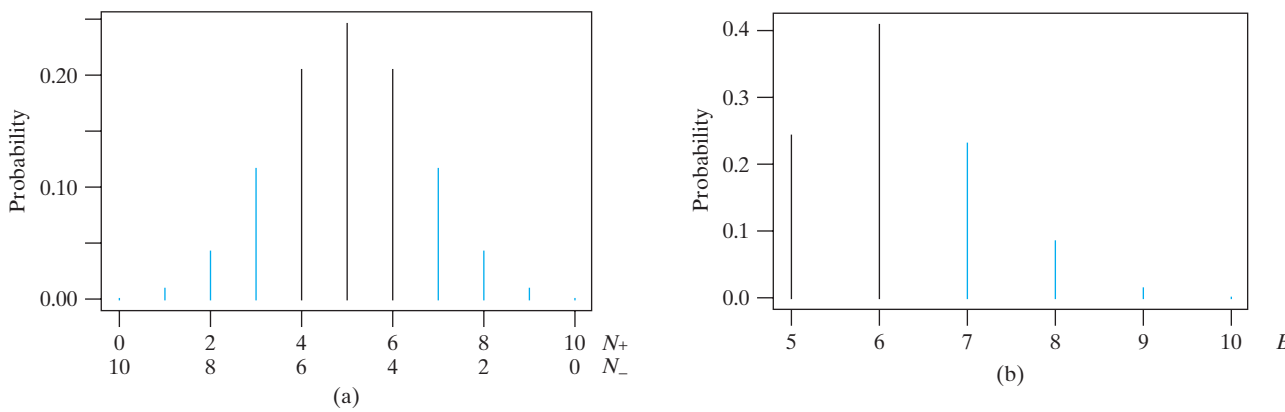

Figure 8.4.1 Null distributions for the sign test when $n_d = 10$. (a) Distribution of N_+ and N_- and (b) distribution of B_s.

that this probability is 0.17188.* This value (0.17188) is the sum of the shaded bars in the right-hand tail in Figure 8.4.1(a). If H_A is nondirectional, then the P-value is the sum of the shaded bars in the left-hand tail and of the right-hand tail of Figure 8.4.1(a). The two shaded areas are both equal to 0.17188; consequently, the total shaded area, which is the P-value, is

$$P = 2(0.17188) = 0.34376 \approx 0.34$$

In terms of the null distribution of B_s, the P-value is an upper-tail probability; thus, the sum of the shaded bars in Figure 8.4.1(b) is equal to 0.34. ■

How Table 7 Is Calculated Throughout your study of statistics you are asked to take on faith the critical values given in various tables. Table 7 is an exception; the following example shows how you could (if you wished to) calculate the critical values yourself. Understanding the example will help you to appreciate how the other tables of critical values have been obtained.

Example 8.4.4

Suppose $n_D = 10$. We saw in Example 8.4.3 that

If $B_s = 7$, the P-value of the data is 0.34376.

Similar calculations using the binomial formula show that

If $B_s = 8$, the P-value of the data is 0.10938.
If $B_s = 9$, the P-value of the data is 0.02148.
If $B_s = 10$, the P-value of the data is 0.00195.

For $n_D = 10$, the critical values from Table 7 are reproduced in Table 8.4.4.

Table 8.4.4 Critical values and P-values for the sign test when $n_D = 10$

n_D	0.20	0.10	0.05	0.02	0.01	0.002	0.001
10	**8** *0.109*	**9** *0.021*	**9** *0.021*	**10** *0.002*	**10** *0.002*	**10** *0.0020*	

The smallest value of B_s that gives a P-value less than 0.20 is $B_s = 8$, so this is the entry in the 0.20 column. For $\alpha = 0.10$ or $\alpha = 0.05$, $B_s = 9$ is needed. The most extreme possibility, $B_s = 10$, gives a P-value of 0.00195, which is rounded to 0.0020 in the table. It is not possible to obtain a nondirectional P-value as small as 0.001, so that entry is left blank. ■

Applicability of the Sign Test

The sign test is valid in any situation where the D's are independent of each other and the null hypothesis can be appropriately translated as

$$H_0\text{: Pr}\{D \text{ is positive}\} = 0.5$$

Thus, the sign test is distribution free; its validity does not depend on any conditions about the form of the population distribution of the D's. This broad validity is bought at a price: If the population distribution of the D's is indeed normal, then the sign test is much less powerful than the t test.

*Applying the binomial formula we have

$$\begin{aligned} &{}_{10}C_7(0.5)^7(0.5)^3 + {}_{10}C_8(0.5)^8(0.5)^2 + {}_{10}C_9(0.5)^9(0.5)^1 + {}_{10}C_{10}(0.5)^{10} \\ &= 0.11719 + 0.04394 + 0.00977 + 0.00098 = 0.17188 \end{aligned}$$

The sign test is useful because it can be applied quickly and in a wide variety of settings. In fact, sometimes the sign test can be applied to data that do not permit a t test at all, as was shown in Example 8.4.1. There is another test for paired data, the Wilcoxon signed-ranks test, which is presented in Section 8.5, that is generally more powerful than the sign test and yet is distribution free. However, the Wilcoxon signed-ranks test is more difficult to carry out than the sign test and, like the t test, there are situations in which it cannot be conducted. The following is another example in which only a sign test is possible.

Example 8.4.5

THC and Chemotherapy Chemotherapy for cancer often produces nausea and vomiting. The effectiveness of THC (tetrahydrocannabinol—the active ingredient of marijuana) in preventing these side effects was compared with the standard drug Compazine. Of the 46 patients who tried both drugs (but were not told which was which), 21 expressed no preference, while 20 preferred THC and 5 preferred Compazine. Since "preference" indicates a sign for the difference, but not a magnitude, a t test is impossible in this situation. For a sign test, we have $n_d = 25$ and $B_s = 20$, so that the P-value is 0.004; even at $\alpha = 0.005$ we would reject H_0 and find that the data provide sufficient evidence to conclude that THC is preferred to Compazine.[14] ■

Exercises 8.4.1–8.4.11

8.4.1 Use Table 7 to find the P-value for a sign test (against a nondirectional alternative), assuming that $n_D = 9$ and

(a) $B_s = 6$ (b) $B_s = 7$

(c) $B_s = 8$ (d) $B_s = 9$

8.4.2 Use Table 7 to find the P-value for a sign test (against a nondirectional alternative), assuming that $n_D = 15$ and

(a) $B_s = 10$ (b) $B_s = 11$

(c) $B_s = 12$ (d) $B_s = 13$

(e) $B_s = 14$ (f) $B_s = 15$

8.4.3 A group of 30 postmenopausal women were given oral conjugated estrogen for one month. Plasma levels of plasminogen-activator inhibitor type 1 (PAI-1) went down for 22 of the women, but went up for 8 women.[15] Use a sign test to test the null hypothesis that oral conjugated estrogen has no effect on PAI-1 level. Use $\alpha = 0.10$ and use a nondirectional alternative.

8.4.4 Can mental exercise build "mental muscle"? In one study of this question, 12 littermate pairs of young male rats were used; one member of each pair, chosen at random, was raised in an "enriched" environment with toys and companions, while its littermate was raised alone in an "impoverished" environment. After 80 days, the animals were sacrificed and their brains were dissected by a researcher who did not know which treatment each rat had received. One variable of interest was the weight of the cerebral cortex, expressed relative to total brain weight. For 10 of the 12 pairs, the relative cortex weight was greater for the "enriched" rat than for his "impoverished" littermate; in the other 2 pairs, the "impoverished" rat had the larger cortex. Use a sign test to compare the environments at $\alpha = 0.05$; let the alternative hypothesis be that environmental enrichment tends to increase the relative size of the cortex.[16]

8.4.5 Twenty institutionalized epileptic patients participated in a study of a new anticonvulsant drug, valproate. Ten of the patients (chosen at random) were started on daily valproate and the remaining 10 received an identical placebo pill. During an eight-week observation period, the numbers of major and minor epileptic seizures were counted for each patient. After this, all patients were "crossed over" to the other treatment, and seizure counts were made during a second eight-week observation period. The numbers of minor seizures are given in the accompanying table.[17] Test for efficacy of valproate using the sign test at $\alpha = 0.05$. Use a directional alternative. (Note that this analysis ignores the possible effect of time—that is, first versus second observation period.)

PATIENT NUMBER	PLACEBO PERIOD	VALPROATE PERIOD	PATIENT NUMBER	PLACEBO PERIOD	VALPROATE PERIOD
1	37	5	11	7	8
2	52	22	12	9	8
3	63	41	13	65	30
4	2	4	14	52	22
5	25	32	15	6	11
6	29	20	16	17	1
7	15	10	17	54	31
8	52	25	18	27	15
9	19	17	19	36	13
10	12	14	20	5	5

8.4.6 An ecological researcher studied the interaction between birds of two subspecies, the Carolina Junco and the Northern Junco. He placed a Carolina male and a Northern male, matched by size, together in an aviary and observed their behavior for 45 minutes beginning at dawn. This was repeated on different days with different pairs of birds. The table shows counts of the episodes in which one bird displayed dominance over the other—for instance, by chasing it or displacing it from its perch.[18] Use a sign test to compare the subspecies. Use a nondirectional alternative and let $\alpha = 0.01$.

	NUMBER OF EPISODES IN WHICH	
PAIR	NORTHERN WAS DOMINANT	CAROLINA WAS DOMINANT
1	0	9
2	0	6
3	0	22
4	2	16
5	0	17
6	2	33
7	1	24
8	0	40

8.4.7

(a) Suppose a paired data set has $n_D = 4$ and $B_s = 4$. Calculate the exact P-value of the data as analyzed by the sign test (against a nondirectional alternative).

(b) Explain why, in Table 7 with $n_D = 3$, no critical values are given in any column.

8.4.8 Suppose a paired data set has $n_D = 15$. Calculate the exact P-value of the data as analyzed by the sign test (against a nondirectional alternative) if $B_s = 15$.

8.4.9 The study described in Example 8.2.4, involving the compound mCPP, included a group of men. The men were asked to rate how hungry they were at the end of each two-week period and differences were computed (hunger rating when taking mCPP–hunger rating when taking the placebo). The distribution of the differences was not normal. Nonetheless, a sign can be conducted using the following information: Out of eight men who recorded hunger ratings, three reported greater hunger on mCPP than on the placebo and five reported lower hunger on mCPP than on the placebo.[2] Conduct a sign test at the $\alpha = 0.10$ level; use a nondirectional alternative.

8.4.10 Refer to Exercise 8.4.9. Calculate the exact P-value of the data as analyzed by the sign test. (*Note:* H_A is nondirectional.)

8.4.11 (Power) A researcher is planning to conduct an experiment to compare two treatments in which matched pairs of subjects will be given the treatments and a sign test will be used, with a nondirectional alternative, to analyze the difference in responses.

Suppose the researcher believes that one treatment will always do better than the other. How many pairs does he need to have in the experiment if he wants to be able to reject H_0 when $\alpha = 0.05$? If one treatment "wins" in every pair, what will be the P-value from the resulting test?

8.5 The Wilcoxon Signed-Rank Test

The **Wilcoxon signed-rank test**, like the sign test, is a nonparametric method that can be used to compare paired samples. Conducting a Wilcoxon signed-rank test is somewhat more complicated than conducting a sign test, but the Wilcoxon test is more powerful than the sign test. Like the sign test, the Wilcoxon signed-rank test does *not* require that the data be a sample from a normally distributed population.

The Wilcoxon signed-rank test is based on the set of differences, $D = Y_1 - Y_2$. It combines the main idea of the sign test—"look at the signs of the differences"—with the main idea of the paired t test—"look at the magnitudes of the differences."

Method

The Wilcoxon signed-rank test proceeds in several steps, which we present here in the context of an example.

Example 8.5.1

Nerve Cell Density For each of nine horses, a veterinary anatomist measured the density of nerve cells at specified sites in the intestine. The results for site I (mid-region of jejunum) and site II (mesenteric region of jejunum) are given in the accompanying table.[19] Each density value is the average of counts of nerve cells in five equal sections of tissue. The null hypothesis of interest is that in the population of all horses there is no difference between the two sites.

1. The first step in the Wilcoxon signed-rank test is to calculate the differences, as shown in Table 8.5.1.

Table 8.5.1 Nerve cell density at each of two sites

Animal	Site I	Site II	Difference
1	50.6	38.0	12.6
2	39.2	18.6	20.6
3	35.2	23.2	12.0
4	17.0	19.0	−2.0
5	11.2	6.6	4.6
6	14.2	16.4	−2.2
7	24.2	14.4	9.8
8	37.4	37.6	−0.2
9	35.2	24.4	10.8

2. Next we find the absolute value of each difference.
3. We then rank these absolute values, from smallest to largest, as shown in Table 8.5.2.

Table 8.5.2

Animal	Difference, d	$\lvert d\rvert$	Rank of $\lvert d\rvert$
1	12.6	12.6	8
2	20.6	20.6	9
3	12.0	12.0	7
4	−2.0	2.0	2
5	4.6	4.6	4
6	−2.2	2.2	3
7	9.8	9.8	5
8	−0.2	0.2	1
9	10.8	10.8	6

4. Next we restore the + and − signs to the ranks of the absolute differences to produce signed ranks, as shown in Table 8.5.3.

Table 8.5.3

Animal	Difference, d	Rank of $\lvert d \rvert$	Signed rank
1	12.6	8	8
2	20.6	9	9
3	12.0	7	7
4	−2.0	2	−2
5	4.6	4	4
6	−2.2	3	−3
7	9.8	5	5
8	−0.2	1	−1
9	10.8	6	6

5. We sum the positive signed ranks to get W_+; we sum the absolute values of the negative signed ranks to get W_-. For the nerve cell data, $W_+ = 8 + 9 + 7 + 4 + 5 + 6 = 39$ and $W_- = 2 + 3 + 1 = 6$. The test statistic, W_s is defined as

$$W_s = \text{Larger of } W_+ \text{ and } W_-$$

   For the nerve cell data, $W_s = 39$.

6. To find the P-value, we consult Table 8 (at the end of the book). Part of Table 8 is reproduced in Table 8.5.4.

Table 8.5.4 Critical values for the Wilcoxon signed-rank test when $n_D = 9$

n	0.20	0.10	0.05	0.02	0.01	0.002	0.001
9	**35** *0.164*	**37** *0.098*	**40** *0.039*	**42** *0.020*	**44** *0.0078*		

From Table 8.5.4, we see that for $W_s = 37$ the P-value is 0.098. There is weak but suggestive evidence ($P = 0.098$) that there is a difference in nerve cell density in the two regions. (We reject H_0 if α is 0.10 or larger.) ■

Bracketing the *P*-Value Like the sign test, the Wilcoxon signed-rank test has a discrete null distribution. Certain critical value entries in Table 8 are blank; this situation is familiar from our study of the Wilcoxon-Mann-Whitney test and the sign test. For example, if $n_D = 9$, then the strongest possible evidence against H_0 occurs when all 9 differences are positive (or when all 9 differences are negative), in which case $W_s = 45$. But the chance that W_s will equal 45 when H_0 is true is $(1/2)^9 + (1/2)^9$, which is approximately 0.0039. Thus, it is not possible to have a two-tailed P-value smaller than 0.002, let alone 0.001. This is why the last two entries are blank in the $n_D = 9$ row of Table 8. Also note that if $W_s = 34$, for example, then the table only tells us that $P > 0.20$.

Directional Alternative To use Table 8 if the alternative hypothesis is directional, we proceed with the familiar two-step procedure:

Step 1. Check directionality (see if the data deviate from H_0 in the direction specified by H_A).

(a) If not, the P-value is greater than 0.50.

(b) If so, proceed to step 2.

Step 2. The P-value is half what it would be if H_A were nondirectional.

Treatment of Zeros If any of the differences $(Y_1 - Y_2)$ are zero, then those data points are deleted and the sample size is reduced accordingly. For example, if one of the 9 differences in Example 8.5.1 had been zero, we would have deleted that point when conducting the Wilcoxon test, so that the sample size would have become 8.

Treatment of Ties If there are ties among the absolute values of the differences (in step 3) we average the ranks of the tied values. If there are ties, then the P-value given by the Wilcoxon signed-rank test is only approximate.

Applicability of the Wilcoxon Signed-Rank Test

The Wilcoxon signed-rank test can be used in any situation in which the D's are independent of each other and come from a symmetric distribution; the distribution need not be normal.* The null hypothesis of "no treatment effect" or "no difference between populations" can be stated as

$$H_0: \mu_D = 0$$

Sometimes the Wilcoxon signed-rank test can be carried out even with incomplete information. For example, a Wilcoxon test is possible for the skin graft data of Example 8.4.1. It is true that an exact value of d cannot be calculated for two of the patients, but for both of these patients the difference is positive and is larger than either of the negative differences. The data in Table 8.5.5 show that there only are two negative differences. The smaller of these is -1, for patient 11. This is the smallest difference in absolute value, so it has signed rank -1. The only other negative signed rank is for patient 7; all of the other signed ranks are positive. (The rest of this example is left as an exercise.)

Table 8.5.5 Skin graft survival times

	HL-A COMPATIBILITY		
Patient	Close y_1	Poor y_2	$d = y_1 - y_2$
1	37	29	8
2	19	13	6
3	57+	15	42+
4	93	26	67
5	16	11	5
6	23	18	5
7	20	26	−6
8	63	43	20
9	29	18	11
10	60+	42	18+
11	18	19	−1

As with the Wilcoxon-Mann-Whitney test for independent samples, there is a procedure associated with the Wilcoxon signed-rank test that can be used to construct a confidence interval for μ_D. The procedure is beyond the scope of this book.

*Strictly speaking, the distribution must be continuous, which means that the probability of a tie is zero.

In summary, when dealing with paired data we have three inference procedures: the paired t test, the Wilcoxon signed-rank test, and the sign test. The t test requires that the data come from a normally distributed population; if this condition is met then the t test is recommended, as it is more powerful than the Wilcoxon test or sign test. The Wilcoxon test does not require normality but does require that the differences come from a symmetric distribution and that they can be ranked; it has more power than the sign test. The sign test is the least powerful of the three methods, but the most widely applicable, since it only requires that we determine whether each difference is positive or negative.

Exercises 8.5.1–8.5.7

8.5.1 Use Table 8 to find the P-value for a Wilcoxon signed-rank test (against a nondirectional alternative), assuming that $n_D = 7$ and

(a) $W_s = 22$
(b) $W_s = 25$
(c) $W_s = 26$
(d) $W_s = 28$

8.5.2 Use Table 8 to find the P-value for a Wilcoxon signed-rank test (against a nondirectional alternative), assuming that $n_D = 12$ and

(a) $W_s = 55$
(b) $W_s = 65$
(c) $W_s = 71$
(d) $W_s = 73$

8.5.3 The study described in Example 8.2.4, involving the compound mCPP, included a group of nine men. The men were asked to rate how hungry they were at the end of each two-week period and differences were computed (hunger rating when taking mCPP—hunger rating when taking the placebo). Data for one of the subjects are not available; the data for the other eight subjects are given in the accompanying table.[2] Analyze these data with a Wilcoxon signed-rank test at the $\alpha = 0.10$ level; use a nondirectional alternative.

	HUNGER RATING		
SUBJECT	MCPP y_1	PLACEBO y_2	DIFFERENCE $d = y_1 - y_2$
1	64	69	−5
2	119	112	7
3	0	28	−28
4	48	95	−47
5	65	145	−80
6	119	112	7
7	149	141	8
8	NA	NA	NA
9	99	119	−20

8.5.4 As part of the study described in Example 8.2.4 (and in Exercise 8.5.3), involving the compound mCPP, weight change was measured for nine men. For each man two measurements were made: weight change when taking mCPP and weight change when taking the placebo. The data are given in the accompanying table.[2] Analyze these data with a Wilcoxon signed-rank test at the $\alpha = 0.05$ level; use a nondirectional alternative.

	WEIGHT CHANGE		
SUBJECT	MCPP y_1	PLACEBO y_2	DIFFERENCE $d = y_1 - y_2$
1	0.0	−1.1	1.1
2	−1.1	0.5	−1.6
3	−1.6	0.5	−2.1
4	−0.3	0.0	−0.3
5	−1.1	−0.5	−0.6
6	−0.9	1.3	−2.2
7	−0.5	−1.4	0.9
8	0.7	0.0	0.7
9	−1.2	−0.8	−0.4

8.5.5 Consider the skin graft data of Example 8.4.1. Table 8.5.5, at the end of Section 8.5, shows the first steps in conducting a Wilcoxon signed-rank test of the null hypothesis that HL-A compatibility has no effect on graft survival time. Complete this test. Use $\alpha = 0.05$ and use the directional alternative that survival time tends to be greater when compatibility score is close.

8.5.6 In an investigation of possible brain damage due to alcoholism, an X-ray procedure known as a computerized tomography (CT) scan was used to measure brain densities in 11 chronic alcoholics. For each alcoholic, a nonalcoholic control was selected who matched the alcoholic on age, sex, education, and other factors. The brain density measurements on the alcoholics and the matched controls are reported in the accompanying table.[20] Use a Wilcoxon signed-rank test to test the null hypothesis of no difference against the alternative that alcoholism reduces brain density. Let $\alpha = 0.01$.

PAIR	ALCOHOLIC	CONTROL	DIFFERENCE
1	40.1	41.3	−1.2
2	38.5	40.2	−1.7
3	36.9	37.4	−0.5
4	41.4	46.1	−4.7
5	40.6	43.9	−3.3
6	42.3	41.9	0.4
7	37.2	39.9	−2.7
8	38.6	40.4	−1.8
9	38.5	38.6	−0.1
10	38.4	38.1	0.3
11	38.1	39.5	−1.4
Mean	39.14	40.66	−1.52
SD	1.72	2.56	1.58

8.5.7 The study described in Example 8.1.1, on the effect of caffeine on myocardial blood flow, had another component in which 10 subjects had their blood flow measured before and after consuming caffeine, but under different environmental conditions than those for the subjects in Example 8.1.1.[21] For this setting the differences do not follow a normal distribution, so a t test would not be valid. Use a Wilcoxon signed-rank test to test the null hypothesis of no difference against the alternative that caffeine has an effect on myocardial blood flow. Let $\alpha = 0.01$.

SUBJECT	BASELINE	CAFFEINE	DIFFERENCE
1	3.43	2.72	0.71
2	3.08	2.94	0.14
3	3.07	1.76	1.31
4	2.65	2.16	0.49
5	2.49	2	0.49
6	2.33	2.37	−0.04
7	2.31	2.35	−0.04
8	2.24	2.26	−0.02
9	2.17	1.72	0.45
10	1.34	1.22	0.12
Mean	2.51	2.15	0.36
SD	0.59	0.50	0.43

8.6 Perspective

In this section we consider some limitations to the analysis of paired data.

Before–After Studies

Many studies in the life sciences compare measurements before and after some experimental intervention, which can present another limitation. These studies can be difficult to interpret, because the effect of the experimental intervention may be confounded with other changes over time. For example, in Example 8.2.3 we found significant evidence for a decrease in myocardial blood flow after taking caffeine, but we noted that it is possible that blood flow would have decreased with the passage of time even if the subjects had not taken caffeine. One way to protect against this difficulty is to use randomized concurrent controls, as in the following example.

Example 8.6.1

Biofeedback and Blood Pressure A medical research team investigated the effectiveness of a biofeedback training program designed to reduce high blood pressure. Volunteers were randomly allocated to a biofeedback group or a control group. All volunteers received health education literature and a brief lecture. In addition, the biofeedback group received eight weeks of relaxation training, aided by biofeedback, meditation, and breathing exercises. The results for systolic blood pressure, before and after the eight weeks, are shown in Table 8.6.1.[22]

Let us analyze the before–after changes by paired t tests at $\alpha = 0.05$. In the biofeedback group, the mean systolic blood pressure fell by 13.8 mm Hg. To evaluate the statistical significance of this drop, the test statistic is

$$t_s = \frac{13.8}{1.34} = 10.3$$

Table 8.6.1 Results of biofeedback experiment

Group	n	Systolic blood pressure (mm Hg)			
		Before Mean	After Mean	Difference Mean	SE
Biofeedback	99	145.2	131.4	13.8	1.34
Control	93	144.2	140.2	4.0	1.30

which is highly significant (P-value $\ll 0.0001$). However, this result alone does not demonstrate the effectiveness of the biofeedback training; the drop in blood pressure might be partly or entirely due to other factors, such as the health education literature or the special attention received by all the participants. Indeed, a paired t test applied to the control group gives

$$t_s = \frac{4.0}{1.30} = 3.08 \quad 0.001 < P\text{-value} < 0.01$$

Thus, the people who received *no* biofeedback training *also* experienced a statistically significant drop in blood pressure.

To isolate the effect of the biofeedback training, we can compare the experience of the two treatment groups, using an independent-samples t test *on the two samples of differences*. We again choose $\alpha = 0.05$. The difference between the mean changes in the two groups is

$$13.8 - 4.0 = 9.8 \text{ mm Hg}$$

and the standard error of this difference is

$$\sqrt{1.34^2 + 1.30^2} = 1.87$$

Thus, the t statistic is

$$t_s = \frac{9.8}{1.87} = 5.24$$

This test provides strong evidence ($P < 0.0001$) that the biofeedback program is effective. If the experimental design had not included the control group, then this last crucial comparison would not have been possible, and the support for efficacy of biofeedback would have been shaky indeed. ■

In analyzing real data, it is wise to keep in mind that the statistical methods we have been considering address only limited questions.

The paired t test is limited in two ways:

1. It is limited to questions concerning $\overline{D}$.
2. It is limited to questions about *aggregate* differences.

The second limitation is very broad; it applies not only to the methods of this chapter but also to those of Chapter 7 and to many other elementary statistical techniques. We will discuss these two limitations separately.

Limitation of $\overline{D}$

One limitation of the paired t test and confidence interval is simple, but too often overlooked: When some of the D's are positive and some are negative, the magnitude of $\overline{D}$ does not reflect the "typical" magnitude of the D's. The following example shows how misleading $\overline{D}$ can be.

Example 8.6.2 Measuring Serum Cholesterol Suppose a clinical chemist wants to compare two methods of measuring serum cholesterol; she is interested in how closely the two methods agree with each other. She takes blood specimens from 400 patients, splits each specimen in half, and assays one half by method A and the other by method B. Table 8.6.2 shows fictitious data, exaggerated to clarify the issue.

Table 8.6.2 Serum cholesterol (mg/dl)

Specimen	Method A	Method B	$d = A - B$
1	200	234	−34
2	284	272	+12
3	146	153	−7
4	263	250	+13
5	258	232	+26
⋮	⋮	⋮	⋮
400	176	190	−14
Mean	215.2	214.5	0.7
SD	45.6	59.8	18.8

In Table 8.6.2, the sample mean difference is small ($\overline{d} = 0.7$). Furthermore, the data indicate that the population mean difference is small (a 95% confidence interval is -1.1 mg/dl $< \mu_D <$ 2.5 mg/dl). But such discussion of $\overline{D}$ or μ_D does not address the central question, which is: How closely do the methods agree? In fact, Table 8.6.2 indicates that the two methods do not agree well; the individual differences between method A and method B are not small *in magnitude*. The mean $\overline{d}$ is small because the positive and negative differences tend to cancel each other. A graph similar to Figure 8.3.1 would be very helpful in visually determining how well the methods agree. We would examine such a graph to see how closely the points cluster around the $y = x$ line as well as to see the spread in the boxplot of differences. To make a numerical assessment of agreement between the methods we should not focus on the mean difference, $\overline{D}$. It would be far more relevant to analyze the absolute (unsigned) magnitudes of the d's (that is, 34, 12, 7,13, 26, and so on). These magnitudes could be analyzed in various ways: We could average them, we could count how many are "large" (say, more than 10 mg/dl), and so on. ■

Limitation of the Aggregate Viewpoint

Consider a paired experiment in which two treatments, say A and B, are applied to the same person. If we apply a t test, a sign test, or a Wilcoxon signed-rank test, we are viewing the people as an ensemble rather than individually. This is appropriate if we are willing to assume that the difference (if any) between A and B is in a consistent direction for all people—or, at least, that the important features of the difference are preserved even when the people are viewed *en masse*. The following example illustrates the issue.

Example 8.6.3

Treatment of Acne Consider a clinical study to compare two medicated lotions for treating acne. Twenty patients participate. Each patient uses lotion A on one (randomly chosen) side of his face and lotion B on the other side. After three weeks, each side of the face is scored for total improvement.

First, suppose that the A side improves more than the B side in 10 patients, while in the other 10 the B side improves more. According to a sign test, this result is in perfect agreement with the null hypothesis. And yet, two very different interpretations are logically possible:

> *Interpretation* 1: Treatments A and B are in fact completely equivalent; their action is indistinguishable. The observed differences between A and B sides of the face were entirely due to chance variation.
>
> *Interpretation* 2: Treatments A and B are in fact completely different. For some people (about 50% of the population), treatment A is more effective than treatment B, whereas in the remaining half of the population treatment B is more effective. The observed differences between A and B sides of the face were biologically meaningful.*

The same ambiguity of interpretation arises if the results favor one treatment over another. For instance, suppose the A side improved more than the B side in 18 of the 20 cases, while B was favored in 2 patients. This result, which is statistically significant ($P < 0.001$), could again be interpreted in two ways. It could mean that treatment A is in fact superior to B for everybody, but chance variation obscured its superiority in two of the patients; or it could mean that A is superior to B for most people, but for about 10% of the population ($2/10 = 0.10$) B is superior to A. ■

The difficulty illustrated by Example 8.6.3 is not confined to experiments with randomized pairs. In fact, it is particularly clear in another type of paired experiment—the measurement of change over time. Consider, for instance, the blood pressure data of Example 8.6.1. Our discussion of that study hinged on an aggregate measure of blood pressure: the mean. If some patients' pressures rose as a result of biofeedback and others fell, these details were ignored in the analysis based on Student's t; only the average change was analyzed.

The difficulties described previously aren't only confined to human experiments either. Suppose, for instance, that two fertilizers, A and B, are to be compared in an agronomic field experiment using a paired design, with the data to be analyzed by a paired t test. If treatment A is superior to B on acid soils, but B is better than A on alkaline soils, this fact would be obscured in an experiment that included soils of both types.

The issue raised by the preceding examples is a very general one. Simple statistical methods such as the sign test and the t test are designed to evaluate treatment effects *in the aggregate*—that is, *collectively*—for a population of people, or of mice, or of plots of ground. The segregation of differential treatment effects in subpopulations requires more delicate handling, both in design and analysis.

This confinement to the aggregate point of view applies to Chapter 7 (independent samples) even more forcefully than to the present chapter. For instance, if treatment A is given to one group of mice and treatment B to another, it is quite impossible to know how a mouse in group A would have responded *if* it had received treatment B; the only possible comparison is an aggregate one. In Section 7.11 we

*This may seem farfetched, but phenomena of this kind do occur; as an obvious example, consider the response of patients to blood transfusions of type A or type B blood.

stated that the statistical comparison of independent samples depends on an "implicit assumption"; essentially, the assumption is that the phenomenon under study can be adequately perceived from an aggregate viewpoint.

In many, perhaps most, biological investigations the phenomena of interest are reasonably universal, so that this issue of submerging the individual in the aggregate does not cause a serious problem. Nevertheless, one should not lose sight of the fact that aggregation may obscure important individual detail.

Reporting of Data

In communicating experimental results, it is desirable to choose a form of reporting that conveys the extra information provided by pairing. With small samples, a graphical approach can be used, as in Figure 8.1.1, where the line segments gave clear visual evidence that blood flow decreased for each subject.

In published reports of biological research, the crucial information related to pairing is often omitted. For instance, a common practice is to report the means and standard deviations of Y_1 and Y_2 but to omit the standard deviation of the difference, D! This is a serious error. It is best to report some description of D, using either a display like Figure 8.1.1, a histogram of the D's, or at least the standard deviation of the D's.

Exercises 8.6.1–8.6.4

8.6.1 Thirty-three men with high serum cholesterol, all regular coffee drinkers, participated in a study to see whether abstaining from coffee would affect their cholesterol level. Twenty-five of the men (chosen at random) drank no coffee for five weeks, while the remaining 8 men drank coffee as usual. The accompanying table shows the serum cholesterol levels (in mg/dl) at baseline (at the beginning of the study) and the change from baseline after five weeks.[23]

	NO COFFEE ($n = 25$)		USUAL COFFEE ($n = 8$)	
	MEAN	SD	MEAN	SD
Baseline	341	37	331	30
Change from baseline	−35	27	+26	56

For the following t tests use nondirectional alternatives and let $\alpha = 0.05$.

(a) The no-coffee group experienced a 35 mg/dl drop in mean cholesterol level. Use a t test to assess the statistical significance of this drop.

(b) The usual-coffee group experienced a 26 mg/dl rise in mean cholesterol level. Use a t test to assess the statistical significance of this rise.

(c) Use a t test to compare the no-coffee mean change (−35) to the usual-coffee mean change (+26).

8.6.2 Eight young women participated in a study to investigate the relationship between the menstrual cycle and food intake. Dietary information was obtained every day by interview; the study was double-blind in the sense that the participants did not know its purpose and the interviewer did not know the timing of their menstrual cycles. The table shows, for each participant, the average caloric intake for the 10 days preceding and the 10 days following the onset of the menstrual period (these data are for one cycle only). For these data, prepare a display like that of Figure 8.1.1.[24]

	FOOD INTAKE (CAL)	
PARTICIPANT	PREMENSTRUAL	POSTMENSTRUAL
1	2,378	1,706
2	1,393	958
3	1,519	1,194
4	2,414	1,682
5	2,008	1,652
6	2,092	1,260
7	1,710	1,239
8	1,967	1,758

8.6.3 For each of 29 healthy dogs, a veterinarian measured the glucose concentration in the anterior chamber of the left eye and the right eye, with the results shown in the table.[25]

ANIMAL NUMBER	GLUCOSE (mg/dl) RIGHT EYE	GLUCOSE (mg/dl) LEFT EYE	ANIMAL NUMBER	GLUCOSE (mg/dl) RIGHT EYE	GLUCOSE (mg/dl) LEFT EYE
1	79	79	16	80	80
2	81	82	17	78	78
3	87	91	18	112	110
4	85	86	19	89	91
5	87	92	20	87	91
6	73	74	21	71	69
7	72	74	22	92	93
8	70	66	23	91	87
9	67	67	24	102	101
10	69	69	25	116	113
11	77	78	26	84	80
12	77	77	27	78	80
13	84	83	28	94	95
14	83	82	29	100	102
15	74	75			

Using the paired t method, a 95% confidence interval for the mean difference is $-1.1\,\text{mg/dl} < \mu_D < 0.7\,\text{mg/dl}$. Does this result suggest that, for the typical dog in the population, the difference in glucose concentration between the two eyes is less than 1.1 mg/dl? Explain.

8.6.4 Tobramycin is a powerful antibiotic. To minimize its toxic side effects, the dose can be individualized for each patient. Thirty patients participated in a study of the accuracy of this individualized dosing. For each patient, the predicted peak concentration of Tobramycin in the blood serum was calculated, based on the patient's age, sex, weight, and other characteristics. Then Tobramycin was administered and the actual peak concentration (μg/ml) was measured. The results were reported as in the table.[26]

	PREDICTED	ACTUAL
Mean	4.52	4.40
SD	0.90	0.85
n	30	30

Does the reported summary give enough information for you to judge whether the individualized dosing is, on the whole, accurate in its prediction of peak concentration? If so, describe how you would make this judgment. If not, describe what additional information you would need and why.

Supplementary Exercises 8.S.1–8.S.23

8.S.1 A volunteer working at an animal shelter conducted a study of the effect of catnip on cats at the shelter. She recorded the number of "negative interactions" each of 15 cats made in 15-minute periods before and after being given a teaspoon of catnip. The paired measurements were collected on the same day within 30 minutes of one another; the data are given in the accompanying table.[27]

(a) Construct a 95% confidence interval for the difference in mean number of negative interactions.

(b) Construct a 95% confidence interval the wrong way, using the independent-samples method. How does this interval differ from the one obtained in part (a)?

CAT	BEFORE (Y_1)	AFTER (Y_2)	DIFFERENCE
Amelia	0	0	0
Bathsheba	3	6	−3
Boris	3	4	−1
Frank	0	1	−1
Jupiter	0	0	0
Lupine	4	5	−1
Madonna	1	3	−2
Michelangelo	2	1	1
Oregano	3	5	−2
Phantom	5	7	−2
Posh	1	0	1
Sawyer	0	1	−1
Scary	3	5	−2
Slater	0	2	−2
Tucker	2	2	0
Mean	1.8	2.8	−1
SD	1.66	2.37	1.20

8.S.2 Refer to Exercise 8.S.1. Compare the before and after populations using a t test at $\alpha = 0.05$. Use a nondirectional alternative.

8.S.3 Refer to Exercise 8.S.1.
Compare the before and after populations using a sign test at $\alpha = 0.05$. Use a nondirectional alternative.

8.S.4 Refer to Exercise 8.S.1. Construct a scatterplot of the data. Does the appearance of the scatterplot indicate that the pairing was effective? Explain.

8.S.5 As part of a study of the physiology of wheat maturation, an agronomist selected six wheat plants at random from a field plot. For each plant, she measured the moisture content in two batches of seeds: one batch from the "central" portion of the wheat head, and one batch from the "top" portion, with the results shown in the following table.[28] Construct a 90% confidence interval for the mean difference in moisture content of the two regions of the wheat head.

	PERCENT MOISTURE	
PLANT	CENTRAL	TOP
1	62.7	59.7
2	63.6	61.6
3	60.9	58.2
4	63.0	60.5
5	62.7	60.6
6	63.7	60.8

8.S.6 Biologists noticed that some stream fishes are most often found in pools, which are deep, slow-moving parts of the stream, while others prefer riffles, which are shallow, fast-moving regions. To investigate whether these two habitats support equal levels of diversity (i.e., equal numbers of species), they captured fish at 15 locations along a river. At each location, they recorded the number of species captured in a riffle and the number captured in an adjacent pool. The following table contains the data.[29] Construct a 90% confidence interval for the difference in mean diversity between the types of habitats.

LOCATION	POOL	RIFFLE	DIFFERENCE
1	6	3	3
2	6	3	3
3	3	3	0
4	8	4	4
5	5	2	3
6	2	2	0
7	6	2	4
8	7	2	5
9	1	2	−1
10	3	2	1
11	4	3	1
12	5	1	4
13	4	3	1
14	6	2	4
15	4	3	1
Mean	4.7	2.5	2.2
SD	1.91	0.74	1.86

8.S.7 Refer to Exercise 8.S.6. What conditions are necessary for the confidence interval to be valid? Are those conditions satisfied? How do you know?

8.S.8 Refer to Exercise 8.S.6. Compare the habitats using a t test at $\alpha = 0.10$. Use a nondirectional alternative.

8.S.9 Refer to Exercise 8.S.6.

(a) Compare the habitats using a sign test at $\alpha = 0.10$. Use a nondirectional alternative.

(b) Use the binomial formula to calculate the exact P-value for part (a).

8.S.10 Refer to Exercise 8.S.6. Analyze these data using a Wilcoxon signed-rank test.

8.S.11 Refer to the Wilcoxon signed-rank test from Exercise 8.S.10. On what grounds could it be argued that the

P-value found in this test might not be accurate? This is, why might it be argued that the Wilcoxon test *P*-value is not a completely accurate measure of the strength of the evidence against H_0 in this case?

8.S.12 In a study of the effect of caffeine on muscle metabolism, nine male volunteers underwent arm exercise tests on two separate occasions. On one occasion, the volunteer took a placebo capsule an hour before the test; on the other occasion he received a capsule containing pure caffeine. (The time order of the two occasions was randomly determined.) During each exercise test, the subject's respiratory exchange ratio (RER) was measured. The RER is the ratio of carbon dioxide produced to oxygen consumed and is an indicator of whether energy is being obtained from carbohydrates or from fats. The results are presented in the accompanying table.[30] Use a *t* test to assess the effect of caffeine. Use a nondirectional alternative and let $\alpha = 0.05$.

	RER (%)	
SUBJECT	PLACEBO	CAFFEINE
1	105	96
2	119	99
3	92	89
4	97	95
5	96	88
6	101	95
7	94	88
8	95	93
9	98	88

8.S.13 For the data of Exercise 8.S.12, construct a display like that of Figure 8.1.1.

8.S.14 Refer to Exercise 8.S.12. Analyze these data using a sign test.

8.S.15 Certain types of nerve cells have the ability to regenerate a part of the cell that has been amputated. In an early study of this process, measurements were made on the nerves in the spinal cord in rhesus monkeys. Nerves emanating from the left side of the cord were cut, while nerves from the right side were kept intact. During the regeneration process, the content of creatine phosphate (CP) was measured in the left and the right portion of the spinal cord. The following table shows the data for the right (control) side (Y_1), and for the left (regenerating) side (Y_2). The units of measurement are mg CP per 100 gm tissue.[31] Use a *t* test to compare the two sides at $\alpha = 0.05$. Use a nondirectional alternative.

ANIMAL	RIGHT SIDE (CONTROL)	LEFT SIDE (REGENERATING)	DIFFERENCE
1	16.3	11.5	4.8
2	4.8	3.6	1.2
3	10.9	12.5	−1.6
4	14.2	6.3	7.9
5	16.3	15.2	1.1
6	9.9	8.1	1.8
7	29.2	16.6	12.6
8	22.4	13.1	9.3
Mean	15.50	10.86	4.64
SD	7.61	4.49	4.89

8.S.16 Aldosterone is a hormone involved in maintaining fluid balance in the body. In a veterinary study, six dogs with heart failure were treated with the drug Captopril, and plasma concentrations of aldosterone were measured before and after the treatment. The results are given in the following table.[32] Use a sign test at $\alpha = 0.10$, and a nondirectional alternative, to investigate the claim that Captopril affects aldosterone level.

ANIMAL	BEFORE	AFTER	DIFFERENCE
1	749	374	375
2	469	300	169
3	343	146	197
4	314	134	180
5	286	69	217
6	223	20	203
Mean	397.3	173.8	223.5
SD	190.5	136.4	76.1

8.S.17 Refer to Exercise 8.S.16. Analyze these data using a Wilcoxon signed-rank test.

8.S.18 Refer to Exercise 8.S.16. Note that the dogs in this study are not compared to a control group. How does this weaken any inference that might be made about the effectiveness of Captopril?

8.S.19 (Computer exercise) For an investigation of the mechanism of wound healing, a biologist chose a paired design, using the left and right hindlimbs of the salamander *Notophthalmus viridescens.* After amputating each limb, she made a small wound in the skin and then kept the limb for 4 hours in either a solution containing benzamil or a control solution. She theorized that the benzamil would impair the healing. The accompanying table shows the amount of healing, expressed as the area (mm^2) covered with new skin after 4 hours.[33]

ANIMAL	CONTROL LIMB	BENZAMIL LIMB	ANIMAL	CONTROL LIMB	BENZAMIL LIMB
1	0.55	0.14	10	0.42	0.21
2	0.15	0.08	11	0.49	0.11
3	0.00	0.00	12	0.08	0.03
4	0.13	0.13	13	0.32	0.14
5	0.26	0.10	14	0.18	0.37
6	0.07	0.08	15	0.35	0.25
7	0.20	0.11	16	0.03	0.05
8	0.16	0.00	17	0.24	0.16
9	0.03	0.05			

(a) Assess the effect of benzamil using a t test at $\alpha = 0.05$. Let the alternative hypothesis be that the researcher's expectation is correct.

(b) Proceed as in part (a) but use a sign test.

(c) Construct a 95% confidence interval for the mean effect of benzamil.

(d) Construct a scatterplot of the data. Does the appearance of the scatterplot indicate that the pairing was effective? Explain.

8.S.20 (Computer exercise) In a study of hypnotic suggestion, 16 male volunteers were randomly allocated to an experimental group and a control group. Each subject participated in a two-phase experimental session. In the first phase, respiration was measured while the subject was awake and at rest. (These measurements were also described in Exercises 7.5.6 and 7.10.4.) In the second phase, the subject was told to imagine that he was performing muscular work, and respiration was measured again.

For subjects in the experimental group, hypnosis was induced between the first and second phases; thus, the suggestion to imagine muscular work was "hypnotic suggestion" for experimental subjects and "waking suggestion" for control subjects. The accompanying table shows the measurements of total ventilation (liters of air per minute per square meter of body area) for all 16 subjects.[34]

EXPERIMENTAL GROUP			CONTROL GROUP		
SUBJECT	REST	WORK	SUBJECT	REST	WORK
1	5.74	6.24	9	6.21	5.50
2	6.79	9.07	10	4.50	4.64
3	5.32	7.77	11	4.86	4.61
4	7.18	16.46	12	4.78	3.78
5	5.60	6.95	13	4.79	5.41
6	6.06	8.14	14	5.70	5.32
7	6.32	11.72	15	5.41	4.54
8	6.34	8.06	16	6.08	5.98

(a) Use a t test to compare the mean resting values in the two groups. Use a nondirectional alternative and let $\alpha = 0.05$. This is the same as Exercise 7.5.6(a).

(b) Use suitable paired and unpaired t tests to investigate (i) the response of the experimental group to suggestion; (ii) the response of the control group to suggestion; (iii) the difference between the responses of the experimental and control groups. Use directional alternatives (suggestion increases ventilation, and hypnotic suggestion increases it more than waking suggestion) and let $\alpha = 0.05$ for each test.

(c) Repeat the investigations of part (b) using suitable nonparametric tests (sign and Wilcoxon-Mann-Whitney tests).

(d) Use suitable graphs to investigate the reasonableness of the normality condition underlying the t tests of part (b). How does this investigation shed light on the discrepancies between the results of parts (b) and (c)?

8.S.21 Suppose we want to test whether an experimental drug reduces blood pressure more than does a placebo. We are planning to administer the drug or the placebo to some subjects and record how much their blood pressures are reduced. We have 20 subjects available.

(a) We could form 10 matched pairs, where we form a pair by matching subjects, as best we can, on the basis of age and sex, and then randomly assign one subject in each pair to the drug and the other subject in the pair to the placebo. Explain why using a matched pairs design might be a good idea.

(b) Briefly explain why a matched pairs design might *not* be a good idea. That is, how might such a design be inferior to a completely randomized design?

8.S.22 A group of 20 postmenopausal women were given transdermal estradiol for one month. Plasma levels of

plasminogen-activator inhibitor type 1 (PAI-1) went down for 10 of the women and went up for the other 10 women.[35] Use a sign test to test the null hypothesis that transdermal estradiol has no effect on PAI-1 level. Use $\alpha = 0.05$ and use a nondirectional alternative.

8.S.23 Six patients with renal disease underwent plasmapheresis. Urinary protein excretion (grams of protein per gram of creatinine) was measured for each patient before and after plasmapheresis. The data are given in the following table.[36] Use these data to investigate whether or not plasmapheresis affects urinary protein excretion in patients with renal disease. (*Hint*: Graph the data and consider whether a t test is appropriate in the original scale.)

PATIENT	BEFORE	AFTER	DIFFERENCE
1	20.3	0.8	19.5
2	9.3	0.1	9.2
3	7.6	3.0	4.6
4	6.1	0.6	5.5
5	5.8	0.9	4.9
6	4.0	0.2	3.8
Mean	8.9	0.9	7.9
SD	5.9	1.1	6.0