Chapter 9

Paired Samples Tests and Confidence Intervals

Chapter Objectives

- Carry out a paired t test for the difference between two population means.
- Compute and interpret a paired t confidence interval for the difference between two population means.
- Carry out a signed rank test on paired samples for a difference between two population means.
- Carry out a sign test on paired samples for a difference between two population means.
- Decide which test (the t test, signed rank test, or sign test) is more appropriate for a given set of data.

Key Takeaways

- Paired samples arise when a variable is measured on the same occasions at two different locations. They also arise when a variable is measured at the same sample locations for each of two time points.
- Paired samples require specialized hypothesis test and confidence interval procedures.
- The paired t test is a parametric paired samples test the difference between two population means that requires that the differences are a sample from a normal population or the sample size is large. A log transformation can make right skewed data more normal prior to conducting a t test.
- The signed rank test is a nonparametric test for the difference between two population means that doesn't require a normality assumption or large sample size.
- The sign test is a nonparametric test for a population median difference that doesn't require a normality assumption or a large sample size.

9.1 Matched Pairs Study Designs

In Chapter 8, the samples used to carry out hypothesis tests and produce confidence intervals were selected *independently* of each other, meaning that the population units selected for one sample had no relationship to the ones selected for the other sample. It's often advantageous, however, to deliberately select the samples in such a way that each unit in the first sample is ***matched*** with one in the second sample. In such ***matched pairs study designs***, the matching is done according to characteristics that are related to the variable that will be compared across the two samples so that within a matched pair, the two values of that variable can be expected to be more similar than they would be for two unmatched units.

Often times, matched pairs studies involve measuring the variable of interest *twice* on each of several units, once under one condition and a second time under a different condition. In this case the *same* units comprise the two samples, but they're measured under different conditions. This is illustrated in the next example, which also highlights the difference between matched pairs study designs and the independent samples designs of Chapter 8.

Example 9.1: Matched Pairs Studies

Dredging a waterway refers to the removal of sediment from its bottom. Waterways are dredged for several reasons, for example to deepen the passageway to a harbor, maintain water quality, improve water circulation, or reduce flood risk by improving downstream flow.

Dredging can have positive environmental and ecological impacts. It can increase the nutrient supply, produce a more even temperature, salinity, and dissolved oxygen distribution (by improving circulation), improve fish spawning grounds or migratory pathways, and remove contaminants residing in the sediments.

But it can have negative impacts too. It can alter sediment composition in ways that are detrimental to benthic (bottom dwelling) organisms and can destroy their refuge, nutrition, and breeding opportunities. It can also destroy aquatic life that relies on tranquil water.

Suppose we are to conduct an impact assessment study of the effects of dredging on a lagoon's sediment quality, but due to the lack of a suitably similar control lagoon, we can't use a BACI design. Instead we must use a before-after design. We're now faced with two study design options:

Independent Samples Study Design: With this study design, a random sample of n sites on the lagoon bottom would be taken and the sediment quality at each of those sites measured before dredging is carried out. Then, after the dredging, *another* random sample of *n different* sites would be taken, independently of the first sample, and the sediment quality measured at these sites after the lagoon has been dredged.

Matched Pairs Study Design: With this study design, a random sample of n sites on the lagoon bottom would be taken and the sediment quality measured at each of these sites before dredging is carried out. Then, after the dredging, the sediment quality would be measured *again* at these *same* n sites. In this study design, each site produces matched pair of before- and after-dredging sediment quality observations.

In the last example, for the independent samples design, the procedures of Chapter 8 would be appropriate. But for the matched pairs design, they wouldn't because the two samples (before and after dredging) aren't selected independently of each other. For the matched pairs design one of the following procedures, described in this chapter, should be used:

1. The paired t test
2. The sign test for paired samples
3. The signed rank test for paired samples

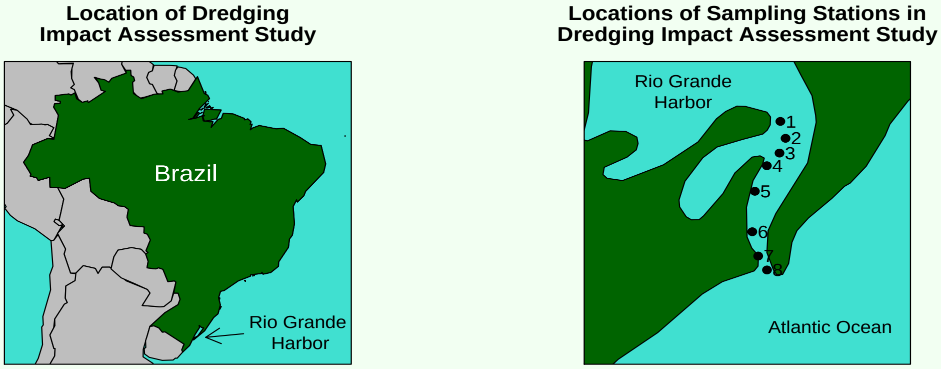
The first is a parametric test, requiring a normality assumption, but the second and third are nonparametric tests, requiring no such assumptions.

Our goal in the lagoon study is to detect an effect of dredging on sediment quality if there is one. Intuitively, it would seem that the matched pairs design would be better able to detect an effect, and it's generally true that matched pairs designs are more sensitive to such effects when they exist. But why? In any lagoon bottom, sediment quality will vary from site to site due to variation in so-called ***extraneous variables*** (variables other than dredging status that affect sediment quality) such as topography, proximity to stream inlets, proximity to human activity, and so on. If the before- and after-dredging sites weren't the same, any observed difference in sediment quality across the two time periods would be a reflection not only of the effect of dredging, but also of the net effect of differences in these extraneous variables for the two sets of sites. But when the before- and after-dredging sites are the same, we can investigate the *site-specific changes* in sediment quality, and because the extraneous variables remain constant at each site, they don't contribute to the observed difference in sediment quality across the two time periods. In this way, the matched pairs study design *controls* for the effects of extraneous variables that vary from site to site, making it easier to detect a dredging effect if there is one.

Hypothesis tests for matched pairs studies are carried out by simply calculating a difference for each pair and then performing a one-sample test using the differences. The next two examples illustrate.

Example 9.2: Pairwise Differences in Matched Pairs Studies

The matched pairs, before-after study design of Example 9.1 was used to assess the impact of dredging on sediment composition in the Patos Lagoon region of the Rio Grande Harbor on the southern coast of Brazil [2]. The percent clay was measured in sediment at $n = 8$ sites in the lagoon the summer before it was dredged and again at the same eight sites the spring after dredging. Fig. 9.1 shows the location of the study region and the eight sample sites within it.


Location of Dredging Impact Assessment Study

Locations of Sampling Stations in Dredging Impact Assessment Study

Figure 9.1: Location of the Rio Grande Harbor, Brazil (left), and the eight sites from which sediment samples were taken before and after dredging (right).

The table below shows the resulting data, including the differences for the eight sites. A negative difference means the clay composition decreased at that site.

Clay Percent

Site	Before Dredging	After Dredging	Difference
1	61.3	53.8	-7.5
2	60.8	38.4	-22.4
3	49.4	54.1	4.7
4	56.2	55.7	-0.5
5	58.6	42.0	-16.6
6	57.1	48.1	-9.0
7	55.4	48.7	-6.7
8	48.3	19.3	-29.0

To decide if there was a statistically significant change in percent clay composition, a one-sample hypothesis test will be carried out on the eight differences. The choice of which test to use will depend on the shape of the distribution of the differences.

Example 9.3: Pairwise Differences in Matched Pairs Studies

The Lake Michigan Mass Balance Study is a U.S. Environmental Protection Agency (EPA) study for monitoring polychlorinated biphenyls (PCBs), mercury, atrazine, and phosphorus in Lake Michigan, and to identify the sediment, air, land, and water pathways by which these pollutants enter the lake. According to the EPA's website:

> Certain toxic substances accumulate or persist in the Great Lakes because, unlike rivers that are constantly flushed with cleaner waters, lakes act as "pollutant sinks". A drop of water entering Lake Michigan today will remain in Lake Michigan for an average of 100 years before it either evaporates or washes into Lake Huron through the Straits of Mackinac. For a particle of soil, the retention time is much, much longer. Unless a pollutant naturally breaks down into harmless components, it persists as a threat to the environment.

As part of the study, sediment cores were sampled at each of four monitoring stations in the lake, the locations of which are shown in Fig. 9.2, and a number of variables were measured at two different depths from the top of the core.

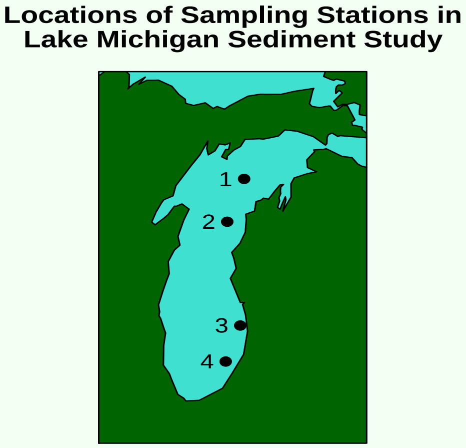

Locations of Sampling Stations in
Lake Michigan Sediment Study
1
2
3
4

Figure 9.2: Locations of sediment sampling stations in the Lake Michigan Mass Balance study.

The table below shows total phosphorus (TP) measurements (mg/g) for the two sediment core depths, 11.5 cm and 33.0 cm. Also shown are the water depths at the stations and the differences between shallow and deep TP concentrations.

Total Phosphorus

	Depth of	11.5 cm	33.0 cm	
Station	Water (m)	Core Depth	Core Depth	Difference
1	172	1.20	0.99	0.21
2	279	1.08	0.80	0.28
3	58	0.75	0.56	0.19
4	52	0.67	0.60	0.07

In Example 9.9, we'll use the four differences to decide if there's statistically significant evidence that the TP concentrations differ for the two sediment depths.

In environmental science, matched pairs studies usually involve pairing observations either by their spatial locations or according to the times at which they were made. The last two examples both used the first approach, with two observations made at each of several locations. In practice, the spatial locations would be selected using one of the random sampling schemes described in Chapter 2. In the next example, pairing is done with respect to time, and two observations are made on each of several occasions. In practice, the set of sampling occasions is often a systematic sample, for example consisting of paired observations made every 5th day.

Regardless of whether pairs correspond to spatial locations or points in time, the central idea is that observations made on matched spatial or temporal units are affected equally by extraneous variables.

Example 9.4: Pairwise Differences in Matched Pairs Studies

In an impact assessment study of the effects of forest clear-cutting on water quality in an adjacent stream, several hydrological variables were measured on each of 11 days upstream and downstream of a newly-completed clear-cutting operation in Southwest Ireland [14]. The table below shows the nitrate concentrations (mg/L) and their differences.

Nitrate Concentration

Date	Upstream	Downstream	Difference
08/15/97	1147.4	995.3	152.1
08/18/97	1412.2	1303.6	108.6
08/31/97	1613.9	1923.3	-309.4
09/18/97	763.3	747.8	15.5
11/04/97	1031.4	1082.9	-51.5
11/07/97	1093.2	1938.7	-845.5
02/27/98	390.8	338.8	52.0
07/14/98	909.8	776.8	133.0
08/25/98	1033.0	676.8	356.2
09/30/98	897.5	1291.0	-393.5
10/29/98	2314.0	1232.9	1081.1

In this study, a pair corresponds to upstream and downstream measurements made on the same day. The idea is that day-to-day fluctuations in extraneous factors that are related to nitrate concentrations, such as rainfall events, temperature shifts, and seasonal streamflow patterns, affect upstream and downstream concentrations equally. Thus these extraneous factors can be *controlled* for by taking measurements in pairs, each pair corresponding to a particular day.

In Problem 9.16, the 11 differences are used to test the hypothesis that the clear-cutting operation had an effect on the nitrate concentrations.

9.2 The Paired t Test

9.2.1 Introduction and Notation

We'll denote observations from a matched pairs study by $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$, where each X_i is paired with the corresponding Y_i. Thus X_1 and Y_1 are a pair, X_2 and Y_2 are a pair, and so on. We'll assume that the X_i's are a random sample from a population whose mean is μ_x and that the Y_i's are a random sample from a population whose mean is μ_y, but now the two samples *aren't* independent. As before, we'll want to test the null hypothesis

$$H_0 : \mu_x - \mu_y = 0$$

that there's no difference between the X and Y population means.

For the paired t test, we'll need some new notation. We'll denote the sample of n pairwise ***differences*** by $\boldsymbol{D_1}, \boldsymbol{D_2}, \ldots, \boldsymbol{D_n}$, that is,

$$\begin{aligned} D_1 &= X_1 - Y_1 \\ D_2 &= X_2 - Y_2 \\ &\vdots \\ D_n &= X_n - Y_n. \end{aligned}$$

These are the values shown in the rightmost columns of data in Examples 9.2, 9.3, and 9.4. We'll think of $D_1, D_2, \ldots, D_n$ as a *single random sample* from a (hypothetical) ***population of differences***, whose mean will be denoted by $\boldsymbol{\mu_d}$, that is

$$\mu_d = \text{The mean of the population of differences.}$$

Example 9.5: Population of Differences

In the Lake Michigan Mass Balance study of Example 9.3, we can think of the entire lake bottom as the population and specific locations as the population units. At each location, there's a value for the *difference* between the TP concentrations at the shallow (11.5 cm) and deep (33.0 cm) sediment depths, and these difference values collectively comprise the *population of differences.* The mean μ_d is the average of the differences over the entire lake bottom, and the *sample of differences* D_1, D_2, D_3, D_4 consists of the four differences shown in the rightmost column of the data table in Example 9.3.

The ***sample mean*** and ***sample mean standard deviation of the differences*** will be denoted by $\boldsymbol{\bar{D}}$ and $\boldsymbol{S_d}$, respectively,

$$\bar{D} = \text{The sample mean of } D_1, D_2, \ldots, D_n$$

and

$$S_d = \text{The sample standard deviation of } D_1, D_2, \ldots, D_n.$$

The next fact says that *the mean of the differences is the difference between the means.*

Fact 9.1 For *any* two data sets $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ (with equal sample sizes),

$$\bar{D} \;=\; \bar{X} - \bar{Y},$$

where $\bar{X}$ and $\bar{Y}$ are the means of the X and Y samples.

This is easy to verify using the properties of summations given in the appendix or by following the next example.

Example 9.6: Mean of the Sample of Differences

For the Lake Michigan Mass Balance study, letting X denote the the shallow (11.5 cm) TP sample and Y the deep (33.0 cm) one, the data from Example 9.3 give

$$\bar{X} \;=\; 0.9250, \qquad \bar{Y} \;=\; 0.7375, \qquad \text{and} \qquad \bar{D} \;=\; 0.1875$$

and so, as guaranteed by Fact 9.1, $\bar{D} = \bar{X} - \bar{Y}$. It's easy to see why by rearranging the data during the computations:

$$\begin{aligned}
\bar{X} - \bar{Y} &= \frac{1}{4}(1.20 + 1.08 + 0.75 + 0.67) - \boldsymbol{\frac{1}{4}}(\mathbf{0.99} + \mathbf{0.80} + \mathbf{0.56} + \mathbf{0.60}) \\
&= \frac{1}{4}[\;(1.20 - \mathbf{0.99}) + (1.08 - \mathbf{0.80}) + (0.75 - \mathbf{0.56}) + (0.67 - \mathbf{0.60})\;] \\
&= \frac{1}{4}(0.21 + 0.28 + 0.19 + 0.07) \\
&= \bar{D}\,.
\end{aligned}$$

(Bold font is used merely to highlight the data rearrangement).

The previous fact has a counterpart for population means.

Fact 9.2 Suppose X and Y are random variables drawn from *any* two populations whose means are μ_x and μ_y. Then the difference $D = X - Y$ can be considered to be a random variable drawn from a population whose mean μ_d is

$$\mu_d \;=\; \mu_x - \mu_y. \tag{9.1}$$

This is actually a direct consequence of Facts 4.2 and 5.3 from Chapters 4 and 5 regarding means of linear functions and sums of random variables.

Example 9.7: Mean of the Population of Differences

In the Lake Michigan Mass Balance study, the X and Y populations are the shallow (11.5 cm) and deep (33.0 cm) sediment layers, respectively. If μ_x is the mean TP concentration over the entire lake bottom for the shallow layer and μ_y is the mean for the deeper layer, then (9.1) says that the mean value of the *difference* in TP concentrations over the lake bottom, μ_d, is the difference between the mean concentrations for the two layers, μ_x and μ_y.

9.2.2 The Paired t Test Procedure

It follows from the Fact 9.2 that any hypothesis about the difference between the X and Y population means μ_x and μ_y can be reformulated as a hypothesis about μ_d:

	Hypothesis **About $\mu_x - \mu_y$**	**Equivalent** **Hypothesis About μ_d**
Null	$H_0 : \mu_x - \mu_y = 0$	$H_0 : \mu_d = 0$
Alternatives	$H_a : \mu_x - \mu_y > 0$	$H_a : \mu_d > 0$
	$H_a : \mu_x - \mu_y < 0$	$H_a : \mu_d < 0$
	$H_a : \mu_x - \mu_y \neq 0$	$H_a : \mu_d \neq 0$

For the ***paired t test***, we're free to choose whichever of these equivalent formulations of the null and alternative hypotheses we want to use. The test is actually just a *one-sample t test* (Chapter 7) for the mean μ_d of the difference population using the sample of differences $D_1, D_2, \ldots, D_n$.

The ***paired t test statistic***, denoted $\boldsymbol{t}$ is defined as follows.

Paired t Test Statistic:

$$t = \frac{\bar{D} - 0}{S_{\bar{D}}} = \frac{\bar{D}}{S_{\bar{D}}}, \tag{9.2}$$

where

$$S_{\bar{D}} = \frac{S_d}{\sqrt{n}}.$$

The sample mean difference $\bar{D}$ is an *estimate* of the true (unknown) population mean difference μ_d, so if the null hypothesis was true, and μ_d equal to zero, we'd expect $\bar{D}$ to be approximately equal to zero too, in which case t would as well. But if the alternative hypothesis was true, and μ_d different from zero, we'd expect $\bar{D}$ to be different from zero in the direction specified by the alternative, in which case t would differ from zero in that same direction. Therefore,

1. *Large positive* values of t provide evidence in favor of $H_a : \mu_d > 0$ (or equivalently $H_a : \mu_x - \mu_y > 0$).
2. *Large negative* values of t provide evidence in favor of $H_a : \mu_d < 0$ (or equivalently $H_a : \mu_x - \mu_y < 0$).
3. *Both large positive and large negative* values of t provide evidence in favor of $H_a : \mu_d \neq 0$ (or equivalently $H_a : \mu_x - \mu_y \neq 0$).

Because its denominator $S_{\bar{D}}$ is the estimated standard error of $\bar{D}$, the observed value of t indicates (approximately) how many standard errors $\bar{D}$ is away from zero, and in what direction (positive or negative).

To decide if an observed t value provides statistically significant evidence in favor of the alternative hypothesis, we'll need to know its sampling distribution under the null. But t is just a one-sample t test statistic using the sample of differences, so from Chapter 7, we have the following.

Sampling Distribution of t Under H_0: Suppose $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ are observations from a matched pairs study, and let $D_1, D_2, \ldots, D_n$ denote the differences. Also let μ_d denote the population mean difference, and suppose either the population of differences is normal or n is large. Then when

$$H_0 : \mu_d \ = \ 0 \qquad \text{(or equivalently } H_0 : \mu_x - \mu_y \ = \ 0\text{)}$$

is true,

$$t \ \sim \ t(n-1).$$

Because values of t that differ from zero in the direction specified by the alternative hypothesis count as evidence in favor of that hypothesis, p-values (and critical values for the rejection region approach) are obtained from the corresponding tail (or tails) of the $t(n-1)$ distribution, as summarized below.

Paired t Test for μ_d

Assumptions: $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ are two random samples that are paired and either the differences $d_1, d_2, \ldots, d_n$ form a single sample from a *normal* population or n is large.

Null hypothesis: $H_0 : \mu_d = 0$.

Test statistic value: $t = \frac{\bar{d}}{s_d/\sqrt{n}}$.

Decision rule: Reject H_0 if p-value $< \alpha$ or t is in rejection region.

Alternative hypothesis	P-value = area under t-distribution with $n-1$ d.f.:	Rejection region = t values such that:*
$H_a : \mu_d > 0$	to the right of t	$t > t_{\alpha,n-1}$
$H_a : \mu_d < 0$	to the left of t	$t < -t_{\alpha,n-1}$
$H_a : \mu_d \neq 0$	to the left of $-\lvert t\rvert$ and right of $\lvert t\rvert$	$t > t_{\alpha/2,n-1}$ or $t < -t_{\alpha/2,n-1}$

* $t_{\alpha,n-1}$ is the $100(1-\alpha)$th percentile of the t distribution with $n-1$ d.f.

9.2.3 Carrying Out the Paired t Test

We'll now look at two examples illustrating the paired t test procedure, the first showing how it's used to decide if dredging impacted the Brazilian lagoon's sediment quality (Example 9.2) and the second showing how it's used to compare phosphorus concentrations at two sediment depths on Lake Michigan's bottom (Example 9.7).

Example 9.8: Paired t Test

For the study of the impact of dredging on the Brazilian lagoon's sediment quality (Examples 9.1 and 9.2), we want to decide if there was any change in the sediment's clay percentage. Thus we'll test the hypotheses

$$\begin{aligned} H_0 : \mu_d \ &= \ 0 \qquad \text{(or equivalently } H_0 : \mu_x - \mu_y = 0\text{)} \\ H_a : \mu_d \ &\neq \ 0 \qquad \text{(or equivalently } H_0 : \mu_x - \mu_y \neq 0\text{)} \end{aligned}$$

where μ_d is the true mean difference in the lagoon's clay percentage (after dredging minus before).

A normal probability plot and a boxplot of the differences (from Example 9.2) are shown below.

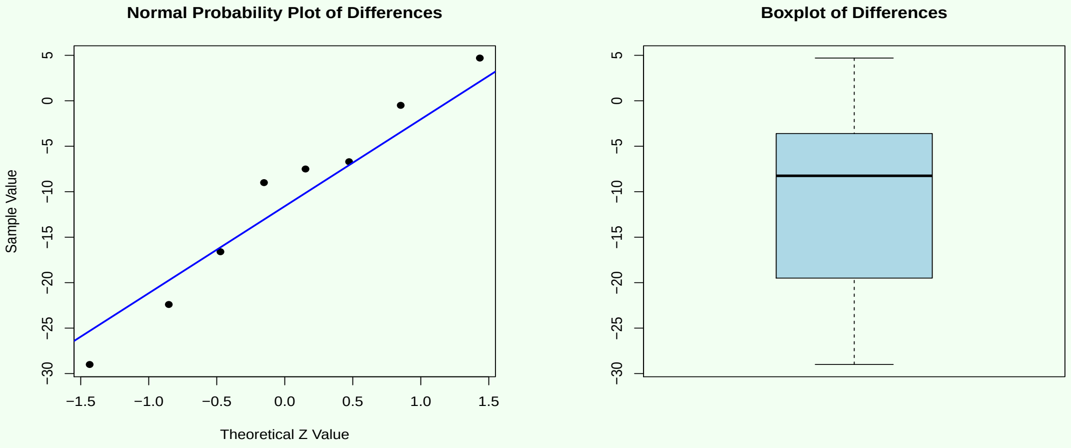


Figure 9.3: Normal probability plot (left) and boxplot (right) of the eight differences in sediment clay percents in the lagoon at the Rio Grande Harbor, Brazil.

The normal probability plot indicates that the normality assumption required for the paired t test appears to be met. The boxplot lies almost entirely below zero, suggesting a decrease in the clay percentage.

The summary statistics for the $n = 8$ differences (clay percentage after dredging minus before) are

$$\begin{aligned} \bar{D} &= -10.9 \\ S_d &= 11.2 \end{aligned}$$

The estimated standard error of $\bar{D}$ is

$$S_{\bar{D}} = \frac{11.2}{\sqrt{8}} = 3.96\,,$$

so the observed test statistic value is

$$\begin{aligned} t &= \frac{-10.9 - 0}{3.96} \\ &= -2.75\,. \end{aligned}$$

Thus the sample mean difference, $\bar{D} = -10.9$, is about 2.75 standard errors below zero.

From the t distribution table, using $n - 1 = 7$ degrees of freedom, the p-value is $2(0.0143) = 0.0286$. Using a level of significance $\alpha = 0.05$, we reject the null hypothesis and conclude that the observed decrease in clay percentage is statistically significant.

Example 9.9: Paired t Test

For the study of total phosphorus (TP) at two sediment depths in Lake Michigan's bottom (Examples 9.3, 9.5, 9.6, and 9.7), we want to decide if there's any difference between shallow (11.5 cm) and deep (33.0 cm) TP concentrations. Thus the hypotheses are

$$\begin{aligned} H_0 : \mu_d &= 0 \qquad \text{(or equivalently } H_0 : \mu_x - \mu_y = 0\text{)} \\ H_a : \mu_d &\neq 0 \qquad \text{(or equivalently } H_0 : \mu_x - \mu_y \neq 0\text{)} \end{aligned}$$

where μ_d is the true mean difference in TP concentrations (shallow minus deep sediment depths).

A normal probability plot of the differences is below.

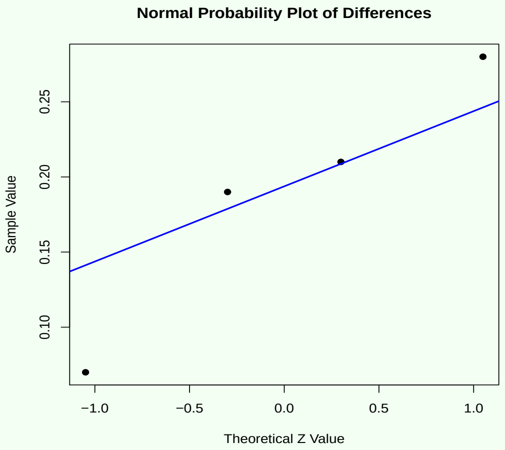


Figure 9.4: Normal probability plot of the differences in total phosphorus concentrations in Lake Michigan sediment.

Although it's difficult to assess normality with such a small sample size, the plot doesn't give any strong indication of non-normality, so we'll proceed with the paired t test.

The summary statistics for the $n = 4$ differences (shallow minus deep) are

$$\begin{aligned} \bar{D} &= 0.19 \\ S_d &= 0.09. \end{aligned}$$

The estimated standard error of $\bar{D}$ is

$$S_{\bar{D}} = \frac{0.09}{\sqrt{4}} = 0.045\,,$$

so the observed test statistic value is

$$\begin{aligned} t &= \frac{0.19 - 0}{0.045} \\ &= 4.22. \end{aligned}$$

From the t distribution table, using $n - 1 = 3$ degrees of freedom, the p-value is $2(0.0122) = 0.0244$, so at the level of significance $\alpha = 0.05$, we reject the null hypothesis. There's statistically significant evidence that the TP concentrations at the shallow sediment depth are *higher* than at the deeper depth.

9.2.4 Some Comments on the Sensitivity of Matched Pairs Designs

As we saw in Example 9.1, when faced with a decision between using an independent samples study design or a matched pairs design, the matched pairs design is usually preferred because it allows us to control for the effects of extraneous variables that would otherwise contribute to random variation between individuals across the two samples.

We'll can compare the test statistics for the two study designs to see directly how the matched pairs design leads to a test that's *more sensitive* to a difference or effect if there is one. For the independent samples design (with equal sample sizes), the two-sample t test statistic (Chapter 8) is

$$t \;=\; \frac{\bar{X}-\bar{Y}}{S_{\bar{X}-\bar{Y}}}\,.$$

For the paired samples design, because $\bar{D}=\bar{X}-\bar{Y}$, we can write the paired t test statistic as

$$t \;=\; \frac{\bar{X}-\bar{Y}}{S_{\bar{D}}}\,.$$

Thus the test statistics have the same numerators but different denominators. The paired t denominator

$$S_{\bar{D}} \;=\; \frac{S_d}{\sqrt{n}}$$

involves S_d, which reflects variation in the pairwise *differences*, but the two-sample t denominator (with equal sample sizes)

$$S_{\bar{X}-\bar{Y}} \;=\; \sqrt{\frac{S_x^2}{n}+\frac{S_y^2}{n}} \;=\; \frac{\sqrt{S_x^2+S_y^2}}{\sqrt{n}}\,.$$

involves S_x and S_y, which reflect variation in the X and Y observations.

If, in a matched pairs study, extraneous variables that contribute to variation in the X and Y observations affect both observations in a pair equally, then their effects will cancel out when we compute the pairwise differences. As a result, those extraneous variables don't contribute to variation in the differences. In this case, the denominator of the paired t statistic will be *smaller* than that of the two-sample t statistic, and so for a given observed difference $\bar{X}-\bar{Y}$, the paired t statistic will tend to be farther away from zero, and result in a smaller p-value, than the two-sample t statistic. The next example illustrates these ideas.

Example 9.10: Power of the Paired t Test

Consider again the study of total phosphorus (TP) in Lake Michigan sediment (Examples 9.3, 9.5, 9.6, 9.7, and 9.9). The left plot below shows the TP concentrations for the two sediment depths (11.5 cm and 33 cm). The right plot shows their differences.

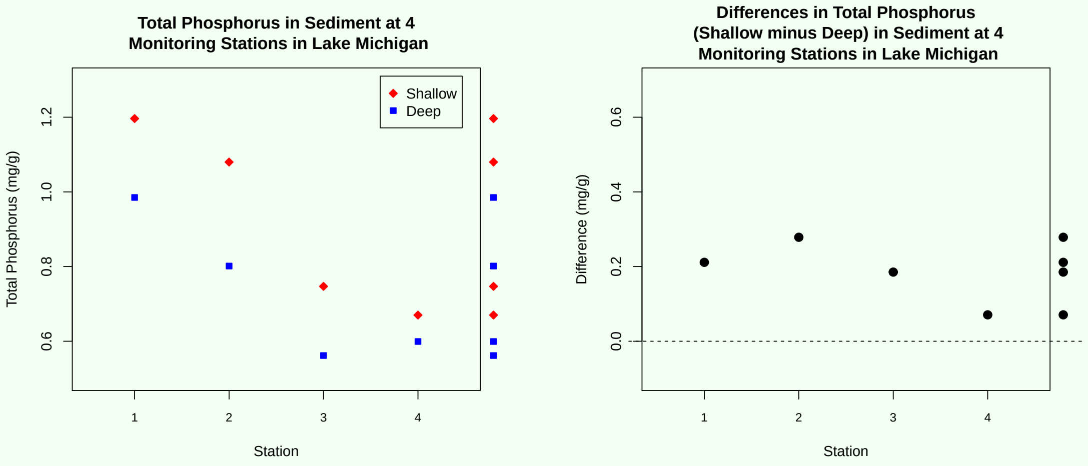


Figure 9.5: TP concentrations at shallow (11.5 cm) and deep (33.0 cm) sediment depths for each of four monitoring stations in Lake Michigan (left); the pairwise differences (right). In both plots, the data have been projected onto the right margin to gauge their variation.

The left plot reveals the substantial station-to-station variation in both the shallow and deep TP observations. Much of this variation is due to extraneous variables that differ from one station to the next, such as proximity to the shore and stream inlets, proximity to human activity such as agriculture, lake depth, and sediment composition. There's much less variation, however, in the differences shown in the right plot because the effects of these extraneous variables cancel out when we compute the differences.

The standard deviation of the differences (from Example 9.9) is

$$S_d \;=\; 0.09$$

which is much smaller than the standard deviations of the shallow and deep TP samples,

$$S_x \;=\; 0.25 \qquad \text{and} \qquad S_y \;=\; 0.20.$$

The result is that the paired t statistic, $t = 4.22$ (from Example 9.9), is substantially farther away from zero than the value we'd get for the two-sample t statistic, $t = 1.16$.

Comment: P-values for the two-sample t and paired t tests are obtained from different t distributions (they have different degrees of freedom), so a direct comparison of the test statistic values isn't entirely valid. But it's usually the case that matched pairs studies lead to smaller p-values.

9.3 Paired t Confidence Intervals

In the impact assessment study of dredging the Rio Grande Harbor (Examples 9.1, 9.2, and 9.8), we may want an *estimate* of *how much* the clay percentage in the sediment decreased. The ***point estimate*** of such an ***effect size*** μ_d (or equivalently $\mu_x - \mu_y$) in a matched pairs study is $\bar{D}$ (or equivalently $\bar{X} - \bar{Y}$). A $\mathbf{100(1-\alpha)\%}$ ***paired t confidence interval*** for μ_d (or equivalently $\mu_x - \mu_y$) attaches a margin of error to the point estimate.

Paired t Confidence Interval for μ_d (or $\mu_x - \mu_y$): Suppose $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ are observations from a matched pairs study and $D_1, D_2, \ldots, D_n$ are the differences. Let μ_x and μ_y denote the X and Y population means and μ_d the population mean difference, and suppose either the population of differences is normal or n is large.

Then a ***100(1 − α)% paired t confidence interval for μ_d (or $\mu_x - \mu_y$)*** is

$$\bar{D} \pm t_{\alpha/2,n-1}\, S_{\bar{D}}\,, \tag{9.3}$$

where

$$S_{\bar{D}} \;=\; \frac{S_d}{\sqrt{n}}\,.$$

We can be $100(1-\alpha)\%$ confident that μ_d (or equivalently $\mu_x - \mu_y$) will be contained in the confidence interval.

The *margin of error* in the paired t confidence interval is

Margin of Error: For the paired t confidence interval (9.3), the margin of error is

$$\text{Margin of Error} \;=\; t_{\alpha/2,n-1}\, S_{\bar{D}} \;=\; t_{\alpha/2,n-1}\, \frac{S_d}{\sqrt{n}},$$

The margin of error measures the degree of precision in the estimate $\bar{D}$ of the true effect size μ_d. A smaller margin of error means a more precise estimate.

Example 9.11: Paired t Confidence Interval

For the impact assessment study of dredging the Rio Grande Harbor (Examples 9.1, 9.2, and 9.8), the summary statistics for the $n = 8$ differences (clay percentage after dredging minus before) are

$$\begin{aligned} \bar{D} &= -10.9 \\ S_d &= 11.2 \end{aligned}$$

so we *estimate* that on average over the lagoon bottom, the clay percentage in the sediment decreased by 10.9 points.

The estimated standard error is

$$S_{\bar{D}} \;=\; \frac{11.2}{\sqrt{8}} \;=\; 3.96\,,$$

so a 95% paired t confidence interval for the true (unknown) mean change in clay percentage, μ_d, is

$$\begin{aligned} \bar{D} \pm t_{\alpha/2,n-1}\, S_{\bar{D}} &= -10.9 \pm 2.36\,(3.96) \\ &= -10.9 \,\pm\, 9.35 \\ &= (-20.25,\; -1.55)\,. \end{aligned}$$

The t critical value $t_{0.025,7} = 2.36$ was obtained from a t distribution table with $n-1 = 7$ degrees of freedom. Thus our *estimate* of the average decrease might be off the mark by up to 9.35 percentage points (the margin of error), and we can be 95% confident that the true mean decrease is between 20.25 and 1.55 points. Notice that this interval lies entirely below zero, which is consistent with the paired t test result of Example 9.8, where $H_0 : \mu_d = 0$ was rejected.

Comment: For a given confidence level, the confidence interval for μ_d (or $\mu_x - \mu_y$) based on a matched pairs study will typically be *narrower* than one based on an independent samples study. The reason is that, as discusses in Subsection 9.2.4, the standard deviation S_d of the differences in a matched pairs study will usually be small relative to the standard deviations S_x and S_y in an independent samples study, so the margin of error for a paired t interval will usually be be smaller than that of a two-sample t interval (even though the t critical values use different degrees of freedom).

9.4 Dealing With Non-Normal Data: Transformations and Nonparametric Procedures

The paired t test rests on an assumption that the sample of differences was drawn from a normal distribution (it's a *parametric* test). If this assumption isn't met (and n isn't large), there are two main courses of action:

1. **Transform the data to normality**: It turns out that if two random variables X and Y both follow normal distributions, their difference $X - Y$ will also follow a normal distribution. Therefore the normality assumption for the differences in a matched pairs study will be met if it's met for both the X and Y samples. If the assumption isn't met for the differences, it's sometimes possible to transform both the X and Y samples, for example by taking their logs or using another transformation in the Ladder of Powers, so that the transformed values in both samples are more normally distributed. Then the paired t test can be carried out using the transformed data.
2. **Carry out a nonparametric test**: We can carry out a *nonparametric* test that doesn't rely on an assumption of normality of the differences. The *signed rank test* and the *sign test for paired samples* described in the next two sections are nonparametric alternatives to the paired t test.

9.5 The Signed Rank Test for Paired Samples

9.5.1 Introduction

The ***signed rank test for paired samples*** (also called the ***Wilcoxon signed rank test***), like the paired t test, is a test for the difference between two population means μ_x and μ_y in a matched pairs study. But unlike it doesn't require the normality assumption for the differences, so it's a *nonparametric* alternative to the paired t test.

9.5.2 The Signed Rank Test Procedure

Consider observations from a matched pairs study, $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$, from two populations whose means are μ_x and μ_y. We'll again denote the differences by $D_1, D_2, \ldots, D_n$ and treat these as a single sample from a population of differences whose mean is μ_d. We assume the difference population is *continuous* and has a ***symmetric*** shape (left and right halves are mirror images), but not necessarily that it's normal.

As for the paired t test, we'll test the null hypothesis

$$H_0 : \mu_d \;=\; 0 \qquad \text{(or equivalently } H_0 : \mu_x - \mu_y = 0)$$

that there's no difference between the population means, versus one of the alternatives

1. $H_a : \mu_d > 0$ (or equivalently $H_0 : \mu_x - \mu_y > 0$) (upper-tailed test)
2. $H_a : \mu_d < 0$ (or equivalently $H_0 : \mu_x - \mu_y < 0$) (lower-tailed test)
3. $H_a : \mu_d \neq 0$ (or equivalently $H_0 : \mu_x - \mu_y \neq 0$) (two-tailed test)

Comment: Because the population of differences is assumed to have a *symmetric* shape, its mean will equal its median. Thus the null hypothesis could be stated in terms of the true (unknown) *median* of the population of differences, $\boldsymbol{\tilde{\mu}_d}$, as

$$H_0 : \tilde{\mu}_d \ = \ 0,$$

In this case, we'd state the alternative hypothesis in terms of the median too.

In the sample, some of the observed differences will be positive and some negative, and each of them will lie a certain distance above or below zero. The ***signed rank test statistic***, denoted $\boldsymbol{W^+}$, test statistic is obtained by taking the absolute values of the differences, sorting and ranking these, and then summing the ranks of the differences that were originally positive.

Signed Rank Test Statistic:

1. If any of the differences $D_1, D_2, \ldots, D_n$ equal zero, discard them prior to computing W^+, and reduce the sample size n by the number of discarded D_i's.
2. Take absolute values of the remaining differences, keeping track of which ones were originally *positive.*
3. *Sort* the absolute differences and *rank* them from smallest to largest. If two or more are tied, assign to each of them the *average* of the ranks they would've been assigned if they hadn't been tied.
4. Sum the *ranks* of the absolute differences that were originally positive. This gives the test statistic:

$$W^+ \ = \ \text{Sum of the ranks of the } |\,D_i\,|\text{'s for which } D_i \text{ is positive.}$$

Example 9.12: Signed Rank Test Statistic

In a study of the particulate-bound dry deposition of atmospheric mercury (Hg) in the state of Indiana, a manual sampling and analysis method was used to measure ground-level atmospheric Hg [12]. The method required holding the air samples for up to 120 h (five days) before analyzing them at a laboratory.

To ensure that the long holding time wouldn't affect the quality of the measurements, a separate quality assurance study was carried out. In this study, air sampling devices were placed in pairs 10 miles from the laboratory. For each pair of sampled air specimens, one was held for 4 hours and the other for 120 hours before being analyzed in the lab.

The table below gives the particulate-bound Hg measurements (pg/m^3) for each of the 10 pairs collected as well as their differences (long holding time minus short).

Particulate-Bound Hg

Air Sample Pair	Long Holding Time	Short Holding Time	Difference
1	4.27	1.33	2.94
2	2.77	3.43	-0.66
3	1.50	1.29	0.21
4	5.70	6.26	-0.56
5	3.80	11.99	-8.19
6	5.64	1.88	3.76
7	3.67	14.58	-10.91
8	0.78	0.54	0.24
9	3.92	3.69	0.23
10	1.85	1.61	0.24

A normal probability plot of the differences is below.

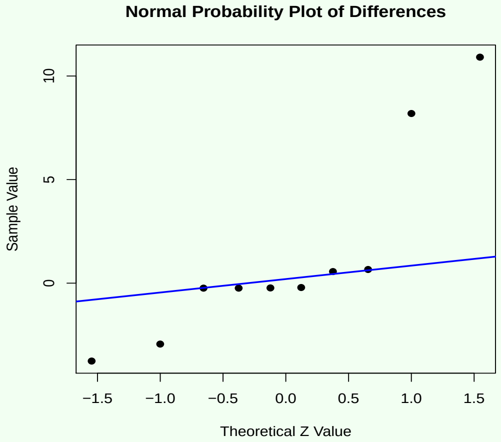


Figure 9.6: Normal probability plot of the differences in Hg measurements for two holding times.

The backward ”S” pattern in the plot suggests a symmetric but ”heavy tailed”, non-normal distribution for the differences, so the researchers carried out a signed rank test.

The hypotheses are

$$\begin{aligned} H_0 : \mu_d &= 0 \qquad \text{(or equivalently } H_0 : \mu_x - \mu_y = 0\text{)} \\ H_a : \mu_d &\neq 0 \qquad \text{(or equivalently } H_0 : \mu_x - \mu_y \neq 0\text{)} \end{aligned}$$

where μ_d is the true (unknown) mean difference in Hg measurements (long holding time minus short) (and μ_x and μ_y are the true means after long and short holding times, respectively).

To compute the test statistic, we first take *absolute values* of the differences, keeping track of which ones were originally positive and which were negative, then we *rank* these from smallest (rank = 1) up to largest (rank = n). If two or more observations are tied, they're each assigned the *average* of the ranks they would've been assigned if they hadn't been tied.

Letting a "+" sign denote a positive difference and a "-" sign a negative one, the sorted and ranked absolute differences are:

Sign	+	+	+	+	-	-	+	+	-	-
|Difference|	**0.21**	**0.23**	**0.24**	**0.24**	0.56	0.66	**2.94**	**3.76**	8.19	10.91
Rank	**1**	**2**	**3.5**	**3.5**	5	6	**7**	**8**	9	10

The test statistic, denoted $\boldsymbol{W^+}$, is the sum of the ranks for the (originally) positive absolute differences,

$$\begin{aligned} W^+ &= 1+2+3.5+3.5+7+8 \\ &= 25. \end{aligned}$$

We'll finish the hypothesis test in Example 9.13.

If the null hypothesis was true, the difference distribution would be centered on zero, and about half of the differences in the sample would be positive and the other half negative. Furthermore, if the difference population was *symmetric*, the absolute values of the positive and negative sample differences would be about the same, and so they'd be "evenly intermingled" when sorted. This can be seen in the bottom of the left plot of Fig. 9.7. It can be shown that in this case, W^+ will approximately equal $\boldsymbol{n(n+1)/4}$.

But if $H_a : \mu_d > 0$ was true, likely more than half of the differences in the sample would be positive, and the positive differences would tend to be larger in absolute value than the negative ones when sorted. This can be seen in the bottom of the right plot of Fig. 9.7. In this case, W^+ would be larger than $n(n+1)/4$.

If $H_a : \mu_d < 0$ was true, fewer than half of the differences in the sample would be positive, and the positive differences would tend to be smaller in absolute value than the negative ones when sorted. In this case, W^+ would be less than $n(n+1)/4$.

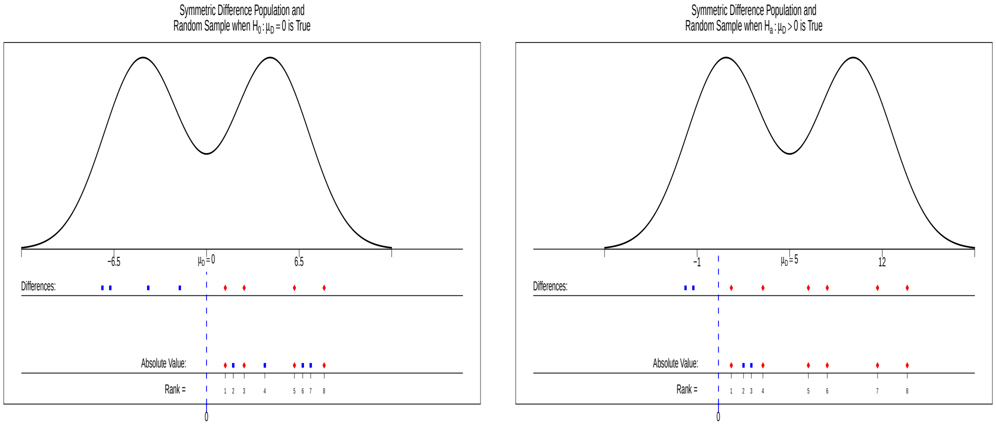


Figure 9.7: A symmetric difference population and random sample from that population. Positive difference values in the sample are depicted as blue squares and negative ones as red diamonds. Their absolute values are shown at the bottom. In the left plot, $H_0 : \mu_d = 0$ is true so the distribution is centered on zero. In the right plot, $H_a : \mu_d > 0$ is true so the distribution is centered to the right of zero.

Comment: The signed rank test statistic reflects both *how many* of the differences are positive as well as *how far* above zero the positive differences are (relative to how far below zero the negative ones are).

1. *Large* values of W^+ (larger than $n(n+1)/4$) provide evidence in favor of $H_a : \mu_d > 0$ (or equivalently $H_a : \mu_x - \mu_y > 0$).
2. *Small* values of W^+ (smaller than $n(n+1)/4$) provide evidence in favor of $H_a : \mu_d < 0$ (or equivalently $H_a : \mu_x - \mu_y < 0$).
3. Both *large and small* values of W^+ (larger or smaller than $n(n+1)/4$) provide evidence in favor of $H_a : \mu_d \neq 0$ (or equivalently $H_a : \mu_x - \mu_y \neq 0$).

To decide if an observed value of W^+ provides statistically significant evidence in support of the alternative hypothesis, we'll need to know its sampling distribution under the null hypothesis.

Sampling Distribution of $\boldsymbol{W^+}$ Under $\boldsymbol{H_0}$: Suppose $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ are *paired* samples from populations whose means are μ_x and μ_y. Consider the differences $D_1, D_2, \ldots, D_n$ to be a single random sample from a population of differences whose mean is μ_d, and suppose that the population of differences is *continuous* and has a *symmetric* shape.

Then when

$$H_0 : \mu_d \;=\; 0 \qquad\qquad \text{(or equivalently } H_0 : \mu_x - \mu_y \;=\; 0\text{)}$$

is true, W_{rs} follows a discrete probability distribution called the ***Wilcoxon signed rank distribution***, which has one parameter, $\boldsymbol{n}$. We write this as

$$W^+ \;\sim\; \text{WilcoxonSR}(n).$$

The *Wilcoxon signed rank distribution* is symmetric and approximately bell-shaped. The distribution is shown below for a sample size $n = 5$.

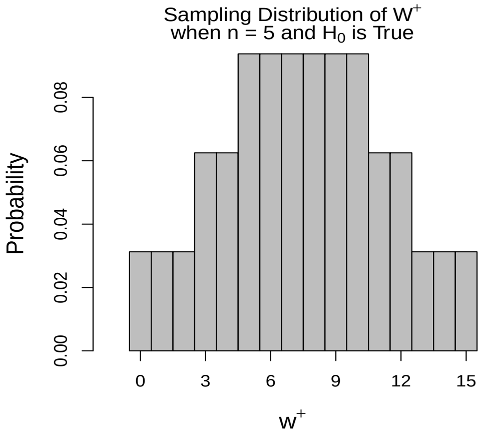


Figure 9.8: Sampling distribution of the signed rank test statistic W^+ when $n = 5$ and H_0 is true.

The mean and standard error (standard deviation) of the Wilcoxon signed rank distribution, denoted $\boldsymbol{\mu_{w^+}}$ and $\boldsymbol{\sigma_{w^+}}$, are

Mean and Standard Error of the Sampling Distribution of $\boldsymbol{W^+}$: The mean $\boldsymbol{\mu_{w^+}}$ and standard error $\boldsymbol{\sigma_{w^+}}$ of the Wilcoxon signed rank distribution are

$$\mu_{w^+} = \frac{n(n+1)}{4}. \tag{9.4}$$

$$\sigma_{w^+} = \sqrt{\frac{n(n+1)(2n+1)}{24}}. \tag{9.5}$$

The mean μ_{w^+} is the value we'd expect to get for W^+, on average, when the differences in a matched pairs study are equally likely to fall a given distance above or below zero (that is, when the null hypothesis is true). Thus if the null hypothesis was true, we'd expect our test statistic to be roughly equal to $n(n+1)/4$. Values of W^+ in the extreme tail of the distribution (in the direction specified by the alternative hypothesis) provide evidence against the null, so the rejection region is comprised of W^+ values in the the extreme $100\alpha\%$ of the distribution, and the p-value is the tail probability outward from the observed W^+ value.

Details about the sampling distribution of W^+ will be given in Subsection 9.5.5. For now, p-values and critical values (for the rejection region approach) will be obtained from a Wilcoxon signed rank distribution table.

The signed rank test procedure for paired samples is summarized in the following table.

Paired Samples Signed Rank Test for $\boldsymbol{\mu_d}$

Assumptions: $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ are two random samples that are paired and the differences $d_1, d_2, \ldots, d_n$ form a single sample from a *continuous* population whose distribution is *symmetric*.

Null hypothesis: $H_0 : \mu_d = 0$.

Test statistic value: $w^+ =$ sum of the ranks of $|\,d_i\,|$'s for which $d_i > 0$.

Decision rule: Reject H_0 if p-value $< \alpha$ or w^+ is in rejection region.

Alternative hypothesis	P-value = tail probability of the W^+ distribution under H_0: *	Rejection region = w^+ values such that: **
$H_a : \mu_d > 0$	to the right of (and including) w^+	$w^+ \geq w_{\alpha,n}$
$H_a : \mu_d < 0$	to the left of (and including) w^+	$w^+ \leq w^*_{\alpha,n}$
$H_a : \mu_d \neq 0$	$2 \cdot$ (the smaller of the tail probabilities to the right of (and including) w^+ and to the left of (and including) w^+)	$w^+ \leq w^*_{\alpha/2,n}$ or $w^+ \geq w_{\alpha/2,n}$

* For a given sample size (after deleting the zero-valued d_i's) n, in Table B6, the p-value can be taken to be less than the smallest α for which H_0 would be rejected using the rejection region approach.

** For a given level of significance α and sample size (after deleting the zero-valued d_i's) n, in Table B6 the upper tail critical value $w_{\alpha,n}$ is the *large* W entry associated with row n, column α. The lower tail critical value $w^*_{\alpha,n}$ is the *small* W entry.

9.5.3 Carrying Out the Signed Rank Test for Paired Samples

In the next example we'll complete the hypothesis test started in Example 9.12.

Example 9.13: Signed Rank Test

In the study of the effect of holding period on particulate-bound mercury (Hg) in air samples (Example 9.12), the sample size is $n = 10$.

The mean of the W^+ distribution under the null hypothesis is $n(n+1)/4 = 27.5$, so the observed value $W^+ = 25$ (from Example 9.12) provides *very little* evidence that the long holding period affects Hg measurements. The sampling distribution of W^+ under the null is shown below along with the critical values that determine the rejection region (left) and the p-value (right).

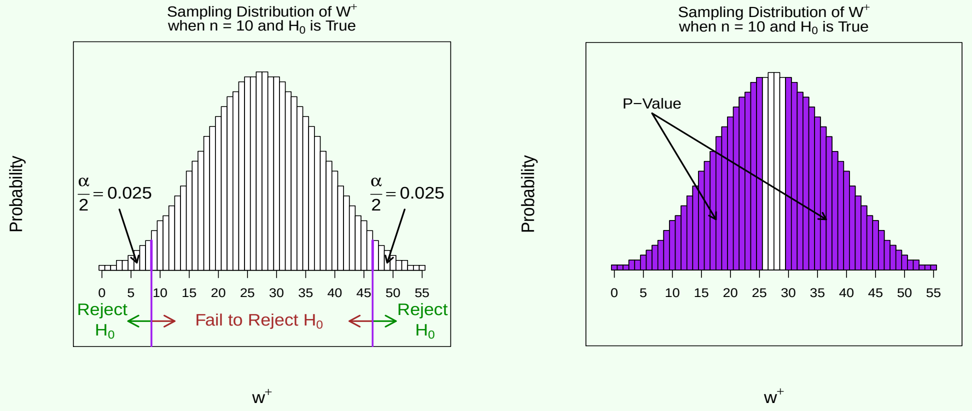


Figure 9.9: Sampling distribution of the signed rank test statistic W^+ when $n = 10$ and H_0 is true. The critical values defining the rejection region are 8 and 47. The p-value is greater than 0.10.

Using a level of significance $\alpha = 0.05$, the critical values for the two-tailed test, $w^*_{\alpha/2,n}$ and $w_{\alpha/2,n}$, from Table B6, are

$$w^*_{0.025,10} \;=\; 8 \qquad \text{and} \qquad w_{0.025,10} \;=\; 47\,.$$

Thus the decision rule is

$$\begin{array}{l} \text{Reject } H_0 \text{ if } \; W^+ \leq 8 \;\text{ or }\; W^+ \geq 47 \\ \text{Fail to reject } H_0 \text{ if } \; 8 \,<\, W^+ \,<\, 47 \end{array}$$

Since $W^+ = 25$ isn't in the rejection region, we fail to reject the null hypothesis. We conclude that there's no statistically significant evidence for any effect of holding time on Hg measurements. (The exact p-value can't be obtained from table B6, but it can be seen from the table to be greater than $2(0.05) = 0.10$.)

Comment: One important scenario in which the differences are guaranteed to follow a continuous, symmetric distribution is if the X and Y populations are continuous and have the *same shape*, as stated by the following fact.

Fact 9.3 Suppose X and Y are random variables drawn from *any* two continuous distributions that have the *same shape*. Then the difference $D = X - Y$ is a random variable that follows a continuous distribution whose shape is *symmetric*.

In practice, this means that we don't have to check the assumption that the differences in a matched pairs study follow a symmetric distribution if we're fairly confident that the two variables X and Y follow same-shaped distributions (with possibly different means).

9.5.4 What if the Difference Population Isn't Symmetric?

The signed ranks test requires that the differences in a matched pairs study follow a *symmetric* distribution. It turns out, though, that we *can* still carry out the test even if the population of differences isn't symmetric. However, in this case it's not testing whether the population mean difference μ_d is zero (or that μ_x and μ_y are equal). Rather, it's testing hypotheses that can be stated in words as:

H_0 : The population of differences has a distribution that's symmetric about zero
H_a : The population of differences has a distribution whose positive values are farther from (or closer to) zero than its negative ones

Here's more formally what's meant by "farther from zero" in the context of Hg measurements made after the two holding times. Each time we measure the Hg in a matched pair of air samples, the difference D is a random variable. It might be positive and it might be negative. There's a probability that it will be greater than 2.0 pg/m^3, $P(D > 2.0)$, and a probability it will be less than -2.0 pg/m^3, $P(D < -2.0)$. If the positive differences are "farther from zero" than the negative ones, then we'd have

$$P(D > 2.0) \ > \ P(D < -2.0).$$

The alternative hypothesis says that this inequality holds not just for 2.0 but for *any* specified difference value. In other words, it says that the difference distribution has more probability to the right of whatever positive value we specify than it has to the left of the negative of that value.

9.5.5 Some Comments on the Sampling Distribution of W^+

For deeper understanding of the signed rank test, it helps to know how the sampling distribution of the test statistics under the null hypothesis is determined. We'll consider the case when $n = 4$ and the difference population is symmetric.

When the null hypothesis is true, the difference population is symmetric about zero, so an observed difference D_i is just as likely to fall a given distance above zero as it is to fall that same distance below zero. Thus when *ranking* the distances of the D_i's away from zero, each rank is equally likely to be associated with a D_i that's positive or one that's negative. In the table below, each "±" is equally likely to be a "+" or a "−".

Sign	±	±	±	±
|Difference|	$\lvert D_i \rvert$	$\lvert D_i \rvert$	$\lvert D_i \rvert$	$\lvert D_i \rvert$
Rank	1	2	3	4

In other words, when the null hypothesis is true, each of the possible sets of ranks for the positive differences, listed below on the left side of the table, are equally likely.

Ranks associated with the positive D_i's	Value of w^+	Frequency of w^+
none	0	1
1	1	1
2	2	1
$1,2;\ \ 3$	3	2
$1,3;\ \ 4$	4	2
$1,4;\ \ 2,3$	5	2
$1,2,3;\ \ 2,4$	6	2
$1,2,4;\ \ 3,4$	7	2
$1,3,4$	8	1
$2,3,4$	9	1
$1,2,3,4$	10	1
		16

Also shown are the values of W^+ (sums of the ranks) and the frequency of each W^+ value. There are a total of 16 possible sets ranks for the positive D_i's, all of which are equally likely when the null hypothesis is true, but W^+ can only take the values $0, 1, \ldots, 10$. Thus the sampling distribution of W^+ is

Value of w^+	0	1	2	3	4	5	6	7	8	9	10
$p(w^+)$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$

and is shown in the probability histogram below.

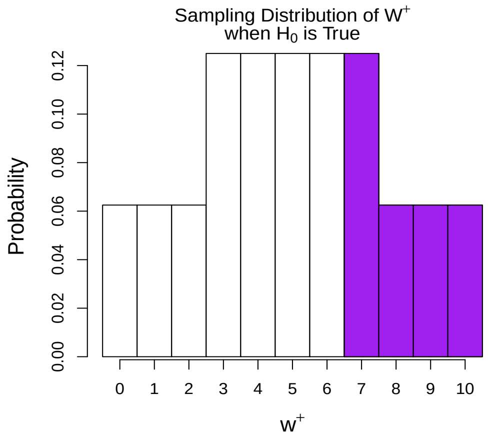


Figure 9.10: Sampling distribution of the signed ranks test statistic W^+ when $n = 4$ and H_0 is true.

If the observed test statistic value was $W^+ = 7$, the p-value for an upper tailed test would be the shaded probability in Fig. 9.10, which, from the table above, is $\frac{2}{16} + \frac{1}{16} + \frac{1}{16} + \frac{1}{16} = 0.3125$.

Properties of the Sampling Distribution of W^+: Here are several properties of W^+ that help us interpret its value.

1. W^+ takes its smallest possible value, zero, when *none* of the pairwise differences are positive.
2. W^+ takes its largest possible value when *all* of the pairwise differences are positive. In this case, W^+ is equal to $1 + 2 + \cdots + n = n(n+1)/2$.
3. Recall that when the null hypothesis is true, the mean of the sampling distribution of W^+ is

$$\mu_{w^+} = \frac{n(n+1)}{4}.$$

This makes sense because it's halfway between the two extreme W^+ values corresponding to *none* of the differences being positive and *all* of them being positive.

9.5.6 Large Sample Version of the Signed Rank Test

Notice from Figs. 9.8 and 9.9 that the sampling distribution of W^+ when the null hypothesis is true is symmetric and roughly bell-shaped. The plots below show that the distribution becomes more and more bell-shaped as the sample size n gets bigger.

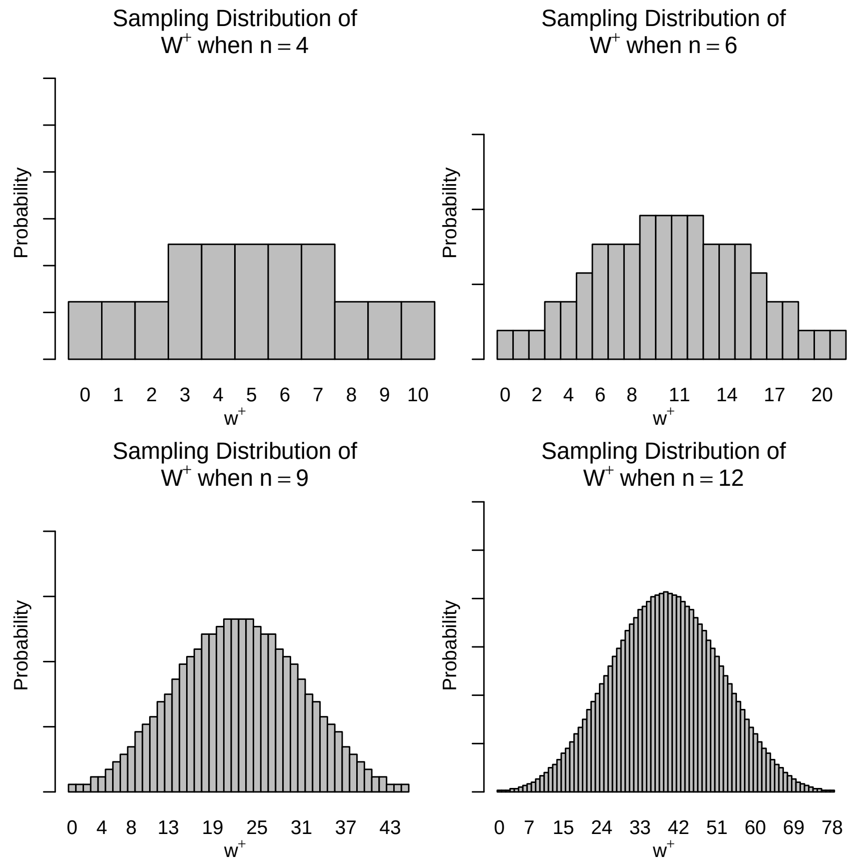


Figure 9.11: Sampling distribution of the signed rank test statistic W^+ for various sample sizes n when H_0 is true.

It turns out that as n increases, the distribution gets closer and closer to a normal distribution, as stated in the following fact.

Fact 9.4 Suppose $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ are *paired* samples, and suppose the differences $D_1, D_2, \ldots, D_n$ can be considered to be a single random sample from a *continuous*, *symmetric* population of differences whose mean is μ_d. Then if n is large ($n \geq 15$ suffices), and $H_0 : \mu_d = 0$ is true,

$$W^+ \sim \mathrm{N}(\mu_{w^+}, \sigma_{w^+}),$$

approximately, where the mean μ_{w^+} and standard error σ_{w^+} are given by (9.4) and (9.5).

It follows that if we standardize W^+, the resulting random variable Z follows a standard normal random distribution, that is,

$$Z \;=\; \frac{W^+ - \mu_{w^+}}{\sigma_{w^+}} \;\sim\; \text{N}(0,\,1)$$

approximately.

When n is large, the appropriate test statistic for carrying out the signed rank test is the ***large sample signed rank test statistic***, denoted $\boldsymbol{Z^+}$, given by the following.

Large Sample Signed Rank Test Statistic:

$$Z^+ \;=\; \frac{W^+ - \mu_{w^+}}{\sigma_{w^+}}, \tag{9.6}$$

where the mean μ_{w^+} and standard error σ_{w^+} are given by (9.4) and (9.5).

P-values (and critical values for the rejection region approach) are obtained from the tails of the N(0, 1) distribution in the direction specified by the alternative hypothesis.

Comment: As was the case for the large sample versions of the sign test of Chapter 7 and the rank sum test of Chapter 8, most statistical software packages use a slightly more accurate ***continuity corrected*** version of Z^+ when computing p-values for the large sample version of the signed rank test. The continuity correction accounts for the fact that a continuous distribution (the standard normal) is approximating a discrete one (the true distribution of Z^+). Details about the continuity correction can be found in many statistics textbooks, including [9].

9.6 The Sign Test for Paired Samples

9.6.1 Introduction

The ***sign test for paired samples***, like the signed rank test, is a *nonparametric* test for paired samples. But unlike the signed rank test, not only does it not required a normality assumption for the differences, it doesn't even require that the difference population be symmetric.

The sign test for paired samples is simply an application of the *one-sample sign test* of Chapter 7 to the sample of differences. Therefore, in contrast to the paired t and signed rank tests, it tests for the value of the *median* difference in the population, not the mean difference.

9.6.2 The Paired Samples Sign Test Procedure

Suppose we have paired samples $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ and that we can consider the differences $D_1, D_2, \ldots, D_n$ to be a *single* random sample from *any continuous* population (not necessarily normal nor even symmetric).

We'll want to test the null hypothesis

$$H_0 : \tilde{\mu}_d \;=\; 0$$

where $\tilde{\mu}_d$ is the true (unknown) median of the population of differences, versus one of the three alternatives

1. $H_a : \tilde{\mu}_d \;>\; 0$ (upper-tailed test)
2. $H_a : \tilde{\mu}_d \;<\; 0$ (lower-tailed test)
3. $H_a : \tilde{\mu}_d \;\neq\; 0$ (two-tailed test)

The null hypothesis says that an observed difference has a 50/50 chance of being positive or negative. In other words, it says that it's just as likely that X will be bigger than Y as the other way around. The alternative hypothesis says either that the difference is more likely to be positive or that it's more likely to be negative, and the direction for the alternative should reflect what we're seeking evidence for in our study.

If the null hypothesis was true then, we'd expect about half of the differences in our sample, or $n/2$ of them, to be positive and the other half negative. The ***paired samples sign test statistic***, denoted $\boldsymbol{S^+}$, is just the number of observed differences that are positive.

Paired Samples Sign Test Statistic: For a sample of differences $D_1, D_2, \ldots, D_n$,

$$S^+ \ = \ \text{Number of } D_i\text{'s that are greater than zero.}$$

If any of the D_i's equal zero, they're discarded prior to computing S^+, and the sample size n is reduced by the number of discarded D_i's.

If substantially more than half or substantially less than half of the differences in the sample are positive, that is, if S^+ differs substantially from $n/2$, it suggests that the true median difference $\tilde{\mu}$ differs from zero. More precisely,

1. *Large* values of S^+ (larger than $n/2$) provide evidence in favor of $H_a : \tilde{\mu}_d > 0$.
2. *Small* values of S^+ (smaller than $n/2$) provide evidence in favor of $H_a : \tilde{\mu}_d < 0$.
3. *Both large and small* values of S^+ (larger or smaller than $n/2$) provide evidence in favor of $H_a : \tilde{\mu}_d \neq 0$.

The paired samples sign test procedure is summarized in the table below.

Paired Samples Sign Test for $\tilde{\mu}_d$

Assumptions: $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ are two random samples that are paired, and the differences $d_1, d_2, \ldots, d_n$ form a single sample from *any continuous* population.

Null hypothesis: $H_0 : \tilde{\mu}_d = 0$.

Test statistic value: s^+ = number of positive d_i's.

Decision rule: Reject H_0 if p-value $< \alpha$ or s^+ is in rejection region.

Alternative hypothesis	P-value = tail probability of the binomial$(n, 0.5)$ distribution: *	Rejection region = s^+ values such that: **
$H_a : \tilde{\mu}_d > 0$	to the right of (and including) s^+	$s^+ \geq s_{\alpha,n}$
$H_a : \tilde{\mu}_d < 0$	to the left of (and including) s^+	$s^+ \leq s^*_{\alpha,n}$
$H_a : \tilde{\mu}_d \neq 0$	2·(the smaller of the tail probabilities to the right of (and including) s^+ and to the left of (and including) s^+)	$s^+ \leq s^*_{\alpha/2,n}$ or $s^+ \geq s_{\alpha/2,n}$

* For a given sample size (after deleting the zero-valued d_i's) n, the p-value for a one-tailed test is obtained from a binomial(n, 0.5) distribution table by locating the upper or lower tail probability (depending on the direction of H_a) associated with the observed S^+ value. For a two-tailed test, locate both the upper and lower tail probabilities and multiply the smaller of these by two.

** For a given sample size (after deleting zero-valued d_i's) n and level of significance α, $s_{\alpha,n}$ is obtained from a binomial(n, 0.5) distribution table by locating the smallest s for which the upper tail probability is less than α. $s^*_{\alpha,n}$ is obtained by locating the largest s for which the lower tail probability is less than α. For the two-tailed test, $s_{\alpha/2,n}$ and $s^*_{\alpha/2,n}$ are defined analogously but with $\alpha/2$ used in place of α. In practice, due to the discreteness of the distribution, it's not always possible obtain a rejection region having exact probability α.

9.6.3 Carrying Out the Paired Samples Sign Test

Example 9.14: Sign Test for Paired Samples

In a study to compare two methods for estimating bird densities, a line transect method and a quadrat-type area search method, densities of LeConte's Sparrow (*Ammodramus leconteii*) were estimated using both methods at each of 16 sites in Comanche County, Oklahoma [13]. The transect method makes an adjustment to account for the fact that birds farther from the transect are less likely to be observed. The quotes below, taken from the cited paper, describe the two methods more fully.

> A line transect is a line traveled by an observer, covering an area of unlimited width, where all birds detected and their perpendicular distances from the line are recorded. A

detection function is calculated based on the probability of detection at a given distance from the line. Methods such as line transects and point counts that adjust through a detection function for birds present but not observed are referred to as *distance sampling*.

An area search is a quadrat-type survey in which an observer moves about within a fixed area and tries to detect all birds within that area.

The table below shows, for each of the 16 sites, the bird density estimates (birds per hectare) made during the winter of 2002-2003. Also shown are the differences.

Bird Density Estimate

Site	Area Search	Transect Method	Difference
1	3.2	2.3	0.9
2	4.4	1.6	2.8
3	8.0	2.9	5.1
4	2.2	0.3	1.9
5	0.4	0.1	0.3
6	0.9	1.0	-0.1
7	2.3	1.0	1.3
8	2.8	2.9	-0.1
9	0.3	0.2	0.1
10	0.6	0.2	0.4
11	0.5	0.2	0.3
12	2.3	1.6	0.7
13	2.7	0.4	2.3
14	0.7	0.6	0.1
15	4.1	0.8	3.3
16	2.0	0.3	1.7

A normal probability plot and a histogram of the differences are below.

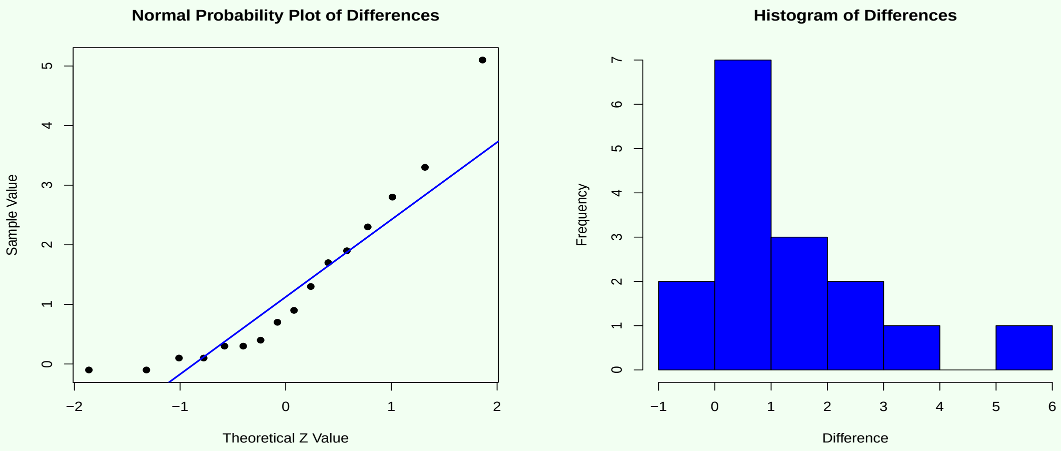


Figure 9.12: Normal probability plot and histogram of the differences in LeConte's Sparrow density estimates.

The plots indicate that the differences follow a right skewed distribution, and since $n = 16$ isn't large a paired t test wouldn't be appropriate. Instead, the researchers used a sign test for paired samples. The hypotheses are

$$\begin{aligned} H_0 : \tilde{\mu}_d &= 0 \\ H_a : \tilde{\mu}_d &\neq 0 \end{aligned}$$

where $\tilde{\mu}_d$ is the true (unknown) median difference between estimates made using the two methods (area search minus transect). Note that 14 of the 16 observed differences are positive, so the test statistic value is

$$s^+ = 14.$$

From a table of binomial distribution tail areas, using $n = 16$, the upper tail probability associated with $S^+ = 14$ is 0.0021, and the lower tail probability is 0.9997. Thus the p-value is two times the smaller of these,

$$\text{P-value} = 2(0.0021) = 0.0042.$$

Using a level of significance $\alpha = 0.05$, we reject the null hypothesis and conclude that the two methods produce different results. In particular, because most of the differences are positive, the area search method tends to produce *higher* density estimates than the line transect method.

9.6.4 Large Sample Version of the Sign Test for Paired Samples

The large sample version of the one-sample sign test described in Chapter 7 can be applied to the differences from a matched pairs study to form a large sample version of the paired samples sign test. Recall from Chapter 7 that when the null hypothesis is true and n is large ($n \geq 30$ is sufficient),

$$S^+ \sim \mathrm{N}(\mu_{s^+}\, \sigma_{s^+})$$

(approximately) and so

$$Z = \frac{S^+ - \mu_{s^+}}{\sigma_{s^+}} \sim \mathrm{N}(0, 1),$$

where the mean and standard error of the sampling distribution of S^+ under the null are

$$\mu_{s^+} = \frac{n}{2} \qquad \text{and} \qquad \sigma_{s^+} = \sqrt{\frac{n}{4}}\,. \tag{9.7}$$

It follows that when n is large, the appropriate test statistic for the paired samples sign test is the ***large sample sign test statistic for paired samples***, denoted $\boldsymbol{Z^+}$ and defined by the following.

Large Sample Paired Sign Test Statistic:

$$Z^+ = \frac{S^+ - \mu_{s^+}}{\sigma_{s^+}},$$

with μ_{s^+} and σ_{s^+} given by (9.7).

When $H_0 : \tilde{\mu}_d = 0$ is true (and n is large), Z^+ follows (approximately) a standard normal distribution. P-values (and critical values for the rejection region approach), therefore, are obtained from the tails of the $\mathrm{N}(0, 1)$ distribution in the direction specified by the alternative hypothesis.

Comment: As mentioned in Chapter 7, most statistical software packages use a more accurate ***continuity corrected*** version of Z^+ when computing p-values for the large-sample version of the sign test. Details about the continuity correction can be found in many statistics textbooks, including [9].

9.7 Nonparametric Confidence Interval for a Median Difference

We saw in Section 6.11 of Chapter 6 how to construct a *nonparametric* confidence interval for a population median based on a single sample from the population. We can applied that procedure to the *differences* in a matched pairs study to obtain a confidence interval for the true (unknown) median $\tilde{\mu}_d$ of the population of differences.

We don't require any assumptions about the population differences other than that it's a *continuous* distribution. The ***nonparametric paired samples confidence interval for*** $\boldsymbol{\tilde{\mu}_d}$ is given by the following.

Nonparametric Paired Samples Confidence Interval for $\boldsymbol{\tilde{\mu}_d}$: Suppose, as for the sign test for paired samples, that $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ are samples in a matched pairs study and that the differences $D_1, D_2, \ldots, D_n$ can be considered to be a sample from *any continuous* distribution. Then for a desired level of confidence
$\mathbf{100(1-\alpha)\%}$, the ***nonparametric confidence interval for*** $\boldsymbol{\tilde{\mu}_d}$ is

$$((k_1+1)\text{th smallest difference}, \quad k_2\text{th smallest difference}) \tag{9.8}$$

where k_1 is the largest value among $0, 1, \ldots, n$ for which

$$P(D \le k_1) \le \frac{\alpha}{2},$$

with $D \sim$ binomial(n, 0.5), and k_2 is the smallest value among $0, 1, \ldots, n$ for which

$$P(D \ge k_2) \le \frac{\alpha}{2}.$$

The values for k_1 and k_2 can be determined using statistical software or obtained from a table of critical values and tail probabilities for the binomial distribution.

Note: Using the above procedure, the *actual* confidence level is guaranteed to be no smaller than $100(1-\alpha)\%$. In other words, we can be *at least* $100(1-\alpha)\%$ confident that $\tilde{\mu}_d$ will be contained in the interval (9.8). The actual confidence level is given by the probability that a binomial(n, 0.5) random variable will fall between k_1+1 and k_2-1, inclusive.

9.8 Which Test Should Be Used, the Paired *t* Test, Signed Rank Test, or the Sign Test for Paired Samples?

When comparing two populations using paired data, our first choice should always be to use the paired t test because it's the *most powerful.* In other words, you'd be more likely to detect a difference or effect, if there is one, using a paired t test than if you used a signed rank test or a sign test for paired samples. But if the normality assumption for the difference population isn't met (and n is small), and we don't want to transform the data, then we have no choice but to use either the signed rank test or the sign test. Of these, the signed rank test is more powerful, so if the assumption of a symmetric difference population appears to be met, this is the test we should use. The sign test should only be used as a last resort, that is, if the assumptions for the other two tests aren't met.

Generally, the more stringent the assumptions that a test requires are, the more powerful the test will be. Tests that require more stringent assumptions, like the paired t test, use the complete information contained in the data (that is, their actual numerical values), whereas those with relaxed assumptions, such as the signed rank test and the sign test, only use the ranks or signs of the data values and therefore ignore some of the information contained in the data.

9.9 Problems

9.1 For each of the following studies, decide whether a matched pairs design or an independent samples design was used.

a) In 1878 Charles Darwin performed an experiment to determine if the height of a *Zea mays* (corn) plant is affected by whether the plant is cross-fertilized or self-fertilized. In each of 15 pots, two plants were grown, one self-fertilized and the other cross-fertilized, and their heights later measured.

b) In a study of the impact of agriculture on surface water quality, phosphorus was measured during each of 33 rainstorm events at the outlets of two adjacent watersheds, one of which contains farms and the other no farms.

c) In a study of the health hazards for workers in two types of swine confinement buildings, dried fecal matter was measured in the air of 12 randomly selected finishing buildings, where 50 - 100 kg animals are housed, and also in a random sample of 11 nursery buildings, which house smaller animals.

d) On April 7, 2000, an oil pipeline owned by the Potomac Electric Power Company ruptured, spilling 126,000 gallons of oil into marsh areas of Swanson Creek, Maryland. To assess the impact on benthic (bottom dwelling) communities, benthos were evaluated at 10 randomly selected locations in Swanson Creek near the spill and at 10 other randomly selected locations in the nearby, undisturbed Hunting Creek.

e) In a study of the long-term effect on shoreline biology of effluent from an oil refinery at Littlewick Bay, Wales, barnacle densities were measured at 10 shoreline locations near the refinery in 1974 and again at the *same* 10 locations in 1981.

f) A study was carried out by the University of Toronto to assess the impact of effluent from a sewage treatment plant near Orangeville, Ontario, Canada into the Credit River. Fecal coliform was measured on each of several days 1.5 km upstream of the plant and on those *same* days 2.5 km downstream.

g) In a study of the effect of pollution on cancer rates, to control for socioeconomic factors that might be related to cancer, researchers matched each of several communities near a source of pollution to one that had similar socioeconomic characteristics, but was located away from the source, and compared their cancer rates.

h) In a laboratory quality assurance study, ten 250 mL certified standard solutions with 100 ppm lead were sent to a lab for analysis and ten others to a different lab.

9.2 To investigate the importance of taking into consideration local wind direction when measuring air pollutants near traffic intersections, nitrogen dioxide (NO_2) was measured (μg/m^3) simultaneously upwind and downwind of an intersection in Patras, Greece on five occasions [15]. Measurements were made at a height of 0.8 to 1 m above ground because at these heights pollution directly impacts the health of drivers and children due to breathing. The data are below.

	$\mathbf{NO_2}$		
Sampling Occasion	Downwind	Upwind	Difference
1	126	75	51
2	103	77	26
3	90	63	27
4	75	67	8
5	75	82	-7

The summary statistics of the differences are

$$\begin{aligned} \bar{D} &= 21 \\ S_d &= 21.9. \end{aligned}$$

a) Carry out a paired t test to decide if the downwind NO_2 is statistically significantly higher than the upwind NO_2. Use a level of significance $\alpha = 0.05$.

b) The estimate of the true (unknown) effect size μ_d of traffic on NO_2 is $\bar{D} = 21$ μg/m^3. Compute and interpret a 95% paired t confidence interval for μ_d.

9.3 In an experiment to study the effect of fertilizers on soil and groundwater chemical properties, each of five farms in the Al-Wafrah region of Kuwait was split into two plots, one of which received fertilizer and the other of which, acting as a control, remained unfertilized [1]. The table below shows the pH levels in soil for the five pairs of plots as well as their differences.

	pH Level		
	Fertilized	Non-fertilized	
Farm	Plot	Plot	Difference
1	7.40	7.40	0.00
2	7.46	7.61	-0.15
3	7.28	7.60	-0.32
4	7.42	7.58	-0.16
5	7.15	7.55	-0.40

Here are the summary statistics for the differences.

$$\begin{aligned} \bar{D} &= -0.21 \\ S_d &= 0.16. \end{aligned}$$

We want to decide if fertilizer has any effect on the pH level, and if so, to estimate the size of that effect. A normal probability plot of the five differences shows no indication of non-normality, so a paired t test is appropriate.

a) Carry out the paired t test to decide if fertilizer has any effect on the pH level. Use a level of significance $\alpha = 0.05$.

b) The estimate of the true (unknown) effect size μ_d of fertilizer on pH level is $\bar{D} = -0.21$. Compute and interpret a 95% paired t confidence interval for μ_d.

9.4 One cause of stomach flu (gastroenteritis) in children is exposure to water contaminated by a virus, the adenovirus type 40. To help prevent stomach flu, it's important to be able to measure concentrations of this virus in water.

An experiment was carried out to compare the results of two methods for measuring the virus in tap water, and also to compare the results of one of the methods when used in tap water with its results when used in sea water [7]. Both methods involve passing the water through fiberglass filters and measuring the virus accumulations on the filters. The first uses the IMDS (electropositive) filter and the other the Filterite (electronegative) filter.

The experiment involved propagating the virus on human liver cells and then separating out the excess liver tissue by centrifugation. This virus propagation and liver separation process was run four times. For each run, virus concentrations were added to two 113 L volumes of tap water, one of which was passed through the IMDS filter and the other the Filterite filter. Virus concentrations were also added to a 113 L volume of sea water which was passed through the Filterite filter.

The table below shows, for each of the four runs, the virus recovery efficiencies, defined as the percent of total virus added to the water that collects on the filter and is able to be removed and measured.

	Virus Recovery Efficiency (%)		
Experimental Run	IMDS with Tap Water	Filterite with Tap Water	Filterite with Sea Water
1	32.0	29.0	35.0
2	33.0	38.0	48.0
3	22.0	37.5	41.0
4	19.0	42.0	29.0

We want to decide if there's any statistically significant difference in recovery efficiencies using IMDS and Filterite filters with tap water, and also if there's any difference using Filterite with tap water and with salt water.

a) Carry out a paired t test to decide if there's any difference in the recovery efficiencies of IMDS and Filterite with tap water. Use a level of significance $\alpha = 0.05$.

b) Carry out a paired t test to decide if there's any difference in the recovery efficiencies of Filterite with tap water and Filterite with sea water. Use a level of significance $\alpha = 0.05$.

9.5 Catalytic converters are devices found on cars that convert hydrocarbons, carbon monoxide, and nitrogen oxides, all toxic byproducts of fuel combustion, into harmless compounds. While their use has resulted in decreased emissions of these toxic gases, it's suspected of increasing emissions of ammonia (NH_3), a gas which contributes to the formation of airborne particles.

A study was carried out to investigate atmospheric NH_3 due to automobile traffic near Rome, Italy [11]. On each of 12 days from spring 2001 to spring 2002, NH_3 concentrations (μg/m^3) were measured at an urban site in Rome and at rural site located about 20 km northeast of Rome. The table below shows the data.

	Atmospheric NH_3		
Date	Urban	Rural	Difference
May 8	3.1	2.8	0.3
May 15	3.0	2.4	0.6
Jun. 26	3.7	2.9	0.8
Jul. 24	3.4	2.6	0.8
Aug. 2	3.6	3.9	-0.3
Sept. 6	3.2	3.1	0.1
Oct. 18	4.2	2.2	2.0
Nov. 13	3.4	2.1	1.3
Dec. 3	5.3	1.5	3.8
Dec. 4	4.9	1.3	3.6
Feb. 27	3.3	1.2	2.1
Mar. 12	2.9	3.0	-0.1

a) Use a normal probability plot and histogram to check the assumption of normality of the differences required for the paired t test.

b) Carry out the paired t test to decide if the urban NH_3 concentrations are statistically significantly higher than the rural concentrations. Use a level of significance $\alpha = 0.05$.

c) Compute and interpret a 95% paired t confidence interval for the true (unknown) mean difference between urban and rural NH_3 concentrations, μ_d.

9.6 Carbon monoxide (CO) and sulfur oxides (SO_x) occur in the atmosphere both naturally and as a result of human combustion of fuel. Exposure to these gases at high levels (or low levels for long periods of time) can be detrimental to human health.

In a study of air quality in the city of San Luis, Argentina, CO and SO_x were measured at 8:30 AM and again at 12:30 PM on Monday through Friday for each of 12 weeks in 1994 [10]. These times of day were chosen because previous studies had found them to have the highest urban activity levels.

A stratified random sample of air quality monitoring sites was used. The city was first divided into five sectors containing about 30 blocks each. Then from each sector, seven, eight, or nine locations were randomly selected giving a total of 38 sites throughout the city.

The tables below show the weekly average CO and SO_x concentrations (ppm) for the two times of day.

CO

Week	8:30 AM	12:30 PM
1	16.45	14.69
2	12.55	13.32
3	5.11	8.12
4	5.89	4.85
5	10.83	11.47
6	5.15	8.25
7	3.15	4.92
8	3.50	10.35
9	7.55	5.30
10	5.31	5.14
11	5.14	15.45
12	10.55	2.99

SO_x

Week	8:30 AM	12:30 PM
1	0.89	0.90
2	0.71	0.77
3	0.00	0.64
4	0.30	0.81
5	0.58	0.80
6	0.40	0.84
7	0.48	0.60
8	0.37	0.80
9	0.48	0.59
10	0.13	0.38
11	0.52	0.93
12	0.59	0.58

In this problem, we'll analyze the CO data. (The SO_x data will be analyzed in Problem 9.7.)

a) Compute the 12 CO concentration differences and plot them in a normal probability plot. Based on the plot, does the normality assumption required for the paired t test appear to be met?

b) Carry out a paired t test to decide if there's any statistically significant difference between CO concentrations at 8:30 AM and 12:30 PM. Use a level of significance $\alpha = 0.05$.

c) Compute and interpret a 95% paired t confidence interval for the true (unknown) mean difference between 8:30 AM and 12:30 PM CO concentrations, μ_d.

9.7 Refer to the study comparing 8:30 AM and 12:30 PM concentrations of carbon monoxide (CO) and sulfur oxides (SO_x) in San Luis, Argentina described in Problem 9.6.

a) Compute the 12 SO_x concentration differences and plot them in a normal probability plot. Based on the plot, does the normality assumption required for the paired t test appear to be met?

b) Carry out a paired t test to decide if there's any statistically significant difference between SO_x concentrations at 8:30 AM and 12:30 PM. Use a level of significance $\alpha = 0.05$.

c) Compute and interpret a 95% paired t confidence interval for the true (unknown) mean difference between 8:30 AM and 12:30 PM SO_x concentrations, μ_d.

9.8 In the quality assurance study described in Example 9.12, the researchers also investigated whether holding air samples for extended periods of time would affect lab measurements of *gaseous* mercury (Hg).

A pair of air specimens was obtained on each of 12 sampling occasions. One specimen in each pair was held for 4 hours and the other for 120 hours hours before being analyzed in the lab for gaseous Hg. The table below shows the data (pg/m^3).

	Gaseous Mercury		
Sampling Occasion	Short Holding Time	Long Holding Time	Difference
1	5.46	7.80	-2.34
2	8.49	9.28	-0.79
3	3.57	2.73	0.84
4	5.35	4.47	0.88
5	6.32	4.73	1.59
6	3.13	5.22	-2.09
7	5.49	4.95	0.54
8	5.05	10.54	-5.49
9	1.93	2.48	-0.55
10	3.48	2.41	1.07
11	1.10	1.14	-0.04
12	1.22	1.65	-0.43

Carry out a signed rank test for paired samples to decide if holding time has any effect on the Hg measurements. Use a level of significance $\alpha = 0.05$.

9.9 A study was carried out to investigate the effectiveness of routine cleaning on reducing bacteria levels in a hospital [8]. For each of several surfaces in the hospital, the failure rate before and after cleaning was recorded. The failure rate is defined as the percentage of times that the bacteria level fails to meet standard specifications when tested.

The table below shows the failure rates (percent) before and after cleaning for ten surfaces in the bedroom and treatment room of one of the hospital's wards. Also shown are the differences.

	Failure Rate (%)		
Surface	Before Cleaning	After Cleaning	Difference
Worktop	90	93	3
Treatment Room Tap Handle	60	64	4
Treatment Trolley	80	86	6
Door Handle	60	43	-17
Fridge Handle	70	57	-13
Treatment Room Bin Lid	80	50	-30
Ward 4 Bed Tap Handle	90	86	-4
Sink	90	78	-12
Bed Rail	80	71	-9
Ward 4 Bed Bin Lid	70	64	-6

Carry out a signed rank test for paired samples to decide whether the routine cleaning reduces the failure rate. Use a level of significance level $\alpha = 0.05$.

9.10 A benefit of social living among animals of a given species is the potential for acquiring information from each other about the location of food or predators. One source of such information is the direction of the other's gaze, and several mammal species have been shown to be able to follow gaze direction.

In an experiment to determine if ravens are able to follow the gaze directions of others, six ravens were examined under each of two experimental conditions defined according to the direction of a human experimenter's gaze [3]. During the treatment condition, the raven was placed beside a barrier and the experimenter looked behind it in an attempt to get the raven to look there too. During the control condition, the experimenter gazed at a location on raven's side of the barrier.

Each raven was examined five times under each condition, and for each condition, the researchers counted how many times the raven looked around the barrier. Thus the maximum possible was five. The data are shown below.

Number of Raven Look-Arounds

	Treatment	Control	
Raven	Condition	Condition	Difference
1	3	0	3
2	1	0	1
3	2	1	1
4	4	1	3
5	2	0	2
6	2	2	0

Carry out a signed rank test to decide if there's statistically significant evidence that the ravens look around the barrier more often when the experimenter does. Use a level of significance $\alpha = 0.05$.

(Note that the signed rank test is often used with discrete data, and in fact it's the test used by the researchers in this study.)

9.11 In 1974 the Bellevue-Stratford Hotel in Philadelphia was the scene of an outbreak of what later became known as Legionnaire's disease. The cause of the disease was finally discovered to be bacteria that thrived on the air conditioning units of the hotel. Owners of the nearby Rip Van Winkle Motel, hearing of the problems at Bellevue-Stratford, immediately replaced their air-conditioning system. Before replacing it, though, the bacteria count was measured in each room. After the system was replaced the bacteria counts were measured again.

The table below shows the bacteria counts (in colonies per cubic foot of air) in the air of eight rooms before and after the new air conditioning system was installed. Also shown are the eight differences (after replacing the system minus before).

	Bacteria Count		
Room Number	Before	After	Difference
121	11.8	10.1	-1.7
163	8.2	7.2	-1.0
125	7.1	3.8	-3.3
264	14.0	12	-2.0
233	10.8	8.3	-2.5
218	10.1	10.5	0.4
324	14.6	12.1	-2.5
325	14.0	13.7	-0.3

A negative difference means the bacteria count decreased in that room. The hope was that replacing the air conditioning system would reduce the bacteria counts.

Carry a signed rank test to decide if there was a statistically significant reduction in bacteria counts. Use a level of significance $\alpha = 0.05$.

9.12 Groundwater is a significant source of drinking water in the San Joaquin Valley, California. But extensive use of fertilizers and pesticides for farming and the generally permeable soils in the region have resulted in problems with groundwater contamination by nitrates and pesticides.

Many studies of groundwater quality make use of water samples from domestic wells (wells that supply water for use in individual homes) rather than monitoring wells (non-pumping wells used specifically for drawing water quality samples). But nitrate and pesticide concentrations might be lower in domestic wells than in monitoring wells, in part because domestic wells are typically deeper than monitoring wells.

A study was carried out to determine if the type of well can affect water quality [5]. Fourteen domestic wells were randomly selected from within the region, and two monitoring wells were installed adjacent to each domestic well, one at the same depth as the domestic well and the other at a shallower depth (less than 20 ft below the groundwater table). Nitrate (mg/L) and various pesticide (ug/L) concentrations were then measured in each well.

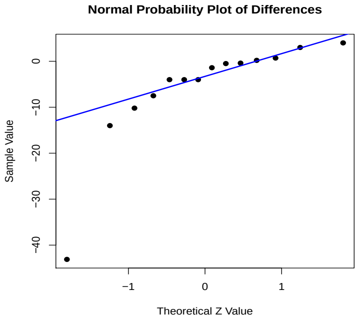
The table below shows the nitrate concentrations for domestic and shallow monitoring wells. (The deep monitoring well data will be analyzed in Problems 9.13 and 9.14.). Also shown are the differences.

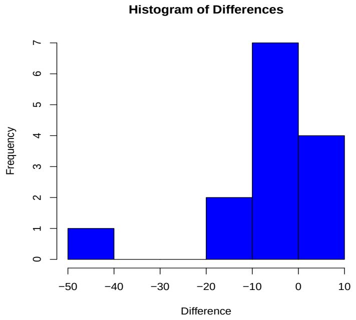
Nitrate Concentration

Site	Domestic Well	Shallow Monitoring Well	Difference
1	2.00	2.40	-0.4
2	0.70	8.20	-7.5
3	9.80	20.00	-10.2
4	6.10	5.40	0.7
5	11.00	7.00	4.0
6	20.00	17.00	3.0
7	10.00	14.00	-4.0
8	9.90	53.00	-43.1
9	2.90	4.30	-1.4
10	29.00	33.00	-4.0
11	4.00	18.00	-14.0
12	14.00	18.00	-4.0
13	4.80	5.30	-0.5
14	1.40	1.20	0.2

The researchers suspected, prior to looking at the data, that domestic wells would have lower nitrate concentrations than shallow monitoring wells. Thus a one-sided test is appropriate. The cited report states that because the normality assumption for the differences isn't met (and the sample size is small), a paired t test isn't appropriate. A normal probability plot and histogram of the differences, shown below, appear to confirm this contention.
Instead, to decide if domestic wells have lower nitrate concentrations, the researchers carried out a sign test for paired samples.

Carry out the sign test for paired samples. Use a level of significance $\alpha = 0.05$.





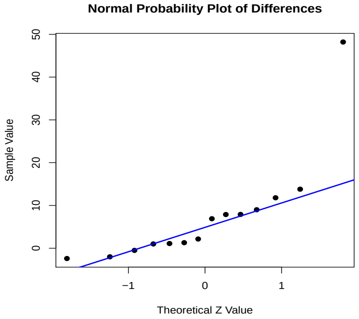
9.13 Refer to the study to decide if the type of well affects water quality in the San Joaquin Valley, described in Problem 9.12.

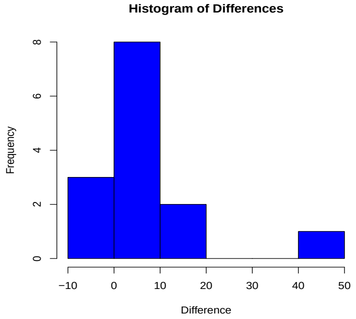
The table below shows the nitrate concentrations for shallow and deep monitoring wells. Also shown are the differences.

Nitrate Concentration

Site	Shallow Monitoring Well	Deep Monitoring Well	Difference
1	2.40	0.24	2.16
2	8.20	0.32	7.88
3	20.00	22.00	-2.00
4	5.40	5.90	-0.50
5	7.00	9.40	-2.40
6	17.00	9.10	7.90
7	14.00	7.10	6.90
8	53.00	4.80	48.20
9	4.30	3.00	1.30
10	33.00	24.00	9.00
11	18.00	4.20	13.80
12	18.00	6.20	11.80
13	5.30	4.20	1.10
14	1.20	0.19	1.01

The researchers suspected, prior to looking at the data, that shallow wells would have higher nitrate concentrations than deep wells. Thus a one-sided test is appropriate. The cited report states that because the normality assumption for the differences isn't met (and the sample size is small), a paired t test isn't appropriate. A normal probability plot and histogram of the differences, shown below, reveal an outlier that appears to confirm this contention.





Instead, to decide if shallow wells have higher nitrate concentrations, the researchers carried out a sign test for paired samples.

Carry out the sign test for paired samples. Use a level of significance $\alpha = 0.05$.

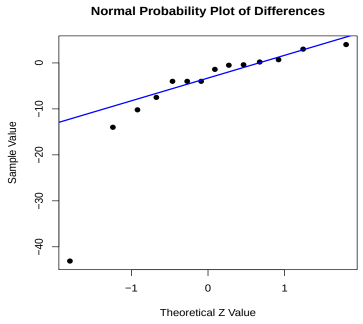
9.14 Refer to the study to decide if the type of well affects water quality in the San Joaquin Valley, described in Problem 9.12.

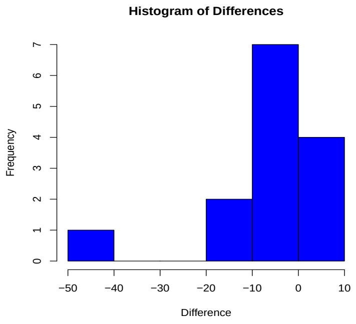
The table below shows the nitrate concentrations for domestic and deep monitoring wells. Also shown are the differences.

Nitrate Concentration

Site	Domestic Well	Deep Monitoring Well	Difference
1	2.00	0.24	1.76
2	0.70	0.32	0.38
3	9.80	22.00	-12.20
4	6.10	5.90	0.20
5	11.00	9.40	1.60
6	20.00	9.10	10.90
7	10.00	7.10	2.90
8	9.90	4.80	5.10
9	2.90	3.00	-0.10
10	29.00	24.00	5.00
11	4.00	4.20	-0.20
12	14.00	6.20	7.80
13	4.80	4.20	0.60
14	1.40	0.19	1.21

Because the domestic and monitoring wells are the same depth, the researchers had no reason to suspect, prior to looking at the data, which type of well, if any, would have higher nitrate concentrations. Thus a two-sided test is appropriate. The cited report states that because the normality assumption for the differences isn't met (and the sample size is small), a paired t test isn't appropriate. A normal probability plot and histogram of the differences, shown below, reveal an outlier that appears to confirm this contention.





Instead, to decide if there's any difference between the nitrate concentrations in domestic and deep monitoring wells, the researchers carried out a sign test for paired samples.

Carry out the sign test for paired samples. Use a level of significance $\alpha = 0.05$.

9.15 Amendments to the Clean Air Act were enacted in 1990 by the U.S. Congress in part to reduce atmospheric acid deposition, which consists mainly of sulfuric and nitric acids derived from the burning of coal and oil. Since then, numerous studies have been carried out to assess the effects of the amendments. Some have found evidence for a reduction in acid deposition, while others have not.

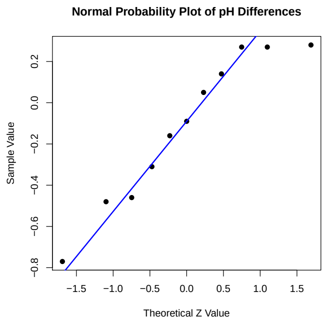
In one study, chemical and biological variables related to acid deposition were measured at 11 sites along streams in the Catskill Mountains, northeastern United States, in 1987 and then again at the same 11 sites in 2003 [4]. Among the variables measured were acidity (pH) and inorganic monomeric aluminum ($\mathrm{Al_{im}}$). Inorganic monomeric aluminum is used as an indicator of acidity. It's more soluble in acidic streams, so its concentration is inversely related to pH (as pH increases, Al decreases). Increased Al concentrations can kill fish, so one potential consequence of atmospheric acid deposition is increased fish deaths.

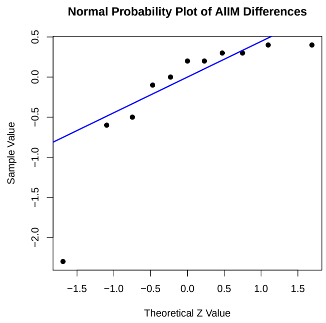
The table below shows the of pH levels and $\mathrm{Al_{im}}$ concentrations (mmol/L) for the two years. Also shown are the differences (2003 minus 1987).

	pH		
Site	1987	2003	Difference
NE-01	4.96	4.65	-0.31
NE-05	5.09	4.63	-0.46
NE-07	5.51	4.74	-0.77
NE-10	5.67	5.19	-0.48
NE-11	5.82	5.66	-0.16
NW-01	4.87	4.78	-0.09
NW-06	5.96	6.23	0.27
NW-04	6.06	6.34	0.28
NW-08	6.24	6.38	0.14
NW-11	6.57	6.62	0.05
N-12	6.22	6.49	0.27

	$\mathbf{Al_{im}}$		
Site	1987	2003	Difference
NE-01	4.9	4.4	-0.5
NE-05	4.9	4.3	-0.6
NE-07	3.3	3.5	0.2
NE-10	1.3	1.7	0.4
NE-11	0.7	1.0	0.3
NW-01	8.2	5.9	-2.3
NW-06	0.8	0.7	-0.1
NW-04	0.7	0.7	0.0
NW-08	0.5	0.7	0.2
NW-11	0.4	0.8	0.4
N-12	0.4	0.7	0.3

Normal probability plots of the differences for these two variables are below.





a) Which test, the paired t test, signed rank test, or sign test, would be most appropriate for deciding if the pH increased between 1987 and 2003?

b) Carry out the test you chose in part a. Use a level of significance $\alpha = 0.05$.

c) Which test, the paired t test, signed rank test, or sign test, would be most appropriate for deciding if $\mathrm{Al_{im}}$ concentrations decreased between 1987 and 2003?

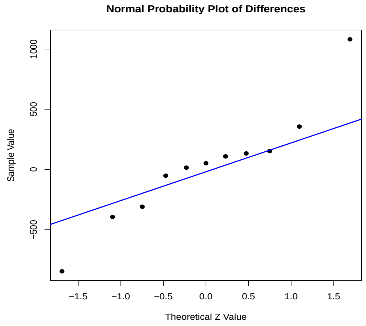
d) Carry out the test you chose in part c. Use a level of significance $\alpha = 0.05$.

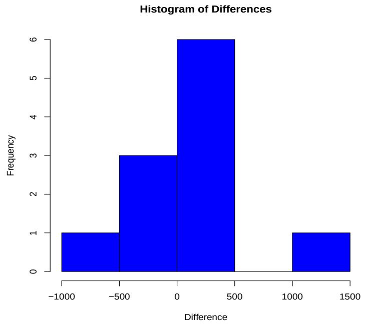
9.16 Here are the data (also shown in Example 9.4) from the study of the effect of a forest clear-cutting operation on a nearby stream's water quality, where nitrate concentrations (mg/L) were measured on each of 11 days upstream and downstream of the clear-cutting operation shortly after it was completed.

Nitrate Concentration

Date	Upstream	Downstream	Difference
08/15/97	1147.4	995.3	152.1
08/18/97	1412.2	1303.6	108.6
08/31/97	1613.9	1923.3	-309.4
09/18/97	763.3	747.8	15.5
11/04/97	1031.4	1082.9	-51.5
11/07/97	1093.2	1938.7	-845.5
02/27/98	390.8	338.8	52.0
07/14/98	909.8	776.8	133.0
08/25/98	1033.0	676.8	356.2
09/30/98	897.5	1291.0	-393.5
10/29/98	2314.0	1232.9	1081.1

We want to decide if there's any statistically significant difference in upstream and downstream nitrate concentrations. A normal probability plot and histogram of the differences are shown below.





a) Which test, the paired t test, signed rank test, or sign test, would be most appropriate for deciding if there's any difference between upstream and downstream nitrate concentrations?

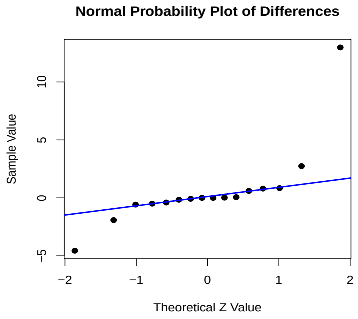
b) Carry out the test you chose in part a. Use a level of significance $\alpha = 0.05$.

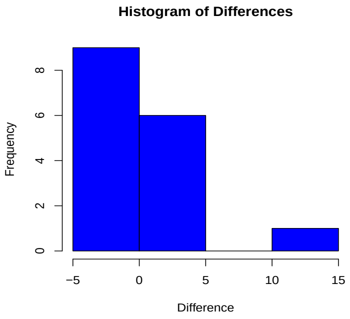
9.17 In the study of the effect of a forest clear-cutting operation on a stream's water quality described in Example 9.4 and Problem 9.16, suspended solids were measured on each of 16 days upstream and downstream of the clear-cutting operation shortly after logging was completed. The table below shows the suspended solids concentrations (mg/L) and their differences.

Suspended Solids

Date	Upstream	Downstream	Difference
06/03/98	8.40	12.96	-4.56
06/09/98	3.24	2.40	0.84
06/15/98	4.30	4.24	0.06
06/22/98	1.96	1.94	0.02
06/29/98	3.72	4.30	-0.58
07/09/98	1.34	0.74	0.60
07/15/98	1.72	1.72	0.00
07/23/98	9.36	11.28	-1.92
07/31/98	21.06	8.08	12.98
08/17/98	0.88	1.38	-0.50
09/15/98	2.38	2.38	0.00
10/08/98	1.08	1.24	-0.16
11/11/98	4.40	3.60	0.80
11/26/98	3.64	3.72	-0.08
04/12/99	23.26	20.52	2.74
04/29/99	1.00	1.40	-0.40

We want to decide if there's any statistically significant difference in upstream and downstream suspended solids concentrations. A normal probability plot and histogram of the differences are shown below.





a) Which test, the paired t test, signed rank test, or sign test, would be most appropriate for deciding if there's any difference between upstream and downstream suspended solids concentrations?

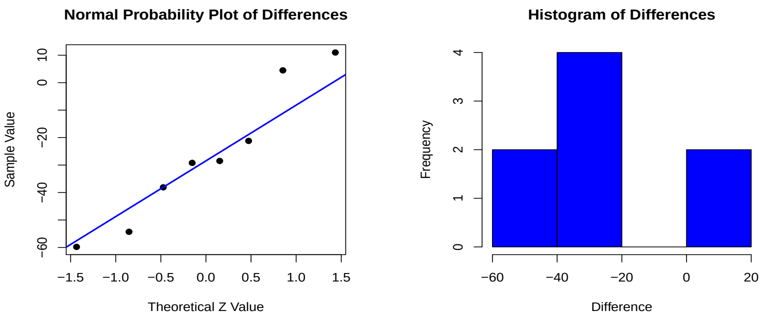
b) Carry out the test you chose in part a. Use a level of significance $\alpha = 0.05$.

9.18 Because coral communities normally grow slowly and live long, detecting changes in their structures, unless the result of a sudden catastrophic event, requires long-term studies. One such study was carried out at a site in Honolua Bay on Maui, Hawaii, to identify long-term changes in coral communities associated with an adjacent resort, golf course, and nearby pineapple agriculture [6].

Eight 50 m long transects were selected from within the bay in 1990. A 1×0.66 m rectangular quadrat frame was placed at 10 randomly selected positions along each transect and a photograph taken of the reef area enclosed by the frame. For each photograph, the percent coral cover was determined by overlaying the photograph with a grid dividing the rectangle into 100 equal-sized sections and counting the number of sections for which coral was present. These were then averaged for each transect. The coral cover was recorded again in July, 2002 along the same eight transects. The table below shows the data and the differences.

	Coral Cover (%)		
Transect	1990	2002	Difference
1	38.4	49.4	11.0
2	77.7	23.4	-54.3
3	39.7	44.2	4.5
4	88.7	60.2	-28.5
5	78.8	19.0	-59.8
6	90.8	61.6	-29.2
7	63.0	41.8	-21.2
8	86.3	48.2	-38.1

We want to carry out a test to decide if there was any statistically significant change in the percent coral cover between 1990 and 2002. A normal probability plot and histogram of the differences are shown below.



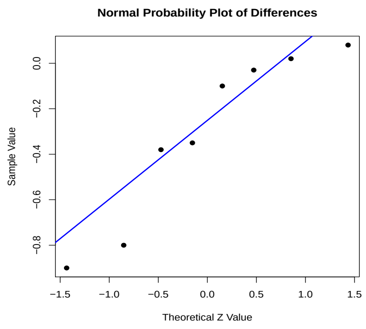
a) Which test, the paired t test, signed rank test, or sign test, would be most appropriate for deciding if there was any change in the percent coral cover between 1990 and 2002?

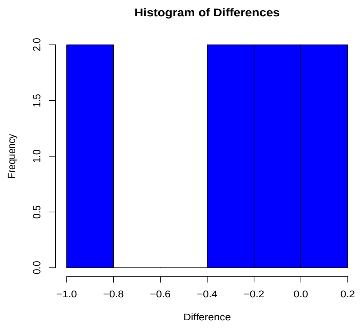
b) Carry out the test you chose in part a. Use a level of significance $\alpha = 0.05$.

9.19 In the study to identify long-term changes in coral communities associated with an adjacent resort, golf course, and pineapple agriculture at a site in Honolua Bay described in Problem 9.18, a measure of coral species diversity known as *Pielou's index* was calculated along each of the eight transects in 1990 and again on the same eight transects in 2002. Larger values of Pielou's index indicate more diversity. The data are below.

	Diversity Index		
Transect	1990	2002	Difference
1	1.73	1.35	-0.38
2	1.53	1.43	-0.10
3	1.85	1.50	-0.35
4	1.25	1.33	0.08
5	1.55	1.57	0.02
6	1.34	1.31	-0.03
7	1.48	0.68	-0.80
8	1.49	0.59	-0.90

We want to carry out a test to decide if there was any statistically significant change in the coral species diversity between 1990 and 2002. A normal probability plot and histogram of the differences are shown below.





a) Which test, the paired t test, signed rank test, or sign test, would be most appropriate for deciding if there was any change in the coral species diversity between 1990 and 2002?

b) Carry out the test you chose in part a. Use a level of significance $\alpha = 0.05$.

Bibliography

[1] Adeeba Al-Hurban. Preliminary assessment of the effects of fertilizers on soil properties in farming areas, southern Kuwait. *Management of Environmental Quality*, 17(3):258 – 27, 2006.

[2] C. E. Bemvenuti et al. Effects of dredging operations on soft bottom macrofauna in a harbor in the Patos Lagoon estuarine region of southern Brazil. *Brazilian Journal of Biology*, 64(4):573 – 581, 2005.

[3] Thomas Bugnyar, Mareike Stöwe, and Bernd Heinrich. Ravens, *Corvus corax*, follow gaze direction of humans around obstacles. *Proceedings of the Royal Society of London B*, 271:1331 – 1336, 2004.

[4] Douglas A. Burns, Karen Riva-Murray, Robert W. Bode, and Sophia Passy. Changes in stream chemistry and biology in response to reduced levels of acid deposition during 1987 - 2003 in the Neversink River basin, Catskill Mountains. *Ecological Indicators*, 8:191 – 203, 2008.

[5] K. R. Burow, J. L. Shelton, and N. M Dubrovsky. Occurrence of nitrate and pesticides in ground water beneath three agricultural land-use settings in the eastern San Joaquin Valley, California, 1993 - 1995. Technical Report USGS Water-Resources Investigations Report 97-4284, U.S. Geological Survey, 1998.

[6] S. J. Dollar and R. W. Grigg. Anthropogenic and natural stresses on selected coral reefs in Hawaii: A multidecade synthesis of impact and recovery. *Pacific Science*, 58(2):281 – 304, 2004.

[7] C. E. Enriques and C. P. Gerba. Concentration of enteric adenovirus 40 from tap, sea and waste water. *Water Research*, 29(11):2254 – 2560, 1995.

[8] C. J. Griffith et al. An evaluation of hospital cleaning regimes and standards. *Journal of Hospital Infection*, 45:19 – 28, 2000.

[9] D.R. Helsel. *Nondetects and Data Analysis, Statistics for Censored Environmental Data.* John Wiley and Sons, Inc., 2005.

[10] R. Lijteroff, V. Cortinez, and J. Raba. Urban development and air quality in San Luis City, Argentina. *Environmental Monitoring and Assessment*, 57:169 – 182, 1999.

[11] C. Perrino, M. Catrambone, A. Di Menno Di Bucchianico, and I. Allegrini. Gaseous ammonia in the urban area of Rome, Italy and its relationship with traffic emissions. *Atmospheric Environment*, 36:5385 – 5394, 2002.

[12] Martin R. Risch, Eric M. Prestbo, and Lucas Hawkins. Measurement of atmospheric mercury species with manual sampling and analysis methods in a case study in Indiana. *Water, Air, and Soil Pollution*, 184:285 – 297, 2007.

[13] Joseph P. Roberts and Gary D. Schnell. Comparison of survey methods for wintering grassland birds. *Journal of Field Ornithology*, 77(1):46 – 60, 2006.

[14] D. R. Thompson, P. H. Becker, and R. W. Furness. A novel approach to assess the impact of landuse activity on chemical and biological parameters in river catchments. *Freshwater Biology*, 50:1273 – 1289, 2005.

[15] Panayotis C. Yannopoulos. Spatial concentration distributions of sulfur dioxide and nitrogen oxides in Patras, Greece, in a winter period. *Environmental Monitoring and Assessment*, 135:163 – 180, 2007.