AN INTRODUCTION TO STATISTICS

WITH

DATA ANALYSIS

by

SHELLEY RASMUSSEN

Department of Mathematical Sciences
Olney 428T
University of Massachusetts/Lowell
Lowell, MA 01854

Originally published by Brooks/Cole Publishing Company,
Division of Wadsworth, Inc.

ISBN 0-534-13578-1

Shelley_Rasmussen@uml.edu

Minitab is a statistical package, a computer program that performs many statistical procedures. The versions of Minitab now available for use on personal computers are menu-driven and much easier to use than the main-frame version originally discussed in this text. Those sections are not included in this online edition of the text. At this time, the most recent version is Minitab 15, available at very reasonable prices for purchase or rental from:

www.e-academy.com/minitab

System Requirements

Processor:	PC with a 1 GHz 32- or 64-bit processor
Memory:	512 MB or more of available RAM
Disk Space:	125 MB free space available
Operating System:	Microsoft Windows 2000, XP, or Vista.
Display:	A display capable of 1024 X 768 or higher resolution
Software:	Adobe Acrobat Reader 5.0 or higher for Meet Minitab

CHAPTER 11

Inferences About Two Measures of Central Tendency

IN THIS CHAPTER

Two-sample comparisons
Two-sample comparisons based on large samples
Two-sample *t* test
Wilcoxon–Mann–Whitney two-sample test
Median test
Paired-sample comparisons
Paired *t* test

Is the weight of tomatoes canned at a factory the same in the morning and the afternoon? Does a chemical treatment retard tumor growth in mice? Can you squeeze more juice from an orange that has been microwaved for 20 seconds or from an orange that has not been microwaved? Will you run a race faster competing against yourself or someone else?

Each of these questions involves comparison of two groups or experimental conditions: morning versus afternoon, chemical treatment versus no treatment, microwave versus no microwave, competition with self versus competition with another. In this chapter we will be concerned with such comparisons of two measures of central tendency.

As with all statistical inference, the method of analysis we select depends on the assumptions we make about the sampling process. Before considering other assumptions, we first distinguish between *two-sample comparisons* and *paired-sample comparisons:*

> We make a **two-sample comparison** when we compare the means of two independent samples.
>
> If we compare the means of paired samples, we are making a **paired-sample comparison.**

We will look at paired-sample comparisons in Section 11-6. For analysis, we take the differences of the observations within pairs. We then make inferences about the center of these differences using the one-sample methods of Chapter 10.

In Sections 11-1 through 11-5, we discuss two-sample comparisons. When the sample sizes are large, we can base our inferences on the standard Gaussian distribution, as we will see in Section 11-1. Large-sample inference about two proportions is a special case, covered in Section 11-2.

For small samples, we first consider the classical approach, in Section 11-3. This requires the most assumptions about the sample, that we have two independent random samples from Gaussian distributions with equal variances. We compare the means of the two distributions using the *two-sample t test* and obtain interval estimates for the difference between two means based on a *t distribution.*

If we have two independent random samples of continuous-type observations from populations with the same shape and variation, then we can base our analyses on the *ranked observations.* The resulting nonparametric inferences are based on a *Wilcoxon–Mann–Whitney distribution,* covered in Section 11-4.

We might assume only that we have two independent random samples of continuous-type observations. If we consider only whether each observation is above or below the median of the combined observations, we are led to the *median test.* This is a nonparametric method of analysis based on a *hypergeometric distribution,* discussed in Section 11-5.

We begin in Section 11-1 with *two-sample comparisons for large samples.* We assume that we have two independent random samples, one from each

of two populations of interest. If the sample sizes are large enough, we can base tests of hypotheses and interval estimates on the *standard Gaussian distribution.*

11-1 Inferences About Two Means When Sample Sizes Are Large

Before we discuss large-sample comparisons of two means in general, let's consider an example.

EXAMPLE 11-1

Wire is wound on plastic spools to make coils for electric motors. When current passes through the wire, the spools heat up. An engineer wants to compare the resulting temperature rise for spools made from two types of plastic. (This example is based on data reported in Nelson, 1986, page 11.) Why do you think the engineer is interested in temperature rise on these plastic spools used in electric motors? Which do you think is preferable: a greater or smaller temperature rise?

The engineer selects a random sample of 30 spools from a large production lot of spools made with an old type of plastic. He selects a separate independent random sample of 30 spools from a large production lot of spools made with a new type of plastic. For each spool, he records the temperature rise after current passes through wire wound around the spool. The results are displayed in stem-and-leaf plots in Figure 11-1.

FIGURE 11-1 Stem-and-leaf plot of temperature rise (in degrees Centigrade) for 30 spools made with an old type of plastic and for 30 spools made with a new type of plastic. The stem shows temperature to the nearest degree. The leaf shows temperature to the nearest tenth of a degree.

New plastic		Old plastic	
Stem	Leaf	Stem	Leaf
44		44	
44		44	7 7
45		45	0 1 3 3 4
45	6 7 9	45	7 8 8 9 9 9
46	0 2 4 4	46	0 0 1 2 2 3 4
46	6 6 7	46	5 5 5 7 7
47	0 0 1 2 2 3 4 4	47	0 2 4
47	6 6 6 7 7 8 9 9	47	
48	0 1 2	48	1
48		48	
49	1	49	
49		49	
50		50	
50		50	6

Note that the scales for the two plots are aligned to make visual comparisons easier. We see there is one large temperature rise for the old plastic. Except for that one extreme value, the spread or variation in observed values is about the same for the two types of plastic. (The interquartile range for both plastics is about 1°C.) In general, however, the distribution of temperature rises for the new plastic is concentrated around higher values than is the distribution for the old plastic. The peak of the distribution for the new plastic is somewhere between 47 and 48 degrees Centigrade, while the peak of the distribution for the old plastic is around 46 degrees Centigrade.

As part of a formal analysis, we would like to test the null hypothesis that the average temperature rise is the same for the two types of plastic. We would also like to estimate the difference in average temperature rise for the two types of plastic. For these inferences, we will use large-sample techniques based on the standard Gaussian distribution. We will outline the approach below and then apply it to this example.

Two-Sample Comparisons of Means Based on Large Samples

Suppose we have two independent random samples, one from each of two populations. Suppose, in addition, that the sample sizes are large. We want to compare the two population means, μ_1 and μ_2.

Let $\bar{X}$ denote the sample mean and SE_1 the standard error of the mean, for the first sample. Let $\bar{Y}$ denote the sample mean and SE_2 the standard error of the mean, for the second sample. A reasonable point estimate for the difference $\mu_1 - \mu_2$ between the two population means is $\bar{X} - \bar{Y}$, the difference between the two sample means. The standard error, or estimated standard deviation, of $\bar{X} - \bar{Y}$ is

$$SE_{\bar{X}-\bar{Y}} = \sqrt{(SE_1)^2 + (SE_2)^2}$$

If the two sample sizes n_1 and n_2 are large enough, the quantity

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{SE_{\bar{X}-\bar{Y}}}$$

has approximately the standard Gaussian distribution. This result, related to the Central Limit Theorem results in Section 8-3, forms the basis of large-sample inference about two population means. The significance level approach to testing hypotheses about the two population means based on large samples is outlined below.

The significance level approach to testing hypotheses about two population means μ_1 and μ_2 based on large samples

1. The hypotheses are H_0: $\mu_1 = \mu_2$ and H_a: $\mu_1 \neq \mu_2$.
2. The test statistic is

$$\frac{\bar{X} - \bar{Y}}{SE_{\bar{X}-\bar{Y}}}$$

3. Assume that we have two independent random samples, one from a population with mean μ_1 and the other from a population with mean μ_2. Also assume that the two sample sizes are large enough to ensure that the sample means $\bar{X}$ and $\bar{Y}$ have approximate Gaussian distributions. Then under the null hypothesis, the test statistic has approximately the standard Gaussian distribution.
4. Select significance level α.
5. The acceptance region is the interval $(-c, c)$. The rejection region includes the intervals $(-\infty, -c]$ and $[c, \infty)$. The number c is chosen so that $P(Z \leq c) = 1 - \alpha/2$, where Z has the standard Gaussian distribution.
6. The decision rule is:

   If $-c <$ test statistic $< c$, say the results are consistent with the null hypothesis.

   If test statistic $\leq -c$ or test statistic $\geq c$, say the results are inconsistent with the null hypothesis.
7. Collect two large independent random samples, one from each of the two populations of interest. Calculate the test statistic based on the sample. Use the decision rule in step 6 to decide whether the observations are consistent with the null hypothesis. Draw conclusions based on analysis of the experimental results.

If we have a one-sided alternative, then in step 5 we select the number c so that $P(Z \leq c) = 1 - \alpha$, where Z has the standard Gaussian distribution. If the one-sided alternative is H_a: $\mu_1 > \mu_2$, then the acceptance region is the interval $(-\infty, c)$; the rejection region is the interval $[c, \infty)$. In step 6 we say values of the test statistic less than c are consistent with the null hypothesis, values greater than or equal to c are inconsistent with the null hypothesis.

If our one-sided alternative is H_a: $\mu_1 < \mu_2$, then the acceptance region is the interval $(-c, \infty)$; the rejection region is the interval $(-\infty, -c]$. In step 6 we say values of the test statistic greater than $-c$ are consistent with the null hypothesis, values less than or equal to $-c$ are inconsistent with the null hypothesis.

Large-Sample Confidence Intervals for the Difference Between Two Population Means

Large-sample confidence intervals for $\mu_1 - \mu_2$ are of the form

$$\bar{X} - \bar{Y} \pm c\text{SE}_{\bar{X}-\bar{Y}}$$

We find the number c from the standard Gaussian distribution. If the area from $-c$ to c under the standard Gaussian curve equals A, then we say the interval is an approximate $100A\%$ confidence interval for $\mu_1 - \mu_2$. For instance, if $c = 2.58$, then we have an approximate 99% confidence interval for $\mu_1 - \mu_2$.

EXAMPLE 11-1 *(continued)*

To apply large-sample inference in Example 11-1, the engineer calculates the following statistics from his sample:

New plastic:	$\bar{X} = 47.163$ °C	$s_1 = .824$ °C	$SE_1 = .150$ °C
Old plastic:	$\bar{Y} = 46.230$ °C	$s_2 = 1.132$ °C	$SE_2 = .207$ °C
		$SE_{\bar{X}-\bar{Y}} = .256$ °C	

He wants to know if the average temperature rise in spools is the same for the two types of plastic. So he compares the hypotheses H_0: $\mu_1 = \mu_2$ and H_a: $\mu_1 \neq \mu_2$, where μ_1 and μ_2 represent the mean temperature rise in spools made of the new and old types of plastic, respectively.

The engineer chooses significance level $\alpha = .01$. If Z has the standard Gaussian distribution, then $P(Z \leq 2.58) = .9951$, which is close to $1 - \alpha/2 = .995$. Therefore, he uses the decision rule:

If $-2.58 <$ test statistic < 2.58, say the results are consistent with the null hypothesis.

If test statistic ≤ -2.58 or test statistic ≥ 2.58, say the results are inconsistent with the null hypothesis.

He calculates the test statistic:

$$\frac{\bar{X} - \bar{Y}}{SE_{\bar{X}-\bar{Y}}} = \frac{47.163 - 46.230}{.256} = 3.6$$

Since 3.6 is in the rejection region, he concludes the results are inconsistent with the null hypothesis that mean temperature rise is the same for the two types of plastic.

The engineer notes that his approximate p-value is less than .0004. If the mean temperature rise were really the same for both types of plastic, there would be less than 4 chances in 10,000 of seeing a test statistic at least as extreme as the one observed. The experimental results strongly suggest that mean temperature rise is not the same for the two plastics.

The engineer then decides to estimate the difference $\mu_1 - \mu_2$ in mean temperature rise for the two types of plastic. His point estimate is $\bar{X} - \bar{Y} = 47.163 - 46.230 \doteq .9$ °C. For an interval estimate, he calculates an approximate 99% confidence interval:

$$(47.163 - 46.230 - 2.58 \times .256,\ 47.163 - 46.230 + 2.58 \times .256) \doteq (.3,\ 1.6)$$

He estimates that the mean temperature rise for the new type of plastic is from .3 °C to 1.6 °C greater than for the old type of plastic. The null hypothesis value of $\mu_1 - \mu_2$ is 0, not in the confidence interval. This agrees with the test of hypotheses.

From his graphical analysis of the data, the engineer knows that he has similar variation in values of temperature rise for the two types of plastic. The distribution of temperature rises for the new plastic is shifted toward larger

values than the distribution for the old plastic. However, the largest temperature rise, which might be considered an outlier because it is so far from the others, was observed for the old plastic. From his formal analysis, the engineer concludes that the difference in average temperature rise is statistically significant, with a p-value less than .0004. Based on a 99% confidence interval for the difference between the two means, he estimates that the mean temperature rise for the new plastic is from .3 °C to 1.6 °C greater than for the old plastic. To evaluate the *practical* significance of these results, the engineer must consider the relative cost of the two types of plastic, the relative lifetimes of spools made from these two plastics, and whether a difference in temperature rise of about 1 °C is of practical concern. Can you think of other issues that should be included in his comparison of the two plastics?

We can use the techniques for large-sample inferences about two population means to compare two proportions when sample sizes are large. We discuss large-sample comparisons of two proportions in Section 11-2.

11-2 Large-Sample Inference About Two Proportions

Suppose we have two independent random samples, one from each of two populations. Each observation has two possible values, say success or failure. We want to compare the proportion of successes in the two populations. Equivalently, we want to compare the probability of success for the two populations.

We can write our null hypothesis as $H_0: p_1 = p_2$, where p_1 represents the proportion of successes for the first population and p_2 the proportion of successes for the second population. If the sample sizes are large, the test statistic is

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where n_1 and n_2 are the two sample sizes, $\hat{p}_1$ and $\hat{p}_2$ are the observed proportions of successes in the two samples, and $\hat{p}$ is the observed proportion of successes in the combined samples. For large sample sizes, this test statistic has approximately the standard Gaussian distribution under the null hypothesis. We test hypotheses using the large-sample techniques discussed in Section 11-1.

Confidence intervals for the difference $p_1 - p_2$ between the two proportions are of the form

$$\hat{p}_1 - \hat{p}_2 \pm c\sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

Here, c comes from the standard Gaussian distribution to give the desired confidence level.

Let's illustrate large-sample inference about two proportions with an example.

EXAMPLE 11-2 Will a vaccine prevent cases of a dread disease? American public health workers in 1954 asked this question about the Salk vaccine for the prevention of polio. Polio is a disease with effects ranging from temporary, mild, flu-like symptoms to permanent disability or death. Parents in the early 1950s lived in fear of polio, often keeping their children away from playgrounds and other public places where polio might be spread. (This dread is perhaps comparable to the fear of AIDS parents today might feel if their child needed a blood transfusion.)

Medical workers must test potential new vaccines for safety and efficacy. A vaccine might not work, or might have unacceptable side effects. A live vaccine, such as the first Salk vaccine, might even cause some cases of the disease it was developed to prevent. To evaluate the effectiveness of the Salk vaccine, American public health workers conducted the largest medical experiment ever. Local health departments chose between two different experimental designs. We will discuss the design that is best from a statistical point of view: the double-blind randomized placebo-controlled experiment. Exercise 11-32 asks you to compare this experimental design with another design, selected by a number of communities.

In one treatment group, children received an injection with the Salk vaccine. In the other treatment group, children received an injection with a biologically inactive (placebo) solution. Public health workers planned to compare the proportions of polio cases for the two groups at the end of the study period. Let p_1 denote the probability of polio in the vaccinated group and p_2 the probability of polio in the placebo control group. Health workers wanted to test the hypotheses

$$H_0: \; p_1 = p_2 \qquad \text{and} \qquad H_a: \; p_1 \neq p_2$$

A two-sided alternative is reasonable here. The vaccine might prevent polio. But it was also possible the vaccine might cause some cases of polio. The two-sided alternative allows for both possibilities.

Health workers asked parents of schoolchildren for their consent to allow their children to participate in the study. They then used a random process to divide the children with parental consent into two groups, a treatment group and a placebo control group. The purpose of the random assignment of children was to balance the groups with respect to extraneous factors that might affect the outcome of the study. Extraneous factors affecting risk of polio included socioeconomic status, age, and geographic location.

The researchers conducted the experiment as a double-blind study. Neither the children, their parents, nor the public health workers treating and examining the children knew how the children had been treated. If parents and children knew the treatment, they might alter their behavior accordingly

(as in avoiding or not avoiding swimming pools or playgrounds where polio might be spread). If physicians examining the children knew the treatment, they might be subtly influenced in their diagnosis, since mild cases of polio were sometimes difficult to distinguish from other illnesses such as colds or flu. Use of the double-blind design avoided the influence of such extraneous factors on the experimental outcome.

Parents of about 400,000 schoolchildren gave consent for their children to participate in this experiment. (Parents of about 350,000 children did not give consent.) Results of the study are shown below (Meier, 1989; Francis et al., 1955):

Group	Number of children	Number of polio cases	Proportion of polio cases
Vaccine	200,745	57	$\hat{p}_1 = \frac{57}{200{,}745} \doteq .000284$
Placebo	201,229	142	$\hat{p}_2 = \frac{142}{201{,}229} \doteq .000706$
Total	401,974	199	$\hat{p} = \frac{199}{401{,}974} \doteq .000495$

To test the hypothesis that the probability of polio was the same for vaccinated and unvaccinated children, we use the test statistic

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$= \frac{.000284 - .000706}{\sqrt{(.000495)(.999505)\left(\frac{1}{200{,}745} + \frac{1}{201{,}229}\right)}} = -6.0$$

Comparing this value of the test statistic with the standard Gaussian distribution, we see that our approximate p-value is less than .0004. This tells us that if the null hypothesis were true and the probability of polio were the same in both groups, there would be less than 4 chances in 10,000 of seeing a test statistic at least this far from 0. The experimental results are inconsistent with the null hypothesis, strongly suggesting that polio incidence is different for vaccinated and unvaccinated children.

A 95% confidence interval for the difference $p_1 - p_2$ between the two probabilities is

$$.000284 - .000706 \pm 1.96\sqrt{\frac{(.000284)(.999716)}{200{,}745} + \frac{(.000706)(.999294)}{201{,}229}}$$

$$\doteq (-.00056, -.00028)$$

It is common in public health to report a proportion as the number of cases of disease per 100,000 people. With this in mind, $\hat{p}_1 \doteq .00028$ tells us there were about 28 polio cases per 100,000 children in the vaccine group. Similarly, $\hat{p}_2 \doteq .00071$ tells us there were about 71 polio cases per 100,000 children in the placebo control group. We can interpret the difference $\hat{p}_1 - \hat{p}_2 \doteq -.00042$ this way: We estimate that use of the polio vaccine under the same conditions would result in about 42 fewer cases per 100,000 children than use of the placebo. Our interval estimate is 28 to 56 fewer cases per 100,000 children, using the polio vaccine.

The results of this experiment strongly suggested that the Salk vaccine prevented polio. Wide use of the Salk vaccine after this study indicated that in addition to preventing polio, the vaccine could cause some cases of polio. Researchers subsequently developed better polio vaccines, which are now in use.

In Section 11-3, we discuss the classical approach to comparing two means when the sample sizes are small.

11-3 Inferences About Two Measures of Central Tendency Based on a t Distribution

The classical test of hypotheses about two means is called the *two-sample t test.* Let's first consider an example. Then we will describe the classical approach to two-sample comparisons and apply it to this example.

EXAMPLE 11-3

Is the weight of tomatoes canned at a factory the same in the morning and the afternoon? Investigators drained and weighed 10 cans of tomatoes one day. Five cans had been filled in the morning and five in the afternoon. The drained weights in ounces of these ten cans are shown below (Duncan, 1974, page 569; from Grant and Leavenworth, 1972, page 41):

Morning:	22.5	24.5	25.5	20.0	21.0
Afternoon:	22.5	19.5	21.5	20.5	20.0

We will use this information to compare the average canned weights for the two times of day.

Suppose you were the manager of this factory. Why would you be interested in comparing morning and afternoon performance at your factory? Here we are considering average canned weights. What other measures of quality would you be interested in evaluating?

A plot of the observations is shown in Figure 11-2. Three morning weights overlap with the range of afternoon weights; the other two morning weights are at least 2 ounces greater than the largest afternoon weight. Just

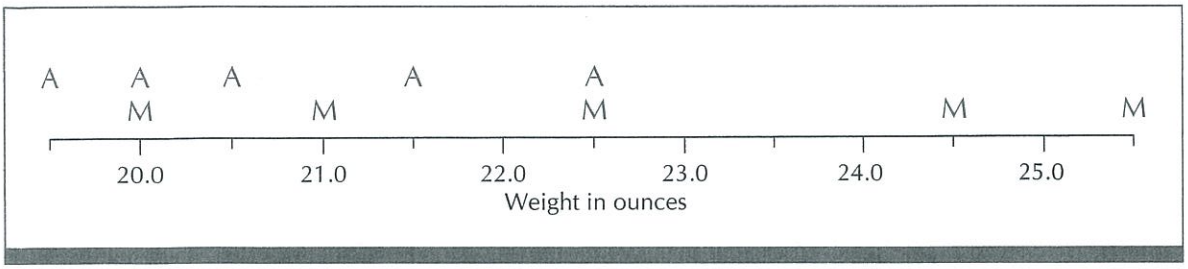

FIGURE 11-2 Plot of drained weights of canned tomatoes in Example 11-3. M represents the weight of a can filled in the morning; A represents afternoon.

from the plot we can see that the average drained weight in the morning sample is larger than the average drained weight in the afternoon sample. The spread or variation in the values also appears somewhat larger in the morning than in the afternoon.

For a formal analysis, we would like to test the null hypothesis that mean drained weight is the same for tomatoes canned in the morning and those canned in the afternoon. We would also like to estimate the difference in mean drained weights for the two times of day.

We will use the two-sample t test to decide whether the apparent difference between the morning and afternoon means is statistically significant. We will also use a t distribution to calculate an interval estimate for the difference between the morning and afternoon means. Before analyzing the experimental results, let's describe the classical approach to two-sample comparisons in general.

Two-Sample Comparisons of Means Based on a t Distribution

Suppose we have two independent random samples, one from each of two populations. Assume that the distribution of values in the two populations is Gaussian and that the two distributions have the same variance, σ^2. Let μ_1 denote the mean of the first population and μ_2 the mean of the second population. We want to test the null hypothesis that μ_1 and μ_2 are equal. We also want to estimate the difference between the means, $\mu_1 - \mu_2$.

The situation is illustrated in Figure 11-3. The two Gaussian distributions have the same variance, but possibly different means, so one distribution is shifted away from the other. The difference $\mu_1 - \mu_2$ between the two means measures the extent of the shift. We want to test whether $\mu_1 - \mu_2 = 0$. If so, the two distributions are identical.

Let $\bar{X}$ denote the mean and s_1 the sample standard deviation of the n_1 observations in the first sample. Similarly, let $\bar{Y}$ and s_2 denote the mean and sample standard deviation of the n_2 observations in the second sample.

The sample variances s_1^2 and s_2^2 are independent estimates (because the samples are independent) of the common variance σ^2 of the two populations.

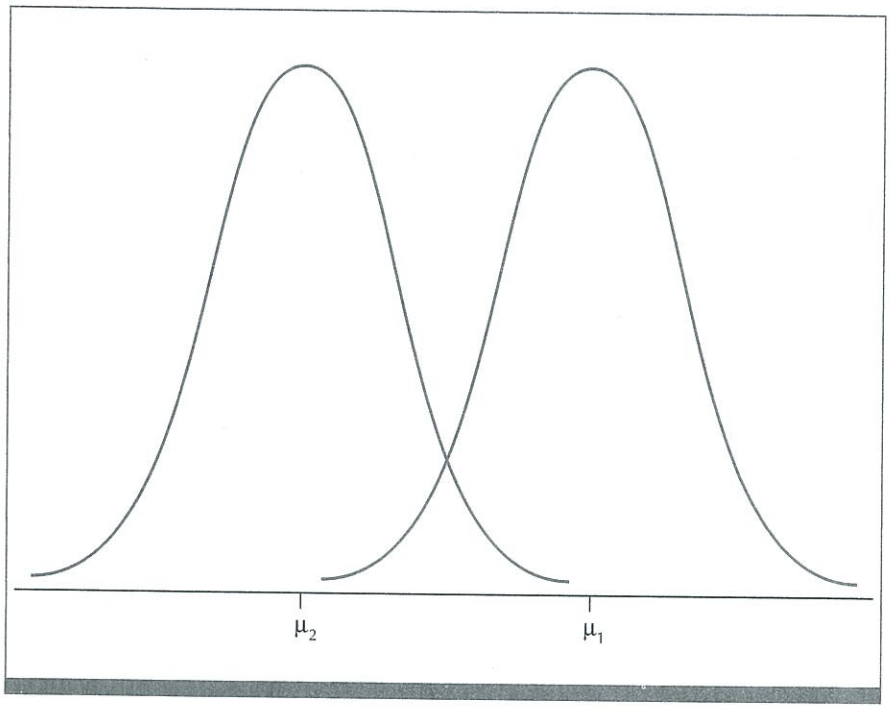

FIGURE 11-3 Probability functions are illustrated for two Gaussian distributions having the same variance.

We get a *pooled* estimate s_p^2 of σ^2 by calculating a weighted average of s_1^2 and s_2^2, with degrees of freedom as weights:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

The standard error $SE_{\bar{X}-\bar{Y}}$ estimates the standard deviation of the difference $\bar{X} - \bar{Y}$ between the two sample means:

$$SE_{\bar{X}-\bar{Y}} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

Under our model assumptions, the quantity

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{SE_{\bar{X}-\bar{Y}}}$$

has the t distribution with $n_1 + n_2 - 2$ degrees of freedom. The number $n_1 + n_2 - 2$ defines the t distribution for the quantity of interest here in the two-sample case. This number is the denominator in the definition of the pooled variance estimator s_p^2; it is the sum of the degrees of freedom for the separate sample variances s_1^2 and s_2^2.

Suppose X_1, X_2, through X_{n_1} are independent observations from a Gaussian distribution with mean μ_1 and variance σ^2. Suppose Y_1, Y_2, through Y_{n_2} are independent observations from a Gaussian distribution with mean μ_2 and variance σ^2. The two samples are independent.

Let $\bar{X}$ denote the sample mean of the first sample and $\bar{Y}$ the sample mean of the second sample. Let $SE_{\bar{X}-\bar{Y}}$ denote the estimated standard deviation of $\bar{X} - \bar{Y}$ as defined above. Then the quantity

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{SE_{\bar{X}-\bar{Y}}}$$

has the t distribution with $n_1 + n_2 - 2$ degrees of freedom.

Now we can outline the significance level approach to testing whether two population means are equal.

The significance level approach to testing for equality of two population means μ_1 and μ_2 based on a t distribution

1. The null and alternative hypotheses are H_0: $\mu_1 = \mu_2$ and H_a: $\mu_1 \neq \mu_2$.
2. The test statistic is

$$\frac{\bar{X} - \bar{Y}}{SE_{\bar{X}-\bar{Y}}}$$

where $SE_{\bar{X}-\bar{Y}}$ is defined above.

3. Assume that we have two independent random samples from Gaussian distributions with means μ_1 and μ_2. The two distributions have the same variance. Then under the null hypothesis, the test statistic has the t distribution with $n_1 + n_2 - 2$ degrees of freedom.
4. Specify the significance level α.
5. Find the number c in Table C such that $P(T \leq c) = 1 - \alpha/2$, where T is a random variable having the t distribution with $n_1 + n_2 - 2$ degrees of freedom. The acceptance region is the interval $(-c, c)$. The rejection region includes the intervals $(-\infty, -c]$ and $[c, \infty)$.
6. The decision rule is:

   If $-c <$ test statistic $< c$, say the results are consistent with the null hypothesis.

   If test statistic $\leq -c$ or test statistic $\geq c$, say the results are inconsistent with the null hypothesis.

7. Collect a random sample satisfying the given assumptions. Calculate the test statistic. Use the decision rule in step 6 to decide whether the observations are consistent with the null hypothesis. Draw conclusions based on the analysis.

If in step 1 we specify a one-sided alternative, then we must alter steps 5 and 6. In step 5, we find the number c in Table C such that $P(T \leq c) = 1 - \alpha$. If our alternative is H_a: $\mu_1 < \mu_2$, then the acceptance region is the interval $(-c, \infty)$; the rejection region is the interval $(-\infty, -c]$. In step 6, we say values

of the test statistic greater than $-c$ are consistent with the null hypothesis; values less than or equal to $-c$ are inconsistent with the null hypothesis.

If our one-sided alternative is H_a: $\mu_1 > \mu_2$, then the acceptance region is the interval $(-\infty, c)$; the rejection region is the interval $[c, \infty)$. In step 6, we say values of the test statistic less than c are consistent with the null hypothesis; values greater than or equal to c are inconsistent with the null hypothesis.

EXAMPLE 11-3 *(continued)*

In Example 11-3, we want to compare the average weight of tomatoes canned in the morning with the average weight of tomatoes canned in the afternoon. In particular, we want to test the hypotheses:

Null hypothesis: There is no difference between the average weight of tomatoes canned in the morning and the average weight of tomatoes canned in the afternoon.

Alternative hypothesis: There is a difference between the average weight of tomatoes canned in the morning and the average weight of tomatoes canned in the afternoon.

Let μ_1 denote the mean weight of cans filled in the morning, μ_2 the mean weight of cans filled in the afternoon. Then we can write our hypotheses as H_0: $\mu_1 = \mu_2$ and H_a: $\mu_1 \neq \mu_2$.

We assume that we have independent random samples from Gaussian distributions with equal variances. Then under the null hypothesis our test statistic has the t distribution with $5 + 5 - 2 = 8$ degrees of freedom.

From the description of the experiment, we do not know whether the independence assumption is reasonable. We would have to know more about how the samples were taken. Looking at Figure 11-2, we see that the variation is somewhat larger among the morning weights. (The two-sample t test is fairly *robust* to small deviations from the equal-variance assumption. This means that actual significance levels and confidence levels are close to the levels we choose, as long as the variances are not too different.) The two sets of observations are roughly symmetrical, so the assumption of Gaussian observations seems reasonable. (The two-sample t test also tends to be robust to deviations from the Gaussian assumption.) We will go ahead with a two-sample t test, aware that the assumptions must be met for the analysis to be valid.

If the significance level α equals .10, then $\alpha/2 = .05$. Referring to Table C, we see that if T has the t distribution with 8 degrees of freedom, then $P(T \leq 1.860) = .95$. The acceptance region is the interval $(-1.860, 1.860)$. The rejection region includes $(-\infty, -1.860]$ and $[1.860, \infty)$. If the test statistic is between -1.860 and 1.860, we will say the experimental results are consistent with the null hypothesis that the mean drained weight of tomatoes is the same for the morning and the afternoon. If the test statistic is less than or equal to -1.860 or else greater than or equal to 1.860, we will say the results are inconsistent with the null hypothesis, suggesting there is a difference between the mean drained weights for tomatoes canned in the morning and the mean for those canned in the afternoon.

To carry out the test, we calculate the following summary statistics for the experimental results:

Morning:	$\bar{X} = 22.7$ ounces	$s_1 = 2.31$ ounces
Afternoon:	$\bar{Y} = 20.8$ ounces	$s_2 = 1.20$ ounces
	$SE_{\bar{X}-\bar{Y}} = 1.16$ ounces	
Test statistic:	$\dfrac{\bar{X} - \bar{Y}}{SE_{\bar{X}-\bar{Y}}} = \dfrac{22.7 - 20.8}{1.16} = 1.6$	

The test statistic is in the acceptance region. We say the results are consistent with the null hypothesis, at the .10 significance level.

The p-value is the probability of seeing a test statistic as extreme as or more extreme than the one observed, if the null hypothesis were true. Since our test statistic equals 1.6, the p-value equals $P(T \leq -1.6) + P(T \geq 1.6)$, where T has the t distribution with 8 degrees of freedom. The p-value is between .1 and .2. Using the two-sample t test to compare the means of the two populations sampled, we say the results are consistent with the "no difference" null hypothesis. Even though the sample morning weights are on average larger than the sample afternoon weights, the difference is not statistically significant.

A reasonable point estimate for the difference $\mu_1 - \mu_2$ between the two population means is $\bar{X} - \bar{Y}$, the difference between the two sample means. Interval estimates for $\mu_1 - \mu_2$ are of the form $\bar{X} - \bar{Y} \pm cSE_{\bar{X}-\bar{Y}}$. If confidence level A is desired, we obtain c from the t distribution with $n_1 + n_2 - 2$ degrees of freedom to satisfy the relationship

$$A = P\left(-c < \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{SE_{\bar{X}-\bar{Y}}} < c\right)$$

Another way to write this is

$$A = P(\bar{X} - \bar{Y} - cSE_{\bar{X}-\bar{Y}} < \mu_1 - \mu_2 < \bar{X} - \bar{Y} + cSE_{\bar{X}-\bar{Y}})$$

Then we say $(\bar{X} - \bar{Y} - cSE_{\bar{X}-\bar{Y}}, \bar{X} - \bar{Y} + cSE_{\bar{X}-\bar{Y}})$ is a $100A\%$ confidence interval for $\mu_1 - \mu_2$.

In our example, we can estimate $\mu_1 - \mu_2$ with the point estimate $\bar{X} - \bar{Y} = 1.9$ ounces, the difference between the average weight in the morning and the average weight in the afternoon. A 90% confidence interval for $\mu_1 - \mu_2$ is

$$(1.9 - 1.860 \times 1.16, 1.9 + 1.860 \times 1.16) = (-.3, 4.1) \text{ ounces}$$

Zero is in this confidence interval, but near the edge.

In our formal analysis, we say the results are consistent with the null hypothesis that the mean drained weight is the same for morning and after-

noon. However, in the sample, the weights of tomatoes canned in the morning are somewhat greater on average than the weights of those canned in the afternoon. This larger average results from the two morning cans that have drained weights at least 2 ounces greater than those recorded for any of the other eight sampled cans. As manager of this factory, you would probably want more information about the canning process, both in the morning and the afternoon. You might decide to look at larger samples of tomatoes canned during these two periods of the day. You would also have to consider such things as the target (labeled) weight of these cans, acceptable variation about this target (both from the company's and the consumer's point of view), and government regulations. How would your evaluation of the canning process differ if the target drained weight for these cans was: 21.0 ounces, 22.0 ounces, 23.0 ounces, 24.0 ounces?

In Section 11-4, we look at two-sample comparisons based on ranks, when we are not willing to assume our observations come from Gaussian distributions.

11-4

Inferences About Two Measures of Central Tendency Based on a Wilcoxon–Mann–Whitney Distribution

Let's assume, as we did in the previous section, that we have two independent random samples of continuous-type observations, one from each of two populations. We assume that the two population distributions have the same shape and variation, but may differ in location. That is, they may be *shifted* away from each other. The difference between the means of the two populations describes the extent of the shift. Such a situation is illustrated in Figure 11-4. A special case is when the distributions are Gaussian with the same variance, as in Figure 11-3. We want to test the null hypothesis that the means of the two populations are equal. This is the same as saying that the two distributions have the same location and are therefore identical. We will test this null hypothesis using the Wilcoxon–Mann–Whitney test based on ranks.

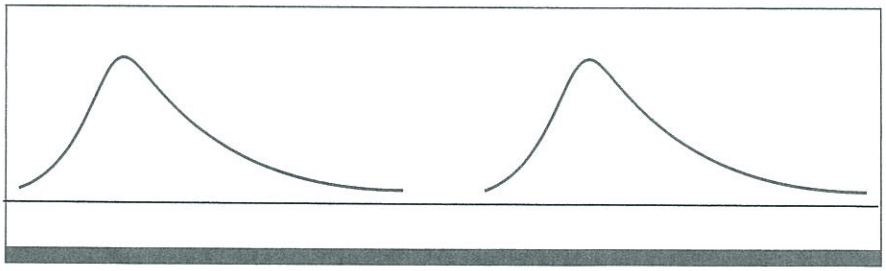

FIGURE 11-4 Illustration of two distributions with the same shape and variation, but different location

the median test is the same as *Fisher's exact test* (discussed in Section 16-5). If the sample sizes are large enough, we can carry out the median test as a chi-square test of homogeneity for a two-way frequency table (Section 16-4).

The median test is crude because we consider only whether each observation is above or below the overall median for the two combined samples. We would expect to do better using either the actual data or their ranks, when the assumptions for the corresponding tests are reasonable.

In Section 11-6, we leave the two-sample comparison of means. Instead, we talk about comparing measures of central tendency based on paired samples.

11-6 Inferences About Measures of Central Tendency Based on Paired Samples

Sometimes we have two measures of central tendency to compare, but we have paired rather than independent samples. When observations occur naturally in pairs and we can consider different pairs to be independent, we have a *paired-sample* problem.

Investigators often use paired-sample experiments in medical research. We might make a measurement on each patient at the beginning of a study, then administer a treatment. We make another measurement after treatment. For each patient, we have a pair of observations, a pretreatment measurement and a posttreatment measurement. We design the experiment so that measurements for different patients are independent. However, the two measurements on a given patient are not independent. This is a paired-sample experiment.

Sometimes the two observations in a pair are not from the same individual. Instead they come from similar individuals (people, animals or objects, depending on the experiment) matched according to characteristics that might affect experimental results. Within each pair, one individual is randomly assigned one of two experimental conditions. The other individual is assigned the other experimental condition. Then our pairs of observations are the responses measured on the paired individuals. This is called a *matched-pairs* experiment. We analyze it as a paired-sample problem.

In a matched-pairs experiment, the individuals within a pair are similar with respect to characteristics we think would affect response. Therefore, if there is no difference between the two treatments, we would expect the individuals within a pair to have similar responses. If their responses differ, it may be chance variation or it may be attributable to treatment differences. A test of hypothesis helps us decide which hypothesis seems more reasonable based on the experimental results.

The general setup for a paired-sample analysis is this. We assume that we have a random sample of pairs of continuous-type observations, (X_1, Y_1) through (X_n, Y_n). We let $d_i = X_i - Y_i$ be the difference between the two observations in the ith pair. Then we think of d_1 through d_n as a random sample of continuous-type observations from some population. We want to

make inferences about the center of this population of differences. Our analysis is based on the differences d_1 through d_n, using the one-sample techniques discussed in Chapter 10. Let's look at an example.

EXAMPLE 11-6

A high school baseball coach wanted to evaluate the effect of competition on base-running time. (This example is adapted from Loynd, 1985). Forty male high school students participated, although they were not told they were involved in an experiment. The coach matched the 40 students according to running speed. He then randomly selected one student within each matched pair to run a rival-competition trial, the other to run a self-competition trial.

A trial consisted of a student running around a regulation baseball diamond (a square 90 feet on each side), making contact with the base at each corner. The coach used a stopwatch to measure the time it took for the student to complete the run. He conducted such time trials regularly at the end of physical education class one day per week.

The coach told a student running a rival-competition trial that if he beat the time of his rival, he would be excused from half the required wind sprints on Friday. (Rivals were those in the same pair; each student knew who his rival was.) The coach told a student running a self-competition trial that if he beat his own best time, he would be excused from half the required wind sprints on Friday.

The experimental results are shown in Table 11-4. For each pair, the time

TABLE 11-4 Times (to the nearest hundredth of a second) to run a regulation baseball diamond, for 20 pairs of students. Within each pair, one student ran a self-competition trial and one student ran a rival-competition trial.

Pair	Self-competition	Rival-competition	Difference
1	16.20	15.95	.25
2	16.78	16.15	.63
3	17.38	17.05	.33
4	17.59	16.99	.60
5	17.37	17.34	.03
6	17.49	17.53	−.04
7	18.18	17.34	.84
8	18.16	17.51	.65
9	18.36	18.10	.26
10	18.53	18.19	.34
11	15.92	16.04	−.12
12	16.58	16.80	−.22
13	17.57	17.24	.33
14	16.75	16.81	−.06
15	17.28	17.11	.17
16	17.32	17.22	.10
17	17.51	17.33	.18
18	17.58	17.82	−.24
19	18.26	18.19	.07
20	17.87	17.88	−.01

FIGURE 11-7 Stem-and-leaf plots of the 20 self-competition running times and the 20 rival-competition times in Example 11-6. Each stem is in seconds and each leaf is in hundredths of a second. The pairings are ignored in this display.

Self-competition		Rival-competition	
Stem	Leaf	Stem	Leaf
15		15	
15	92	15	95
16	20	16	04 15
16	58 75 78	16	80 81 99
17	28 32 37 38 49	17	05 11 22 24 33 34 34
17	51 57 58 59 87	17	51 53 82 88
18	16 18 26 36	18	10 19 19
18	53	18	

FIGURE 11-8 Stem-and-leaf plot of the difference between the self-competition and rival-competition running times for the 20 pairs of students in Example 11-6. Each stem is in tenths of a second and each leaf is in hundredths of a second.

Difference in running times

Stem	Leaf
−.2	2 4
−.1	2
−.0	1 4 6
.0	3 7
.1	0 7 8
.2	5 6
.3	3 3 4
.4	
.5	
.6	0 3 5
.7	
.8	4

for the self-competition student, the time for the rival-competition student, and the difference between these two times are shown. The times are reported to the nearest hundredth of a second.

First let's look at some graphical displays. Figure 11-7 shows a stem-and-leaf plot of the 20 self-competition running times and a similar plot for the 20 rival-competition times, with pairings ignored. From these plots, we cannot easily compare the running times under the two experimental conditions.

In Figure 11-8, we take into account the pairing of the runners. This is a stem-and-leaf plot of the difference between the self-competition and rival-competition running times for the 20 pairs of students. The distribution of differences is fairly symmetrical, centered about some positive value. Now we can easily see that within the pairs, there is a tendency for the self-competition students to have longer running times than the rival-competition students.

The null hypothesis of interest is that type of competition (self versus

rival) does not affect running time, on average. Our analysis will be based on the 20 differences listed in Table 11-4. Let μ_d denote the mean difference in running times under the self-competition and rival-competition conditions, for a hypothetical population of pairs of students similar to those in the study. For our formal analysis, the null hypothesis states that the mean difference in running times under the two competition conditions equals 0. The alternative states that the mean difference is not 0. We can write these hypotheses as H_0: $\mu_d = 0$ and H_a: $\mu_d \neq 0$.

If we want to use a t test (Section 10-3), then we must assume that the 20 differences represent independent observations from a Gaussian distribution. The stem-and-leaf plot of differences in Figure 11-8 does not provide strong evidence either against or in favor of the Gaussian assumption. The reasonableness of the independence assumption depends on how the experiment was conducted. We will proceed with our analysis, aware that we must be cautious in interpreting results.

We calculate the sample average and standard error based on the 20 differences in running times:

$$\bar{d} = .2045 \text{ second} \quad \text{and} \quad \text{SE}_{\bar{d}} = .0672 \text{ second}$$

From Section 10-3, we know that for a t test our test statistic is

$$\frac{\bar{d} - 0}{\text{SE}_{\bar{d}}} = \frac{.2045 - 0}{.0672} = 3.04$$

since $\mu_d = 0$ under the null hypothesis. Our p-value is $P(T \leq -3.04 \text{ or } T \geq 3.04)$ where T is a random variable having the t distribution with 19 degrees of freedom. From Table C, we see that the p-value is less than .01. The experimental results are inconsistent with the null hypothesis, with a p-value less than .01.

Again using the t distribution with 19 degrees of freedom, we find a 95% confidence interval for the mean difference between self-competition and rival-competition running times: $.2045 \pm (2.093)(.0672)$ or (.06, .35) second. We estimate that students run the bases an average of .06 to .35 second faster under rival-competition than under self-competition.

What do these experimental results suggest to you about a runner's performance when trying to beat his or her own best time compared to his or her performance when trying to beat a rival's best time? Would you be willing to guess how a runner would perform when trying to beat his or her own best time, as compared with trying to beat a well-matched rival in an actual race? In general, what might this experiment suggest about human performance under different types of competition?

The Paired t Test

When we use a t test to ask whether an average difference is 0, we say we are doing a paired t test.

> A **paired t test** is a t test applied to the differences in a paired-sample problem.

If we make fewer assumptions about the sample, then we may choose to do a Wilcoxon signed rank test or a sign test based on the differences (Exercise 11-25).

How do we choose between paired-sample and two-sample experimental designs? When we plotted the 20 self-competition running times separately from the 20 rival-competition times in Figure 11-7, it was difficult to see differences between the two distributions. The variation among running times within a single experimental condition was so large that differences between the experimental conditions were hard to see. Suppose we *incorrectly* ignore the pairing and analyze the experimental results as a two-sample problem. If we assume that we have two independent random samples from Gaussian distributions with equal variances, then we use a two-sample t test. To test the null hypothesis that the self-competition and rival-competition means are equal, our test statistic would be

$$\frac{17.434 - 17.229}{\sqrt{.469\left(\frac{1}{20} + \frac{1}{20}\right)}} = .9$$

Here, 17.434 is the sample average running time for the self-competition condition, 17.229 is the sample average for the rival-competition condition, and .469 is the pooled estimate of the common variance assumed for the two conditions. With our assumptions, the test statistic would have the t distribution with $20 + 20 - 2 = 38$ degrees of freedom under the null hypothesis. The p-value is greater than .30. This is consistent with the null hypothesis that there is no difference in average base-running times under the two experimental conditions. What we saw in examining Figure 11-7 is borne out by the test of hypotheses. The variation within experimental groups is so large that it obscures any differences between experimental groups.

We *correctly* took the pairing into account in Figure 11-8. We saw that within a pair, the self-competition running time tended to be longer than the corresponding rival-competition running time. The results of the paired-sample t test supported this observation.

The experimenter had a choice between a two-sample design and a paired-sample design. If he had chosen a two-sample design, he would have randomly selected 20 students to run under the self-competition condition, the other 20 to run under the rival-competition condition. His total sample size would be 40. For a two-sample t test, he would have 38 degrees of freedom.

Instead of a two-sample design, the experimenter opted for a paired design. He matched students by running speed, to control for differences in running ability, an extraneous variable that could affect the results of the experiment. He hoped that he would then be able to get a clearer picture of any differences between experimental conditions that might exist.

The experimenter pays a price for the increased power he expects from a paired design. The paired-sample analyses were based on 20 differences,

effectively half the sample size he would have with a two-sample design. (For the paired-sample t test we had 19 degrees of freedom.)

The paired-sample design is a good idea when variation within groups is large relative to variation between groups. But the two-sample design offers twice the effective sample size for analysis. So, we may prefer the two-sample design when within-group variation is not so large.

There are other times when a two-sample design is preferable. An investigator may not be able to find suitable pairs among the participants in a study, or may not know all the factors that should form the basis for pairing. In these situations, he or she may choose a two-sample design. The investigator hopes that randomization will approximately balance the two experimental groups with respect to extraneous factors.

Summary of Chapter 11

This chapter is about comparing two measures of central tendency. We distinguished between two-sample designs and paired-sample designs. We have a two-sample design if we have two independent random samples, one from each of the two populations of interest; we use the samples to compare the centers of the two populations. In a paired-sample design, we have independent pairs of observations; we make inferences about the mean (or median) difference between observations within pairs. For two-sample and paired-sample designs, the method of analysis we choose depends on sample size and on the assumptions we are willing to make about the sample(s).

With a paired-sample design, we assume that we have a random sample of pairs of continuous-type observations (X_1, Y_1) through (X_n, Y_n). We define the difference between the observations within the ith pair by $d_i = X_i - Y_i$. We want to make inferences about the center of the distribution of these differences. To carry out an analysis of such a paired-sample experiment, we use one-sample techniques on the differences d_1 through d_n.

For a two-sample design, if we have two independent random samples and the sample sizes are large, we base inferences about the population means on the standard Gaussian distribution. A special case is when sample sizes are large and we wish to compare two proportions.

If we have two independent random samples from Gaussian distributions with equal variances, we can compare the two population means using the two-sample t test. Tests of hypotheses and confidence intervals for the difference between the two population means are based on a t distribution.

Suppose we have two independent random samples of continuous-type observations from distributions that have the same shape and variation, but possibly different locations. Then we can test for equality of the two population means using ranked data. We base tests of hypotheses and confidence intervals for the difference between the population means on a Wilcoxon–Mann–Whitney distribution. The Wilcoxon–Mann–Whitney distribution for sample

sizes n_1 and n_2 is derived from the probability model for an experiment in which ranks 1 through $n_1 + n_2$ are randomly divided into two groups, one of size n_1 and one of size n_2.

If the assumptions for the two-sample t test are met, it is more powerful (more likely to reject the null hypothesis when the alternative is true) than the Wilcoxon–Mann–Whitney test.

If we have two independent random samples of continuous-type observations, we can test for equality of the two population medians using the median test. We base inferences on a hypergeometric distribution for small samples. For large samples, we can use the chi-square distribution with 1 degree of freedom (see Section 16-4). If the assumptions for the Wilcoxon–Mann–Whitney test are met, it is much more powerful (more likely to reject the null hypothesis when the alternative is true) than the median test.

Procedures based on a t distribution are parametric because we assume that the observations follow Gaussian distributions. Inferences based on a hypergeometric distribution or on a Wilcoxon–Mann–Whitney distribution, as well as those based on the standard Gaussian distribution for large samples, are nonparametric because we assume no particular type of probability distribution for the observations.

For a two-sample design, a confidence interval provides an interval estimate or range of reasonable values for the difference between two population means, $\mu_1 - \mu_2$. If the confidence level is A, we say we have a $100A\%$ confidence interval for the difference between the two population means. The interpretation of such a confidence interval is that if we repeated the sampling process in the same way many times and calculated a $100A\%$ confidence interval for $\mu_1 - \mu_2$ each time, we would expect about $100A\%$ of these confidence intervals to contain $\mu_1 - \mu_2$ and about $100(1 - A)\%$ not to contain $\mu_1 - \mu_2$. When we calculate a specific confidence interval, it either contains the difference between the two population means or it does not. The confidence level refers to what we would expect if we repeated the sampling process many times.

```
TEST OF MU = 0.0000 VS MU N.E. 0.0000

              N      MEAN     STDEV   SE MEAN        T     P VALUE
diff         20    0.2045    0.3004    0.0672     3.04      0.0067
```

FIGURE M11-3 Paired *t* test for Example 11-6

Carrying Out a Two-Sample Comparison of Means Based on Large Samples

For a two-sample test based on large samples, we can use the TWOT command, without the POOLED subcommand. Minitab does not have a command for the median test. To use Minitab for a comparison of two proportions based on large samples, see the Minitab Appendix for Chapter 16 for the chi-square test of homogeneity.

Carrying Out a Paired-Sample Comparison of Means

We can carry out paired-sample analyses using the appropriate one-sample commands. Suppose we have the data for Example 11-6 in our worksheet. Column 1 contains a code for the pair of runners; column 2, the self-competition time; column 3, the rival-competition time. For a paired-sample *t* test that the mean difference in times under the two conditions is 0, we can use these commands:

```
MTB>  let c4=c2-c3
MTB>  name c4 'diff'
SUBC> ttest 'diff'
```

If we specify no value for the mean under the null hypothesis in the TTEST command, Minitab uses 0. The output is in Figure M11-3.

We can get a confidence interval for the mean difference in times under the two conditions using the TINTERVAL command on the differences in column 4.

Exercises for Chapter 11

In each exercise, describe the population(s) sampled, whether real or hypothetical. Graph the data in any way that seems helpful. For each statistical procedure, state the assumptions that make the analysis appropriate. Do the assumptions seem reasonable? What additional information about the experiment would you like to have? Discuss the results of your analysis.

EXERCISE 11-1 In 10 trials, the number of female mosquitos collected coming to bite a human and the number killed in an electrocuting device (over a 2-hour period in the same yard) were recorded (Nasci et al., 1983). The results are given below.

Trial	Number of female mosquitos captured: Electrocuting device	Number of female mosquitos captured: Human bait
1	31	94
2	44	146
3	129	194
4	15	54
5	11	39
6	49	90
7	151	172
8	30	219
9	12	60
10	17	21

a. Plot these observations.

b. How does the insect electrocuting device compare with human bait in attracting female mosquitos? State and test appropriate hypotheses.

c. Calculate an interval estimate for the mean difference in number of female mosquitos captured by the two methods. Discuss your findings.

EXERCISE 11-2

Do drugs that suppress abnormal heart rhythms increase the likelihood of sudden cardiac death? A multicenter clinical trial enrolled patients who had had a heart attack and developed cardiac arrhythmias no more than 2 years before start of the study. Seven hundred thirty patients received either encainide or flecainide for cardiac arrhythmia and 730 patients received placebo (*Science News,* April 29, 1989, volume 135, page 260).

The trial began in June 1987. Early review of results by a safety monitoring board found that 33 of the 730 patients receiving encainide or flecainide had experienced either sudden cardiac death or a nonfatal heart attack. Nine of the 730 patients in the placebo group had suffered a nonfatal heart attack or died suddenly from cardiac problems. The trial was stopped based on this early review.

a. Compare the proportion of patients suffering heart attack or sudden cardiac death in the two groups (treatment and placebo). State and test appropriate hypotheses.

b. Calculate a confidence interval for the difference between the proportions for the two groups.

c. Discuss your results.

EXERCISE 11-3

A study was designed to compare the effects of two types of ammunition (.22 Long Rifle versus .22 Magnum) on target-shooting accuracy (Snow, 1986). An experienced shooter used a revolver with interchangeable cylinders to shoot at a circular target 25 yards away. A trial consisted of five shots. The score for a

trial was the total score for the five shots. Sixteen trials were run, eight with each type of ammunition. The order in which the two types of ammunition were used was determined by a random process. The results of the experiment are listed here.

Total score of five shots	
.22 Magnum	.22 Long Rifle
42	41
43	43
46	41
47	41
46	40
47	40
39	45
47	47

a. Plot these observations.

b. Is the average score the same for the two types of ammunition? State and test appropriate hypotheses.

c. Calculate an interval estimate for the difference in mean score for the two types of ammunition.

EXERCISE 11-4

Two analysts each made eight independent determinations of the melting point of hydroquinone (Duncan, 1974, pages 575–576; from Wernimont, 1947, page 8).

Analyst	Melting point determination (°C)				
1	174.0	173.5	173.0	173.5	171.5
	172.5	173.5	173.5		
2	173.0	173.0	172.0	173.0	171.0
	172.0	171.0	172.0		

a. Plot the observations.

b. Test the null hypothesis that the mean determination of melting point is the same for the two analysts.

c. Calculate a confidence interval for the difference in mean melting point determinations for the two analysts.

EXERCISE 11-5

Researchers wanted to study effects of regular alcohol consumption (Jerome Hojnacki, 1986, personal communication). The participants in the experiment were 20 adult male squirrel monkeys, of similar age and good health. The

researchers randomly divided the monkeys into two equal sized groups. Monkeys in the alcohol group consumed a steady diet of 12% ethyl alcohol (ethyl alcohol constituted 12% of their total calories each meal). Monkeys in the control group did not consume alcohol. At the end of the treatment period, the researchers measured plasma estrogen (in nanograms/deciliter) for each monkey. The results are shown below.

Alcohol:	3.17	2.52	2.59	4.25	3.27	4.92
	5.46	2.83	4.80	2.26		
Control:	6.57	5.81	5.63	5.75	4.54	5.35
	4.16	5.12	4.69	4.52		

a. Plot the observations.

b. Use the median test to test the null hypothesis that the median plasma estrogen level is the same for the two groups of monkeys.

c. Use the Wilcoxon–Mann–Whitney test to test the null hypothesis that the median plasma estrogen level is the same for the two groups of monkeys.

d. Use the two-sample t test to test the null hypothesis that the mean plasma estrogen level is the same for the two groups of monkeys.

e. Compare the results of your analyses in parts (b), (c), and (d). Discuss your findings.

EXERCISE 11-6

In the experiment described in Exercise 11-5, researchers also measured plasma testosterone level (nanograms/deciliter) for each monkey at the end of the treatment period (Jerome Hojnacki, 1986, personal communication). The results are shown below.

Alcohol:	313.99	152.06	145.64	128.86	262.16
	251.29	505.55	94.79	157.49	171.81
Control:	632.92	308.56	1,239.68	440.38	233.02
	142.67	84.91	342.63	1,005.66	735.61

a. Plot the observations.

b. Use the median test to test the null hypothesis that the median testosterone level is the same for the two groups of monkeys.

c. Use the Wilcoxon–Mann–Whitney test to test the null hypothesis that the median testosterone level is the same for the two groups of monkeys.

d. Use the two-sample t test to test the null hypothesis that the mean testosterone level is the same for the two groups.

e. The equal-variance assumption of the two-sample t test is violated for these samples. Take the logarithm base-10 of each observation. Plot these trans-

formed observations, for the two groups. Are the variances for the two groups more similar for these transformed observations? When we transform observations to get more similar variances for different groups, we say we are making a *variance-stabilizing transformation.*

f. Use the two-sample t test to test the null hypothesis that the mean of the logarithm base-10 of testosterone level is the same for the two groups.

g. Compare the results of your tests of hypotheses in parts (b), (c), (d), and (f). Discuss your findings. Note that the test in part (f) applies to the *transformed* observations.

EXERCISE 11-7

Sputum histamine levels (μg/g dry weight sputum) are shown below for 9 allergic and 13 nonallergic people, all smokers (Hollander and Wolfe, 1973, page 74; a subset of data in Thomas and Simmons, 1969).

Allergics:	31.0	39.6	64.7	65.9	67.9	100.0
	102.4	1,112.0	1,651.0			
Nonallergics:	4.7	5.2	6.6	18.9	27.3	29.1
	32.4	34.3	35.4	41.7	45.5	48.0
	48.1					

a. Plot the observations. Because of the wide range for the allergic people, you may wish to take the logarithm base-10 of each value (for both groups) before plotting.

b. Use the median test to test the null hypothesis that the median sputum histamine level is the same for the allergics and nonallergics.

c. Use the Wilcoxon–Mann–Whitney test to test the null hypothesis that the median sputum histamine level is the same for the two groups.

d. The equal-variance assumption of the two-sample t test is violated for these two groups. When you plot the logarithm base-10 of the observations, does the variation seem similar in the two groups? If so, we call this transformation a *variance-stabilizing transformation.* Use the two-sample t test to test the null hypothesis that the mean of the logarithm base-10 of sputum histamine levels is the same for the allergics and nonallergics.

e. Discuss the results of your tests of hypotheses in parts (b), (c), and (d). Note that your test in part (d) applies to the *transformed* observations.

EXERCISE 11-8

This study compared the effectiveness of a bronchodilating aerosol administered by hand and by an automatic inhalation device (Box, Hunter, and Hunter, 1978, page 158; from a larger study reported by F. J. McIlneath and B. M. Cohen in *J. Med.,* 1970, volume 1, page 229). Specific airway resistance 30 minutes after administration is shown below for each of 12 patients using hand administration and 12 patients using an automatic inhalation device.

Hand:	17.00	22.80	21.60	20.40	11.20	14.00	52.25
	7.50	12.20	18.85	6.05	4.05		
Automatic:	11.60	11.60	13.65	17.22	8.25	6.20	41.50
	6.96	8.40	9.00	5.18	3.00		

a. Plot the observations.

b. Test the null hypothesis that mean airway resistance is the same 30 minutes after administration of the aerosol, for the two methods.

c. Calculate a confidence interval for the difference in mean airway resistance after 30 minutes for the two methods.

EXERCISE 11-9

Does diet restriction prolong life? In this experiment, researchers examined the influence of different diets on the aging process in rats (Berger, Boos, and Guess, 1988; from Yu et al., 1982). Lifetimes (in days) of rats on a restricted diet and rats on an unrestricted (*ad libitum*) diet are shown below.

Restricted diet (sample size = 106, mean = 968.7 days, sample standard deviation = 284.6 days)										
105	193	211	236	302	363	389	390	391	403	530
604	605	630	716	718	727	731	749	769	770	789
804	810	811	833	868	871	875	893	897	901	906
907	919	923	931	940	957	958	961	962	974	979
982	1,001	1,008	1,010	1,011	1,012	1,014	1,017	1,032	1,039	1,045
1,046	1,047	1,057	1,063	1,070	1,073	1,076	1,085	1,090	1,094	1,099
1,107	1,119	1,120	1,128	1,129	1,131	1,133	1,136	1,138	1,144	1,149
1,160	1,166	1,170	1,173	1,181	1,183	1,188	1,190	1,203	1,206	1,209
1,218	1,220	1,221	1,228	1,230	1,231	1,233	1,239	1,244	1,258	1,268
1,294	1,316	1,327	1,328	1,369	1,393	1,435				
Unrestricted diet (sample size = 90, mean = 682.3 days, sample standard deviation = 134.3 days)										
89	104	387	465	479	494	496	514	532	533	536
545	547	548	582	606	609	619	620	621	630	635
639	648	652	653	654	660	665	667	668	670	675
677	678	678	681	684	688	694	695	697	698	702
704	710	711	712	715	716	717	720	721	730	731
732	733	735	736	738	739	741	743	746	749	751
753	764	765	768	770	773	777	779	780	788	791
794	796	799	801	806	807	815	836	838	850	859
894	963									

a. Plot these observations.

b. Test the null hypothesis that mean lifetime is the same for rats on the two diets.

c. Calculate a confidence interval for the difference in mean lifetimes for rats on the two diets.

d. Discuss your findings.

EXERCISE 11-10 Cotton dust in textile factories can lead to a respiratory problem known as byssinosis. A survey of cotton textile workers was carried out to evaluate byssinosis problems (Higgins and Koch, 1977). Of 2,916 men surveyed, 128 had byssinosis complaints. Of 2,503 women surveyed, 37 had byssinosis complaints.

a. Is there a difference between the proportions of men and women with byssinosis complaints? State and test appropriate hypotheses.

b. Calculate a confidence interval for the difference between the proportions of men and women with byssinosis complaints.

c. Calculate a confidence interval for the proportion of men with byssinosis complaints.

d. Calculate a confidence interval for the proportion of women with byssinosis complaints.

EXERCISE 11-11 Cirrhotic patients with bleeding problems were randomly divided into two groups. Patients in one group underwent a standard operation (nonselective shunt). Patients in the other group underwent a new operation (selective shunt). The response is maximal rate of urea synthesis, a measure of liver function; low values correspond to poor liver function. Responses before and after surgery are shown below (Brogan and Kutner, 1980; from *Annals of Surgery,* 1978, volume 188, pages 271–282).

	Maximal rate of urea synthesis (mg urea N/hr/kg $BW^{3/4}$)			Maximal rate of urea synthesis (mg urea N/hr/kg $BW^{3/4}$)	
Patient	Before surgery	After surgery	Patient	Before surgery	After surgery
New operation (selective shunt)			**Standard operation (nonselective shunt)**		
1	51	48	9	34	16
2	35	55	10	40	36
3	66	60	11	34	16
4	40	35	12	36	18
5	39	36	13	38	32
6	46	43	14	32	14
7	52	46	15	44	20
8	42	54	16	50	43
			17	60	45
			18	63	67
			19	50	36
			20	42	34
			21	43	32

a. Plot the observations in any way that seems helpful.

b. Compare average liver function before surgery for the two groups of patients. State and test appropriate hypotheses.

c. Compare before- and after-surgery liver function for patients undergoing the new operation. State and test appropriate hypotheses. Calculate a confidence interval for the mean difference in liver function for this group of patients.

d. Compare before- and after-surgery liver function for patients undergoing the standard operation. State and test appropriate hypotheses. Calculate a confidence interval for the mean difference in liver function for this group of patients.

e. Compare the change in liver function for the two groups of patients. State and test appropriate hypotheses.

f. Discuss your findings.

EXERCISE 11-12 Change in pupil diameter following treatment is shown below for 11 volunteers (Box, Hunter, and Hunter, 1978, page 160; from H. W. Elliott, G. Navarro, and N. Nomof, *J. Med.,* 1970, volume 1, page 77). Six volunteers received several doses of morphine. Five volunteers received several doses of nalbuphine.

Treatment	Change in pupil diameter (millimeters)					
Morphine	.08	.8	1.0	1.9	2.0	2.4
Nalbuphine	−.3	.0	.2	.4	.8	

a. Plot these observations.

b. Test the null hypothesis that the mean change in pupil diameter is the same for the two drugs, using a t distribution.

c. Repeat part (b), using a Wilcoxon–Mann–Whitney distribution.

d. Test the null hypothesis that the median change in pupil diameter is the same for the two drugs, using a hypergeometric distribution.

e. Compare the results of parts (b), (c), and (d).

EXERCISE 11-13 Scientists have developed a new method of determining serum iron concentration that is faster and requires smaller samples than an older method. To compare the accuracy of the two methods, researchers made replicate analyses of control sera containing a concentration of 105 μg serum iron per 100 milliliters. The measurements of serum iron concentration (μg/100 ml) are shown below (Hollander and Wolfe, 1973, pages 85–86; a portion of the data of Jung and Parekh, 1970).

New:	107	108	106	98	105	103	110	105
	104	100	96	108	103	104	114	114
	113	108	106	99				
Old:	111	107	100	99	102	106	109	108
	104	99	101	96	97	102	107	113
	116	113	110	98				

a. Plot the observations.

b. Test the null hypothesis that average serum iron determination is the same using both methods.

c. Calculate a confidence interval for the difference in mean serum iron determination for the two methods.

d. Calculate a confidence interval for the mean serum iron determination using the new method.

e. Calculate a confidence interval for the mean serum iron determination using the old method.

f. Discuss your findings.

EXERCISE 11-14 Investigators tested the effect of a drug with antiarrhythmic properties on patients with frequent premature ventricular contractions (PVCs). For each of 12 patients, the researchers recorded the number of PVCs during a 1-minute electrocardiograph, both before and after treatment with the drug (Berry, 1987).

Patient	Pre-treatment	Post-treatment	Pre – Post (decrease in PVCs)
1	6	5	1
2	9	2	7
3	17	0	17
4	22	0	22
5	7	2	5
6	5	1	4
7	5	0	5
8	14	0	14
9	9	0	9
10	7	0	7
11	9	13	−4
12	51	0	51

a. Plot the observations in any way that seems helpful.

b. Test the null hypothesis of no average change in PVC count before and after treatment, based on a t distribution. Calculate a confidence interval for the average decrease in PVC count.

c. Repeat part (b), based on a Wilcoxon signed rank distribution.

d. Test the null hypothesis of zero median change in PVC count before and after treatment, based on a binomial distribution. Calculate a confidence interval for the median decrease in PVC count.

e. Compare the results of parts (b), (c), and (d).

f. Patient 12 is an outlier in the sense that his pretreatment PVC count is far

from the others. Delete the observations for patient 12 and repeat parts (b), (c) and (d). Compare your results.

EXERCISE 11-15 Survival times (units not given) are shown here for skin grafts on 11 burn patients. Each patient received both a closely matched graft and a poorly matched graft (O'Brien and Fleming, 1987; from Woolson and Lachenbruch, 1980; slightly modified from original data in Batchelor and Hackett, 1970). A + indicates a graft still viable at the recorded time.

	Survival time for		
Patient	Closely matched graft	Poorly matched graft	Difference
1	37	29	8
2	19	13	6
3	57+	15	42+
4	93	26	67
5	16	11	5
6	22	17	5
7	20	26	−6
8	18	21	−3
9	63	43	20
10	29	15	14
11	60+	40	20+

a. Plot the observations in any way that seems helpful.

b. Test the null hypothesis of no median difference in survival for the two types of grafts.

c. Calculate a confidence interval for the median difference in survival for the two types of grafts.

d. Calculate a confidence interval for median survival of closely matched grafts.

e. Calculate a confidence interval for median survival of poorly matched grafts.

EXERCISE 11-16 Experimenters wanted to see if practice and training can affect hypnotic susceptibility. Each of six volunteers was evaluated with Stanford profile scales of hypnotic susceptibility by a hypnotist (not one of the experimenters). Then each volunteer went through hypnotic training with one of the experimenters. After this training, each volunteer was evaluated by a different hypnotist (again not one of the experimenters) using equivalent scales of hypnotic susceptibility. Results are shown below (Hollander and Wolfe, 1973, page 45; part of a larger data set of Cooper et al., 1967). A low score indicates low hypnotic susceptibility.

Volunteer	Average score before training	Average score after training
1	10.5	18.5
2	19.5	24.5
3	7.5	11.0
4	4.0	2.5
5	4.5	5.5
6	2.0	3.5

a. Plot these observations.

b. Why was the evaluation of hypnotic susceptibility done by hypnotists other than the experimenters?

c. Does this experiment suggest that training affects hypnotic susceptibility? State and test appropriate hypotheses.

d. Calculate a confidence interval for the average difference in scores before and after training. Discuss your findings.

EXERCISE 11-17 In a study of antibodies associated with malaria, investigators treated adult volunteers for malaria, achieving what is known as a radical cure (Hoffman et al., 1987). By 98 days after radical cure, 60 of the 83 volunteers completing the study were infected with the malaria parasite and 23 were not. One measure of antibody activity is ISI (inhibition of sporozoite invasion of hepatoma cells). Based on ISI, investigators classified the volunteers as having either high antibody levels ($> 75\%$ ISI) or low antibody levels ($\leq 75\%$ ISI). With this measure, 26 of the 60 infected volunteers and 9 of the 23 uninfected volunteers had high levels of antibodies at the end of the study.

a. Is there any difference between the infected and uninfected volunteers with respect to this measure of antibody levels? State and test appropriate hypotheses.

b. Calculate a confidence interval for the difference in proportions with high antibody levels for the infected and uninfected volunteers. Discuss your findings.

EXERCISE 11-18 Do the two anesthetics enflurane and halothane have the same effects on cardiovascular performance? To answer this question, investigators studied 19 children undergoing elective noncardiac surgery. Ten children received enflurane and nine received halothane. Echocardiographs provided values for shortening fraction and mean blood pressure, two measures of cardiovascular performance. Values of shortening fraction and mean blood pressure before and during low dose of the anesthetic are listed below for each child (Hui and Rosenberg, 1985; from Barash et al., 1979).

	Shortening fraction (percent)		Blood pressure (mm Hg)	
Child	Before	At low dose	Before	At low dose
Enflurane				
1	25	23	79	74
2	39	24	71	76
3	27	30	92	87
4	30	28	90	79
5	31	28	82	69
6	35	37	83	73
7	23	23	77	67
8	27	29	69	63
9	27	21	90	86
10	25	17	87	76
Halothane				
1	34	21	90	89
2	30	29	83	77
3	33	31	68	63
4	23	26	70	69
5	27	26	76	72
6	26	27	64	56
7	28	22	75	67
8	27	29	68	67
9	33	29	79	77

a. Plot these observations in any way that seems helpful.

b. Is there a difference on average between the before and low-dose measurements of shortening fraction for children on enflurane? State and test appropriate hypotheses.

c. Is there a difference on average between the before and low-dose measurements of shortening fraction for children on halothane? State and test appropriate hypotheses.

d. Is the change in shortening fraction for children on enflurane the same as for children on halothane, on average?

e. Is there a difference on average between the before and low-dose measurements of blood pressure for children on enflurane?

f. Is there a difference on average between the before and low-dose measurements of blood pressure for children on halothane?

g. Is the change in blood pressure for children on enflurane the same as for children on halothane, on average?

h. Summarize your findings. Do the two anesthetics appear to have the same effects on cardiovascular performance as measured by shortening fraction and mean blood pressure?

i. In answering parts (b)–(g) you use six tests of hypotheses. Should you be concerned that the overall significance level of the tests taken together is not the same as the significance level(s) you used for the separate tests? We address this question in the next chapter.

EXERCISE 11-19 Eleven healthy young volunteers participated in a study to compare nasal clearance times before and after jogging (Cederlund, Camner, and Svartengren, 1987). Nasal clearance refers to the process by which inhaled particles are transported to the pharynx and then swallowed. People born with impaired nasal clearance tend to be at increased risk of respiratory problems.

This study was designed to investigate the effects of jogging, with associated increased rate of respiration, on nasal clearance time. Nasal clearance was recorded as a transport time: the time between placement of a saccharin particle in the nasal passage and first indication of a sweet taste. Nasal clearance was measured in this way for each volunteer immediately before jogging and approximately 1 hour later after jogging from 8 to 10 kilometers. Personal characteristics and experimental results for the 11 volunteers are shown here.

					Transport time in minutes	
Volunteer	Sex	Age (years)	Smoker	History of allergy	Before jogging	After jogging
1	F	22	No	No	7.5	33.0
2	F	20	No	No	9.5	28.0
3	F	21	No	No	19.5	>45
4	M	30	No	No	4.5	10.5
5	M	19	Yes	Yes	18.0	19.0
6	M	19	Yes	No	13.0	10.0
7	M	19	No	Yes	12.0	17.5
8	M	20	No	No	26.0	29.0
9	M	21	No	No	6.5	17.0
10	F	19	No	No	10.5	40.0
11	F	19	No	No	8.5	13.0

Volunteer 3 reported no sweet taste within 45 minutes of receiving the saccharin particle. The investigators considered this to be an experimental failure and chose to exclude her results from the analysis. Therefore, we are left with the data on the remaining 10 volunteers. In answering parts (a)–(f), ignore the personal information given for the volunteers, considering only the before- and after-jogging transport times.

a. Plot the observations in any way that seems helpful.

b. State null and alternative hypotheses.

c. Test the hypotheses in part (b) and calculate an interval estimate, using a binomial distribution. Discuss the relationship between the test of hypotheses and the confidence interval.

d. Test the hypotheses in part (b) and calculate an interval estimate, using a Wilcoxon signed rank distribution. Discuss the relationship between the test of hypotheses and the confidence interval.

e. Test the hypotheses in part (b) and calculate an interval estimate, using a t distribution. Discuss the relationship between the test of hypotheses and the confidence interval.

f. Compare the results of the tests of hypotheses in parts (c), (d), and (e). Discuss your findings.

g. Discuss the results of the experiment, taking into consideration now the personal characteristics of the volunteers.

h. The investigators chose to exclude the results for volunteer 3 from the analysis because she did not report a sweet taste within 45 minutes. We must always be very cautious when excluding cases. Data values that are different from the others may represent extreme ends of the probability distribution for the observations, or they may represent unusual conditions. Either way, it is unwise to exclude such values without careful thought. In this experiment, the investigators believed that if a volunteer had not detected the sweet taste within 45 minutes, the results should be excluded. (Because of the solubility of saccharin, it would be difficult to interpret responses after 45 minutes.) How does the exclusion of results for this volunteer affect your interpretations of the experimental results?

EXERCISE 11-20 Eight healthy young volunteers participated in a control study of nasal clearance times (Cederlund, Camner, and Svartengren, 1987). The original study was designed to compare nasal transport times before jogging and approximately 1 hour later (see Exercise 11-19). In this control study, nasal transport times were measured while subjects were at rest, the second measurement 1 hour after the first.

	Transport time (minutes)	
Volunteer	First measurement	Second measurement
2	18.5	20.5
4	12.5	20.5
5	25.0	20.0
6	16.0	12.0
7	12.5	10.5
9	15.5	14.0
10	10.0	11.0
11	15.5	15.0

a. Plot the observations in any way that seems helpful.

b. Is there evidence of differences in nasal transport times taken 1 hour apart, on average? State and test appropriate hypotheses.

c. Calculate an interval estimate for the mean difference in transport times taken 1 hour apart. Discuss your findings.

EXERCISE 11-21 Investigators wanted to determine physical characteristics of urine with and without calcium oxalate crystals. They measured calcium concentrations (in millimoles/liter) from urine specimens of 34 men with calcium oxalate crystals and 45 men without crystals (data were obtained from the laboratory of Dr. James S. Elliot and contributed by D. P. Byar to a collection of problems in Andrews and Herzberg, 1985, pages 249–251). The results are shown below.

	Calcium concentrations in urine (millimoles/liter)					
Men with crystals	6.96	13.00	5.54	6.19	7.31	14.34
	4.74	2.50	1.27	4.18	3.10	3.01
	6.81	8.28	2.33	7.18	5.67	12.68
	8.94	3.16	3.30	6.99	.65	4.18
	4.45	.27	7.64	6.63	8.53	9.04
	.58	7.82	12.20	9.39		
	(sample size = 34, mean = 6.143, sample standard deviation = 3.637)					
Men without crystals	2.45	4.49	2.36	2.15	1.16	3.34
	1.40	8.48	1.16	2.21	1.93	1.27
	1.03	1.47	1.53	5.09	1.05	2.03
	7.68	1.45	5.16	.81	1.32	1.55
	1.52	.77	2.17	.17	.83	3.04
	1.06	3.93	5.38	3.53	4.54	3.98
	1.02	3.46	1.19	5.64	2.66	1.22
	2.64	2.31	4.49			
	(sample size = 45, mean = 2.625, sample standard deviation = 1.863)					

a. Graph the calcium concentrations for the two groups of men. Describe and compare the two distributions.

b. State and test appropriate hypotheses to compare mean calcium concentrations for the two groups.

c. Calculate an interval estimate for the difference between the two population means.

d. Discuss your results.

EXERCISE 11-22 In a study comparing a cost-reduced product with a current product, experimenters asked 200 judges their preference on a pair of samples, A and B. Half of the judges tasted sample A first and the other half tasted sample B first. Ignoring the order of tasting, the preference results are shown below (contributed by M. B. Carroll of General Foods Corporation to a collection of problems in Andrews and Herzberg, 1985, pages 189–193):

Preference	Number of judges
Sample A	98
Sample B	88
No preference	14

a. Is there evidence that one sample is preferred over the other? State and test appropriate hypotheses.

b. Calculate an interval estimate for the proportion of judges preferring sample A. Discuss your findings.

EXERCISE 11-23 Investigators studied a method of measuring aflatoxin concentration in contaminated peanuts. (In sampling inspection, inspectors want to protect consumers while not rejecting too many good peanuts.) They ground the peanuts into meal and divided the meal into separate samples. They blended each sample in a chemical solution. For each sample, the experimenters divided the blend equally among 16 centrifuge bottles. The determination of aflatoxin concentration for each bottle (units not given) is shown below (Quesenberry, Whitaker, and Dickens, 1976; from Waltking, Bleffert, and Kiernan, 1968).

Sample 1:	95.33	55.94	72.01	58.96	114.62	41.64
	98.76	53.62	90.23	91.92	66.88	91.81
	100.37	77.26	91.56	66.25		
Sample 2:	20.04	20.30	24.69	22.26	24.92	21.45
	19.44	24.04	24.40	19.85	11.88	24.34
	15.32	14.85	23.10	22.21		

Compare the aflatoxin determinations for the two samples. You may find it useful to take the logarithm of each value for part of your analysis.

EXERCISE 11-24 Recall that in Example 9-2, the student wished to compare the median distance pedaled on an exercise cycle under two experimental conditions: with a Walkman and without a Walkman.

a. Make the comparison using the Wilcoxon–Mann–Whitney test.

b. Make the comparison using the two-sample t test.

c. Compare the results in parts (a) and (b) with what we found using the median test in Example 9-2.

EXERCISE 11-25 Refer to the base-running experiment in Example 11-6.

a. Test the hypotheses and calculate a confidence interval based on ranks.

b. Test the hypotheses using the sign test. Calculate a confidence interval based on a binomial distribution.

c. Compare the tests of hypotheses and confidence intervals in parts (a) and (b) with what we found using a t distribution in Example 11-6.

EXERCISE 11-26 In the appendix on the Wilcoxon–Mann–Whitney distributions at the end of the book, we find the Wilcoxon–Mann–Whitney distribution of T_1 for sample sizes 2 and 4. Show that T_2 has this same probability distribution.

EXERCISE 11-27 Find the Wilcoxon–Mann–Whitney distribution for samples of size 3 and 3.

EXERCISE 11-28 Refer to Example 11-3, comparing weights of tomatoes canned in the morning and afternoon.

a. Test the null hypothesis that the median weight is the same in the morning and the afternoon, using the median test.

b. Repeat part (a) using the Wilcoxon–Mann–Whitney test.

c. Compare the results in parts (a) and (b) with what we found using the two-sample t test in Example 11-3.

EXERCISE 11-29 Refer to Example 11-4, comparing tumor weights in treated and untreated mice.

a. Test the null hypothesis that the median tumor weight is the same for treated and untreated mice, using the median test.

b. Repeat part (a) using the two-sample t test.

c. Compare the results in parts (a) and (b) with what we found using the Wilcoxon–Mann–Whitney test in Example 11-4.

EXERCISE 11-30 Refer to Example 11-5, comparing yields of microwaved and unmicrowaved oranges.

a. Test the null hypothesis that the mean yield is the same for microwaved and unmicrowaved oranges, using the Wilcoxon–Mann–Whitney test.

b. Repeat part (a) using the two-sample t test.

c. Compare the results in parts (a) and (b) with what we found using the median test in Example 11-5.

EXERCISE 11-31 Discuss the sampling situations in which the two-sample t test, the Wilcoxon–Mann–Whitney two-sample test, and the median test are appropriate. Which test is preferred in each of these situations?

EXERCISE 11-32 In Example 11-2 we mentioned that local health departments chose between two different experimental designs for the polio field trials. We discussed the double-blind randomized placebo-controlled design. Under this design, parents of first-, second-, and third-grade children in participating school districts were asked for permission to include their children in the study. Children with parental consent were randomly divided into two groups. The children in the treatment group received a vaccine injection; children in the control group received a placebo (saline solution) injection. The experiment was double-blind, meaning that neither the children nor the immediate caregivers nor the workers making the diagnosis knew the treatment received by any inoculated child.

The other experimental design offered may be termed the NFIP design because it was favored by the National Foundation for Infantile Paralysis. Under this design, parents of second-graders were asked for permission to include their children in the study. All second-grade children with parental consent received a vaccine injection. The second-graders without parental consent, as well as all first- and third-grade children, received no injection, and served as control groups.

a. Discuss the NFIP design. What problems do you see with this experimental design? Under what conditions would this design be likely to bias the results in favor of the polio vaccine? Under what conditions would this design be likely to bias the results against the polio vaccine? What are the advantages of the double-blind randomized placebo-controlled design over the NFIP design? In particular, discuss the control groups to be compared with the vaccinated group in each study.

The results of the two studies can be summarized as shown below, with numbers rounded for this discussion (Freedman, Pisani, and Purves, 1978, page 6; from Francis, 1955). Instead of listing proportion diagnosed with polio in each group, we show the rate, or number of cases of polio per 100,000 children. Rates are often used instead of proportions by workers reporting results in public health.

Group	Number of children	Polio cases per 100,000 children
Double-blind randomized placebo-controlled design		
Vaccinated	200,000	28
Placebo control	200,000	71
No parental consent	350,000	46
NFIP design		
Vaccinated second-graders	225,000	25
No-consent second-graders	125,000	44
All first- and third-graders	725,000	54

b. Discuss the results of these two sets of experiments. Which design do you think gave the clearer comparison of vaccination versus no vaccination? Consider again your discussion of the two experimental designs in light of these results.

c. Show the relationship between the proportion of polio cases in a group and the rate or number of cases per 100,000 children.

Note: In fact, children of parents in higher socioeconomic levels were more likely to receive permission to participate in the study and also were more likely to develop polio. How does this information fit in with your discussion?