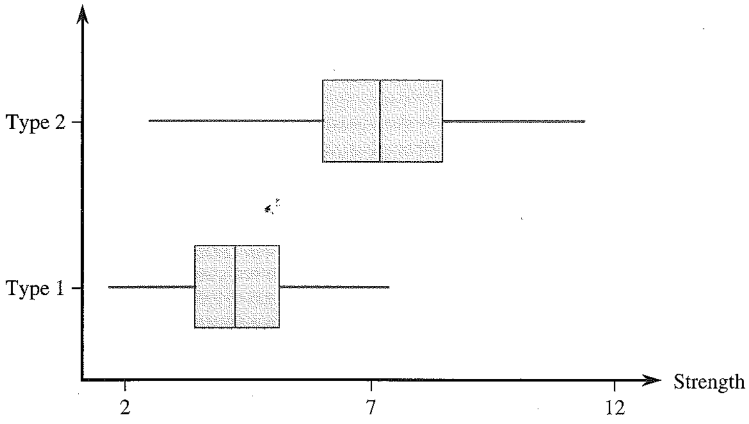


Figure 9.1 A comparative boxplot of the shear strength data

Let's now calculate a confidence interval for the difference between true average shear strength for 3/8-in. bolts (μ_1) and true average shear strength for 1/2-in. bolts (μ_2) using a confidence level of 95%:

$$4.25 - 7.14 \pm (1.96)\sqrt{\frac{(1.30)^2}{78} + \frac{(1.68)^2}{88}} = -2.89 \pm (1.96)(.2318)$$
$$= -2.89 \pm .45 = (-3.34, -2.44)$$

That is, with 95% confidence, $-3.34 < \mu_1 - \mu_2 < -2.44$. We can therefore be highly confident that the true average shear strength for the 1/2-in. bolts exceeds that for the 3/8-in. bolts by between 2.44 kip and 3.34 kip. Notice that if we relabel so that μ_1 refers to 1/2-in. bolts and μ_2 to 3/8-in. bolts, the confidence interval is now centered at +2.89 and the value .45 is still subtracted and added to obtain the confidence limits. The resulting interval is (2.44, 3.34), and the interpretation is identical to that for the interval previously calculated. ■

If the variances σ_1^2 and σ_2^2 are at least approximately known and the investigator uses equal sample sizes, then the common sample size n that yields a $100(1 - \alpha)\%$ interval of width w is

$$n = \frac{4z_{\alpha/2}^2(\sigma_1^2 + \sigma_2^2)}{w^2}$$

which will generally have to be rounded up to an integer.

EXERCISES Section 9.1 (1–16)

1. An article in the November 1983 *Consumer Reports* compared various types of batteries. The average lifetimes of Duracell Alkaline AA batteries and Eveready Energizer Alkaline AA batteries were given as 4.1 hours and 4.5 hours, respectively. Suppose these are the population average lifetimes.

a. Let $\overline{X}$ be the sample average lifetime of 100 Duracell batteries and $\overline{Y}$ be the sample average lifetime of 100 Eveready batteries. What is the mean value of $\overline{X} - \overline{Y}$ (i.e., where is the distribution of $\overline{X} - \overline{Y}$ centered)? How does your answer depend on the specified sample sizes?

b. Suppose the population standard deviations of lifetime are 1.8 hours for Duracell batteries and 2.0 hours for Eveready batteries. With the sample sizes given in part (a), what is the variance of the statistic $\overline{X} - \overline{Y}$, and what is its standard deviation?

c. For the sample sizes given in part (a), draw a picture of the approximate distribution curve of $\overline{X} - \overline{Y}$ (include a measurement scale on the horizontal axis). Would the shape of the curve necessarily be the same for sample sizes of 10 batteries of each type? Explain.

2. The National Health Statistics Reports dated Oct. 22, 2008, included the following information on the heights (in.) for non-Hispanic white females:

Age	Sample Size	Sample Mean	Std. Error Mean
20–39	866	64.9	.09
60 and older	934	63.1	.11

a. Calculate and interpret a confidence interval at confidence level approximately 95% for the difference between population mean height for the younger women and that for the older women.

b. Let μ_1 denote the population mean height for those aged 20–39 and μ_2 denote the population mean height for those aged 60 and older. Interpret the hypotheses H_0: $\mu_1 - \mu_2 = 1$ and H_a: $\mu_1 - \mu_2 > 1$, and then carry out a test of these hypotheses at significance level .001 using the rejection region approach.

c. What is the *P*-value for the test you carried out in (b)? Based on this *P*-value, would you reject the null hypothesis at any reasonable significance level? Explain.

d. What hypotheses would be appropriate if μ_1 referred to the older age group, μ_2 to the younger age group, and you wanted to see if there was compelling evidence for concluding that the population mean height for younger women exceeded that for older women by more than 1 in.?

3. Let μ_1 denote true average tread life for a premium brand of P205/65R15 radial tire, and let μ_2 denote the true average tread life for an economy brand of the same size. Test H_0: $\mu_1 - \mu_2 = 5000$ versus H_a: $\mu_1 - \mu_2 > 5000$ at level .01, using the following data: $m = 45, \bar{x} = 42{,}500$, $s_1 = 2200, n = 45, \bar{y} = 36{,}800$, and $s_2 = 1500$.

4. a. Use the data of Example 9.4 to compute a 95% CI for $\mu_1 - \mu_2$. Does the resulting interval suggest that $\mu_1 - \mu_2$ has been precisely estimated?

b. Use the data of Exercise 3 to compute a 95% upper confidence bound for $\mu_1 - \mu_2$.

5. Persons having Reynaud's syndrome are apt to suffer a sudden impairment of blood circulation in fingers and toes. In an experiment to study the extent of this impairment, each subject immersed a forefinger in water and the resulting heat output (cal/cm^2/min) was measured. For $m = 10$ subjects with the syndrome, the average heat output was $\bar{x} = .64$, and for $n = 10$ nonsufferers, the average output was 2.05. Let μ_1 and μ_2 denote the true average heat outputs for the two types of subjects. Assume that the two distributions of heat output are normal with $\sigma_1 = .2$ and $\sigma_2 = .4$.

a. Consider testing H_0: $\mu_1 - \mu_2 = -1.0$ versus H_a: $\mu_1 - \mu_2 < -1.0$ at level .01. Describe in words what H_a says, and then carry out the test.

b. Compute the *P*-value for the value of *Z* obtained in part (a).

c. What is the probability of a type II error when the actual difference between μ_1 and μ_2 is $\mu_1 - \mu_2 = -1.2$?

d. Assuming that $m = n$, what sample sizes are required to ensure that $\beta = .1$ when $\mu_1 - \mu_2 = -1.2$?

6. An experiment to compare the tension bond strength of polymer latex modified mortar (Portland cement mortar to which polymer latex emulsions have been added during mixing) to that of unmodified mortar resulted in $\bar{x} = 18.12$ kgf/cm^2 for the modified mortar ($m = 40$) and $\bar{y} = 16.87$ kgf/cm^2 for the unmodified mortar ($n = 32$). Let μ_1 and μ_2 be the true average tension bond strengths for the modified and unmodified mortars, respectively. Assume that the bond strength distributions are both normal.

a. Assuming that $\sigma_1 = 1.6$ and $\sigma_2 = 1.4$, test H_0: $\mu_1 - \mu_2 = 0$ versus H_a: $\mu_1 - \mu_2 > 0$ at level .01.

b. Compute the probability of a type II error for the test of part (a) when $\mu_1 - \mu_2 = 1$.

c. Suppose the investigator decided to use a level .05 test and wished $\beta = .10$ when $\mu_1 - \mu_2 = 1$. If $m = 40$, what value of *n* is necessary?

d. How would the analysis and conclusion of part (a) change if σ_1 and σ_2 were unknown but $s_1 = 1.6$ and $s_2 = 1.4$?

7. Is there any systematic tendency for part-time college faculty to hold their students to different standards than do full-time faculty? The article "Are There Instructional Differences Between Full-Time and Part-Time Faculty?" (*College Teaching*, 2009: 23–26) reported that for a sample of 125 courses taught by full-time faculty, the mean course GPA was 2.7186 and the standard deviation was .63342, whereas for a sample of 88 courses taught by part-timers, the mean and standard deviation were 2.8639 and .49241, respectively. Does it appear that true average course GPA for part-time faculty differs from that for faculty teaching full-time? Test the appropriate hypotheses at significance level .01 by first obtaining a *P*-value.

8. Tensile-strength tests were carried out on two different grades of wire rod ("Fluidized Bed Patenting of Wire Rods," *Wire J.*, June 1977: 56–61), resulting in the accompanying data.

Grade	Sample Size	Sample Mean (kg/mm^2)	Sample SD
AISI 1064	$m = 129$	$\bar{x} = 107.6$	$s_1 = 1.3$
AISI 1078	$n = 129$	$\bar{y} = 123.6$	$s_2 = 2.0$

a. Does the data provide compelling evidence for concluding that true average strength for the 1078 grade exceeds that for the 1064 grade by more than 10 kg/mm^2? Test the appropriate hypotheses using the *P*-value approach.

b. Estimate the difference between true average strengths for the two grades in a way that provides information about precision and reliability.

9. The article "Evaluation of a Ventilation Strategy to Prevent Barotrauma in Patients at High Risk for Acute Respiratory Distress Syndrome" (*New Engl. J. of Med.*, 1998: 355–358) reported on an experiment in which 120 patients with similar clinical features were randomly divided into a control group

However, the UCLA Statistics Department homepage (http://www.stat.ucla.edu) permits access to a power calculator that will do this. For example, we specified $m = 10, n = 8, \sigma_1 = 300, \sigma_2 = 225$ (these are the sample sizes for Example 9.7, whose sample standard deviations are somewhat smaller than these values of σ_1 and σ_2) and asked for the power of a two-tailed level .05 test of H_0: $\mu_1 - \mu_2 = 0$ when $\mu_1 - \mu_2 = 100, 250$, and 500. The resulting values of the power were .1089, .4609, and .9635 (corresponding to $\beta = .89$, .54, and .04), respectively. In general, β will decrease as the sample sizes increase, as α increases, and as $\mu_1 - \mu_2$ moves farther from 0. The software will also calculate sample sizes necessary to obtain a specified value of power for a particular value of $\mu_1 - \mu_2$.

EXERCISES Section 9.2 (17–35)

17. Determine the number of degrees of freedom for the two-sample t test or CI in each of the following situations:
a. $m = 10, n = 10, s_1 = 5.0, s_2 = 6.0$
b. $m = 10, n = 15, s_1 = 5.0, s_2 = 6.0$
c. $m = 10, n = 15, s_1 = 2.0, s_2 = 6.0$
d. $m = 12, n = 24, s_1 = 5.0, s_2 = 6.0$

18. Let μ_1 and μ_2 denote true average densities for two different types of brick. Assuming normality of the two density distributions, test H_0: $\mu_1 - \mu_2 = 0$ versus H_a: $\mu_1 - \mu_2 \neq 0$ using the following data: $m = 6, \bar{x} = 22.73, s_1 = .164$, $n = 5, \bar{y} = 21.95$, and $s_2 = .240$.

19. Suppose μ_1 and μ_2 are true mean stopping distances at 50 mph for cars of a certain type equipped with two different types of braking systems. Use the two-sample t test at significance level .01 to test H_0: $\mu_1 - \mu_2 = -10$ versus H_a: $\mu_1 - \mu_2 < -10$ for the following data: $m = 6$, $\bar{x} = 115.7, s_1 = 5.03, n = 6, \bar{y} = 129.3$, and $s_2 = 5.38$.

20. Use the data of Exercise 19 to calculate a 95% CI for the difference between true average stopping distance for cars equipped with system 1 and cars equipped with system 2. Does the interval suggest that precise information about the value of this difference is available?

21. Quantitative noninvasive techniques are needed for routinely assessing symptoms of peripheral neuropathies, such as carpal tunnel syndrome (CTS). The article "A Gap Detection Tactility Test for Sensory Deficits Associated with Carpal Tunnel Syndrome" (*Ergonomics,* 1995: 2588–2601) reported on a test that involved sensing a tiny gap in an otherwise smooth surface by probing with a finger; this functionally resembles many work-related tactile activities, such as detecting scratches or surface defects. When finger probing was not allowed, the sample average gap detection threshold for $m = 8$ normal subjects was 1.71 mm, and the sample standard deviation was .53; for $n = 10$ CTS subjects, the sample mean and sample standard deviation were 2.53 and .87, respectively. Does this data suggest that the true average gap detection threshold for CTS subjects exceeds that for normal subjects? State and test the relevant hypotheses using a significance level of .01.

22. The slant shear test is widely accepted for evaluating the bond of resinous repair materials to concrete; it utilizes cylinder specimens made of two identical halves bonded at 30°. The article "Testing the Bond Between Repair Materials and Concrete Substrate" (*ACI Materials J.,* 1996: 553–558) reported that for 12 specimens prepared using wire-brushing, the sample mean shear strength (N/mm^2) and sample standard deviation were 19.20 and 1.58, respectively, whereas for 12 hand-chiseled specimens, the corresponding values were 23.13 and 4.01. Does the true average strength appear to be different for the two methods of surface preparation? State and test the relevant hypotheses using a significance level of .05. What are you assuming about the shear strength distributions?

23. Fusible interlinings are being used with increasing frequency to support outer fabrics and improve the shape and drape of various pieces of clothing. The article "Compatibility of Outer and Fusible Interlining Fabrics in Tailored Garments" (*Textile Res. J.,* 1997: 137–142) gave the accompanying data on extensibility (%) at 100 gm/cm for both high-quality (H) fabric and poor-quality (P) fabric specimens.

H	1.2	.9	.7	1.0	1.7	1.7	1.1	.9	1.7
	1.9	1.3	2.1	1.6	1.8	1.4	1.3	1.9	1.6
	.8	2.0	1.7	1.6	2.3	2.0			
P	1.6	1.5	1.1	2.1	1.5	1.3	1.0	2.6	

a. Construct normal probability plots to verify the plausibility of both samples having been selected from normal population distributions.
b. Construct a comparative boxplot. Does it suggest that there is a difference between true average extensibility for high-quality fabric specimens and that for poor-quality specimens?
c. The sample mean and standard deviation for the high-quality sample are 1.508 and .444, respectively, and those

for the poor-quality sample are 1.588 and .530. Use the two-sample *t* test to decide whether true average extensibility differs for the two types of fabric.

24. Damage to grapes from bird predation is a serious problem for grape growers. The article "Experimental Method to Investigate and Monitor Bird Behavior and Damage to Vineyards" (*Amer. J. of Enology and Viticulture,* 2004: 288–291) reported on an experiment involving a bird-feeder table, time-lapse video, and artificial foods. Information was collected for two different bird species at both the experimental location and at a natural vineyard setting. Consider the following data on time (sec) spent on a single visit to the location.

Species	Location	n	$\bar{x}$	SE mean
Blackbirds	Exptl	65	13.4	2.05
Blackbirds	Natural	50	9.7	1.76
Silvereyes	Exptl	34	49.4	4.78
Silvereyes	Natural	46	38.4	5.06

a. Calculate an upper confidence bound for the true average time that blackbirds spend on a single visit at the experimental location.

b. Does it appear that true average time spent by blackbirds at the experimental location exceeds the true average time birds of this type spend at the natural location? Carry out a test of appropriate hypotheses.

c. Estimate the difference between the true average time blackbirds spend at the natural location and true average time that silvereyes spend at the natural location, and do so in a way that conveys information about reliability and precision.

[*Note:* The sample medians reported in the article all seemed significantly smaller than the means, suggesting substantial population distribution skewness. The authors actually used the distribution-free test procedure presented in Section 2 of Chapter 15.]

25. Low-back pain (LBP) is a serious health problem in many industrial settings. The article "Isodynamic Evaluation of Trunk Muscles and Low-Back Pain Among Workers in a Steel Factory" (*Ergonomics,* 1995: 2107–2117) reported the accompanying summary data on lateral range of motion (degrees) for a sample of workers without a history of LBP and another sample with a history of this malady.

Condition	Sample Size	Sample Mean	Sample SD
No LBP	28	91.5	5.5
LBP	31	88.3	7.8

Calculate a 90% confidence interval for the difference between population mean extent of lateral motion for the two conditions. Does the interval suggest that population mean lateral motion differs for the two conditions? Is the message different if a confidence level of 95% is used?

26. The article "The Influence of Corrosion Inhibitor and Surface Abrasion on the Failure of Aluminum-Wired Twist-On Connections" (*IEEE Trans. on Components, Hybrids, and Manuf. Tech.,* 1984: 20–25) reported data on potential drop measurements for one sample of connectors wired with alloy aluminum and another sample wired with EC aluminum. Does the accompanying SAS output suggest that the true average potential drop for alloy connections (type 1) is higher than that for EC connections (as stated in the article)? Carry out the appropriate test using a significance level of .01. In reaching your conclusion, what type of error might you have committed? [*Note:* SAS reports the *P*-value for a two-tailed test.]

```
Type    N          Mean       Std Dev     Std Error
1      20   17.49900000    0.55012821    0.12301241
2      20   16.90000000    0.48998389    0.10956373

         Variances          T        DF      Prob>|T|
           Unequal     3.6362      37.5        0.0008
             Equal     3.6362      38.0        0.0008
```

27. Anorexia Nervosa (AN) is a psychiatric condition leading to substantial weight loss among women who are fearful of becoming fat. The article "Adipose Tissue Distribution After Weight Restoration and Weight Maintenance in Women with Anorexia Nervosa" (*Amer. J. of Clinical Nutr.,* 2009: 1132–1137) used whole-body magnetic resonance imagery to determine various tissue characteristics for both an AN sample of individuals who had undergone acute weight restoration and maintained their weight for a year and a comparable (at the outset of the study) control sample. Here is summary data on intermuscular adipose tissue (IAT; kg).

Condition	Sample Size	Sample Mean	Sample SD
AN	16	.52	.26
Control	8	.35	.15

Assume that both samples were selected from normal distributions.

a. Calculate an estimate for true average IAT under the described AN protocol, and do so in a way that conveys information about the reliability and precision of the estimation.

b. Calculate an estimate for the difference between true average AN IAT and true average control IAT, and do so in a way that conveys information about the reliability and precision of the estimation. What does your estimate suggest about true average AN IAT relative to true average control IAT?

28. As the population ages, there is increasing concern about accident-related injuries to the elderly. The article "Age and

Gender Differences in Single-Step Recovery from a Forward Fall" (*J. of Gerontology,* 1999: M44–M50) reported on an experiment in which the maximum lean angle—the furthest a subject is able to lean and still recover in one step—was determined for both a sample of younger females (21–29 years) and a sample of older females (67–81 years). The following observations are consistent with summary data given in the article:

YF: 29, 34, 33, 27, 28, 32, 31, 34, 32, 27
OF: 18, 15, 23, 13, 12

Does the data suggest that true average maximum lean angle for older females is more than 10 degrees smaller than it is for younger females? State and test the relevant hypotheses at significance level .10 by obtaining a *P*-value.

29. The article "Effect of Internal Gas Pressure on the Compression Strength of Beverage Cans and Plastic Bottles" (*J. of Testing and Evaluation,* 1993: 129–131) includes the accompanying data on compression strength (lb) for a sample of 12-oz aluminum cans filled with strawberry drink and another sample filled with cola. Does the data suggest that the extra carbonation of cola results in a higher average compression strength? Base your answer on a *P*-value. What assumptions are necessary for your analysis?

Beverage	Sample Size	Sample Mean	Sample SD
Strawberry drink	15	540	21
Cola	15	554	15

30. The article "Flexure of Concrete Beams Reinforced with Advanced Composite Orthogrids" (*J. of Aerospace Engr.,* 1997: 7–15) gave the accompanying data on ultimate load (kN) for two different types of beams.

Type	Sample Size	Sample Mean	Sample SD
Fiberglass grid	26	33.4	2.2
Commercial carbon grid	26	42.8	4.3

a. Assuming that the underlying distributions are normal, calculate and interpret a 99% CI for the difference between true average load for the fiberglass beams and that for the carbon beams.

b. Does the upper limit of the interval you calculated in part (a) give a 99% upper confidence bound for the difference between the two μ's? If not, calculate such a bound. Does it strongly suggest that true average load for the carbon beams is more than that for the fiberglass beams? Explain.

31. Refer to Exercise 33 in Section 7.3. The cited article also gave the following observations on degree of polymerization for specimens having viscosity times concentration in a higher range:

429	430	430	431	436	437
440	441	445	446	447	

a. Construct a comparative boxplot for the two samples, and comment on any interesting features.

b. Calculate a 95% confidence interval for the difference between true average degree of polymerization for the middle range and that for the high range. Does the interval suggest that μ_1 and μ_2 may in fact be different? Explain your reasoning.

32. The degenerative disease osteoarthritis most frequently affects weight-bearing joints such as the knee. The article "Evidence of Mechanical Load Redistribution at the Knee Joint in the Elderly when Ascending Stairs and Ramps" (*Annals of Biomed. Engr.,* 2008: 467–476) presented the following summary data on stance duration (ms) for samples of both older and younger adults.

Age	Sample Size	Sample Mean	Sample SD
Older	28	801	117
Younger	16	780	72

Assume that both stance duration distributions are normal.

a. Calculate and interpret a 99% CI for true average stance duration among elderly individuals.

b. Carry out a test of hypotheses at significance level .05 to decide whether true average stance duration is larger among elderly individuals than among younger individuals.

33. The article "The Effects of a Low-Fat, Plant-Based Dietary Intervention on Body Weight, Metabolism, and Insulin Sensitivity in Postmenopausal Women" (*Amer. J. of Med.,* 2005: 991–997) reported on the results of an experiment in which half of the individuals in a group of 64 postmenopausal overweight women were randomly assigned to a particular vegan diet, and the other half received a diet based on National Cholesterol Education Program guidelines. The sample mean decrease in body weight for those on the vegan diet was 5.8 kg, and the sample SD was 3.2, whereas for those on the control diet, the sample mean weight loss and standard deviation were 3.8 and 2.8, respectively. Does it appear the true average weight loss for the vegan diet exceeds that for the control diet by more than 1 kg? Carry out an appropriate test of hypotheses at significance level .05 based on calculating a *P*-value.

34. Consider the pooled *t* variable

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S_p\sqrt{\dfrac{1}{m} + \dfrac{1}{n}}}$$

which has a t distribution with $m + n - 2$ df when both population distributions are normal with $\sigma_1 = \sigma_2$ (see the Pooled t Procedures subsection for a description of S_p).

a. Use this t variable to obtain a pooled t confidence interval formula for $\mu_1 - \mu_2$.

b. A sample of ultrasonic humidifiers of one particular brand was selected for which the observations on maximum output of moisture (oz) in a controlled chamber were 14.0, 14.3, 12.2, and 15.1. A sample of the second brand gave output values 12.1, 13.6, 11.9, and 11.2 ("Multiple Comparisons of Means Using Simultaneous Confidence Intervals," *J. of Quality Technology,* 1989: 232–241). Use the pooled t formula from part (a) to estimate the difference between true average outputs for the two brands with a 95% confidence interval.

c. Estimate the difference between the two μ's using the two-sample t interval discussed in this section, and compare it to the interval of part (b).

35. Refer to Exercise 34. Describe the pooled t test for testing H_0: $\mu_1 - \mu_2 = \Delta_0$ when both population distributions are normal with $\sigma_1 = \sigma_2$. Then use this test procedure to test the hypotheses suggested in Exercise 33.

9.3 Analysis of Paired Data

In Sections 9.1 and 9.2, we considered making an inference about a difference between two means μ_1 and μ_2. This was done by utilizing the results of a random sample $X_1, X_2, \ldots X_m$ from the distribution with mean μ_1 and a completely independent (of the X's) sample $Y_1, \ldots, Y_n$ from the distribution with mean μ_2. That is, either m individuals were selected from population 1 and n different individuals from population 2, or m individuals (or experimental objects) were given one treatment and another set of n individuals were given the other treatment. In contrast, there are a number of experimental situations in which there is only one set of n individuals or experimental objects; making two observations on each one results in a natural pairing of values.

Example 9.8 Trace metals in drinking water affect the flavor, and unusually high concentrations can pose a health hazard. The article "Trace Metals of South Indian River" (*Envir. Studies,* 1982: 62–66) reports on a study in which six river locations were selected (six experimental objects) and the zinc concentration (mg/L) determined for both surface water and bottom water at each location. The six pairs of observations are displayed in the accompanying table. Does the data suggest that true average concentration in bottom water exceeds that of surface water?

	Location					
	1	**2**	**3**	**4**	**5**	**6**
Zinc concentration in bottom water (x)	.430	.266	.567	.531	.707	.716
Zinc concentration in surface water (y)	.415	.238	.390	.410	.605	.609
Difference	.015	.028	.177	.121	.102	.107

Figure 9.4(a) displays a plot of this data. At first glance, there appears to be little difference between the x and y samples. From location to location, there is a great deal of variability in each sample, and it looks as though any differences between the samples can be attributed to this variability. However, when the observations are identified by location, as in Figure 9.4(b), a different view emerges. At each location, bottom concentration exceeds surface concentration. This is confirmed by the fact that all $x - y$ differences displayed in the bottom row of the data table are positive. A correct analysis of this data focuses on these differences.

a. Is it plausible that the population distribution of differences is normal?

b. Does it appear that the true average difference between intake values measured by the two methods is something other than zero? Determine the *P*-value of the test, and use it to reach a conclusion at significance level .05.

40. Lactation promotes a temporary loss of bone mass to provide adequate amounts of calcium for milk production. The paper "Bone Mass Is Recovered from Lactation to Postweaning in Adolescent Mothers with Low Calcium Intakes" (*Amer. J. of Clinical Nutr.,* 2004: 1322–1326) gave the following data on total body bone mineral content (TBBMC) (g) for a sample both during lactation (L) and in the postweaning period (P).

	Subject									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
L	1928	2549	2825	1924	1628	2175	2114	2621	1843	2541
P	2126	2885	2895	1942	1750	2184	2164	2626	2006	2627

a. Does the data suggest that true average total body bone mineral content during postweaning exceeds that during lactation by more than 25 g? State and test the appropriate hypotheses using a significance level of .05. [*Note:* The appropriate normal probability plot shows some curvature but not enough to cast substantial doubt on a normality assumption.]

b. Calculate an upper confidence bound using a 95% confidence level for the true average difference between TBBMC during postweaning and during lactation.

c. Does the (incorrect) use of the two-sample *t* test to test the hypotheses suggested in (a) lead to the same conclusion that you obtained there? Explain.

41. Antipsychotic drugs are widely prescribed for conditions such as schizophrenia and bipolar disease. The article "Cardiometabolic Risk of Second-Generation Antipsychotic Medications During First-Time Use in Children and Adolescents" (*J. of the Amer. Med. Assoc.,* 2009) reported on body composition and metabolic changes for individuals who had taken various antipsychotic drugs for short periods of time.

a. The sample of 41 individuals who had taken aripiprazole had a mean change in total cholesterol (mg/dL) of 3.75, and the estimated standard error $s_D/\sqrt{n}$ was 3.878. Calculate a confidence interval with confidence level approximately 95% for the true average increase in total cholesterol under these circumstances (the cited article included this CI).

b. The article also reported that for a sample of 36 individuals who had taken quetiapine, the sample mean cholesterol level change and estimated standard error were 9.05 and 4.256, respectively. Making any necessary assumptions about the distribution of change in cholesterol level, does the choice of significance level impact your conclusion as to whether true average cholesterol level increases? Explain. [*Note*: The article included a *P*-value.]

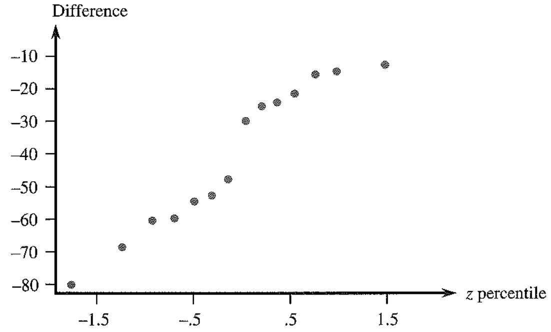
c. For the sample of 45 individuals who had taken olanzapine, the article reported (7.38, 9.69) as a 95% CI for true average weight gain (kg). What is a 99% CI?

42. It has been estimated that between 1945 and 1971, as many as 2 million children were born to mothers treated with diethylstilbestrol (DES), a nonsteroidal estrogen recommended for pregnancy maintenance. The FDA banned this drug in 1971 because research indicated a link with the incidence of cervical cancer. The article "Effects of Prenatal Exposure to Diethylstilbestrol (DES) on Hemispheric Laterality and Spatial Ability in Human Males" (*Hormones and Behavior,* 1992: 62–75) discussed a study in which 10 males exposed to DES, and their unexposed brothers, underwent various tests. This is the summary data on the results of a spatial ability test: $\bar{x} = 12.6$ (exposed), $\bar{y} = 13.7$, and standard error of mean difference $= .5$. Test at level .05 to see whether exposure is associated with reduced spatial ability by obtaining the *P*-value.

43. Cushing's disease is characterized by muscular weakness due to adrenal or pituitary dysfunction. To provide effective treatment, it is important to detect childhood Cushing's disease as early as possible. Age at onset of symptoms and age at diagnosis (months) for 15 children suffering from the disease were given in the article "Treatment of Cushing's Disease in Childhood and Adolescence by Transphenoidal Microadenomectomy" (*New Engl. J. of Med.,* 1984: 889). Here are the values of the differences between age at onset of symptoms and age at diagnosis:

−24	−12	−55	−15	−30	−60	−14	−21
−48	−12	−25	−53	−61	−69	−80	

a. Does the accompanying normal probability plot cast strong doubt on the approximate normality of the population distribution of differences?



b. Calculate a lower 95% confidence bound for the population mean difference, and interpret the resulting bound.

c. Suppose the (age at diagnosis) – (age at onset) differences had been calculated. What would be a 95% upper confidence bound for the corresponding population mean difference?

of the two samples and then replace the $\hat{p}$'s and $\hat{q}$'s in the foregoing formula by $\tilde{p}$'s and $\tilde{q}$'s where $\tilde{p}_1 = (x + 1)/(m + 2)$, etc. This modified interval can also be used when sample sizes are quite small.

Example 9.13 The authors of the article "Adjuvant Radiotherapy and Chemotherapy in Node-Positive Premenopausal Women with Breast Cancer" (*New Engl. J. of Med.,* 1997: 956–962) reported on the results of an experiment designed to compare treating cancer patients with chemotherapy only to treatment with a combination of chemotherapy and radiation. Of the 154 individuals who received the chemotherapy-only treatment, 76 survived at least 15 years, whereas 98 of the 164 patients who received the hybrid treatment survived at least that long. With p_1 denoting the proportion of all such women who, when treated with just chemotherapy, survive at least 15 years and p_2 denoting the analogous proportion for the hybrid treatment, $\hat{p}_1 = 76/154 = .494$ and $98/164 = .598$. A confidence interval for the difference between proportions based on the traditional formula with a confidence level of approximately 99% is

$$.494 - .598 \pm (2.58)\sqrt{\frac{(.494)(.506)}{154} + \frac{(.598)(.402)}{164}} = -.104 \pm .143$$
$$= (-.247, .039)$$

At the 99% confidence level, it is plausible that $-.247 < p_1 - p_2 < .039$. This interval is reasonably wide, a reflection of the fact that the sample sizes are not terribly large for this type of interval. Notice that 0 is one of the plausible values of $p_1 - p_2$, suggesting that neither treatment can be judged superior to the other. Using $\tilde{p}_1 = 77/156 = .494$, $\tilde{q}_1 = 79/156 = .506$, $\tilde{p}_2 = .596$, $\tilde{q}_2 = .404$ based on sample sizes of 156 and 166, respectively, the "improved" interval here is identical to the earlier interval. ■

Small-Sample Inferences

On occasion an inference concerning $p_1 - p_2$ may have to be based on samples for which at least one sample size is small. Appropriate methods for such situations are not as straightforward as those for large samples, and there is more controversy among statisticians as to recommended procedures. One frequently used test, called the Fisher–Irwin test, is based on the hypergeometric distribution. Your friendly neighborhood statistician can be consulted for more information.

EXERCISES Section 9.4 (49–58)

49. Is someone who switches brands because of a financial inducement less likely to remain loyal than someone who switches without inducement? Let p_1 and p_2 denote the true proportions of switchers to a certain brand with and without inducement, respectively, who subsequently make a repeat purchase. Test H_0: $p_1 - p_2 = 0$ versus H_a: $p_1 - p_2 < 0$ using $\alpha = .01$ and the following data:

$m = 200$ number of success $= 30$
$n = 600$ number of success $= 180$

(Similar data is given in "Impact of Deals and Deal Retraction on Brand Switching," *J. of Marketing,* 1980: 62–70.)

50. Recent incidents of food contamination have caused great concern among consumers. The article "How Safe Is That Chicken?" (*Consumer Reports,* Jan. 2010: 19–23) reported that 35 of 80 randomly selected Perdue brand broilers tested positively for either campylobacter or salmonella (or both), the leading bacterial causes of food-borne disease, whereas 66 of 80 Tyson brand broilers tested positive.

a. Does it appear that the true proportion of non-contaminated Perdue broilers differs from that for the Tyson brand? Carry out a test of hypotheses using a significance level .01 by obtaining a P-value.

b. If the true proportions of non-contaminated chickens for the Perdue and Tyson brands are .50 and .25, respectively, how likely is it that the null hypothesis of equal proportions will be rejected when a .01 significance level is used and the sample sizes are both 80?

51. It is thought that the front cover and the nature of the first question on mail surveys influence the response rate. The article "The Impact of Cover Design and First Questions on Response Rates for a Mail Survey of Skydivers" (*Leisure Sciences,* 1991: 67–76) tested this theory by experimenting with different cover designs. One cover was plain; the other used a picture of a skydiver. The researchers speculated that the return rate would be lower for the plain cover.

Cover	Number Sent	Number Returned
Plain	207	104
Skydiver	213	109

Does this data support the researchers' hypothesis? Test the relevant hypotheses using $\alpha = .10$ by first calculating a P-value.

52. Do teachers find their work rewarding and satisfying? The article "Work-Related Attitudes" (*Psychological Reports,* 1991: 443–450) reports the results of a survey of 395 elementary school teachers and 266 high school teachers. Of the elementary school teachers, 224 said they were very satisfied with their jobs, whereas 126 of the high school teachers were very satisfied with their work. Estimate the difference between the proportion of all elementary school teachers who are very satisfied and all high school teachers who are very satisfied by calculating and interpreting a CI.

53. Olestra is a fat substitute approved by the FDA for use in snack foods. Because there have been anecdotal reports of gastrointestinal problems associated with olestra consumption, a randomized, double-blind, placebo-controlled experiment was carried out to compare olestra potato chips to regular potato chips with respect to GI symptoms ("Gastrointestinal Symptoms Following Consumption of Olestra or Regular Triglyceride Potato Chips," *J. of the Amer. Med. Assoc.,* 1998: 150–152). Among 529 individuals in the TG control group, 17.6% experienced an adverse GI event, whereas among the 563 individuals in the olestra treatment group, 15.8% experienced such an event.

a. Carry out a test of hypotheses at the 5% significance level to decide whether the incidence rate of GI problems for those who consume olestra chips according to the experimental regimen differs from the incidence rate for the TG control treatment.

b. If the true percentages for the two treatments were 15% and 20%, respectively, what sample sizes ($m = n$) would be necessary to detect such a difference with probability .90?

54. Teen Court is a juvenile diversion program designed to circumvent the formal processing of first-time juvenile offenders within the juvenile justice system. The article "An Experimental Evaluation of Teen Courts" (*J. of Experimental Criminology,* 2008: 137–163) reported on a study in which offenders were randomly assigned either to Teen Court or to the traditional Department of Juvenile Services method of processing. Of the 56 TC individuals, 18 subsequently recidivated (look it up!) during the 18-month follow-up period, whereas 12 of the 51 DJS individuals did so. Does the data suggest that the true proportion of TC individuals who recidivate during the specified follow-up period differs from the proportion of DJS individuals who do so? State and test the relevant hypotheses by obtaining a P-value and then using a significance level of .10.

55. In medical investigations, the ratio $\theta = p_1/p_2$ is often of more interest than the difference $p_1 - p_2$ (e.g., individuals given treatment 1 are how many times as likely to recover as those given treatment 2?). Let $\hat{\theta} = \hat{p}_1/\hat{p}_2$. When m and n are both large, the statistic $\ln(\hat{\theta})$ has approximately a normal distribution with approximate mean value $\ln(\theta)$ and approximate standard deviation $[(m - x)/(mx) + (n - y)/(ny)]^{1/2}$.

a. Use these facts to obtain a large-sample 95% CI formula for estimating $\ln(\theta)$, and then a CI for θ itself.

b. Return to the heart-attack data of Example 1.3, and calculate an interval of plausible values for θ at the 95% confidence level. What does this interval suggest about the efficacy of the aspirin treatment?

56. Sometimes experiments involving success or failure responses are run in a paired or before/after manner. Suppose that before a major policy speech by a political candidate, n individuals are selected and asked whether (S) or not (F) they favor the candidate. Then after the speech the same n people are asked the same question. The responses can be entered in a table as follows:

		After	
		S	F
Before	S	x_1	x_2
	F	x_3	x_4

where $x_1 + x_2 + x_3 + x_4 = n$. Let p_1, p_2, p_3, and p_4 denote the four cell probabilities, so that $p_1 = P(S$ before and S after), and so on. We wish to test the hypothesis that the true proportion of supporters (S) after the speech has not increased against the alternative that it has increased.

a. State the two hypotheses of interest in terms of p_1, p_2, p_3, and p_4.

b. Construct an estimator for the after/before difference in success probabilities.

c. When n is large, it can be shown that the rv $(X_i - X_j)/n$ has approximately a normal distribution with variance given by $[p_i + p_j - (p_i - p_j)^2]/n$. Use this to construct a test statistic with approximately a standard normal distribution when H_0 is true (the result is called McNemar's test).

d. If $x_1 = 350$, $x_2 = 150$, $x_3 = 200$, and $x_4 = 300$, what do you conclude?