Typos in 6th Edition of

Introdution to Statistical Methods and Data Analysis

by R. Lyman Ott and Michael Longnecker

The following pages contain corrections to the first printing of the 6th Edition of *Introdution to Statistical Methods and Data Analysis.* We apologize for any inconvenience that may have resulted from the errors found in the first printing of the 6th Edition of our book. Many of the errors resulted from the 6th Edition being an entirely new typing of the book.

provided $\mu \neq 0$. Thus, the coefficient of variation is the standard deviation of the population or process expressed in units of μ. The two filling processes would have equivalent degrees of variability if the two processes had the same CV. For the fertilizer process, the CV = 1.2/80 = .015. The cornflakes process has CV = 0.4/24 = .017. Hence, the two processes have very similar variability relative to the size of their means. The CV is a unit-free number because the standard deviation and mean are measured using the same units. Hence, the CV is often used as an index of process or population variability. In many applications, the CV is expressed as a percentage: CV = $100(\sigma/|\mu|)\%$. Thus, if a process has a CV of 15%, the standard deviation of the output of the process is 15% of the process mean. Using sampled data from the population, we estimate CV with $100(s/|\bar{y}|)\%$.

3.6 The Boxplot

boxplot

As mentioned earlier in this chapter, a stem-and-leaf plot provides a graphical representation of a set of scores that can be used to examine the shape of the distribution, the range of scores, and where the scores are concentrated. The **boxplot,** which builds on the information displayed in a stem-and-leaf plot, is more concerned with the symmetry of the distribution and incorporates numerical measures of central tendency and location to study the variability of the scores and the concentration of scores in the tails of the distribution.

Before we show how to construct and interpret a boxplot, we need to introduce several new terms that are peculiar to the language of exploratory data analysis (EDA).We are familiar with the definitions for the first, second (median), and third quartiles of a distribution presented earlier in this chapter. The boxplot uses the median and **quartiles** of a distribution.

quartiles

We can now illustrate a *skeletal boxplot* using an example.

Replace with attached data →

EXAMPLE 3.13

A criminologist is studying whether there are wide variations in violent crime rates across the United States. Using Department of Justice data from 2000, the crime rates in 90 cities selected from across the United States were obtained. Use the data given in Table 3.12 to construct a skeletal boxplot to demonstrate the degree of variability in crime rates.

TABLE 3.12
Violent crime rates for 90 standard metropolitan statistical areas selected from around the United States

South	Rate	North	Rate	West	Rate
Albany, GA	876	Allentown, PA	189	Abilene, TX	570
Anderson, SC	578	Battle Creek, MI	661	Albuquerque, NM	928
Anniston, AL	718	Benton Harbor, MI	877	Anchorage, AK	516
Athens, GA	388	Bridgeport, CT	563	Bakersfield, CA	885
Augusta, GA	562	Buffalo, NY	647	Brownsville, TX	751
Baton Rouge, LA	971	Canton, OH	447	Denver, CO	561
Charleston, SC	698	Cincinnati, OH	336	Fresno, CA	1,020
Charlottesville, VA	298	Cleveland, OH	526	Galveston, TX	592
Chattanooga, TN	673	Columbus, OH	624	Houston, TX	814
Columbus, GA	537	Dayton, OH	605	Kansas City, MO	843

(*continued*)

TABLE 3.12
Violent crime rates for 90 standard metropolitan statistical areas selected from around the United States (*continued*)

South	Rate	North	Rate	West	Rate
Dothan, AL	642	Des Moines, IA	496	Lawton, OK	466
Florence, SC	856	Dubuque, IA	296	Lubbock, TX	498
Fort Smith, AR	376	Gary, IN	628	Merced, CA	562
Gadsden, AL	508	Grand Rapids, MI	481	Modesto, CA	739
Greensboro, NC	529	Janesville, WI	224	Oklahoma City, OK	562
Hickery, NC	393	Kalamazoo, MI	868	Reno, NV	817
Knoxville, TN	354	Lima, OH	804	Sacramento, CA	690
Lake Charles, LA	735	Madison, WI	210	St. Louis, MO	720
Little Rock, AR	811	Milwaukee, WI	421	Salinas, CA	758
Macon, GA	504	Minneapolis, MN	435	San Diego, CA	731
Monroe, LA	807	Nassau, NY	291	Santa Ana, CA	480
Nashville, TN	719	New Britain, CT	393	Seattle, WA	559
Norfolk, VA	464	Philadelphia, PA	605	Sioux City, IA	505
Raleigh, NC	410	Pittsburgh, PA	341	Stockton, CA	703
Richmond, VA	491	Portland, ME	352	Tacoma, WA	809
Savannah, GA	557	Racine, WI	374	Tucson, AZ	706
Shreveport, LA	771	Reading, PA	267	Victoria, TX	631
Washington, DC	685	Saginaw, MI	684	Waco, TX	626
Wilmington, DE	448	Syracuse, NY	685	Wichita Falls, TX	639
Wilmington, NC	571	Worcester, MA	460	Yakima, WA	585

Note: Rates represent the number of violent crimes (murder, forcible rape, robbery, and aggravated assault) per 100,000 inhabitants, rounded to the nearest whole number.

Source: Department of Justice, Crime Reports and the United States, 2000.

replace with numbers on next page

Solution The data were summarized using a stem-and-leaf plot as depicted in Figure 3.23. Use this plot to construct a skeletal boxplot.

FIGURE 3.23
Stem-and-leaf plot of crime data

```
 1  40  45
 1  63
 2  11  20  21
 2  60  61  64  85
 3  12  26  30  39  43  44
 3  50  54  57  58  67  68  97
 4  03  05  13  15  18  26  27  28  47
 4  51  53  63  65  65  70  77  77  86  90  94  96  97  98
 5  21  22  23  28  28  49
 5  55  64  68  71  74  76  79  84
 6  10  25  37  45  46  49
 6  57  60  76  89  92
 7  17
 7  54  54  61  72  73  92  98
 8
 8  71  72  87  97
 9  25  44  46
 9  73  98
10
10  64  94
```

When the scores are ordered from lowest to highest, the median is computed by averaging the 45th and 46th scores. For these data, the 45th score (counting

Replacement for Example 3.13 on pages 97-97

EXAMPLE 3.13

A criminologist is studying whether there are wide variations in violent crime rates across the United States. Using Department of Justice data from 2000, the crime rates in 90 cities selected from across the United States were obtained. Use the data given in Table 3.12 to construct a skeletal boxplot to demonstrate the degree of variability in crime rates.

TABLE 3.12
Violent crime rates for 90 standard metropolitan statistical areas selected from around the United States

South	Rate	North	Rate	West	Rate
Albany, GA	498	Allentown, PA	285	Abilene, TX	343
Anderson, SC	676	Battle Creek, MI	490	Albuquerque, NM	946
Anniston, AL	344	Benton Harbor, MI	528	Anchorage, AK	584
Athens, GA	368	Bridgeport, CT	427	Bakersfield, CA	494
Augusta, GA	772	Buffalo, NY	413	Brownsville, TX	463
Baton Rouge, LA	497	Canton, OH	220	Denver, CO	357
Charleston, SC	415	Cincinnati, OH	163	Fresno, CA	761
Charlottesville, VA	925	Cleveland, OH	428	Galveston, TX	717
Chattanooga, TN	555	Columbus, OH	625	Houston, TX	1094
Columbus, GA	260	Dayton, OH	339	Kansas City, MO	637
Dothan, AL	528	Des Moines, IA	211	Lawton, OK	692
Florence, SC	649	Dubuque, IA	451	Lubbock, TX	522
Fort Smith, AR	571	Gary, IN	358	Merced, CA	397
Gadsden, AL	470	Grand Rapids, MI	660	Modesto, CA	521
Greensboro, NC	897	Janesville, WI	330	Oklahoma City, OK	610
Hickery, NC	973	Kalamazoo, MI	145	Reno, NV	477
Knoxville, TN	486	Lima, OH	326	Sacramento, CA	453
Lake Charles, LA	447	Madison, WI	403	St. Louis, MO	798
Little Rock, AR	689	Milwaukee, WI	523	Salinas, CA	646
Macon, GA	754	Minneapolis, MN	312	San Diego, CA	645
Monroe, LA	465	Nassau, NY	576	Santa Ana, CA	549
Nashville, TN	496	New Britain, CT	261	Seattle, WA	568
Norfolk, VA	871	Philadelphia, PA	221	Sioux City, IA	465
Raleigh, NC	1064	Pittsburgh, PA	754	Stockton, CA	350
Richmond, VA	579	Portland, ME	140	Tacoma, WA	574
Savannah, GA	792	Racine, WI	418	Tucson, AZ	944
Shreveport, LA	367	Reading, PA	657	Victoria, TX	426
Washington, DC	998	Saginaw, MI	564	Waco, TX	477
Wilmington, DE	773	Syracuse, NY	405	Wichita Falls, TX	354
Wilmington, NC	887	Worcester, MA	872	Yakima, WA	264

97-98

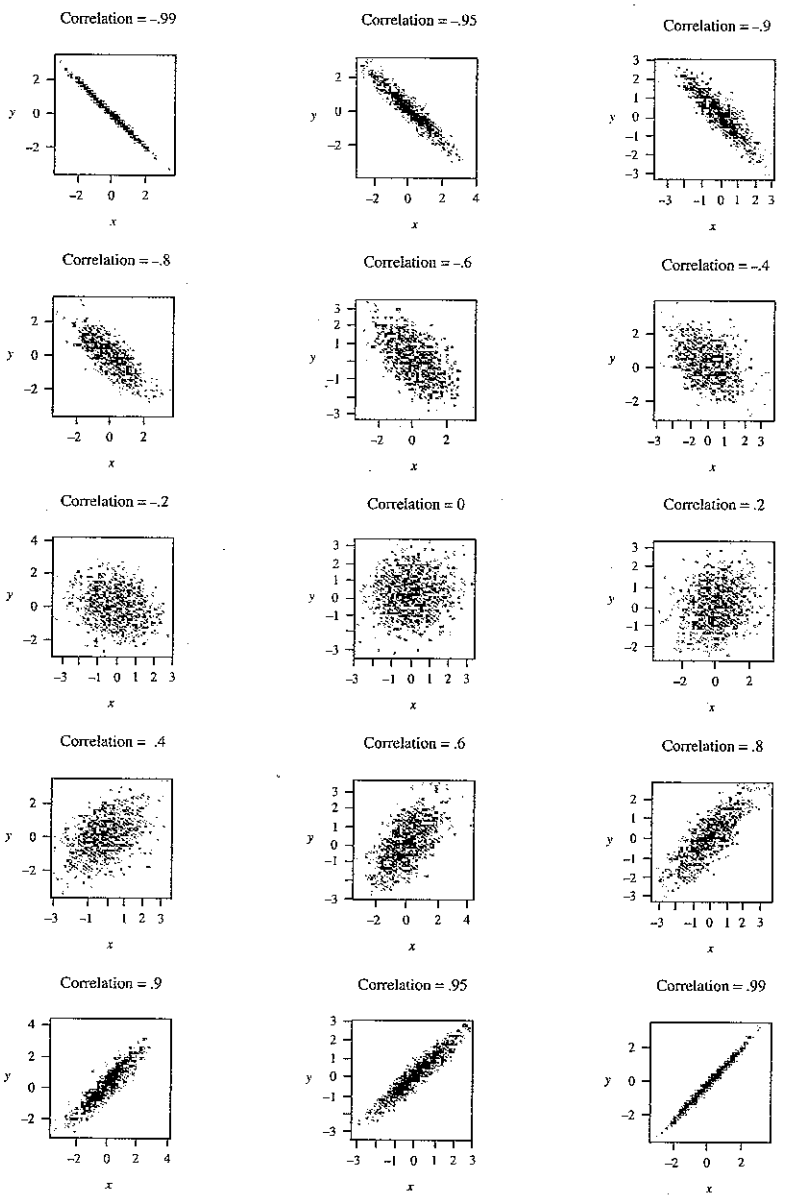


FIGURE 3.31
Scatterplots showing various values for r

27.98

observation. The sample standard deviation is 26.10 for the placebo group and 27.37 for the treatment group. This seemingly inconsistent result occurs due to the large base count for a single patient in the treatment group. The median number of base seizures is higher for the treatment group than for the placebo group. The means are nearly identical for the two groups. The means are in greater agreement than are the medians due to the skewed-to-the-right distribution of the middle 50% of the data for the placebo group, whereas the treatment group is nearly symmetric for the middle 50% of its data. Figure 3.32(b) displays the nearly identical distribution of age for the two treatment groups; the only difference is that the treatment group has a slightly smaller median age and is slightly more variable than the placebo group. Thus, the two groups appear to have similar age and base-seizure distributions prior to the start of the clinical trials.

Brand	Price per Roll	Number of Sheets per Roll	Cost per Sheet
7	0.79	52	.0152
8	0.75	72	.0104
9	0.72	80	.0090
10	0.53	52	.0102
11	0.59	85	.0069
12	0.89	80	.0111
13	0.67	85	.0079
14	0.66	80	.0083
15	0.59	80	.0074
16	0.76	80	.0095
17	0.85	85	.0100
18	0.59	85	.0069
19	0.57	78	.0073
20	1.78	180	.0099
21	1.98	180	.0011
22	0.67	100	.0067
23	0.79	100	.0079
24	0.55	90	.0061

.0100

a. Compute the standard deviation for both the price per roll and the price per sheet.
b. Which is more variable, price per roll or price per sheet?
c. In your comparison in part (b), should you use s or CV? Justify your answer.

3.46 Refer to Exercise 3.45. Use a scatterplot to plot the price per roll and number of sheets per roll.
a. Do the 24 points appear to fall on a straight line?
b. If not, is there any other relation between the two prices?
c. What factors may explain why the ratio of price per roll to number of sheets is not a constant?

3.47 Construct boxplots for both price per roll and number of sheets per roll. Are there any "unusual" brands in the data?

Env. **3.48** The paper "Conditional simulation of waste-site performance" [*Technometrics* (1994) 36: 129–161] discusses the evaluation of a pilot facility for demonstrating the safe management, storage, and disposal of defense-generated, radioactive, transuranic waste. Researchers have determined that one potential pathway for release of radionuclides is through contaminant transport in groundwater. Recent focus has been on the analysis of transmissivity, a function of the properties and the thickness of an aquifer that reflects the rate at which water is transmitted through the aquifer. The following table contains 41 measurements of transmissivity, T, made at the pilot facility.

9.354	6.302	24.609	10.093	0.939	354.81	15399.27	88.17	1253.43	0.75	312.10
1.94	3.28	1.32	7.68	2.31	16.69	2772.68	0.92	10.75	0.000753	
1.08	741.99	3.23	6.45	2.69	3.98	2876.07	12201.13	4273.66	207.06	
2.50	2.80	5.05	3.01	462.38	5515.69	118.28	10752.27	956.97	20.43	

a. Draw a relative frequency histogram for the 41 values of T.
b. Describe the shape of the histogram.
c. When the relative frequency histogram is highly skewed to the right, the Empirical Rule may not yield very accurate results. Verify this statement for the data given.
d. Data analysts often find it easier to work with mound-shaped relative frequency histograms. A transformation of the data will sometimes achieve this shape. Replace the given 41 T values with the logarithm base 10 of the values and reconstruct the relative frequency histogram. Is the shape more mound-shaped than the original data? Apply

Definition 4.9 leads to a special case of $P(A \cap B)$. When events A and B are independent, it follows that

$$P(A \cap B) = P(A)P(B|A) = P(A)P(B)$$

The concept of independence is of particular importance in sampling. Later in the text, we will discuss drawing samples from two (or more) populations to compare the population means, variances, or some other population parameters. For most of these applications, we will select samples in such a way that the observed values in one sample are independent of the values that appear in another sample. We call these **independent samples.**

independent samples

4.5 Bayes' Formula

In this section, we will show how Bayes' Formula can be used to update conditional probabilities by using sample data when available. These "updated" conditional probabilities are useful in decision making. A particular application of these techniques involves the evaluation of diagnostic tests. Suppose a meat inspector must decide whether a randomly selected meat sample contains *E. coli* bacteria. The inspector conducts a diagnostic test. Ideally, a positive result (Pos) would mean that the meat sample actually has *E. coli*, and a negative result (Neg) would imply that the meat sample is free of *E. coli*. However, the diagnostic test is occasionally in error. The results of the test may be a **false positive,** for which the test's indication of *E. coli* presence is incorrect, or a **false negative,** for which the test's conclusion of *E. coli* absence is incorrect. Large-scale screening tests are conducted to evaluate the accuracy of a given diagnostic test. For example, *E. coli* (E) is placed in 10,000 meat samples, and the diagnostic test yields a positive result for 9,500 samples and a negative result for 500 samples; that is, there are 500 false negatives out of the 10,000 tests. Another 10,000 samples have all traces of *E. coli* (NE) removed, and the diagnostic test yields a positive result for 100 samples and a negative result for 9,900 samples. There are 100 false positives out of the 10,000 tests. We can summarize the results in Table 4.3.

false positive
false negative

Evaluation of test results is as follows:

$$\text{True positive rate} = P(\text{Pos}|\text{E}) = \frac{9{,}500}{10{,}000} = .95$$

$$\text{False positive rate} = P(\text{Pos}|\text{NE}) = \frac{100}{10{,}000} = .01$$

$$\text{True negative rate} = P(\text{Neg}|\text{NE}) = \frac{9{,}900}{10{,}000} = .99$$

$$\text{False negative rate} = P(\text{Neg}|\text{NE}) = \frac{500}{10{,}000} = .05$$

$P(Neg|E)$

TABLE 4.3
E. coli test data

Diagnostic Test Result	Meat Sample Status	
	E	NE
Positive	9,500	100
Negative	500	9,900
Total	10,000	10,000

b. The quality control section of a large chemical manufacturing company has undertaken an intensive process-validation study. From this study, the QC section claims that the probability that the shelf life of a newly released batch of chemical will exceed the minimal time specified is .998.

c. A new blend of coffee is being contemplated for release by the marketing division of a large corporation. Preliminary marketing survey results indicate that 550 of a random sample of 1,000 potential users rated this new blend better than a brandname competitor. The probability of this happening is approximately .001, assuming that there is actually no difference in consumer preference for the two brands.

d. The probability that a customer will receive a package the day after it was sent by a business using an "overnight" delivery service is .92.

e. The sportscaster in College Station, Texas, states that the probability that the Aggies will win their football game against the University of Florida is .75.

f. The probability of a nuclear power plant having a meltdown on a given day is .00001.

g. If a customer purchases a single ticket for the Texas lottery, the probability of that ticket being the winning ticket is 1/15,890,700.

4.2 A study of the response time for emergency care for heart attack victims in a large U.S. city reported that there was a 1 in 200 chance of the patient surviving the attack. That is, for a person suffering a heart attack in the city, $P(\text{survival}) = 1/200 = .05$. The low survival rate was attributed to many factors associated with large cities, such as heavy traffic, misidentification of addresses, and the use of phones for which the 911 operator could not obtain an address. The study documented the 1/200 probability based on a study of 20,000 requests for assistance by victims of a heart attack.

a. Provide a relative frequency interpretation of the .05 probability.

b. The .05 was based on the records of 20,000 requests for assistance from heart attack victims. How many of the 20,000 in the study survived? Explain your answer.

4.3 A casino claims that every pair of dice in use are completely fair. What is the meaning of the term *fair* in this context?

4.4 A baseball player is in a deep slump, having failed to obtain a base hit in his previous 20 times at bat. On his 21st time at bat, he hits a game-winning home run and proceeds to declare that "he was due to obtain a hit." Explain the meaning of his statement.

4.5 In advocating the safety of flying on commercial airlines, the spokesperson of an airline stated that the chance of a fatal airplane crash was 1 in 10 million. When asked for an explanation, the spokesperson stated that you could fly daily for the next 27,000 years (27,000(365) = 9,855,000 days) before you would experience a fatal crash. Discuss why this statement is misleading.

4.2 Finding the Probability of an Event

Edu. **4.6** Suppose an exam consists of 20 true-or-false questions. A student takes the exam by guessing the answer to each question. What is the probability that the student correctly answers 15 or more of the questions? [*Hint:* Use a simulation approach. Generate a large number (2,000 or more sets) of 20 single-digit numbers. Each number represents the answer to one of the questions on the exam, with even digits representing correct answers and odd digits representing wrong answers. Determine the relative frequency of the sets having 15 or more correct answers.]

Med. **4.7** The example in Section 4.1 considered the reliability of a screening test. Suppose we wanted to simulate the probability of observing at least 15 positive results and 5 negative results in a set of 20 results, when the probability of a positive result was claimed to be .75. Use a random number generator to simulate the running of 20 screening tests.

a. Let a two-digit number represent an individual running of the screening test. Which numbers represent a positive outcome of the screening test? Which numbers represent a negative outcome?

b. If we generate 2,000 sets of 20 two-digit numbers, how can the outcomes of this simulation be used to approximate the probability of obtaining at least 15 positive results in the 20 runnings of the screening test?

4.8 The state consumers affairs office provided the following information on the frequency of automobile repairs for cars 2 years old or older: 20% of all cars will require repairs once

	Number Surveyed	Number Responding "Poor"
Site 1	192	48
Site 2	248	80

Let A be the event the worker comes from Site 1 and B be the event the response is "poor." Compute $P(A)$, $P(B)$, and $P(A \cap B)$.

4.24

4.25 Refer to Exercise 4.23

a. Are events A and B independent?

b. Find $P(B|A)$ and $P(B|\overline{A})$. Are they equal?

H.R. **4.26** A large corporation has spent considerable time developing employee performance rating scales to evaluate an employee's job performance on a regular basis, so major adjustments can be made when needed and employees who should be considered for a "fast track" can be isolated. Keys to this latter determination are ratings on the ability of an employee to perform to his or her capabilities and on his or her formal training for the job.

	Formal Training			
Workload Capacity	**None**	**Little**	**Some**	**Extensive**
Low	.01	.02	.02	.04
Medium	.05	.06	.07	.10
High	.10	.15	.16	.22

The probabilities for being placed on a fast track are as indicated for the 12 categories of workload capacity and formal training. The following three events (A, B, and C) are defined:

A: An employee works at the high-capacity level

B: An employee falls into the highest (extensive) formal training category

C: An employee has little or no formal training and works below high capacity

a. Find $P(A)$, $P(B)$, and $P(C)$.

b. Find $P(A|B)$, $P(B|\overline{B})$, and $P(\overline{B}|C)$.

c. Find $P(A \cup B)$, $P(A \cap C)$, and $P(B \cap C)$.

Bus. **4.27** The utility company in a large metropolitan area finds that 70% of its customers pay a given monthly bill in full.

a. Suppose two customers are chosen at random from the list of all customers. What is the probability that both customers will pay their monthly bill in full?

b. What is the probability that at least one of them will pay in full?

4.28 Refer to Exercise 4.27. A more detailed examination of the company records indicates that 95% of the customers who pay one monthly bill in full will also pay the next monthly bill in full; only 10% of those who pay less than the full amount one month will pay in full the next month.

a. Find the probability that a customer selected at random will pay two consecutive months in full.

b. Find the probability that a customer selected at random will pay neither of two consecutive months in full.

c. Find the probability that a customer chosen at random will pay exactly one month in full.

4.5 Bayes' Formula

Bus. **4.29** Of a finance company's loans, 1% are defaulted (not completely repaid). The company routinely runs credit checks on all loan applicants. It finds that 30% of defaulted loans went to poor risks, 40% to fair risks, and 30% to good risks. Of the nondefaulted loans, 10% went to poor risks, 40% to fair risks, and 50% to good risks. Use Bayes' Formula to calculate the probability that a poor-risk loan will be defaulted.

4.30 Refer to Exercise 4.29. Show that the posterior probability of default, given a fair risk, equals the prior probability of default. Explain why this is a reasonable result.

c. Use a normal approximation to compute the probability that less than six sales are made.
d. Use a Poisson approximation to compute the probability that less than six sales are made.
e. Use a computer program (if available) to compute the exact probability that less than six sales are made. Compare this result with your calculations in (c) and (d).

4.52 A certain birth defect occurs in 1 of every 10,000 births. In the next 5,000 births at a major hospital, what is the probability that at least one baby will have the defect? What assumptions are required to calculate this probability?

4.10 A Continuous Probability Distribution: The Normal Distribution

4.53 Use Table 1 of the Appendix to find the area under the normal curve between these values:
a. $z = 0$ and $z = 1.6$
b. $z = 0$ and $z = 2.3$

4.54 Repeat Exercise 4.53 for these values:
a. $z = .7$ and $z = 1.7$
b. $z = -1.2$ and $z = 0$

4.55 Repeat Exercise 4.53 for these values:
a. $z = -1.29$ and $z = 0$
b. $z = -.77$ and $z = 1.2$

4.56 Repeat Exercise 4.53 for these values:
a. $z = -1.35$ and $z = -.21$
b. ~~$z = -.37$ and $z = -1.20$~~ $z = -1.20$ and $z = -.37$

4.57 Find the probability that z is greater than 1.75.

4.58 Find the probability that z is less than 1.14.

4.59 Find a value for z, say z_0, such that $P(z > z_0) = .5$.

4.60 Find a value for z, say z_0, such that $P(z > z_0) = .025$.

4.61 Find a value for z, say z_0, such that $P(z > z_0) = .0089$.

4.62 Find a value for z, say z_0, such that $P(z > z_0) = .05$.

4.63 Find a value for z, say z_0, such that $P(-z_0 < z < z_0) = .95$.

4.64 Let y be a normal random variable with mean equal to 100 and standard deviation equal to 8. Find the following probabilities:
a. $P(y > 100)$
b. $P(y > 105)$
c. $P(y < 110)$
d. $P(88 < y < 120)$
e. $P(100 < y < 108)$

4.65 Let y be a normal random variable with $\mu = 500$ and $\sigma = 100$. Find the following probabilities:
a. $P(500 < y < 665)$
b. $P(y > 665)$
c. $P(304 < y < 665)$
d. k such that $P(500 - k < y < 500 + k) = .60$

4.66 Suppose that y is a normal random variable with $\mu = 100$ and $\sigma = 15$.
a. Show that $y < 115$ is equivalent to $z < 1$.
b. Convert $y > 85$ to the z-score equivalent.
c. Find $P(y < 115)$ and $P(y > 85)$.
d. Find $P(y > 106)$, $P(y < 94)$, and $P(94 < y < 106)$.
e. Find $P(y < 70)$, $P(y > 130)$, and $P(70 < y < 130)$.

4.67 Find the value of z for these areas.
a. an area .025 to the right of z
b. an area .05 to the left of z

first person on the list and 1,000 to the last person. You need to next obtain a random sample of 50 numbers from the numbers 1 to 1,000. The names on the sampling frame corresponding to these 50 numbers will be the 50 persons selected for the poll. A Minitab program is shown here for purposes of illustration. Note that you would need to run this program 230 separate times to obtain a new random sample for each of the 230 precincts.

Follow these steps:

Click on **Calc.**
Click on **Random Data.**
Click on **Integer.**
Type **5** in the **Generate rows of data** box.
Type **c1–c10** in the **Store in Column(s):** box.
Type **1** in the **Minimum value:** box.
Type **1000** in the **Maximum value:** box.
Click on **OK.**
Click on **File.**
Click on **Print Worksheet.**

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
1	340	701	684	393	313	312	834	596	321	739
2	783	877	724	498	315	282	175	611	725	571
3	862	625	971	30	766	256	40	158	444	546
4	974	402	768	593	980	536	483	244	51	201
5	232	742	1	861	335	129	409	724	340	218

a. Using either a random number table or a computer program, generate a second random sample of 50 numbers from the numbers 1 to 1,000.
b. Give several reasons why you need to generate a different set of random numbers for each of the precincts. Why not use the same set of 50 numbers for all 230 precincts?

4.12 Sampling Distributions

4.77 A random sample of 16 measurements is drawn from a population with a mean of 60 and a standard deviation of 5. Describe the sampling distribution of $\bar{y}$, the sample mean. Within what interval would you expect $\bar{y}$ to lie approximately 95% of the time?

4.78 Refer to Exercise 4.77. Describe the sampling distribution for the sample sum Σy_i. Is it unlikely (improbable) that Σy_i would be more than 70 units away from 960? Explain.

Psy. **4.79** Psychomotor retardation scores for a large group of manic-depressive patients were approximately normal, with a mean of 930 and a standard deviation of 130.

a. ~~What fraction of the patients scored between 800 and 1,100?~~ Between 900 and 960?
b. Less than ~~800~~? 900?
c. Greater than ~~1,200~~? 1000?

A mean retardation score was computed for a random sample of 20 patients. What is the probability that their mean score was

Soc. **4.80** Federal resources have been tentatively approved for the construction of an outpatient clinic. In order to design a facility that will handle patient load requirements and stay within a limited budget, the designers studied patient demand. From studying a similar facility in the area, they found that the distribution of the number of patients requiring hospitalization during a week could be approximated by a normal distribution with a mean of 125 and a standard deviation of 32.

a. Use the Empirical Rule to describe the distribution of y, the number of patients requesting service in a week.
b. If the facility was built with a 160-patient capacity, what fraction of the weeks might the clinic be unable to handle the demand?

4.88 Use the binomial distribution with $n = 20$, $\pi = .5$ to compare accuracy of the normal approximation to the binomial.

- **a.** Compute the exact probabilities and corresponding normal approximations for $y < 5$.
- **b.** The normal approximation can be improved slightly by taking $P(y \leq 4.5)$. Why should this help? Compare your results.
- **c.** Compute the exact probabilities and corresponding normal approximations with the continuity correction for $P(8 < y < 14)$.

4.89 Let y be a binomial random variable with $n = 10$ and $\pi = .5$.

- **a.** Calculate $P(4 \leq y \leq 6)$.
- **b.** Use a normal approximation without the continuity correction to calculate the same probability. Compare your results. How well did the normal approximation work?

4.90 Refer to Exercise 4.89. Use the continuity correction to compute the probability $P(4 \leq y \leq 6)$. Does the continuity correction help?

Bus. **4.91** A marketing research firm believes that approximately 12.5% of all persons mailed a sweepstakes offer will respond if a preliminary mailing of 10,000 is conducted in a fixed region.

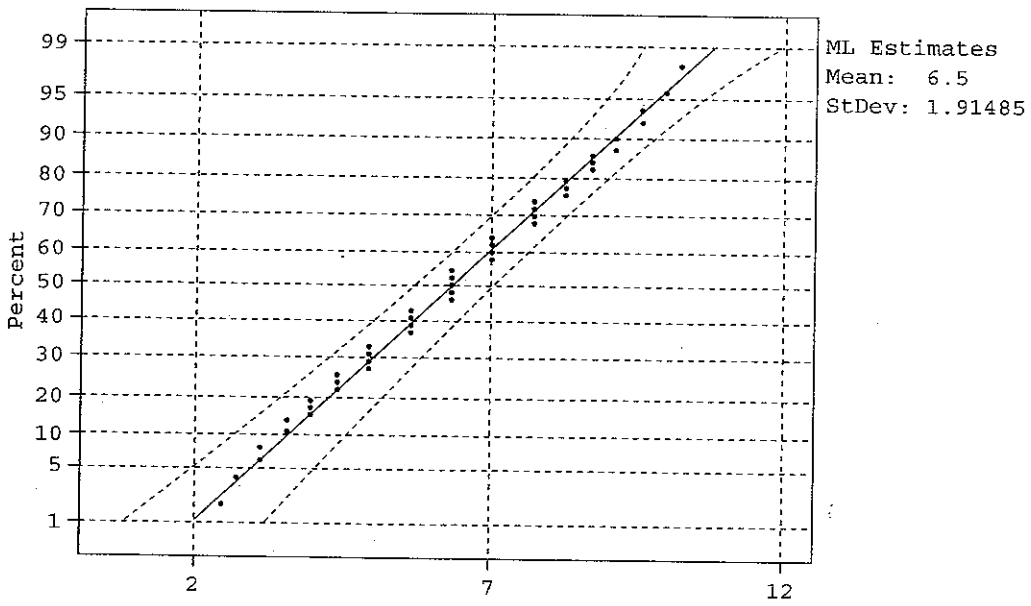
- **a.** What is the probability that 1,000 or fewer will respond?
- **b.** What is the probability that 3,000 or more will respond?

4.14 Evaluating Whether or Not a Population Distribution Is Normal

4.92 In Figure 4.19, we visually inspected the relative frequency histogram for sample means based on two measurements and noted its bell shape. Another way to determine whether a set of measurements is bell-shaped (normal) is to construct a **normal probability plot** of the sample data. If the plotted points are nearly a straight line, we say the measurements were selected from a normal population. We can generate a normal probability plot using the following Minitab code. If the plotted points fall within the curved dotted lines, we consider the data to be a random sample from a normal distribution.

Minitab code:

1. Enter the 45 measurements into C1 of the data spreadsheet.
2. Click on **Graph,** then **Probability Plot.**
3. Type **c1** in the box labeled **Variables.**
4. Click on **OK.**



a. Does it appear that the 45 data values appear to be a random sample from a normal distribution?
b. Compute the correlation coefficient and p-value to assess whether the data appear to be sampled from a normal distribution.
c. Do the results in part (b) confirm your conclusion from part (a)?

4.93 Suppose a population consists of the 10 measurements (2, 3, 6, 8, 9, 12, 25, 29, 39, 50). Generate the 45 possible values for the sample mean based on a sample of $n = 2$ observations per sample.

a. Use the 45 sample means to determine whether the sampling distribution of the sample mean is approximately normally distributed by constructing a boxplot, relative frequency histogram, and normal quantile plot of the 45 sample means.
b. Compute the correlation coefficient and p-value to assess whether the 45 means appear to be sampled from a normal distribution.
c. Do the results in part (b) confirm your conclusion from part (a)?

4.94 The fracture toughness in concrete specimens is a measure of how likely blocks used in new home construction may fail. A construction investigator obtains a random sample of 15 concrete blocks and determines the following toughness values:

.47, .58, .67, .70, .77, .79, .81, .82, .84, .86, .91, .95, .98, 1.01, 1.04

a. Use a normal quantile plot to assess whether the data appear to fit a normal distribution.
b. Compute the correlation coefficient and p-value for the normal quantile plot. Comment on the degree of fit of the data to a normal distribution.

Supplementary Exercises

Bus. **4.95** One way to audit expense accounts for a large consulting firm is to sample all reports dated the last day of each month. Comment on whether such a sample constitutes a random sample.

Engin. **4.96** The breaking strengths for 1-foot-square samples of a particular synthetic fabric are approximately normally distributed with a mean of 2,250 pounds per square inch (psi) and a standard deviation of 10.2 psi.

a. Find the probability of selecting a 1-foot-square sample of material at random that on testing would have a breaking strength in excess of 2,265 psi.
b. Describe the sampling distribution for $\bar{y}$ based on random samples of 15 1-foot sections.

4.97 Refer to Exercise 4.96. Suppose that a new synthetic fabric has been developed that may have a different mean breaking strength. A random sample of 15 one-foot sections is obtained and each section is tested for breaking strength. If we assume that the population standard deviation for the new fabric is identical to that for the old fabric, give the standard deviation for the sampling distribution of $\bar{y}$ using the new fabric.

4.98 Refer to Exercise 4.97. Suppose that the mean breaking strength for the sample of 15 one-foot sections of the new synthetic fabric is 2,268 psi. What is the probability of observing a value of $\bar{y}$ equal to or greater than 2,268, assuming that the mean breaking strength for the new fabric is 2,250, the same as that for the old?

4.99 Based on your answer in Exercise 4.98, do you believe the new fabric has the same mean breaking strength as the old? (Assume $\sigma = 10.2$.)

Gov. **4.100** Suppose that you are a regional director of an IRS office and that you are charged with sampling 1% of the returns with gross income levels above $15,000. How might you go about this? Would you use random sampling? How?

Med. **4.101** Experts consider high serum cholesterol levels to be associated with an increased incidence of coronary heart disease. Suppose that the natural logarithm of cholesterol levels for males in a given age bracket is normally distributed with a mean of 5.35 and a standard deviation of .12.

a. What percentage of the males in this age bracket could be expected to have a serum cholesterol level greater than 250 mg/ml, the upper limit of the clinical normal range?

5.5 Choosing the Sample Size for Testing μ

The quantity of information available for a statistical test about μ is measured by the magnitudes of the Type I and II error probabilities, α and $\beta(\mu)$, for various values of μ in the alternative hypothesis H_a. Suppose that we are interested in testing H_0: $\mu \leq \mu_0$ against the alternative H_a: $\mu > \mu_0$. First, we must specify the value of α. Next we must determine a value of μ in the alternative, μ_1, such that if the actual value of the mean is larger than μ_1, then the consequences of making a Type II error would be substantial. Finally, we must select a value for $\beta(\mu_1)$, β. Note that for any value of μ larger than μ_1, the probability of Type II error will be smaller than $\beta(\mu_1)$; that is,

$$\beta(\mu) < \beta(\mu_1), \text{ for all } \mu > \mu_1$$

Let $\Delta = \mu_1 - \mu_0$. The sample size necessary to meet these requirements is

$$n = \sigma^2 \frac{(z_\alpha + z_\beta)^2}{\Delta^2}$$

Note: If σ^2 is unknown, substitute an estimated value from previous studies or a pilot study to obtain an approximate sample size.

The same formula applies when testing H_0: $\mu \geq \mu_0$ against the alternative H_a: $\mu < \mu_0$, with the exception that we want the probability of a Type II error to be of magnitude β or less when the actual value of μ is *less* than μ_1, a value of the mean in H_a; that is,

$$\beta(\mu) < \beta, \text{ for all } \mu < \mu_1$$

with $\Delta = \mu_0 - \mu_1$.

16

EXAMPLE 5.11

A cereal manufacturer produces cereal in boxes having a labeled weight of 12 ounces. The boxes are filled by machines that are set to have a mean fill per box of 16.37 ounces. Because the actual weight of a box filled by these machines has a normal distribution with a standard deviation of approximately .225 ounces, the percentage of boxes having weight less than 16 ounces is 5% using this setting. The manufacturer is concerned that one of its machines is underfilling the boxes and wants to sample boxes from the machine's output to determine whether the mean weight μ is less than 16.37—that is, to test

$$H_0: \ \mu \geq 16.37$$
$$H_a: \ \mu < 16.37$$

with $\alpha = .05$. If the true mean weight is 16.27 or less, the manufacturer needs the probability of failing to detect this underfilling of the boxes with a probability of at most .01, or risk incurring a civil penalty from state regulators. Thus, we need to determine the sample size n such that our test of H_0 versus H_a has $\alpha = .05$ and $\beta(\mu)$ less than .01 whenever μ is less than 16.27 ounces.

Solution We have $\alpha = .05$, $\beta = .01$, $\Delta = 16.37 - 16.27 = .1$, and $\sigma = .225$. Using our formula with $z_{.05} = 1.645$ and $z_{.01} = 2.33$, we have

$$n = \frac{(.225)^2(1.645 + 2.33)^2}{(.1)^2} = 79.99 \approx 80$$

Thus, the manufacturer must obtain a random sample of $n = 80$ boxes to conduct this test under the specified conditions.

Because the confidence limits are computed using the binomial distribution, which is a discrete distribution, the level of confidence of (M_L, M_U) will generally be somewhat larger than the specified $100(1-\alpha)\%$. The exact level of confidence is given by

$$\text{Level} = 1 - 2P[\text{Bin}(n, .5) \leq C_{\alpha(2),n}]$$

The following example will demonstrate the construction of the interval.

EXAMPLE 5.20

The sanitation department of a large city wants to investigate ways to reduce the amount of recyclable materials that are placed in the city's landfill. By separating the recyclable material from the remaining garbage, the city could prolong the life of the landfill site. More important, the number of trees needed to be harvested for paper products and the aluminum needed for cans could be greatly reduced. From an analysis of recycling records from other cities, it is determined that if the average weekly amount of recyclable material is more than 5 pounds per household, a commercial recycling firm could make a profit collecting the material. To determine the feasibility of the recycling plan, a random sample of 25 households is selected. The weekly weight of recyclable material (in pounds/week) for each household is given here.

14.2	5.3	2.9	4.2	1.2	4.3	1.1	2.6	6.7	7.8	25.9	43.8	2.7
5.6	7.8	3.9	4.7	6.5	29.5	2.1	34.8	3.6	5.8	4.5	6.7	

Determine an appropriate measure of the amount of recyclable waste from a typical household in the city.

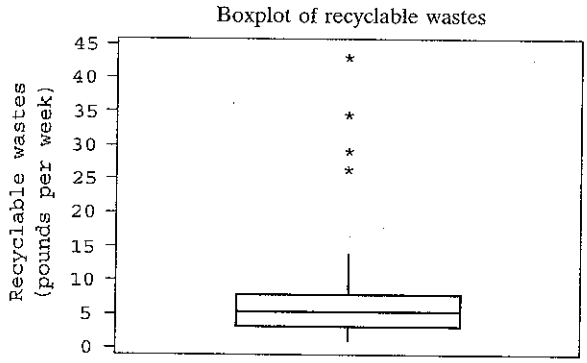


FIGURE 5.22(a) Boxplot for waste data

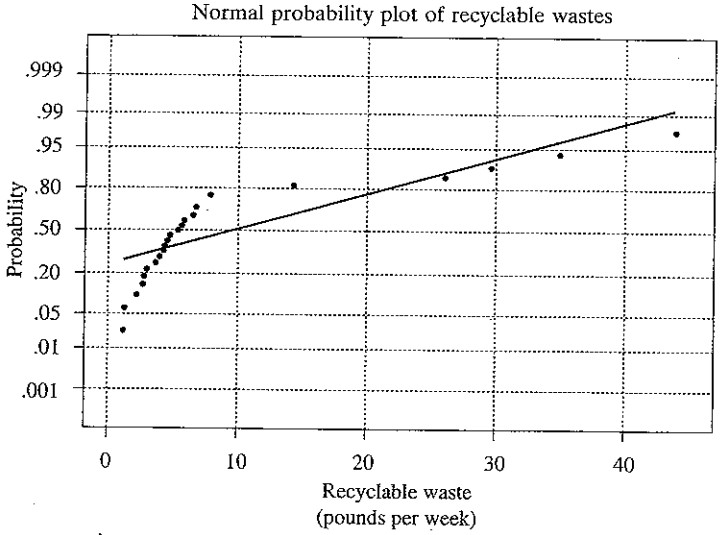


FIGURE 5.22(b) Normal probability plot for waste data

Engin. **5.3** Face masks used by firefighters often fail by having their lenses fall out when exposed to very high temperatures. A manufacturer of face masks claims that for their masks the average temperature at which pop out occurs is 550°F. A sample of 75 masks are tested and the average temperature at which the lense popped out was 470°F. Based on this information is the manufacturer's claim valid?

a. Identify the population of interest to us in this problem.

b. Would an answer to the question posed involve estimation or testing a hypothesis?

5.4 Refer to Exercise 5.3. How might you select a sample of face masks from the manufacturer to evaluate the claim?

5.2 Estimation of μ

Engin. **5.5** A company that manufacturers coffee for use in commercial machines monitors the caffeine content in its coffee. The company selects 50 samples of coffee every hour from its production line and determines the caffeine content. From historical data, the caffeine content (in milligrams, mg) is known to have a normal distribution with $\sigma = 7.1$ mg. During a 1-hour time period, the 50 samples yielded a mean caffeine content of $\bar{y} = 110$ mg.

a. Calculate a 95% confidence interval for the mean caffeine content μ of the coffee produced during the hour in which the 50 samples were selected.

b. Explain to the CEO of the company in nonstatistical language, the interpretation of the constructed confidence interval.

5.6 Refer to Exercise 5.5. The engineer in charge of the coffee manufacturing process examines the confidence intervals for the mean caffeine content calculated over the past several weeks and is concerned that the intervals are too wide to be of any practical use. That is, they are not providing a very precise estimate of μ.

a. What would happen to the width of the confidence intervals if the level of confidence of each interval is increased from 95% to 99%?

b. What would happen to the width of the confidence intervals if the number of samples per hour was increased from 50 to 100?

5.7 Refer to Exercise 5.5. Because the company is sampling the coffee production process every hour, there are 720 confidence intervals for the mean caffeine content μ constructed every month.

a. If the level of confidence remains at 95% for the 720 confidence intervals in a given month, how many of the confidence intervals would you expect to fail to contain the value of μ and hence provide an incorrect estimation of the mean caffeine content?

b. If the number of samples is increased from 50 to 100 each hour, how many of the 95% confidence intervals would you expect to fail to contain the value of μ in a given month?

c. If the number of samples remains at 50 each hour but the level of confidence is increased from 95% to 99% for each of the intervals, how many of the 95% confidence intervals would you expect to fail to contain the value of μ in a given month? 99%

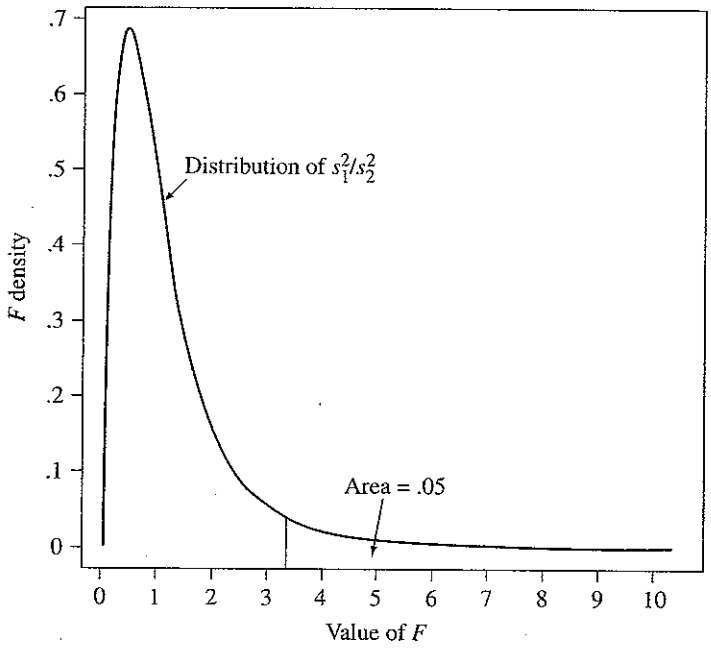
Bus. **5.8** As part of the recruitment of new businesses in their city, the economic development department of the city wants to estimate the gross profit margin of small businesses (under one million dollars in sales) currently residing in their city. A random sample of the previous years annual reports of 15 small businesses shows the mean net profit margins to be 7.2% (of sales) with a standard deviation of 12.5%.

a. Construct a 99% confidence interval for the mean gross profit margin of μ of all small businesses in the city.

b. The city manager reads the report and states that the confidence interval for μ constructed in part (a) is not valid because the data are obviously not normally distributed and thus the sample size is too small. Based on just knowing the mean and standard deviation of the sample of 15 businesses, do you think the city manager is valid in his conclusion about the data? Explain your answer.

Soc. **5.9** A social worker is interested in estimating the average length of time spent outside of prison for first offenders who later commit a second crime and are sent to prison again. A random sample of $n = 150$ prison records in the county courthouse indicates that the average length of prison-free life between first and second offenses is 3.2 years, with a standard deviation of 1.1 years. Use the

FIGURE 7.8
Critical value for the F distributions ($df_1 = 5$, $df_2 = 10$)



A statistical test comparing σ_1^2 and σ_2^2 utilizes the test statistic s_1^2/s_2^2. When $\sigma_1^2 = \sigma_2^2$, $\sigma_1^2/\sigma_2^2 = 1$ and s_1^2/s_2^2 follows an F distribution with $df_1 = n_1 - 1$ and $df_2 = n_2 - 1$. For a one-tailed alternative hypothesis, the designation of which population is 1 and which population is 2 is made such that H_a is of the form $\sigma_1^2 > \sigma_2^2$. Then the rejection region is located in the upper-tail of the F distribution.

We summarize the test procedure next.

A Statistical Test Comparing σ_1^2 and σ_2^2

H_0: **1.** $\sigma_1^2 \leq \sigma_2^2$ $\quad H_a$: **1.** $\sigma_1^2 > \sigma_2^2$

2. $\sigma_1^2 = \sigma_2^2$ $\quad H_a$: **2.** $\sigma_1^2 \neq \sigma_2^2$

T.S.: $F = s_1^2/s_2^2$

R.R.: For a specified value of α and with $df_1 = n_1 - 1$, $df_2 = n_2 - 1$,

1. Reject H_0 if $F \geq F_{\alpha,df_1,df_2}$.

2. Reject H_0 if $F \leq F_{1-\alpha/2,df_1,df_2}$ or if $F \geq F_{\alpha/2,df_1,df_2}$.

F

Table 8 in the Appendix provides the upper percentiles of the F distribution. The lower percentiles are obtained from the upper percentiles using the following relationship. Let F_{α,df_1,df_2} be the upper α percentile and $F_{1-\alpha,df_1,df_2}$ be the lower α percentile of an F distribution with df_1 and df_2. Then,

$$F_{1-\alpha,df_1,df_2} = \frac{1}{F_{\alpha,df_2,df_1}}$$

Note that the degrees of freedom have been reversed for the upper F percentile on the right-hand side of the equation.

populations. Hence, we should not use Hartley's F_{max} test for evaluating differences in the variances in this example. The information in Table 7.6 will assist us in calculating the value of the BFL test statistic. The medians of the percentage increase in mileage, y_{ij}s, for the three additives are 5.80, 7.55, and 9.15. We then calculate the absolute deviations of the data values about their respective medians—namely, $z_{1j} = |y_{1j} - 5.80|$, $z_{2j} = |y_{2j} - 7.55|$, and $z_{3j} = |y_{3j} - 9.15|$ for $j = 1, \ldots, 10$. These values are given in column 3 of the table. Next, we calculate the three means of these values, $\bar{z}_{1.} = 4.07$, $\bar{z}_{2.} = 8.88$, and $\bar{z}_{3.} = 2.23$. Next, we calculate the squared deviations of the z_{ij}s about their respective means, $(z_{ij} - \bar{z}_{i.})^2$; that is, $(z_{1j} - 4.07)^2$, $(z_{2j} - 8.88)^2$, and $(z_{3j} - 2.23)^2$. These values are contained in column 6 of the table. Then we calculate the squared deviations of the z_{ij}s about the overall

TABLE 7.6
Percentage increase in mpg from cars driven using three additives

Additive	y_{1j}	$\tilde{y}_1$	$z_{1j} = \lvert y_{1j} - 5.80 \rvert$	$\bar{z}_{1.}$	$(z_{1j} - 4.07)^2$	$(z_{ij} - 5.06)^2$
1	4.2	5.80	1.60	4.07	6.1009	11.9716
1	2.9		2.90		1.3689	4.6656
1	0.2		5.60		2.3409	0.2916
1	25.7		19.90		250.5889	220.2256
1	6.3		0.50		12.7449	20.7936
1	7.2		1.40		7.1289	13.3956
1	2.3		3.50		0.3249	2.4336
1	9.9		4.10		0.0009	0.9216
1	5.3		0.50		12.7449	20.7936
1	6.5		0.70		11.3569	19.0096
Additive	y_{2j}	$\tilde{y}_2$	$z_{2j} = \lvert y_{2j} - 7.55 \rvert$	$\bar{z}_{2.}$	$(z_{1j} - 8.88)^2$	$(z_{2j} - 5.06)^2$
2	0.2	7.55	7.35	8.88	2.3409	5.2441
2	11.3		3.75		26.3169	1.7161
2	0.3		7.25		2.6569	4.7961
2	17.1		9.55		0.4489	20.1601
2	51.0		43.45		1,195.0849	1,473.7921
2	10.1		2.55		40.0689	6.3001
2	0.3		7.25		2.6569	4.7961
2	0.6		6.95		3.7249	3.5721
2	7.9		0.35		72.7609	22.1841
2	7.2		0.35		72.7609	22.1841
Additive	y_{3j}	$\tilde{y}_3$	$z_{3j} = \lvert y_{3j} - 9.15 \rvert$	$\bar{z}_{3.}$	$(z_{1j} - 2.23)^2$	$(z_{3j} - 5.06)^2$
3	7.2	9.15	1.95	2.23	0.0784	9.6721
3	6.4		2.75		0.2704	5.3361
3	9.9		0.75		2.1904	18.5761
3	3.5		5.65		11.6964	0.3481
3	10.6		1.45		0.6084	13.0321
3	10.8		1.65		0.3364	11.6281
3	10.6		1.45		0.6084	13.0321
3	8.4		0.75		2.1904	18.5761
3	6.0		3.15		0.8464	3.6481
3	11.9		2.75		0.2704	5.3361
Total				5.06	1,742.6	1,978.4

TABLE 8.6
An example of an AOV table for a completely randomized design

Source	Sum of Squares	Degrees of Freedom	Mean Square	*F* Test
Between samples	SSB	$t-1$	$s_B^2 = \text{SSB}/(t-1)$	s_B^2/s_W^2
Within samples	SSW	$n_T - t$	$s_W^2 = \text{SSW}/(n_T - t)$	
Totals	TSS	$n_T - 1$		

AOV table

After we complete the *F* test, we then summarize the results of a study in an *analysis of variance table.* The format of an **AOV table** is shown in Table 8.6. The AOV table lists the sources of variability in the first column. The second column lists the sums of squares associated with each source of variability. We showed that the total sum of squares (TSS) can be partitioned into two parts, so SSB and SSW must add up to TSS in the AOV table. The third column of the table gives the degrees of freedom associated with the sources of variability. Again, we have a check; $(t-1) + (n_T - t)$ must add up to $n_T - 1$. The mean squares are found in the fourth column of Table 8.6, and the *F* test for the equality of the t population means is given in the fifth column.

EXAMPLE 8.1

A large body of evidence shows that soy has health benefits for most people. Some of these benefits come largely from isoflavones, plant compounds that have estrogen-like properties. The amount of isoflavones varies widely depending on the type of food processing. A consumer group purchased various soy products and ran laboratory tests to determine the amount of isoflavones in each product. There were three major categories of soy products: cereals and snacks (1), energy bars (2), and veggie burgers (3). Five different products from each of the three categories were selected and the amount of isoflavones (in mg) was determined for an adult serving of the product. The consumer group wanted to determine if the average amount of isoflavones was different for the three sources of soy products. The data are given in Table 8.7. Use these data to test the research hypothesis of a difference in the mean isoflavones level for the three categories. Use $\alpha = .05$.

TABLE 8.7
Isoflavones content from three sources of soy

Source of Soy	Isoflavones Content (mg)					Sample Sizes	Sample Means	Sample Variances
1	5	17	12	10	4	5	9.20	33.7000
2	19	10	9	7	5	5	10.00	29.0000
3	25	15	12	9	8	5	13.80	46.7000
Total						15	11.00	

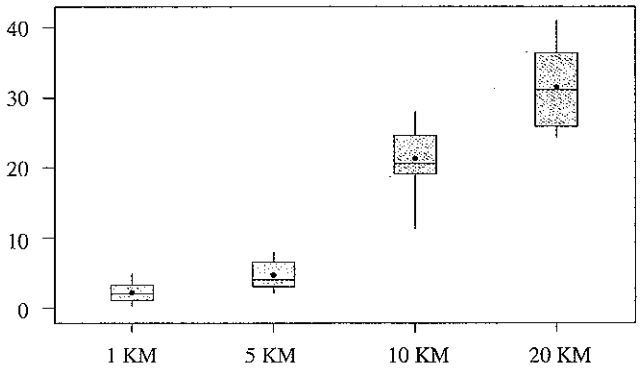
Solution The null and alternative hypotheses for this example are

H_0: $\mu_1 = \mu_2 = \mu_3$

H_a: At least one of the three population means is different from the rest.

The sample sizes are $n_1 = n_2 = n_3 = 5$, which yields $n_T = 15$. Using the sample means and sample variances, the sum of squares within and between are given here with

FIGURE 8.8
Boxplots of 1–20 KM (means are indicated by solid circles)



b. We next examine the relationship between the sample means $\bar{y}_{i.}$ and sample variances s_i^2.

$$\frac{s_1^2}{\bar{y}_{1.}} = .99 \qquad \frac{s_2^2}{\bar{y}_{2.}} = .97 \qquad \frac{s_3^2}{\bar{y}_{3.}} = 1.06 \qquad \frac{s_4^2}{\bar{y}_{4.}} = .97$$

Thus, it would appear that $\sigma_i^2 = k\mu_i$, with $k \approx 1$. From Table 8.15, the suggested transformation is $y_T = \sqrt{y + .375}$. The values of y_T appear in Table 8.17 along with their means and standard deviations. Although the original data had heterogeneous variances, the sample variances are all approximately .25, as indicated in Table 8.17.

TABLE 8.17
Transformation of data in Table 8.16: $y_T = \sqrt{y + .375}$

	Distance to Mouth			
Sample	**1 KM**	**5 KM**	**10 KM**	**20 KM**
1	1.173	2.092	4.514	6.114
2	2.318	2.894	5.136	5.511
3	1.541	1.541	4.937	5.136
4	1.173	1.837	3.373	4.937
5	1.541	2.894	5.327	6.432
6	1.541	2.318	4.514	5.037
7	2.092	2.525	4.402	6.031
8	1.837	2.092	4.402	5.601
9	0.612	1.837	4.623	5.601
10	1.541	1.837	4.937	5.777
Mean	1.54	2.19	4.62	5.62
Variances	.24	.22	.29	.24

$y_T = \log y$

The second transformation indicated in Table 8.15 (↓) is for an experimental situation in which the population variance is proportional to the square of the population mean, or equivalently, where $\sigma = \mu$. That is, the logarithmic transformation is appropriate any time the **coefficient of variation** σ_i/μ_i is constant across the populations of interest.

coefficient of variation

$\sigma^2 = K\mu^2$

a. Do the conditions necessary for conducting the AOV F test appear to be satisfied by these data?
b. Because the data are counts of number of successes for the EDGs, the Poisson model may be an alternative to the normal based analysis. Apply a transformation to the data and then apply the AOV F test to the transformed data.
c. As a second alternative analysis which has fewer restrictions, answer the agency's question by applying the Kruskal–Wallis test to the reliability data.
d. Compare your conclusions to parts (a)–(c). Which of the three procedures do you feel more confident with its conclusion?

Envir. **8.24** Refer to Example 8.4.
a. Apply the Kruskal–Wallis test to determine if there is a difference in the distributions of oxygen content for the various distances to the mouth of the Mississippi River.
b. Does your conclusion differ from the conclusion reached in Exercise 8.16?

Med. **8.25** Refer to Example 8.5.
a. Apply the Kruskal–Wallis test to determine if there is a difference in the distributions of pain reduction for the three analgesics.
b. Does your conclusion differ from the conclusion reached in Exercise 8.22?

Med. **8.26** Refer to Example 8.6.
a. Apply the Kruskal–Wallis test to determine if there is a difference in the distributions of opinions across the four geographical regions.
b. Does your conclusion differ from the conclusion reached in Exercise 8.17?

Engin. **8.27** In the manufacture of soft contact lenses, the actual strength (power) of the lens needs to be very close to the target value for the lenses to properly fit the customer's needs. In the paper, "An ANOM-type test for variances from normal populations," *Technometrics* (1997), 39: 274–283, a comparison of several suppliers is made relative to the consistency of the power of the lenses. The following table contains the deviations from the target power of lenses produced using materials from three different suppliers:

	Lens								
Supplier	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
A	189.9	191.9	190.9	183.8	185.5	190.9	192.8	188.4	189.0
B	156.6	158.4	157.7	154.1	152.3	161.5	158.1	150.9	156.9
C	218.6	208.4	187.1	199.5	202.0	211.1	197.6	204.4	206.8

a. Using the appropriate tests and plots given here, assess whether the data meet the necessary conditions to use an AOV to determine whether there is a significant difference in the mean deviations for the three suppliers. using an F-test.
b. Conduct an AOV with $\alpha = .05$ ~~and compare your results with the conclusions from (a)~~.
c. Apply the Kruskal–Wallis test to evaluate the research hypothesis that the three suppliers have different distributions of deviations.
d. Suppose that a difference in mean deviation of 20 units would have commercial consequences for the manufacture of the lenses. Does there appear to be a *practical* difference in the three suppliers?

When computing the confidence interval for π in those situations where $y = 0$ or $y = 1$ the confidence intervals using the normal approximation would not be valid. We can use the following confidence intervals, which are derived from using the binomial distribution.

100(1 − α)% Confidence Interval for π, when y = 0 or y = n

When $y = 0$, the confidence interval is $(0, 1 - (\alpha/2)^{1/n})$.

When $y = n$, the confidence interval is $((\alpha/2)^{1/n}, 1)$.

EXAMPLE 10.3

A new PC operating system is being developed. The designer claims the new system will be compatible with nearly all computer programs currently being run on Microsoft Windows operating system. A sample of 50 programs are run and all 50 programs perform without error. Estimate π, the proportion of all Microsoft Windows–compatible programs that would run without change on the new operating system. Compute a 95% confidence interval for π.

Solution If we used the standard estimator of π, we would obtain

$$\hat{\pi} = \frac{50}{50} = 1.0$$

Thus, we would conclude that 100% of all programs that are Microsoft Windows–compatible programs would run without alteration on the new operating system. Would this conclusion be valid? Probably not, since we have only investigated a tiny fraction of all Microsoft Windows–compatible programs. Thus, we will use the alternative estimators and confidence interval procedures. The point estimator would be given by

$$\hat{\pi}_{\text{Adj.}} = \frac{\left(n + \frac{3}{8}\right)}{\left(n + \frac{3}{4}\right)} = \frac{\left(50 + \frac{3}{8}\right)}{\left(50 + \frac{3}{4}\right)} = .993$$

A 95% confidence interval for π would be

$$((\alpha/2)^{1/n}, 1) = ((.05/2)^{1/50}, 1) = ((.025)^{.02}, 1) = (.929, 1.0)$$

We would now conclude that we are reasonably confident (95%) a high proportion (between 92.9% and 100%) of all programs that are Microsoft Windows–compatible would run without alteration on the new operating system.

Keep in mind, however, that a sample size that is sufficiently large to satisfy the rule *does not* guarantee that the interval will be informative. It only judges the adequacy of the normal approximation to the binomial—the basis for the confidence level.

Sample size calculations for estimating π follow very closely the procedures we developed for inferences about μ. The required sample size for a $100(1 - \alpha)\%$ confidence interval for π of the form $\hat{\pi} \pm E$ (where E is specified) is found by solving the expression

$$z_{\alpha/2}\sigma_{\hat{\pi}} = E$$

for n. The result is shown here.

Sample Size Required for a 100(1− α)% Confidence Interval for π of the Form $\hat{\pi} \pm E$

$$n = \frac{z_{\alpha/2}^2 \pi(1 - \pi)}{E^2}$$

$\pi = .5$.

Note: Since π is not known, either substitute an educated guess or use $\pi = .5$. Use of $\pi = .5$ will generate the largest possible sample size for the specified confidence interval width, $2E$, and thus will give a conservative answer to the required sample size.

EXAMPLE 10.4

In Example 10.3, the designer of the new operating system has decided to conduct a more extensive study. She wants to determine how many programs to randomly sample in order to estimate the proportion of Microsoft Windows–compatible programs that would perform adequately using the new operating system. The designer wants the estimator to be within .03 of the true proportion using a 95% confidence interval as the estimator.

Solution The designer wants the 95% confidence interval to be of the form $\hat{\pi} \pm .03$. The sample size necessary to achieve this accuracy is given by

$$n = \frac{z_{\alpha/2}^2 \pi(1 - \pi)}{E^2}$$

where the specification of 95% yields $z_{\alpha/2} = z_{.025} = 1.96$ and $E = .03$. If we did not have any prior information about π, then $\pi = .5$ must be used in the formula yielding

$$n = \frac{(1.96)^2 .5(1 - .5)}{(.03)^2} = 1{,}067.1$$

That is, 1,068 programs would need to be tested in order to be 95% confident that the estimate of π is within .03 of the actual value of π. The lower bound of the estimate of π obtained in Example 10.3 was .929. Suppose the designer is not too confident in this value but fairly certain that π is greater than .80. Using $\pi = .8$ as a lower bound then the value of n is given by

$$n = \frac{(1.96)^2 .8(1 - .8)}{(.03)^2} = 682.95$$

Thus, if the designer was fairly certain that the actual value of π was at least .80, then the required sample size can be greatly reduced.

A statistical test about a binomial parameter π is very similar to the large-sample test concerning a population mean presented in Chapter 5. These results are summarized next, with three different alternative hypotheses along with their corresponding rejection regions. Recall that only one alternative is chosen for a particular problem.

Summary of a Statistical Test for π, π_0 Is Specified

H_0:	H_a:
1. $\pi \leq \pi_0$	**1.** $\pi > \pi_0$
2. $\pi \geq \pi_0$	**2.** $\pi < \pi_0$
3. $\pi = \pi_0$	**3.** $\pi \neq \pi_0$

Notation for Comparing Two Binomial Proportions

	Population 1	Population 2
Population proportion	π_1	π_2
Sample size	n_1	n_2
Number of successes	y_1	y_2
Sample proportion	$\hat{\pi}_1 = \frac{y_1}{n_1}$	$\hat{\pi}_2 = \frac{y_2}{n_2}$

Inferences about two binomial proportions are usually phrased in terms of their difference $\pi_1 - \pi_2$, and we use the difference in sample proportions $\hat{\pi}_1 - \hat{\pi}_2$ as part of a confidence interval or statistical test. The sampling distribution for $\hat{\pi}_1 - \hat{\pi}_2$ can be approximated by a normal distribution with mean and standard error given by

$$\mu_{\hat{\pi}_1 - \hat{\pi}_2} = \pi_1 - \pi_2$$

and

$$\sigma_{\hat{\pi}_1 - \hat{\pi}_2} = \sqrt{\frac{\pi_1(1 - \pi_1)}{n_1} + \frac{\pi_2(1 - \pi_2)}{n_2}}$$

This approximation is appropriate, if we apply the same requirements to both binomial populations that we applied in recommending a normal approximation to a binomial (see Chapter 4). Thus, the normal approximation to the distribution of $\hat{\pi}_1 - \hat{\pi}_2$ is appropriate if both $n_i\pi_i$ and $n_i(1 - \pi_i)$ are 5 or more for $i = 1, 2$. Since π_1 and π_2 are not known, the validity of the approximation is made by examining $n_i\hat{\pi}_i$ and $n_i(1 - \hat{\pi}_i)$ for $i = 1, 2$.

Confidence intervals and statistical tests about $\pi_1 - \pi_2$ are straightforward and follow the format we used for comparisons using $\mu_1 - \mu_2$. Interval estimation is summarized here; it takes the usual form, point estimate $\pm$ z (standard error).

100(1 − α)% Confidence Interval for $\pi_1 - \pi_2$

$$\hat{\pi}_1 - \hat{\pi}_2 \pm z_{\alpha/2}\hat{\sigma}_{\hat{\pi}_1 - \hat{\pi}_2},$$

where

$$\hat{\sigma}_{\hat{\pi}_1 - \hat{\pi}_2} = \sqrt{\frac{\hat{\pi}_1(1 - \hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1 - \hat{\pi}_2)}{n_2}}$$

EXAMPLE 10.6

A company test-markets a new product in the Grand Rapids, Michigan, and Wichita, Kansas, metropolitan areas. The company's advertising in the Grand Rapids area is based almost entirely on television commercials. In Wichita, the company spends a roughly equal dollar amount on a balanced mix of television,

	Blood Pressure		
Dietary Cholesterol	**High**	**Low**	**Total**
High	159	91	250
Low	78	172	250
Total	237	263	500

a. Compute the difference in the estimated risk of having high blood pressure $(\hat{\pi}_1 - \hat{\pi}_2)$ for the two groups (low versus high dietary cholesterol intake).
b. Compute the estimated relative risk of having high blood pressure $\left(\frac{\hat{\pi}_1}{\hat{\pi}_2}\right)$ for the two groups (low versus high dietary cholesterol intake).
c. Compute the estimated odds ratio of having high blood pressure for the two groups (low versus high dietary cholesterol intake).
d. Based on your results from (a)–(c), how do the two groups compare?

Med. **10.53** Refer to Exercise 10.52.
a. Is there a significant difference between the low and high dietary cholesterol intake groups relative to their risk of having high blood pressure? Use $\alpha = .05$.
b. Place a 95% confidence interval on the odds ratio of having high blood pressure. What can you conclude about the odds of having high blood pressure for the two groups?
c. Are your conclusions in (a) and (b) consistent?

for low versus high cholestrol intake.

Safety **10.54** The article "Who Wants Airbags" in *Chance* 18 (2005): 3–16 discusses whether air bags should be mandatory equipment in all new automobiles. Using data from the National Highway Traffic Safety Administration (NHTSA), they obtain the following information about fatalities and the usage of air bags and seat belts. All passenger cars sold in the U.S. starting in 1998 are required to have air bags. NHTSA estimates that air bags have saved 10,000 lives as of January 2004. The authors examined accidents in which there was a harmful event (personal or property), and from which at least one vehicle was towed. After some screening of the data, they obtained the following results. (The authors detail in their article the types of screening of the data that was done.)

	Air Bag Installed		
	Yes	**No**	**Total**
Killed	19,276	27,924	47,200
Survived	5,723,539	4,826,982	10,550,521
Total	5,742,815	4,854,906	10,597,721

a. Calculate the odds of being killed in a harmful event car accident for a vehicle with and without air bags. Interpret the two odds.
b. Calculate the odds ratio of being killed in a harmful event car accident with and without air bags. What does this ratio tell you about the importance of having air bags in a vehicle?
c. Is there significant evidence of a difference between vehicles with and without air bags relative to the proportion of persons killed in a harmful event vehicle accident? Use $\alpha = .05$.
d. Place a 95% confidence interval on the odds ratio. Interpret this interval.

10.55 Refer to Exercise 10.54. The authors also collected information about accidents concerning seat belt usage. The article compared fatality rates for occupants using seat belts properly with those for occupants not using seat belts. The data are given here.

the orchard. Thus, the Poisson distribution was suggested as a possible model. Based on the data given here, does the Poisson distribution appear to be a plausible model for the concentration of European red mites on apple trees?

Mites per Leaf	0	1	2	3	4	5	6	7
Frequency	233	127	57	33	30	10	7	3

10.86 A sample of 1,200 individuals arrested for driving under the influence of alcohol was obtained from police records. The research recorded the gender, socioeconomic status (from occupation information), and the number of previous alcohol-related arrests. These data are shown here:

Socioeconomic Status	Number of Previous Alcohol-Related Arrests	Male	Female
Low	0	110	130
	1 or more	90	70
Medium	0	105	101
	1 or more	95	99
High	0	90	80
	1 or more	110	120

Separately for each socioeconomic status group answer the following questions.

a. Is there significant evidence of a difference between males and females with respect to the number of previous alcohol-related arrests?

b. Compute the odds of having a previous alcohol-related arrest for both males and females. Interpret these values.

c. Compute the odds ratio of males to females and place a 95% confidence interval on the odds ratio. Interpret the interval.

d. Compare the results for the three socioeconomic statues.

having a previous alcohol-related arrest for males versus females

10.87 Run the Mantel–Haenszel test for the above data and interpret your results

10.88 A study was conducted to determine the relationship between annual income and number of children per family. Compute percentages for each of the income categories; then run a chi-square test of independence and draw conclusions. Use $\alpha = .10$.

Region	Number of Children per Family	Annual Income	
		<$20,000	≥$20,000
East	≤ 2 children	38	67
	>2 children	220	125
South	≤ 2 children	25	78
	>2 children	120	77
West	≤ 2 children	36	66
	>2 children	95	103

Separately for each region, answer the following questions.

a. Is there significant evidence of an association between annual income and number of children?

b. Compute the odds ratio of low versus high income and place a 95% confidence interval on the odds ratio.

having more than 2 children for

The estimate of σ_ε^2 based on the sample data is the sum of squared residuals divided by $n - 2$, the degrees of freedom. The estimated variance is often shown in computer output as MS(Error) or MS(Residual). Recall that MS stands for "mean square" and is always a sum of squares divided by the appropriate degrees of freedom:

$$s_e^2 = \frac{\Sigma_i (y_i - \hat{y}_i)^2}{n - 2} = \frac{SSE}{n - 2}$$

In the computer output for Example 11.3, SS(Residual) is shown to be 0.9498.

Just as we divide by $n - 1$ rather than by n in the ordinary sample variance s^2 (in Chapter 3), we divide by $n - 2$ in s_ε^2, the estimated variance around the line. The reduction from n to $n - 2$ occurs because in order to estimate the variability around the regression line, we must first estimate the two parameters β_0 and β_1 to obtain the estimated line. The effective sample size for estimating σ_ε^2 is thus $n - 2$. In our definition, s_ε^2 is undefined for $n = 2$, as it should be. Another argument is that dividing by $n - 2$ makes s_ε^2 an unbiased estimator of σ_ε^2. In the computer output of Example 11.3, $n - 2 = 10 - 2 = 8$ is shown as DF (degrees of freedom) for RESIDUAL and $s_\varepsilon^2 = 0.1187$ is shown as MS for RESIDUAL.

residual standard deviation

The square root s_ε of the sample variance is called the **sample standard deviation around the regression line,** the **standard error of estimate,** or the **residual standard deviation.** Because s_ε estimates σ_ε, the standard deviation of y_i, σ_ε estimates the standard deviation of the population of y values associated with a given value of the independent variable x. The output in Example 11.3 labels s_ε as S with $S = 0.344566$.

Like any other standard deviation, the residual standard deviation may be interpreted by the Empirical Rule. About 95% of the prediction errors will fall within ± 2 standard deviations of the mean error; the mean error is always 0 in the least-squares regression model. Therefore, a residual standard deviation of 0.345 means that about 95% of prediction errors will be less than $\pm 2(0.345) = \pm 0.690$.

The estimates $\hat{\beta}_0, \hat{\beta}_1$, and s_ε are basic in regression analysis. They specify the regression line and the probable degree of error associated with y values for a given value of x. The next step is to use these sample estimates to make inferences about the true parameters.

EXAMPLE 11.4

Forest scientists are concerned with the decline in forest growth throughout the world. One aspect of this decline is the possible effect of emissions from coal-fired power plants. The scientists in particular are interested in the pH level of the soil and the resulting impact on tree growth retardation. The scientists study various forests which are likely to be exposed to these emissions. They measure various aspects of growth associated with trees in a specified region and the soil pH in the same region. The forest scientists then want to determine impact on tree growth as the soil becomes more acidic. An index of growth retardation is constructed from the various measurements taken on the trees with a high value indicating greater retardation in tree growth. A higher value of soil pH indicates a more acidic soil. Twenty tree stands which are exposed to the power plant emissions are selected for study. The values of the growth retardation index and average soil pH are recorded in Table 11.3.

EXAMPLE 11.7

Compute a 95% confidence interval for the slope β_1 using the output from Example 11.4.

Solution In the output, $\hat{\beta}_1 = -7.859$ and the estimated standard error of $\hat{\beta}_1$ is shown in the column labelled **SE Coef** as 1.090. Because n is 20, there are $20 - 2 = 18$ df for error. The required table value for $\alpha/2 = .05/2 = .025$ is 2.101. The corresponding confidence interval for the true value of β_1 is then

$$-7.859 \pm 2.101(1.090) \quad \text{or} \quad -10.149 \text{ to } -5.569$$

The predicted decrease in growth retardation for a unit increase in soil pH ranges from -10.149 to -5.569. The large width of this interval is mainly due to the small sample size.

There is an alternative test, an F test, for the null hypothesis of no predictive value. It was designed to test the null hypothesis that *all* predictors have no value in predicting y. This test gives the same result as a two-sided t test of H_0: $\beta_1 = 0$ in simple linear regression; to say that all predictors have no value is to say that the (only) slope is 0. The F test is summarized next.

F Test for H_0: $\beta_1 = 0$

H_0: $\beta_1 = 0$

H_a: $\beta_1 \neq 0$

T.S.: $F = \dfrac{\text{SS(Regression)}/1}{\text{SS(Residual)}/(n-2)} = \dfrac{\text{MS(Regression)}}{\text{MS(Residual)}}$

R.R.: With $df_1 = 1$ and $df_2 = n - 2$, reject H_0 if $F > F_\alpha$.

Check assumptions and draw conclusions.

SS(Regression) is the sum of squared deviations of predicted y values from the y mean. SS(Regression) $= \Sigma(\hat{y}_i - \bar{y})^2$. SS(Residual) is the sum of squared deviations of actual y values from predicted y values. SS(Residual) $= \Sigma(\hat{y}_i - \bar{y}_i)^2$.

Virtually all computer packages calculate this F statistic. In Example 11.3, the output shows $F = 54.03$ with a p-value given by 0.000 (in fact, p-value $= .00008$). Again, the hypothesis of no predictive value can be rejected. It is always true for simple linear regression problems that $F = t^2$; in the example, $54.03 = (7.35)^2$, to within round-off error. The F and two-sided t tests are equivalent in simple linear regression; they serve different purposes in multiple regression.

EXAMPLE 11.8

For the output of Example 11.4, reproduced here, use the F test for testing H_0: $\beta_1 = 0$. Show that $t^2 = F$ for this data set.

independent variable. Then, if there are n_i observations at the ith level of the independent variable, the quantity

$$\sum_j (y_{ij} - \bar{y}_i)^2$$

provides a measure of what we will call pure experimental error. This sum of squares has $n_i - 1$ degrees of freedom.

Similarly, for each of the other levels of x, we can compute a sum of squares due to pure experimental error. The pooled sum of squares

$$\text{SSP}_{\text{exp}} = \sum_{ij} (y_{ij} - \bar{y}_i)^2$$

called the sum of squares for pure experimental error, has $\Sigma_i(n_i - 1)$ degrees of freedom. With SS_{Lack} representing the remaining portion of SSE, we have

$$\text{SSE} \quad \text{SS(Residuals)} = \begin{matrix}\text{SSP}_{\text{exp}} \\ \text{due to pure} \\ \text{experimental} \\ \text{error}\end{matrix} + \begin{matrix}\text{SS}_{\text{Lack}} \\ \text{due to lack} \\ \text{to fit}\end{matrix}$$

SSE

If SS(Residuals) is based on $n - 2$ degrees of freedom in the linear regression model, then SS_{Lack} will have $\text{df} = n - 2 - \Sigma_i(n_i - 1)$.

Under the null hypothesis that our model is correct, we can form independent estimates of σ_ε^2, the model error variance, by dividing SSP_{exp} and SS_{Lack} by their respective degrees of freedom; these estimates are called **mean squares** and are denoted by MSP_{exp} and MS_{Lack}, respectively.

mean squares

The test for lack of fit is summarized here.

A Test for Lack of Fit in Linear Regression

H_0: A linear regression model is appropriate.
H_a: A linear regression model is not appropriate.
T.S.: $F = \dfrac{\text{MS}_{\text{Lack}}}{\text{MSP}_{\text{exp}}}$,

where

$$\text{MS}_{\text{exp}} = \frac{\text{SSP}_{\text{exp}}}{\Sigma_i(n_i - 1)} = \frac{\Sigma_{ij}(y_{ij} - \bar{y}_i)^2}{\Sigma_i(n_i - 1)}$$

and

SSE

$$\text{MS}_{\text{Lack}} = \frac{\text{SS(Residual)} - \text{SSP}_{\text{exp}}}{n - 2 - \Sigma(n_1 - 1)}$$

R.R.: For specified value of α, reject H_0 (the adequacy of the model) if the computed value of F exceeds the table value for $\text{df}_1 = n - 2 - \Sigma_i(n_i - 1)$ and $\text{df}_2 = \Sigma_i(n_1 - 1)$.

Conclusion: If the F test is significant, this indicates that the linear regression model is inadequate. A nonsignificant result indicates that there is insufficient evidence to suggest that the linear regression model is inappropriate.

EXAMPLE 11.11

Refer to the data of Example 11.10. Conduct a test for lack of fit of the linear regression model.

Solution It is easy to show that the contributions to experimental error for the differential levels of x are as given in Table 11.5.

TABLE 11.5
Pure experimental error calculation

Level of x	$\bar{y}_i$	Contribution to Pure Experimental Error $\Sigma_i(y_{ij} - \bar{y}_i)^2$	$n_i - 1$
20	81	42	2
40	79	42	2
60	38	50	2
Total		134	6

Summarizing these results, we have

$$\text{SSP}_{\text{exp}} = \sum_{ij} (y_{ij} - \bar{y}_i)^2 = 134$$

The calculation of SSP_{exp} can be obtained by using the One-Way ANOVA command in a software package. Using the theory from Chapter 8, designate the levels of the independent variable x as the levels of a treatment. The sum of squares error from this output is the value of SSP_{exp}. This concept is illustrated using the output from Minitab given here.

```
One-way ANOVA: HeatLoss versus OutTemp

Source      DF      SS      MS      F      P
OutTemp      2  3534.0  1767.0  79.12  0.000
Error        6   134.0    22.3
Total        8  3668.0

S = 4.726  R-Sq = 96.35%  R-Sq(adj) = 95.13%
```

Note that the value of sum of square error from the ANOVA is exactly the value that was computed above. Also, the degrees of freedom are given as 6, the same as in our calculations.

The output shown for Example 11.10 gives ~~SS(Residual)~~ SSE = 894.5; hence, by subtraction,

$$\text{SS}_{\text{Lack}} = \text{~~SS(Residual)~~ SSE} - \text{SSP}_{\text{exp}} = 894.5 - 134 = 760.5$$

The sum of squares due to pure experimental error has $\Sigma_i(n_i - 1) = 6$ degrees of freedom; it therefore follows that with $n = 9$, SS_{Lack} has $n - 2 - \Sigma_i(n_i - 1) = 1$ degree of freedom. We find that

$$\text{MSP}_{\text{exp}} = \frac{\text{SSP}_{\text{exp}}}{6} = \frac{134}{6} = 22.33$$

and

$$\text{MS}_{\text{Lack}} = \frac{\text{SS}_{\text{Lack}}}{1} = 760.5$$

The solution to the second inverse prediction problem is summarized next.

Case 2: Predicting x Based on m y-Values

Predicting the value of x corresponding to $100P\%$ of the mean of m independent y values. For $0 \le P \le 1$,

$$\text{Predictor of } x\text{: } \hat{x} = \frac{P\bar{y}_m - \hat{\beta}_0}{\hat{\beta}_1}$$

$$\hat{x}_U = \bar{x} + \frac{1}{1 - c^2}[(\hat{x} - \bar{x}) + g]$$

$$\hat{x}_L = \bar{x} + \frac{1}{1 - c^2}[(\hat{x} - \bar{x}) - g]$$

where

$$g = \frac{t_{\alpha/2}}{\hat{\beta}_1}\sqrt{\left(s_{\bar{y}}^2 P^2 + \frac{s_\varepsilon^2}{n}\right)(1 - c^2) + \frac{(\hat{x} - \bar{x})^2 s_\varepsilon^2}{S_{xx}}}$$

and $\bar{y}_m$ and $s_{\bar{y}}$ are the mean and standard error, respectively, of m independent y-values.

11.7 Correlation

Once we have found the prediction line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, we need to measure how well it predicts actual values. One way to do so is to look at the size of the residual standard deviation in the context of the problem. About 95% of the prediction errors will be within $\pm 2s_\varepsilon$. For example, suppose we are trying to predict the yield of a chemical process, where yields range from .50 to .94. If a regression model had a residual standard deviation of .01, we could predict most yields within $\pm .02$—fairly accurate in context. However, if the residual standard deviation were .08, we could predict most yields within $\pm .16$, which is not very impressive given that the yield range is only $.94 - .50 = .44$. This approach, though, requires that we know the context of the study well; an alternative, more general approach is based on the idea of correlation.

Suppose that we compare the squared prediction error for two prediction methods: one using the regression model, the other ignoring the model and always predicting the mean y value. In the road resurfacing example of Section 11.2, if we are given the mileage values x_i, we could use the prediction equation $\hat{y}_i = 2.0 + 3.0x_i$ to predict costs. The deviations of actual values from predicted values, the residuals, measure prediction errors. These errors are summarized by the sum of squared residuals, ~~SS(Residual)~~ $= \Sigma(y_i - \hat{y}_i)^2$, which is 44 for these data. For comparison, if we were not given the x_i values, the best squared error predictor of y would be the mean value $\bar{y} = 14$, and the sum of squared prediction errors would, in this case, be $\Sigma_i(y_i - \bar{y})^2 = \text{SS(Total)} = 224$. The proportionate reduction in error would be

$$\frac{\text{SS(Total)} - \text{SS(Residual)}}{\text{SS(Total)}} = \frac{224 - 44}{224} = .804$$

$$\sum_{i=1}^{30} x_i = 109.5 \Rightarrow \bar{x} = 3.65, \qquad \sum_{i=1}^{30} y_i = 2065 \Rightarrow \bar{y} = 68.8333$$

$$S_{xx} = \sum_{i=1}^{30} (x_i - \bar{x})^2$$

$$= (2.1 - 3.65)^2 + (2.3 - 3.65)^2 + \cdots + (5.1 - 3.65)^2 = 17.615$$

$$S_{yy} = \sum_{i=1}^{30} (y_i - \bar{y})^2$$

$$= (27 - 68.8333)^2 + (32 - 68.8333)^2 + \cdots + (65 - 68.8333)^2 = 6{,}066.1667$$

$$S_{xy} = \sum_{i=1}^{30} (x_i - \bar{x})(y_i - \bar{y})$$

$$= (2.1 - 3.65)(27 - 68.8333) + (2.3 - 3.65)(32 - 68.8333) + \cdots + (5.1 - 3.65)(65 - 68.8333) = 198.05$$

$$r_{xy} = \frac{198.05}{\sqrt{(17.615)(6066.1667)}} = 0.606$$

The correlation is indeed a positive number.

coefficient of determination

Correlation and regression predictability are closely related. The proportionate reduction in error for regression we defined earlier is called the **coefficient of determination**. The coefficient of determination is simply the square of the correlation coefficient,

$$r_{yx}^2 = \frac{\text{SS(Total)} - \text{SS(Residual)}}{\text{SS(Total)}} = \frac{SST - SSE}{SST}$$

which is the proportionate reduction in error. In the resurfacing example, $r_{yx} = .896$ and $r_{yx}^2 = .803$.

A correlation of zero indicates no predictive value in using the equation $y = \hat{\beta}_0 + \hat{\beta}_1 x$; that is, one can predict y as well without knowing x as one can knowing x. A correlation of 1 or -1 indicates perfect predictability—a 100% reduction in error attributable to knowledge of x. A correlation coefficient should routinely be interpreted in terms of its squared value, the coefficient of determination. Thus, a correlation of $-.3$, say, indicates only a 9% reduction in squared prediction error. Many books and most computer programs use the equation

$$\text{SS(Total)} = \text{SS(Residual)} + \text{SS(Regression)} \qquad SST = SS_{Reg} + SSE$$

where

$$\text{SS(Regression)} = \sum_i (\hat{y}_i - \bar{y})^2$$

Because the equation can be expressed as $\text{SS(Residual)} = (1 - r_{yx}^2)\text{SS(Total)}$, it follows that $\text{SS(Regression)} = r_{yx}^2 \text{SS(Total)}$, which again says that regression on x explains a proportion r_{yx}^2 of the total squared error of y.

EXAMPLE 11.14

For the grasshopper data in Example 11.13, compute SS(Total), SS(Regression), and SS(Residual).

Solution $\text{SS(Total)} = S_{yy}$, which we computed to be 6,066.1667 in Example 11.13. We also found that $r_{yx} = 0.606$, so $r_{xy}^2 = (0.606)^2 = 0.367236$. Using the fact that $\text{SS(Regression)} = r_{yx}^2 \text{SS(Total)}$, we have

$$\text{SS(Regression)} = (0.367236)\,(6{,}066.1667) = 2{,}227.7148.$$

From the equation ~~SS(Residual) = SS(Total) − SS(Regression)~~, we obtain

SSE ~~SS(Residual)~~ = 6,066.1667 − 2,227.7148 = 3,838.45

Note that $r_{yx}^2 = (.606)^2 = 0.37$ indicates that a regression line predicting the number of eggs as a linear function of the weight of the female grasshopper would only explain about 37% of the variation in the number of eggs laid. This suggests that weight of the female is not a good predictor of the number of eggs. An examination of the scatterplot in Figure 11.21 shows a strong relationship between x and y but the relation is extremely nonlinear. A *linear* equation in x does not predict y very well, but a nonlinear equation would provide an excellent fit.

What values of r_{yx} indicate a "strong" relationship between y and x? Figure 11.22 displays 15 scatterplots obtained by randomly selecting 1,000 pairs (x_i, y_i)

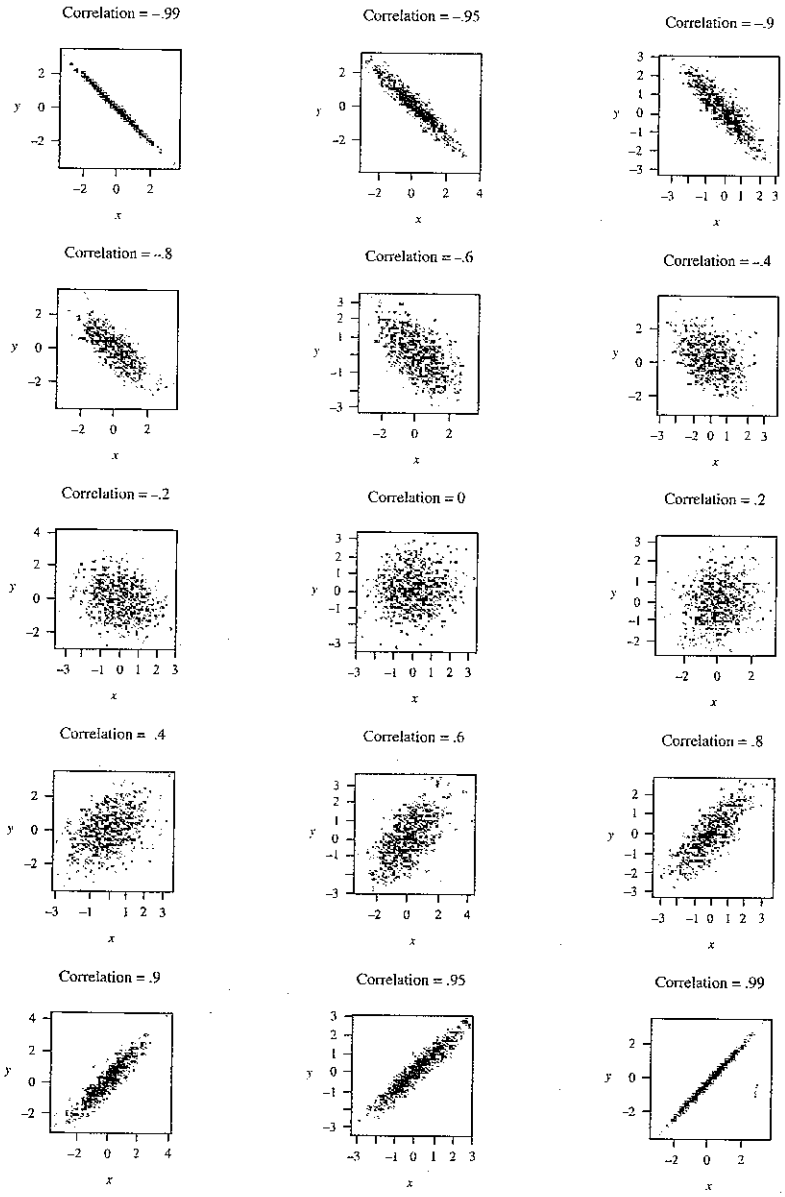


FIGURE 11.22
Samples of size 1,000 from the bivariate normal distribution

In the regression approach, past data on the relevant variables are used to develop and evaluate a prediction equation. The variable that is being predicted by this equation is the dependent variable. A variable that is being used to make the prediction is an independent variable. In this chapter, we discuss regression methods involving a single independent variable. In Chapter 12, we extend these methods to multiple regression, the case of several independent variables.

A number of tasks can be accomplished in a regression study:

1. The data can be used to obtain a prediction equation.
2. The data can be used to estimate the amount of variability or uncertainty around the equation.
3. The data can be used to identify unusual points far from the predicted value, which may represent unusual problems or opportunities.
4. Because the data are only a sample, inferences can be made about the true (population) values for the regression quantities.
5. The prediction equation can be used to predict a reasonable range of values for future values of the dependent variable.
6. The data can be used to estimate the degree of correlation between dependent and independent variables, a measure that indicates how strong the relation is.

Key Formulas

1. Least-squares estimates of slope and intercept

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where

$$S_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

and

$$S_{xx} = \sum_i (x_i - \bar{x})^2$$

2. Estimate of σ_ε^2

$$s_\varepsilon^2 = \frac{\Sigma_i (y_i - \hat{y}_i)^2}{n - 2} = \frac{\text{SS(Residual)}}{n - 2} \quad \text{SSE}$$

3. Statistical test for β_1

H_0: $\beta_1 = 0$ (two-tailed)

$$\text{T.S.:} \quad t = \frac{\hat{\beta}_1}{\frac{s_\varepsilon}{\sqrt{S_{xx}}}}$$

4. Confidence interval for β_1

$$\hat{\beta}_1 \pm t_{\alpha/2} s_\varepsilon \sqrt{\frac{1}{S_{xx}}}$$

5. F test for H_0: $\beta_1 = 0$ (two-tailed)

$$\text{T.S.:} \quad F = \frac{\text{MS(Regression)}}{\text{MS(Residual)}} = \frac{MS_{Reg}}{MSE}$$

6. Confidence interval for $E(y_{n+1})$

$$\hat{y}_{n+1} \pm t_{\alpha/2} s_\varepsilon \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}}}$$

7. Prediction interval for y_{n+1}

$$\hat{y}_{n+1} \pm t_{\alpha/2} s_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}}}$$

8. Test for lack of fit in linear regression

$$\text{T.S.:} \quad F = \frac{\text{MS}_{\text{Lack}}}{\text{MSP}_{\text{exp}}}$$

where

$$\text{MSP}_{\text{exp}} = \frac{\text{SSP}_{\text{exp}}}{\Sigma_i (n_i - 1)} = \frac{\Sigma_{ij} (y_{ij} - \bar{y}_i)^2}{\Sigma_i (n_i - 1)}$$

and

$$MS_{Lack} = \frac{\text{SS(Residual)} - SSP_{exp}}{(n-2) - \Sigma_i (n_i - 1)}$$

9. Prediction limits for x based on a single y value

$$\hat{x} = \frac{y - \hat{\beta}_0}{\hat{\beta}_1}$$

$$\hat{x}_U = \bar{x} + \frac{1}{1-c^2}[(\hat{x} - \bar{x}) + d]$$

$$\hat{x}_L = \bar{x} + \frac{1}{1-c^2}[(\hat{x} - \bar{x}) - d]$$

where

$$c^2 = \frac{t^2_{\alpha/2} s^2_\varepsilon}{\hat{\beta}_1^2 S_{xx}}$$

and

$$d = \frac{t_{\alpha/2} s_\varepsilon}{\hat{\beta}_1} \sqrt{\frac{n+1}{n}(1-c^2) + \frac{(\hat{x} - \bar{x})^2}{S_{xx}}}$$

10. Prediction interval for x based on m y-values

$$\hat{x}_U = \bar{x} + \frac{1}{1-c^2}[(\hat{x} - \bar{x}) + g]$$

$$\hat{x}_L = \bar{x} + \frac{1}{1-c^2}[(\hat{x} - \bar{x}) - g]$$

where

$$\hat{x} = \frac{P\bar{y}_m - \hat{\beta}_0}{\hat{\beta}_1}$$

and

$$g = \frac{t_{\alpha/2}}{\hat{\beta}_1} \sqrt{\left(s^2_{\hat{y}} P^2 + \frac{s^2_\varepsilon}{n}\right)(1-c^2) + \frac{(\hat{x} - \bar{x})^2 s^2_\varepsilon}{S_{xx}}}$$

11. Correlation coefficient

$$r_{yx} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \hat{\beta}_1 \sqrt{\frac{S_{xx}}{S_{yy}}}$$

12. Coefficient of determination

$$r^2_{yx} = \frac{\text{SS(Total)} - \text{SS(Residual)}}{\text{SS(Total)}}$$

13. Confidence interval for ρ_{yx}

$$\left(\frac{e^{2z_1} - 1}{e^{2z_1} + 1}, \frac{e^{2z_2} - 1}{e^{2z_2} + 1}\right)$$

14. Statistical test for ρ_{yx}

H_0: $\rho_{yx} = 0$ (two-tailed)

T.S.: $t = r_{yx} \dfrac{\sqrt{n-2}}{\sqrt{1 - r^2_{yx}}}$

11.10 Exercises

11.2 Estimating Model Parameters

Basic **11.1** Plot the data shown here in a scatter diagram and sketch a line through the points.

x	5	10	15	20	25	30
y	14	28	43	62	79	87

Basic **11.2** Use the equation $\hat{y} = 2.3 + 1.8x$ to answer the following questions.

a. Predict y for $x = 6$.

b. Plot the equation on a graph with the horizontal axis scaled from 0 to 8 and the vertical axis scaled from 0 to 18.

Basic **11.3** Use the data given here to answer the following questions.

x	7	12	14	22	27	33
y	14	28	43	62	79	87

a. Plot the data values in a scatter diagram.

b. Determine the least-squares prediction equation.

Bus. **11.10** In the JMP output of Exercise 11.9, the residual standard deviation is called "Root Mean Square Error." Locate and interpret this number.

Bus. **11.11** In the preceding exercises, why can the residual standard deviation for the transformed data be compared to the residual standard deviation for the original data?

Engin. **11.12** A manufacturer of cases for sound equipment requires drilling holes for metal screws. The drill bits wear out and must be replaced; there is expense not only in the cost of the bits but also for lost production. Engineers varied the rotation speed of the drill and measured the lifetime y (thousands of holes drilled) of four bits at each of five speeds x. The data were:

x:	60	60	60	60	80	80	80	80	100	100
y:	4.6	3.8	4.9	4.5	4.7	5.8	5.5	5.4	5.0	4.5
x:	100	100	120	120	120	120	140	140	140	140
y:	3.2	4.8	4.1	4.5	4.0	3.8	3.6	3.0	3.5	3.4

a. Create a scatterplot of the data. Does there appear to be a relation? Does it appear to be linear?

b. Is there any evident outlier? If so, does it have high influence?

Engin. **11.13** The data of Exercise 11.12 were analyzed yielding the following output.

```
Regression Analysis: Lifetime versus DrillSpeed

The regression equation is
Lifetime = 6.03 - 0.0170 DrillSpeed


Predictor        Coef    SE Coef       T      P
Constant       6.0300     0.5195   11.61  0.000
DrillSpeed  -0.017000   0.004999   -3.40  0.003


S = 0.632368   R-Sq = 39.1%   R-Sq(adj) = 35.7%


Analysis of Variance

Source           DF      SS      MS     F      P
Regression        1  4.6240  4.6240  11.56  0.003
Residual Error   18  7.1980  0.3999
Total            19 11.8220

Unusual Observations

Obs  DrillSpeed  LifeTime    Fit  SE Fit  Residual  St Resid
  2         60     3.800  5.010   0.245    -1.210    -2.08R

R denotes an observation with a large standardized residual.
```

a. Find the least-squares estimates of the slope and intercept in the output.

b. What does the sign of the slope indicate about the relation between the speed of the drill and bit lifetime?

c. Locate the residual standard deviation. What does this value indicate about the fitted regression line?

Engin. **11.14** Refer to the data of Exercise 11.12.

a. Use the regression line of Exercise 11.13 to calculate predicted values for x = 60, 80, 100, 120, 140.

b. For which x values are most of the actual y values larger than the predicted values? For which x values are most of the actual y values smaller than the predicted values? What does this pattern indicate about whether there is a linear relation between drill speed and the lifetime of the drill?

c. Suggest a transformation of the data to obtain a linear relation between lifetime of the drill and the transformed values of the drill speed.

Cloud	Time, x (in minutes)	Biological Recovery (%)
11	50	7.9
12	55	7.7
13	60	7.7

Bio. **11.22** Refer to Exercise 11.21.
- **a.** Fit the linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$, where y is the log biological recovery.
- **b.** Compute an estimate of σ_ε.
- **c.** Identify the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$.

Bio. **11.23** Refer to Exercise 11.21. Conduct a test of the null hypothesis that $\beta_1 = 0$. Use $\alpha = .05$.

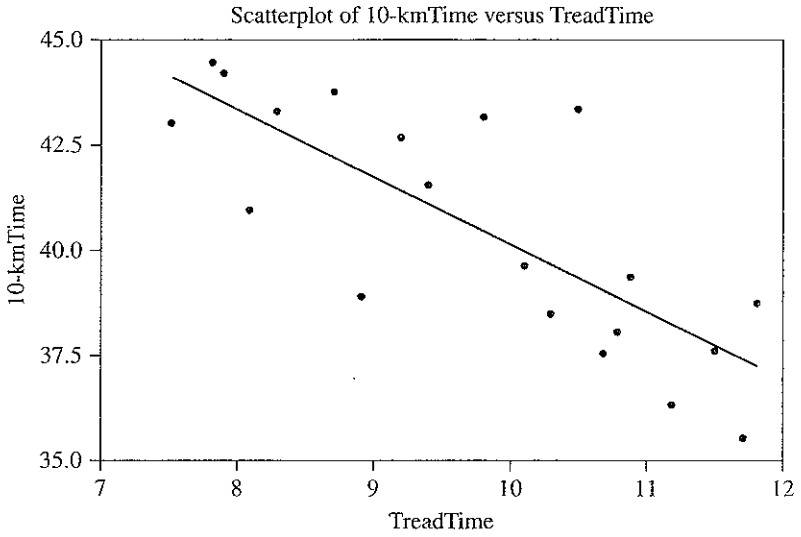
Bio. **11.24** Refer to Exercise 11.21. Place a 95% confidence interval on β_0, the mean log biological recovery percentage at time zero. Interpret your findings. (*Note:* $E(y) = \beta_0$ when $x = 0$.)

Med. **11.25** Athletes are constantly seeking measures of the degree of their cardiovascular fitness prior to a major race. Athletes want to know when their training is at a level which will produce a peak performance. One such measure of fitness is the time to exhaustion from running on a treadmill at a specified angle and speed. The important question is then "Does this measure of cardiovascular fitness translate into performance in a 10-km running race?" Twenty experienced distance runners who professed to be at top condition were evaluated on the treadmill and then had their times recorded in a 10-km race. The data are given here.

38.7

Treadmill Time (minutes)	7.5	7.8	7.9	8.1	8.3	8.7	8.9	9.2	9.4	9.8
10-km Time (minutes)	43.5	45.2	44.9	41.1	43.8	44.4	42.7	43.1	41.8	43.7
Treadmill Time (minutes)	10.1	10.3	10.5	10.7	10.8	10.9	11.2	11.5	11.7	11.8
10-km Time (minutes)	39.5	38.2	43.9	37.1	37.7	39.2	35.7	37.2	34.8	38.5

```
Minitab Output

Regression Analysis: 10-kmTime versus TreadTime

The regression equation is
10-kmTime = 58.8 - 1.87 TreadTime

Predictor     Coef   SE Coef      T      P
Constant    58.816     3.410  17.25  0.000
TreadTime   -1.8673    0.3462  -5.39  0.000

S = 2.10171  R-Sq = 61.8%  R-Sq(adj) = 59.7%

Analysis of Variance

Source          DF      SS      MS      F      P
Regression       1  128.49  128.49  29.09  0.000
Residual Error  18   79.51    4.42
Total           19  208.00

Unusual Observations

Obs  TreadTime  10-kmTime     Fit  SE Fit  Residual  St Resid
 13       10.5     43.900  39.209   0.536     4.691     2.31R

R denotes an observation with a large standardized residual.
```



a. Refer to the output. Does a linear model seem appropriate?
b. From the output, obtain the estimated linear regression model $\hat{y} = \hat{\beta}_0 + \beta_1 x$.

11.26 Refer to the output of Exercise 11.25.
a. Estimate σ_ε^2.
b. Identify the standard error of $\hat{\beta}_1$.
c. Place a 95% confidence interval on β_1.
d. Test the hypothesis that there is a linear relationship between the amount of time needed to run a 10 km race and the time to exhaustion on a treadmill. Use $\alpha = .05$.

11.27 The focal point of an agricultural research study was the relationship between when a crop is planted and the amount of crop harvested. If a crop is planted too early or too late farmers may fail to obtain optimal yield and hence not make a profit. An ideal date for planting is set by the researchers, and the farmers then record the number of days either before or after the designated date. In the following data set, D is the deviation (in days) from the ideal planting date and Y is the yield (in bushels per acre) of a wheat crop:

D	−11	−10	−9	−8	−7	−6	−4	−3	−1	0
Y	43.8	44.0	44.8	47.4	48.1	46.8	49.9	46.9	46.4	53.5
D	1	3	6	8	12	13	15	16	18	19
Y	55.0	46.9	44.1	50.2	41.0	42.8	36.5	35.8	32.2	33.3

a. Plot the above data. Does a linear relation appear to exist between yield and deviation from ideal planting date?
b. Plot yield versus absolute deviation from ideal planting date. Does a linear relation seem more appropriate in this plot than the plot in (a)?

11.28 Refer to Exercise 11.27. The following computer output compares yield to the absolute deviation from the ideal planting date.

```
Regression Analysis: Yield versus AbsDevIdeal

The regression equation is
Yield = 52.8 - 0.983 AbsDevIdeal


Predictor        Coef   SE Coef       T      P
Constant       52.819     1.101   47.98  0.000
AbsDevIdeal   -0.9834    0.1083   -9.08  0.000


S = 2.69935   R-Sq = 82.1%   R-Sq(adj) = 81.1%


Analysis of Variance

Source           DF       SS       MS      F      P
Regression        1   600.57   600.57  82.42  0.000
Residual Error   18   131.16     7.29
Total            19   731.73

Unusual Observations

Obs  AbsDevIdeal   Yield      Fit  SE Fit  Residual  St Resid
  9          1.0  46.400   51.836   1.012    -5.436    -2.17R

R denotes an observation with a large standardized residual.
```

$\hat{\beta}_1 x.$

a. From the output, obtain the estimated linear regression model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

b. Estimate σ_ε^2.

c. Identify the standard error of $\hat{\beta}_1$.

d. Place a 95% confidence interval on β_1.

e. Test the hypothesis that there is a linear relationship between yield per acre and the absolute deviation from the ideal planting date. Use $\alpha = .05$.

11.29 Refer to Exercise 11.27.

a. For this study, would it make sense to give any physical interpretation to β_0?

b. Place a 95% confidence interval on β_0 and give an interpretation to the interval relative to this particular study.

c. The output in Exercise 11.28 provides a test of the hypotheses $H_0: \beta_0 = 0$ versus $H_a: \beta_0 \neq 0$. Does this test have any practical importance in this particular study?

Bus. **11.30** A firm that prints automobile bumper stickers conducts a study to investigate the relation between the direct cost of producing an order of bumper stickers and the number of stickers (thousands of stickers) in a particular order. The data are given here along with the relevant output from Minitab.

RunSize	2.6	5.0	10.0	2.0	.8	4.0	2.5	.6	0.8	1.0
TOTCOST	230	341	629	187	159	327	206	124	155	147
RunSize	2.0	3.0	.4	.5	5.0	20.0	5.0	2.0	1.0	1.5
TOTCOST	209	247	135	125	366	1146	339	208	150	179
RunSize	.5	1.0	1.0	.6	2.0	1.5	3.0	6.5	2.2	1.0
TOTCOST	128	155	143	131	219	171	258	415	226	159

Obs	RunSize	TotalCost	Fit	SE Fit	Residual	St Resid	
1	2.6	230.00	234.76	2.24	-4.76	-0.40	
2	5.0	341.00	359.37	2.53	-18.37	-1.54	
3	10.0	629.00	618.96	4.69	10.04	0.89	
4	2.0	187.00	203.61	2.30	-16.61	-1.39	
5	0.8	159.00	141.31	2.57	17.69	1.48	
6	4.0	327.00	307.45	2.31	19.55	1.63	
7	2.5	206.00	229.57	2.25	-23.57	-1.96	
8	0.6	124.00	130.93	2.63	-6.93	-0.58	
9	0.8	155.00	141.31	2.57	13.69	1.15	
10	1.0	147.00	151.69	2.51	-4.69	-0.39	
11	2.0	209.00	203.61	2.30	5.39	0.45	
12	3.0	247.00	255.53	2.23	-8.53	-0.71	
13	0.4	135.00	120.54	2.69	14.46	1.21	
14	0.5	125.00	125.74	2.66	-0.74	-0.06	
15	5.0	366.00	359.37	2.53	6.63	0.56	
16	20.0	1146.00	1138.13	10.23	7.87	1.18	X
17	5.0	339.00	359.37	2.53	-20.37	-1.71	
18	2.0	208.00	203.61	2.30	4.39	0.37	
19	1.0	150.00	151.69	2.51	-1.69	-0.14	
20	1.5	179.00	177.65	2.39	1.35	0.11	
21	0.5	128.00	125.74	2.66	2.26	0.19	
22	1.0	155.00	151.69	2.51	3.31	0.28	
23	1.0	143.00	151.69	2.51	-8.69	-0.73	
24	0.6	131.00	130.93	2.63	0.07	0.01	
25	2.0	219.00	203.61	2.30	15.39	1.28	
26	1.5	171.00	177.65	2.39	-6.65	-0.56	
27	3.0	258.00	255.53	2.23	2.47	0.21	
28	6.5	415.00	437.24	3.04	-22.24	-1.88	
29	2.2	226.00	214.00	2.27	12.00	1.00	
30	1.0	159.00	151.69	2.51	7.31	0.61	

X denotes an observation whose X value gives it large influence.

a. Examine the plot of the data. Do you detect any difficulties with using a linear regression model? Can you find any blatant violations of the regression assumptions?
b. Write the estimated regression line as given in the output.
c. Locate the residual standard deviation in the output.
d. Construct a 95% confidence interval for the true slope.
e. What are the interpretations of the intercept and slope in this study?

11.31 Refer to the output in Exercise 11.30.
a. Test the hypothesis $H_0: \beta_0 = 0$ using a t-test with $\alpha = .05$.
b. Locate the p-value for this test. Is the p-value one-tailed or two-tailed? If necessary, calculate the p-value for the appropriate number of tails.

11.32 Refer to the output in Exercise 11.30.
a. Locate the value of the F statistic and the associated p-value.
b. How do the p-values for this F statistic and the t test of Exercise 11.31 compare? Why should this relation hold?

11.4 Predicting New *y* Values Using Regression

Bio. **11.33** Refer to Exercise 11.21. Using the least-squares line obtained in Exercise 11.21

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

estimate the mean log biological recovery percentage at 30 minutes using a 95% confidence interval.

Bio. **11.34** Use the data from Exercise 11.21 to answer the following questions.
a. Construct a 95% prediction interval for the log biological recovery percentage at 30 minutes.
b. Compare your results to the confidence interval on $E(y)$ from Exercise 11.33.
c. Explain the different interpretation for the two intervals.

```
Correlations: Male/Verbal, Female/Verbal, Male/Math, Female/Math

                Male/Verbal  Female/Verbal   Male/Math
Female/Verbal         0.981
                      0.000

Male/Math             0.417          0.496
                      0.157          0.085

Female/Math           0.218          0.322       0.960
                      0.474          0.284       0.000

Cell Contents: Pearson correlation
               P-Value
```

b. Are the correlations between Verbal and Math scores higher for Males than for Females? Use the confidence intervals from part a. in answer this question.

a. Which, if any, of the six correlations are significantly different from 0 at the 5% level?

b. Do the plots reflect the size of the correlations between the four variables?

c. Are male verbal scores more correlated with male or female math scores?

11.59 Refer to Exercise 11.58.

a. Place a 95% confidence interval on the six correlations.

b. ~~Using the confidence intervals from (b) are there any differences in the degree~~ of correlation between male and female math scores?

c. ~~Using the confidence intervals from (b) are there any differences in the degree of correlation between male and female verbal scores?~~

c. **d.** Are your answers to part (b) ~~and (c)~~ different from your answer to part (c) in Exercise 11.58?

Supplementary Exercises

11.60 A construction science class project was to compare the daily gas consumption of 20 homes with a new form of insulation to 20 similar homes with standard insulation. They set up instruments to record the temperature both inside and outside of the homes over a six-month period of time (October–March). The average differences in these values are given below. They also obtained the average daily gas consumption (in kilowatt hours). All the homes were heated with gas. The data are given here:

Data for Homes with Standard Form of Insulation:

TempDiff(°F)	20.3	20.7	20.9	22.8	23.1	24.8	25.9	26.1	27.0	27.2
GasConsumption(kWh)	70.3	70.7	72.9	77.6	79.3	86.5	90.6	91.9	94.5	92.7
TempDiff(°F)	29.8	30.2	30.6	31.8	33.2	33.4	34.2	35.1	36.2	36.5
GasConsumption(kWh)	104.8	103.2	91.2	89.6	116.2	116.9	105.1	106.1	117.8	120.3

Data for Homes with New Form of Insulation:

TempDiff(°F)	20.1	21.1	21.9	22.6	23.4	24.2	24.9	25.1	26.0	27.2
GasConsumption(kWh)	65.3	66.5	67.8	73.2	75.3	81.1	82.2	85.7	90.9	87.4
TempDiff(°F)	28.8	29.2	30.6	30.8	32.6	32.4	34.8	35.9	36.0	36.5
GasConsumption(kWh)	94.9	93.9	87.1	84.2	106.6	111.3	100.9	101.9	110.1	119.1

```
Unusual Observations

Obs  Age      BP      Fit   SE Fit  Residual  St Resid
  8  6.00  110.000  104.820   0.837     5.180      2.46R
 16  6.00  103.000  104.452   1.982    -1.452     -1.33 X

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large influence.

New
Obs  Age  Weight    Fit     SE Fit        95% CI              95% PI
  1    4       3   90.854   0.737  (89.316,  92.393)  (85.891,  95.817)
  2    8       5  107.870   5.661  (96.062, 119.679)  (95.154, 120.587)XX

XX denotes a point that is an extreme outlier in the predictors.
```

a. Provide an estimate for the mean systolic blood pressure for an infant of age 4 days weighing 3 kg.

b. Provide a 95% confidence interval for the mean systolic blood pressure for an infant of age 4 days weighing 3 kg.

12.34 Refer to Exercise 12.33 and the accompanying Minitab output.

a. Provide an estimate for the mean systolic blood pressure for an infant of age 8 days weighing 5 kg.

b. Provide a 95% confidence interval for the mean systolic blood pressure for an infant of age 8 days weighing 5 kg.

12.35 The following artificial data are designed to illustrate the effect of correlated and uncorrelated independent variables:

explanatory

y:	17	21	26	22	27	25	28	34	29	37	38	38
x:	1	1	1	1	2	2	2	2	3	3	3	3
w:	1	2	3	4	1	2	3	4	1	2	3	4
v:	1	1	2	2	3	3	4	4	5	5	6	6

Here is relevant Minitab output:

```
MTB > Correlation 'y' 'x' 'w' 'v'.

                    y       x       w
x               0.856
w               0.402   0.000
v               0.928   0.956   0.262

MTB > Regress 'y' 3 'x' 'w' 'v';
SUBC>   Predict at x 3 w 1 v 6.

The regression equation is
y = 10.0 + 5.00 x + 2.00 w + 1.00 v

s = 2.646     R-sq = 89.5%     R-sq(adj) = 85.6%

    Fit  Stdev.Fit         95% C.I.              95% P.I.
 33.000      4.077   ( 23.595,  42.405)   ( 21.788,  44.212) XX

X  denotes a row with X values away from the center
XX denotes a row with very extreme X values
```

Locate the 95% prediction interval. Explain why Minitab gave the "very extreme X values" warning.

12.9 Some Multiple Regression Theory (Optional)

12.46 Suppose that we have 10 observations on the response variable, y, and two explanatory variables, x_1 and x_2, which are given below in matrix form.

$$\mathbf{Y} = \begin{bmatrix} 25 \\ 31 \\ 26 \\ 38 \\ 18 \\ 27 \\ 29 \\ 17 \\ 35 \\ 21 \\ 36 \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & 1.7 & 10.8 \\ 1 & 6.3 & 9.4 \\ 1 & 6.2 & 7.2 \\ 1 & 6.3 & 8.5 \\ 1 & 10.5 & 9.4 \\ 1 & 1.2 & 5.4 \\ 1 & 1.3 & 3.6 \\ 1 & 5.7 & 10.5 \\ 1 & 4.2 & 8.2 \\ 1 & 6.1 & 7.2 \end{bmatrix}$$

delete

a. Compute $\mathbf{X}'\mathbf{X}$, $(\mathbf{X}'\mathbf{X})^{-1}$, and $\mathbf{X}'\mathbf{Y}$,

b. Compute the least squares estimators of the prediction equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$

12.46

12.47 Using the data given in Exercise 12.43, display the $\mathbf{X}$ matrix for the following two prediction models:

a. $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2$

b. $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2 + \hat{\beta}_4 x_1^2 + \hat{\beta}_5 x_2^2$

12.48 Refer to Exercise 12.10. Display the $\mathbf{Y}$ and $\mathbf{X}$ matrices for the following two prediction models:

a. $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1$ AGE $+ \hat{\beta}_2$ Weight

b. $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1$ AGE $+ \hat{\beta}_2$ Weight $+ \hat{\beta}_3$ AGE^2 $+ \hat{\beta}_4$ Weight^2 $+ \hat{\beta}_5$ AGE $\cdot$ Weight

Supplementary Exercises

Bus.

12.49 One of the functions of bank branch offices is to arrange profitable loans to small businesses and individuals. As part of a study of the effectiveness of branch managers, a bank collected data from a sample of branches on current total loan volumes (the dependent variable), the total deposits held in accounts opened at that branch, the number of such accounts, the average number of daily transactions, and the number of employees at the branch. Correlations and a scatterplot matrix are shown in the figure.

a. Which independent variable is the best predictor of loan volume?

b. Is there a substantial collinearity problem?

c. Do any points seem extremely influential?

Correlations

Variable	Loan volume (millions)	Deposit volume (millions)	Number of accounts	Transactions	Employees
Loan volume (millions)	1.0000	0.9369	0.9403	0.8766	0.6810
Deposit volume (millions)	0.9369	1.0000	0.9755	0.9144	0.7377
Number of accounts	0.9403	0.9755	1.0000	0.9299	0.7487
Transactions	0.8766	0.9144	0.9299	1.0000	0.8463
Employees	0.6810	0.7377	0.7487	0.8463	1.0000

in greater detail in *Applied Linear Regression Models* by Kutner, Nachtsheim, and Neter (2004). The BP procedure involves the following steps:

Step 1: Fit the regression model, $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i$, to the data and obtain the residuals, e_is and the sum of squared residuals, SS(Residuals).

Step 2: Regress e_i^2 on the explanatory variables: Fit the model $e_i^2 = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \eta_i$ and obtain SS(Regression)*, the regression sum of squares from fitting the model with e_i^2 as the response variable.

Step 3: Compute the BP statistic:

$$\text{BP} = \frac{\text{SS(Regression)*}/2}{(\text{SS(Residuals)}/n)^2}$$

where SS (Regression)* is the regression sum of squares from fitting the model with e_i^2 as the response variable and SS(Residuals) is the sum of square residuals from fitting the regression model with y as the response variable.

Step 4: Reject the null hypothesis of homogeneous variance if BP $> \chi^2_{\alpha,k-1}$ the upper α percentile from a squared distribution with degrees of freedom $k-1$.

Note: The residuals referred to in the BP procedure are the unstandardized residuals: $e_i = y_i - \hat{y}_i$.

Warning: The Breusch–Pagan test should only be used after it has been confirmed that the residuals have a normal distribution.

EXAMPLE 13.15

Refer to the data of Example 13.14, where the residual plots seemed to indicate a violation of the constant variance condition. Apply the Breusch–Pagan test to this data set and determine if there is significant evidence of nonconstant variance.

Solution We will discuss methods for detecting whether or not the residuals appear to have a normal distribution at the end of this section. After that discussion, we will demonstrate in Example 13.17 that the residuals from the data in Example 13.14 appear to have a normal distribution. Thus, we can validly proceed to apply the BP test. Minitab output is given here.

```
Regression Analysis: AGE versus DIAMETER, DIA_SQ

Analysis of Variance

Source           DF       SS       MS      F      P
Regression        2   358412   179207   6.76  0.004
Residual Error   27   715958    26517
Total            29  1074371


Regression Analysis: RESID_SQ versus DIAMETER, DIA_SQ

Analysis of Variance

Source           DF          SS          MS     F      P
Regression        2  1234178.7513  617089.0757  1.62  0.002
Residual Error   27  2185902849.1  80959365.5
Total            29  3420076600.4
```

From the first analysis of variance table we obtain SS(Residual) = 715,958 and from the second analysis of variance table we obtain SS(Regression)* = 12,341,737,513. We then compute

$$BP = \frac{SS(Regression)^*/2}{(SS(Residuals)/n)^2} = \frac{12{,}341{,}737{,}513/2}{(715{,}958/30)^2} = 10.83$$

The critical chi-squared value is $\chi^2_{\alpha,k-1} = \chi^2_{.05,1} = 3.84$. Because $BP = 10.83 > 3.84 = \chi^2_{.05,1}$ we reject H_0: homogenous variances and conclude that there is significant evidence that there is nonconstant variance in this situation.

weighted least squares

What are the consequences of having a nonconstant variance problem in a regression model? First, if the variance about the regression line is not constant, the least-squares estimates may not be as accurate as possible. A technique called **weighted least squares** [see Draper and Smith (1997)] will give more accuracy. Perhaps more important, however, the weighted least-squares technique improves the statistical tests (F and t tests) on model parameters and the interval estimates for parameter because they are, in general, based on smaller standard errors.

The more serious pitfall involved with inferences in the presence of nonconstant variance seems to be for estimates $E(y)$ and predictions of y. For these inferences, the point estimate y is sound but the width of the interval may be too large or too small depending on whether we're predicting in a low or high variance section of the experimental region.

The best remedy for nonconstant variance is to use weighted least squares. We will not cover this technique in the text. However, when the nonconstant variance possesses a pattern related to y, a reexpression (transformation) of y may resolve the problem. Several transformations for y were discussed in Chapter 11; ones that help to stabilize the variance when there is a pattern to the nonconstant variance were discussed in Chapter 8 for the analysis of variance. They can also be applied in certain regression situations.

An excellent discussion of transformations is given in the book *Introduction to Regression Modeling* by Abraham and Ledolter (2006). A special class of transformations is called the **Box–Cox** transformations. The general form of the Box–Cox transformation is

$$g(y_i) = (y_i^\lambda - 1)/\lambda$$

where λ is a constant to be determined from the data. From the form of $g(y_i)$ we can observe the following special cases:

- If $\lambda = 1$, then no transformation is needed. The original data should be modeled.
- If $\lambda = 2$, then the Box–Cox transformation is the square of the original response variable and y_i^2 should be modeled.
- If $\lambda = -1$, then the Box–Cox transformation is the reciprocal of the original response variable and $1/y_i$ should be modeled.
- If $\lambda = 1/2$, then the Box–Cox transformation is the reciprocal of the original response variable and $\sqrt{y_i}$ should be modeled.
- If $\lambda = 0$, then in the limit as λ converges to 0, the Box–Cox transformation is the natural logarithm of the original response variable and $\log(y_i)$ should be modeled.
- If $\lambda = -1/2$, then the Box–Cox transformation is the reciprocal of the square root of the original response variable and $1/\sqrt{y_i}$ should be modeled.

with the brightness of finished paper. The article, "Advantages of CE-HDP bleaching for high brightness kraft pulp production," *Tappi* 47 (1964): 170A–175A, contains the following data on the variables: y = brightness of finished paper, x_1 = hydrogen peroxide (% by weight), x_2 = sodium hydroxide (% by weight), x_3 = silicate (% by weight), x_4 = process temperature (in °F). There were 31 runs in the study.

Run	x_1	x_2	x_3	x_4	y	Run	x_1	x_2	x_3	x_4	y
1	.2	.2	1.5	145	83.9	17	.1	.3	2.5	160	82.9
2	.4	.2	1.5	145	84.9	18	.5	.3	2.5	160	85.5
3	.2	.4	1.5	145	83.4	19	.3	.1	2.5	160	85.2
4	.4	.4	3.5	145	84.2	20	.3	.5	2.5	160	84.5
5	.2	.2	3.5	145	83.8	21	.3	.3	2.5	160	84.7
6	.4	.2	3.5	145	84.7	22	.3	.3	2.5	160	85.0
7	.2	.4	3.5	145	84.0	23	.3	.3	2.5	160	84.9
8	.4	.4	1.5	175	84.8	24	.3	.3	2.5	160	84.0
9	.2	.2	1.5	175	84.5	25	.3	.3	2.5	160	84.5
10	.4	.2	1.5	175	86.0	26	.3	.3	2.5	160	84.7
11	.2	.4	1.5	175	82.6	27	.3	.3	2.5	160	84.6
12	.4	.4	3.5	175	85.1	28	.3	.3	2.5	160	84.9
13	.2	.2	3.5	175	84.5	29	.3	.3	2.5	160	84.9
14	.4	.2	3.5	175	86.0	30	.3	.3	2.5	160	84.5
15	.2	.4	3.5	175	84.0	31	.3	.3	2.5	160	84.6
16	.4	.4	3.5	175	85.4						

a. Use scatterplots and VIF to determine if there is evidence of collinearity in the explanatory variables.

b. This was a designed experiment with non-random explanatory variables. Was it really necessary to investigate collinearity in this type of study?

c. Use a variable selection procedure with maximum R^2 as the criterion to formulate a model.

d. Use a variable selection procedure with maximum R^2_{adj} as the criterion to formulate a model.

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

13.7 Refer to Exercise 13.6. Include the square of each of the explanatory variables and all crossproduct terms in your model selection procedure.

a. Use a variable selection procedure with maximum R^2_{adj} as the criterion to formulate a model.

b. Use a variable selection procedure C_p as the criterion to formulate a model.

c. Use a variable selection procedure with minimum PRESS statistic as the criterion to formulate a model.

d. Compare the included terms from the models formulated with the three criteria in (a)–(c).

13.3 Formulating the Model (Step 2)

13.8 The cotton aphid is pale to dark green in cool seasons and yellow in hot, dry summers. Generally distributed throughout temperate, subtropic, and tropic zones, the cotton aphid occurs in all cotton-producing areas of the world. These insects congregate on lower leaf surfaces and on terminal buds, extracting plant sap. If weather is cool during the spring, populations of natural enemies will be slow in building up and heavy infestations of aphids may result. When this occurs, leaves begin to curl and pucker; seedling plants become stunted and may die. Most aphid damage is of this type. If honeydew resulting from late season aphid infestations falls onto open cotton, it can act as a growing medium for sooty mold. Cotton stained by this black fungus is reduced in quality and brings a low price for the grower. Entomologists studied the aphids to determine

b. Using your results from part (a), obtain separate prediction equations for varieties Hallertau and Saaz.
c. Interpret the values of the coefficients (βs) in the model.
d. Using your prediction equations in part (b), estimate the mean alpha acid percentage when the atmospheric conditions are a mean temperature of 19°C and a mean sunshine of 6.5. How different are the two estimates?
e. Place 95% confidence intervals on your estimates.

13.18

13.20 Refer to Exercise 13.17.
a. Using the model fit in part (a) of Exercise 13.17, is there significant evidence ($\alpha = .05$) that the mean sunshine partial slope coefficients are different?
b. Using the model fit in part (a) of Exercise 13.17, is there significant evidence ($\alpha = .05$) that the mean temperature partial slope coefficients are different?
c. Interpret the values of the coefficients (βs) in the model.

13.19

Bus. **13.21** A supermarket chain analyzed data on sales of a particular brand of snack cracker at 104 stores in the chain for a certain 1-week period. The analyst tried to predict sales based on the total sales of all brands in the snack cracker category, the price charged for the particular brand in question, and whether or not there was a promotion for a competing brand at a given store (promotion = 1 if there was such a promotion, 0 if not). (There were no promotions for the brand in question.) A portion of the JMP multiple regression output is shown in the figure.
a. Interpret the coefficient of the promotion variable.
b. Should a promotion by a competing product increase or decrease sales of the brand in question? According to the coefficient, does it?
c. Is the coefficient significantly different from 0 at usual α values?

Response: Sales

Summary of Fit

RSquare	0.434158
RSquare Adj	0.417182
Root Mean Square Error	17.9913
Mean of Response	356.7692
Observations (or Sum Wgts)	104

Lack of Fit

Parameter Estimates

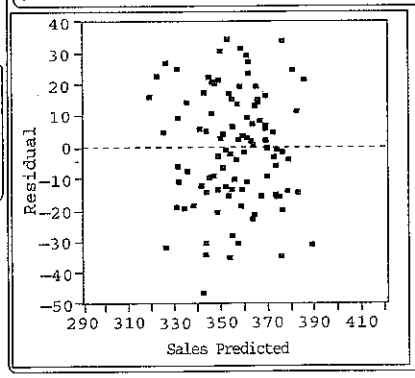
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	129.85375	80.66628	1.61	0.1106
Price	44.849952	39.93534	1.12	0.2641
Category sales	0.1214871	0.018249	6.66	0.0000
Promotion by other?	-19.95964	3.702304	-5.39	0.0000

Effect Test

Whole-Model Test

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	24835.761	8278.59	25.5759
Error	100	32368.701	323.69	Prob>F
C Total	103	57204.462		0.0000



13.22 In the previous question, how accurately can sales be predicted for one particular week, with 95% confidence?

Bus. **13.23** An additional regression model for the snack cracker data is run, incorporating products of the promotion variable with price and with category sales. The output for this model is given in the figure. What effect do the product term coefficients have in predicting sales when there is a promotion by a competing brand? In particular, do these coefficients affect the intercept of the model or the slopes?

variance

13.78 Use the model you selected in Exercise 13.77, to answer the following questions.
- **a.** Do the residuals appear to have a normal distribution? Justify your answer.
- **b.** Does the condition of constant appear to be satisfied? Justify your answer.
- **c.** Obtain the Box–Cox transformation of this data set.

13.79 Use the model you selected in Exercise 13.77, to answer the following questions.
- **a.** Do any of the data points appear to have high influence? Leverage? Justify your answer.
- **b.** If you identified any high leverage or high influence points in part (a), compare the estimated models with and without these points.
- **c.** What is your final model describing sulfur dioxide air pollution?
- **d.** Display any other explanatory variables which may improve the fit of your model.

13.80 Use the model you selected in Exercise 13.79 to answer the following questions.
- **a.** Estimate the average level of sulfur dioxide content of the air in a city having the following values for the six explanatory variables:

$$x_1 = 60 \quad x_2 = 150 \quad x_3 = 600 \quad x_4 = 10 \quad x_5 = 40 \quad x_6 = 100$$

- **b.** Place a 95% confidence interval on your estimated sulfur dioxide level.
- **c.** List any major limitations in your estimation of this mean.

TABLE 14.3
Analysis of variance table for a completely randomized design

Source	SS	df	MS	F
Treatments	SST	$t-1$	MST = SST/$(t-1)$	MST/MSE
Error	SSE	$N-t$	MSE = SSE/$(N-t)$	
Total	TSS	$N-1$		

Recall from Chapter 8 that we summarized this information in an analysis of variance (AOV) table, as represented in Table 14.3, with $N = \Sigma_i n_i$.

unbiased estimates

When H_0: $\tau_1 = \cdots = \tau_t = 0$ is true, both MST and MSE are **unbiased estimates** of σ_ε^2, the variance of the experimental error. That is, when H_0 is true, both MST and MSE have a mean value in repeated sampling, called the **expected mean squares,** equal to σ_ε^2. We express these terms as

expected mean squares

$$E(\text{MST}) = \sigma_\varepsilon^2 \quad \text{and} \quad E(\text{MSE}) = \sigma_\varepsilon^2$$

Thus, we would expect $F = \text{MST/MSE}$ to be near 1 when H_0 is true. When H_a is true and there is a difference in the treatment means, the mean of MSE is still an unbiased estimate of σ_ε^2,

$$E(\text{MSE}) = \sigma_\varepsilon^2$$

However, MST is no longer unbiased for σ_ε^2. In fact, the expected mean square for treatments can be shown to be

$$E(\text{MST}) = \sigma_\varepsilon^2 + n\theta_T$$

where $\theta_T = 1/(t-1)(\Sigma_i \tau_i^2)$. When H_a is true, some of the τ_i's are not zero, and θ_T is positive. Thus, MST will tend to overestimate σ_ε^2. Hence, under H_a, the ratio $F = \text{MST/MSE}$ will tend to be greater than 1, and we will reject H_0 in the upper tail of the distribution of F.

In particular, for selected values of the probability of Type I error α, we will reject H_0: $\tau_1 = \cdots = \tau_t = 0$ if the computed value of F exceeds $F_{\alpha,t-1,N-t}$, the critical value of F found in Table 8 in the Appendix with Type I error probability, α, $\text{df}_1 = t-1$, and $\text{df}_2 = N-t$. Note that df_1 and df_2 correspond to the degrees of freedom for MST and MSE, respectively, in the AOV table.

The completely randomized design has several advantages and disadvantages when used as an experimental design for comparing t treatment means.

Advantages and Disadvantages of a Completely Randomized Design

Advantages

1. The design is extremely easy to construct.
2. The design is easy to analyze even though the sample sizes might not be the same for each treatment.
3. The design can be used for any number of treatments.

Disadvantages

1. Although the completely randomized design can be used for any number of treatments, it is best suited for situations in which there are relatively few treatments.
2. The experimental units to which treatments are applied must be as homogeneous as possible. Any extraneous sources of variability will tend to inflate the error term, making it more difficult to detect differences among the treatment means.

On each farm, five 1-acre plots are selected,

15.8 Exercises

15.2 Randomized Complete Block Design

Ag.

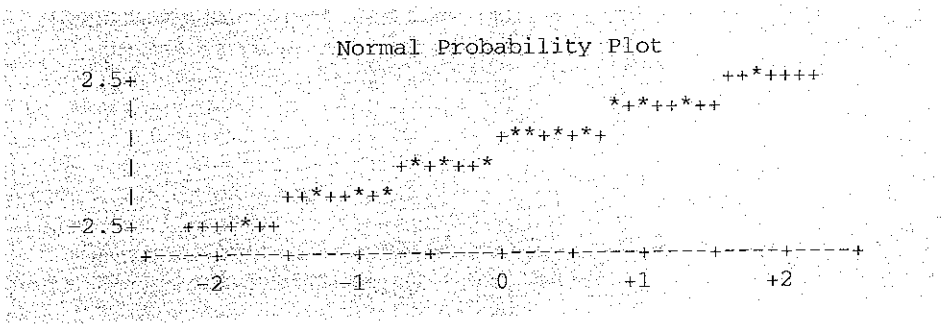
15.1 A horticulturist is designing a study to investigate the effectiveness of five methods for the irrigation of blueberry shrubs. The methods are surface, trickle, center pivot, lateral move, and subirrigation. There are 10 blueberry farms available for the study representing a wide variety of types of soil, terrains, and wind gradients. The horticulturist wants to use each of the five methods of irrigation on all 10 farms to moderate the effect of the many extraneous sources of variation that may impact the blueberry yields. Each farm is divided into five and the response variable will be the weight of the harvested fruit from each of the plots of blueberry shrubs.

- **a.** Show the details of how you would randomly assign the five methods of irrigation to the plots.
- **b.** How many different arrangements of the five methods of irrigation are possible in each of the farms?
- **c.** How many different arrangements are possible for the whole study of 10 farms?

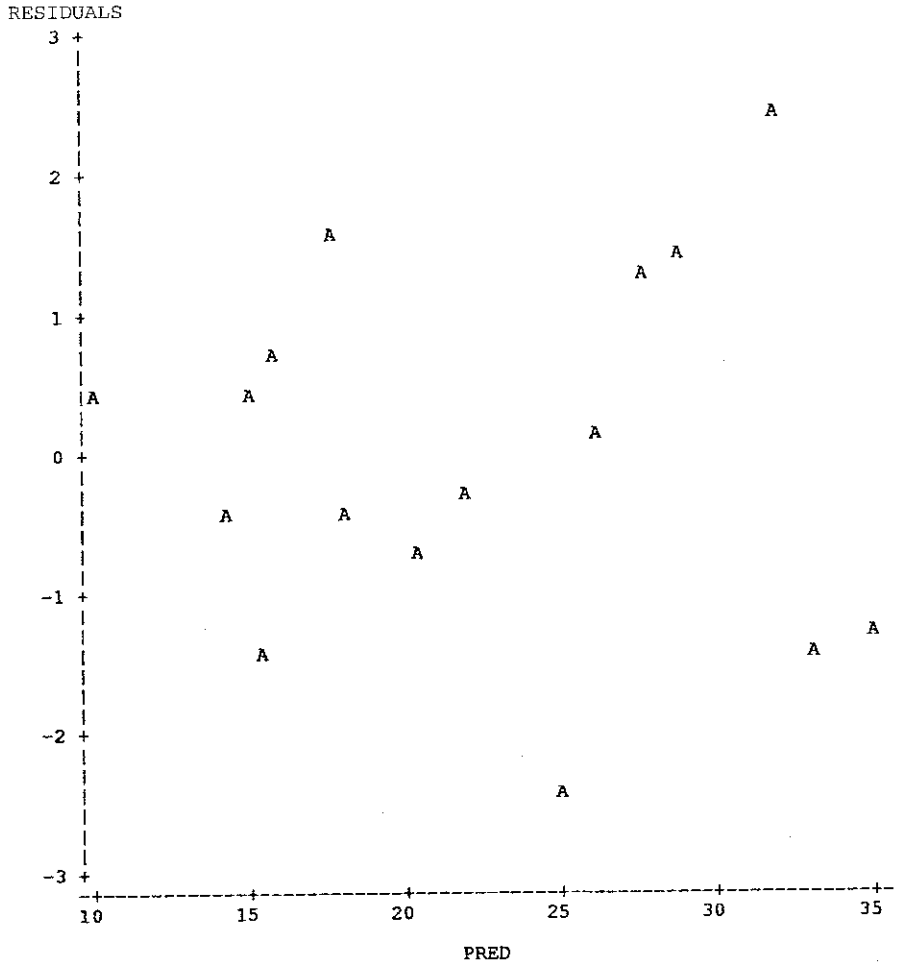
15.2 Refer to Exercise 15.1. The study was conducted and the yields in pounds of blueberries over a growing season are given here along with the Minitab output for the analysis.

	Method of Irrigation					
Farm	**Surface**	**Trickle**	**Center Point**	**Lateral**	**Subirrigation**	**Farm Mean**
1	597	248	391	423	350	401.9
2	636	382	434	461	370	456.6
3	591	348	492	504	460	478.9
4	603	366	468	580	452	493.9
5	649	258	457	449	343	430.9
6	512	321	406	464	340	408.7
7	588	423	466	550	327	470.8
8	689	406	502	526	378	500.0
9	690	400	559	469	419	507.3
10	608	380	469	550	458	493.2
Method Mean	616.3	353.2	464.3	497.6	389.6	464.2

```
Two-way ANOVA: Yield versus Method, Farm

Source   DF      SS     MS      F      P
Method    4  421560 105390  61.10  0.000
Farm      9   66318   7369   4.27  0.001
Error    36   62097   1725
Total    49  549975

S = 41.53   R-Sq = 88.71%   R-Sq(adj) = 84.63%
```



PLOT OF RESIDUALS VERSUS ESTIMATED TREATMENT MEANS FOR EXERCISE 15.11



15.4 Factorial Treatment Structure in a Randomized Complete Block Design

15.12 A researcher decides to design an experiment consisting of 3 replications of a treatment structure created by combining the 4 levels of factor A with 5 levels of factor B. The experimenter is concerned about the heterogeneity in the 40 experimental units and hence decides to block the experimental units into 3 groups.

60

585.0002400

Source	DF	Type III SS	Mean Square	F Value	Pr > F
XI (AGE)	1	188.8171093	188.8171093	4.27	0.0423
X2 (C1)	1	554.5949725	554.5949725	12.55	0.0007
X3 (C2)	1	170.5036575	170.5036575	3.86	0.0533
X4 (C3)	1	31.7644655	31.7644655	0.72	0.3993
X5 (X1*X2)	1	2.3592730	2.3592730	0.05	0.8179
X6 (X1*X3)	1	13.4041339	13.4041339	0.30	0.5835
X7 (X1*X4)	1	69.8470840	69.8470840	1.58	0.2127

(C1)

Parameter	Estimate	Standard Error	t Value	Pr> |t|
Intercept	34.70618792	5.34237764	6.50	<.0001
XI (AGE)	0.35901225	0.17365438	2.07	0.0423
X2 (C1)	-23.90173188	6.74587910	-3.54	0.0007
X3 (C2)	-12.60819010	6.41775371	-1.96	0.0533
X4 (C3)	-6.14139936	7.24258684	-0.85	0.3993
X5 (X1*X2)	0.05050567	0.21854869	0.23	0.8179
X6 (X1*X2)	-0.11399476	0.20694867	-0.55	0.5835
X7 (X1*X3)	-0.29938948	0.23810008	-1.26	0.2127

II

MODEL II: SAME SLOPES BUT TREATMENT DIFFERENCES

The GLK Procedure

Dependent Variable: Y VSKILL

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	6201.828754	1550.457189	35.06	<.0001
Error	75	3316.828121	44.224375		
Corrected Total	79	9518.656875			

(C1)

Source	DF	Type III SS	Mean Square	F Value	Pr > F
XI (AGE)	1	688.689379	688.689379	15.57	0.0002
X2 (C1)	1	5056.080561	5056.080561	114.33	<.0001
X3 (C2)	1	2542.485319	2542.485319	57.49	<.0001
X4 (C3)	1	2183.510378	2183.510378	49.37	<.0001

Parameter	Estimate	Standard Error	t Value	Pr> |t|
Intercept	37.19688646	2.53835518	14.65	<.0001
XI (AGE)	0.27472465	0.06961724	3.95	0.0002
X2 (C1)	-22.49016521	2.10337402	-10.69	<.0001
X3 (C2)	-15.95148404	2.10379161	-7.58	<.0001
X4 (C3)	-14.78401158	2.10399893	-7.03	<-0001

MODEL III: SAME SLOPES AND NO TREATMENT DIFFERENCES

The GLM Procedure

Dependent Variable: Y VSKILL

Source	DF	Sum of Squares	Mean Square	F Value	P > F
Model	1	793.871687	793.871687	7.10	0.0094
Error	78	8724.785188	111.856220		
Corrected Total	79	9518.656875			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
XI (AGE)	1	793.8716867	793.8716867	7.10	0.0094

Parameter	Estimate	Error	t Value	Pr> |t|
Intercept	23.30979334	3.41463015	6.83	<.0001
XI (AGE)	0.29478261	0.11065137	2.66	0.0094

$$\text{SSL} = \sum_j 6(\bar{y}_{.j.} - \bar{y}_{...})^2 = 6\{(11.3 - 10.4)^2 + (10.1 - 10.4)^2 + (9.8 - 10.4)^2\}$$
$$= 7.56$$
$$\text{SSAL} = \sum_{ij} 2(\bar{y}_{ij.} - \bar{y}_{...})^2 = 2\{(10.9 - 10.4)^2 + (10.15 - 10.4)^2 + (9.85 - 10.4)^2$$
$$+ \cdots + (10.55 - 10.4)^2\} = 1.64$$
$$\text{SSE} = \text{TSS} - \text{SSA} - \text{SSL} - \text{SSAL} = 12.00 - 1.56 - 7.56 - 1.64 = 1.24$$

Our results are summarized in an analysis of variance table in Table 17.15.

TABLE 17.15
AOV table for Example 17.3 experiment

Source	SS	df	MS	EMS
Assay	1.56	2	.78	$\sigma_\varepsilon^2 + 2\sigma_{\tau\beta}^2 + 6\sigma_\tau^2$
Lab	7.56	2	3.78	$\sigma_\varepsilon^2 + 2\sigma_{\tau\beta}^2 + 6\sigma_\beta^2$
Assay*Lab	1.64	4	.41	$\sigma_\varepsilon^2 + 2\sigma_{\tau\beta}^2$
Error	1.24	9	.1378	σ_ε^2
Total	12.00	17		

We can proceed with appropriate statistical tests, using the results presented in the AOV table. For the AL interaction we have

H_0: $\sigma_{\tau\beta}^2 = 0$

H_a: $\sigma_{\tau\beta}^2 > 0$

T.S.: $F = \dfrac{\text{MSAL}}{\text{MSE}} = \dfrac{.41}{.1378} = 2.98$

R.R.: For $\alpha = .05$, we will reject H_0 if F exceeds 3.63, the critical value for F with $\alpha = .05$, $df_1 = 4$, and $df_2 = 9$.

Conclusion: There is insufficient evidence to reject H_0. There does not appear to be a significant interaction between the levels of factors A and L.

For factor B we have

H_0: $\sigma_\beta^2 = 0$

H_a: $\sigma_\beta^2 > 0$

T.S.: $F = \dfrac{\text{MSL}}{\text{MSAL}} = \dfrac{3.78}{.41} = 9.22$

R.R.: For $\alpha = .05$, we will reject H_0 if F exceeds 6.94, the critical value based on $\alpha = .05$, $df_1 = 2$, and $df_2 = 4$.

Conclusion: Because the observed value of F is much larger than 6.94, we reject H_0 and conclude that there is a significant variability in calcium concentrations from lab to lab.

The test for factor A follows:

H_0: $\sigma_\tau^2 = 0$

H_a: $\sigma_\tau^2 > 0$

T.S.: $F = \dfrac{\text{MSA}}{\text{MSAL}} = \dfrac{.78}{.41} = 1.90$

b. Display a partial AOV table including df and expected mean squares for all sources of variation.

c. Provide the ratio of mean squares for all appropriate F tests for determining the significance of variability.

Env. **17.16** Refer to Exercise 17.10. Suppose the four chemicals were randomly selected from the hundreds of different chemicals used to control fire ants. The researchers were interested in whether the effectiveness of a chemical to control fire ants varied across different environments.

a. Write an appropriate model for this situation. Indicate how the conditions placed on the terms in the model differ from the conditions placed on the model used when the chemicals were the only chemicals of interest to the researchers.

b. Construct the AOV table and test all relevant hypotheses.

c. Compare the conclusions and inferences in this problem to those of Exercise 17.10.

17.17 Refer to Exercise 17.16.

a. Which model and analysis seem to be most appropriate? Explain your answer.

b. Under what circumstances would a fixed-effects model be appropriate?

Eng. **17.18** The civil engineering department at a university was awarded a large grant to study the campus traffic problems and to recommend alternative solutions. One small phase of the study involved obtaining daily counts on the number of cars crossing, but not making use of, the campus facilities. To do this, a team of volunteers was stationed at each entrance to monitor simultaneously the license number and the time of entrance or exit for each car passing through the checkpoint. By comparing lists for all checkpoints and allowing a reasonable time for cars to traverse the campus, the teams were able to determine the number of cars crossing but not using the campus facilities during the 8:00 A.M. to 5:00 P.M. time period. A random sample of 6 weeks throughout the academic year was used, with 2 midweek days selected for study in the weeks sampled. The traffic volume data appear next.

Week 1	Week 2	Week 3	Week 4	Week 5	Week 6
680	438	539	264	693	530
618	520	600	198	646	575

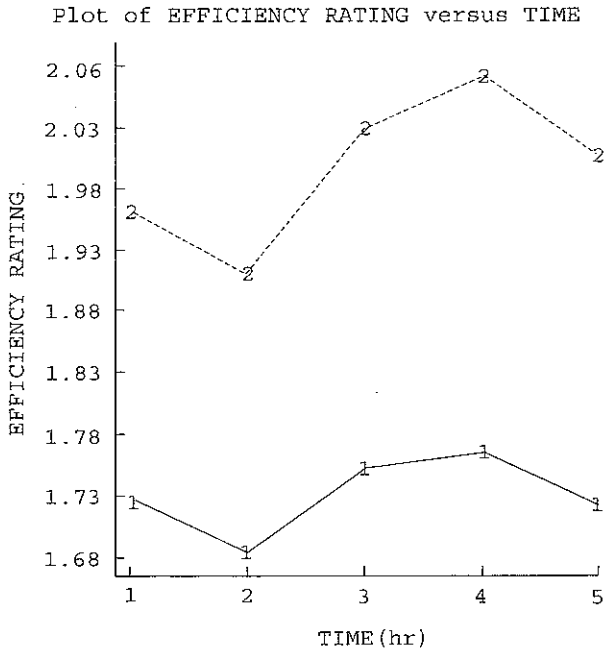
a. Write an appropriate linear statistical model. Identify all terms in the model.

b. Perform an analysis of variance, indicating expected mean squares. Use $\alpha = .05$.

Gov. **17.19** The public safety department at a large urban university is concerned about criminal activities involving nonstudents stealing bicycles and laptops from students. The campus police design a study to investigate the number of automobiles entering the campus that do not have a campus parking sticker or do not enter a campus parking facility. The police are suspicious that such individuals may be involved in criminal activities. A team of criminal justice students was stationed at each entrance to the campus to monitor simultaneously the license number of all cars and to determine if the car had a campus parking sticker. By utilizing the computer records of all campus parking facilities which record the license number of all cars upon their entrance to a parking facility, the teams were able to determine the number of cars entering the campus but not using campus facilities. Data were collected during a random sample of 10 weeks throughout the academic year. The counts of "suspicious" cars are recorded on the five business days during the selected 10 weeks and appear here.

12

	Week											
Day	1	2	3	4	5	6	7	8	9	10	11	12
Mon	52	51	52	54	56	54	51	56	51	48	52	53
Tue	47	50	50	51	55	51	49	54	49	46	51	50
Wed	49	50	50	52	54	51	49	54	49	47	52	50
Thu	49	50	49	52	54	50	48	54	49	46	51	51
Fri	44	48	48	50	53	50	48	52	48	45	50	51



Psy. **18.29** An experimenter is designing an experiment in which she plans to compare nine different formulations of a meat product. One factor, F, is percent fat (10%, 15%, 20%) in the meat. The other factor, C, is cooking method (broil, bake, fry). She will prepare samples of each of the nine combinations and present them to tasters who will score the samples based on various criteria. Four tasters are available for the study. Each taster will taste nine samples. There are taster-to-taster differences, but the order in which the samples are tasted will not influence the taste scores. The samples will be prepared in the following manner so that the meat samples can be prepared and kept warm for the tasters. A portion of meat containing 15% fat will be divided into three equal portions. Each of the three methods of cooking will then be randomly assigned to one of the three portions. This procedure will be repeated for meat samples having 15% and 20% fat. The nine meat samples will then be tasted and scored by the taster. The whole process is repeated for the other three tasters. The taste scores (0 to 100) are given here.

	10% Fat			15% Fat			20% Fat		
	Broil	**Bake**	**Fry**	**Broil**	**Bake**	**Fry**	**Broil**	**Bake**	**Fry**
Taster 1	75	79	82	78	82	81	81	85	87
Taster 2	74	78	81	78	81	83	84	87	88
Taster 3	75	78	79	80	82	83	87	88	92
Taster 4	91	88	83	80	76	73	81	77	74

a. Identify the design.
b. Give an appropriate model with assumptions.
c. Give the sources of variability and degrees of freedom for an AOV.

(*continued*)

		Time since treatment (Hours)									Time since treatment (Hours)						
CCl_4	$CHCl_3$	0	.01	.25	.5	1.0	2.0	3.0	CCl_4	$CHCl_3$	0	.01	.25	.5	1.0	2.0	3.0
0	10	.08	.14	.24	.27	.29	.32	.34	0	10	.05	.05	.15	.16	.19	.22	.23
0	25	.07	.10	.25	.51	.65	.66	.70	0	25	.07	.07	.17	.24	.34	.37	.41
0	25	.11	.11	.33	.39	.48	.52	.55	0	25	.07	.06	.16	.24	.31	.36	.41
1	0	.06	.11	.13	.09	.10	.11	.11	1	0	.05	.08	.10	.10	.11	.12	.13
1	0	.08	.14	.15	.14	.16	.19	.21	1	0	.05	.09	.08	.09	.11	.12	.13
1	5	.05	.13	.18	.37	.41	.42	.46	1	5	.06	.10	.14	.16	.16	.20	.18
1	5	.10	.16	.22	.22	.29	.30	.21	1	5	.05	.08	.15	.18	.19	.21	.21
1	10	.06	.10	.25	.61	.57	.60	.63	1	10	.05	.07	.24	.27	.29	.32	.32
1	10	.11	.14	.26	.30	.30	.35	.29	1	10	.05	.06	.16	.21	.24	.27	.27
1	25	.07	.09	.23	.39	.58	.53	.67	1	25	.06	.06	.15	.22	.30	.44	.56
1	25	.08	.11	.28	.40	.42	.75	.72	1	25	.06	.05	.15	.27	.36	.43	.55
2.5	0	.06	.09	.19	.56	.64	.33	.34	2.5	0	.05	.08	.18	.19	.19	.21	.20
2.5	0	.10	.10	.19	.21	.23	.28	.23	2.5	0	.05	.10	.21	.23	.28	.29	.31
2.5	5	.07	.10	.22	.57	.62	.66	.70	2.5	5	.06	.08	.19	.23	.24	.27	.31
2.5	5	.07	.11	.24	.28	.30	.35	.30	2.5	5	.06	.07	.21	.25	.28	.30	.32
2.5	10	.05	.12	.28	.33	.43	.49	.58	2.5	10	.06	.09	.33	.26	.31	.34	.36
2.5	10	.08	.14	.23	.37	.43	.47	.40	2.5	10	.06	.09	.19	.23	.29	.34	.34
2.5	25	.05	.07	.22	.59	.65	.67	.67	2.5	25	.04	.05	.21	.29	.36	.54	.72
2.5	25	.09	.09	.24	.31	.35	.46	.45	2.5	25	.05	.04	.15	.25	.36	.40	.48
5	0	.06	.09	.52	.77	.78	.73	.76	5	0	.06	.08	.45	.50	.49	.60	.71
5	0	.08	.09	.60	.60	.57	.73	.79	5	0	.06	.10	.42	.44	.62	.62	.73
5	5	.05	.11	.21	.27	.30	.36	.41	5	5	.05	.10	.20	.22	.24	.28	.33
5	5	.09	.12	.21	.22	.27	.32	.28	5	5	.05	.08	.17	.21	.26	.27	.32
5	10	.04	.10	.24	.26	.33	.39	.47	5	10	.06	.09	.25	.29	.33	.37	.40
5	10	.11	.11	.23	.27	.31	.36	.31	5	10	.05	.05	.12	.16	.22	.27	.29
5	25	.07	.07	.21	.55	.60	.66	.66	5	25	.05	.05	.23	.31	.35	.53	.66
5	25	.08	.09	.23	.31	.41	.58	.67	5	25	.06	.04	.12	.20	.31	.41	.57

a. Plot the mean percentage LDH leakage by time for the 16 treatments. Does there appear to be an effect due to increasing the levels of CCl_4 or $CHCl_3$?

b. From the plot, does there appear to be an increase in the mean percentage leakage as time after treatment increases?

c. Plot a profile plot of the mean percentage LDH leakage separately for each time period. Does there appear to be a difference in the profile plots?

18.31 Refer to Exercise 18.30.

a. Run a repeated measures analysis of variance and determine if there are significant interaction and/or main effects due to CCl_4 and $CHCl_3$. Is there a significant time effect?

b. Do the conditions necessary for using a split-plot analysis of repeated-measures data appear to be valid?

18.32 Refer to Exercise 18.30. Consider as your response variable the proportional change in the mean percentage leakage at time 3 hours and at time 0. That is,

$$y = \frac{P_3 - P_1}{P_0}$$

where P_0 and P_3 are the percentage leakage values at times 0 and 3 hours, respectively. Run an analysis of variance on y and test for significant interaction and/or main effects due to CCl_4 and $CHCl_3$. Do you reach similar conclusions to those obtained in Exercise 18.31?

19.2 A Randomized Block Design with One or More Missing Observations

unbalanced design

Any time the number of observations is not the same for all factor–level combinations, we call the design **unbalanced.** Thus, a randomized block design or a Latin square design with one or more missing observations is an unbalanced design. We will begin our examination by considering a simple case, a randomized block design with one missing observation.

value of missing observation
estimation bias

The analysis of variance for a randomized block design with one missing observation can be performed rather easily by using the formulas for a randomized complete block design, after we have estimated the **value of the missing observation** and corrected for the **estimation bias.**

Let y_{ij} be the response from the experimental unit observed under treatment i in block j. Suppose that the missing observation occurs in cell (k, h), the observation on treatment k in block h. The formula for estimating the missing observation y_{kh} is given by

$$\hat{y}_{kh} = \frac{ty_{k.} + by_{.h} - y_{..}}{(t-1)(b-1)}$$

where t is the number of treatments, b is the number of blocks, $y_{k.}$ is sum of all observations on treatment k, the treatment which has the missing observation, $y_{.h}$ is the sum of all measurements in block h, the block which has the missing observation, and $y_{..}$ is the sum of all the observations.

The sums of squares for the analysis of variance table are obtained by replacing the missing value, y_{kh}, with its estimate $\hat{y}_{kh}$ and then applying the formulas for a balanced design to the data set that now has no missing cells:

$$\text{TSS} = \sum_{i=1}^{t}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{..})^2$$

$$\text{SST} = b\sum_{i=1}^{t}(\bar{y}_{i.} - \bar{y}_{..})^2$$

$$\text{SSB} = t\sum_{j=1}^{b}(\bar{y}_{.j} - \bar{y}_{..})^2$$

$$\text{SSE} = \text{TSS} - \text{SST} - \text{SSB}$$

The value of SST has a bias in its estimation given by

$$\text{Bias} = \frac{(y_{.h} - (t-1)\hat{y}_{kh})^2}{t(t-1)}$$

The corrected treatment sum of squares is $\text{SST}_C = \text{SST} - \text{Bias}$. The other sums of squares are given in their uncorrected form.

Another difference in the analysis of variance table for the unbalanced block designs is a change in the entries for degrees of freedom for total and error. Because n in the unbalanced design refers to the number of actual observations, the value of n is given by $n = tb - 1$ due to the missing data point. Therefore, the degrees of freedom for *Error* will be decreased by one to $n - t - b + 1 = tb - t - b + 2$ as compared to $tb - t - b + 1$ for the corresponding balanced design. The AOV table for an unbalanced design with t treatments, b blocks, and one missing value is shown in Table 19.1.

We illustrate the analysis of variance for this design with an example.

+1 +1 delete

TABLE 19.1
AOV table for testing the effects of treatments with one missing observation

Source	SS	df	MS	F
$\text{Blocks}_{\text{unadj}}$	$\text{SSB}_{\text{unadj}}$	$b-1$	$\text{MSB}_{\text{unadj}}$	
Treatments_C	SST_C	$t-1$	MST_C	MST_C/MSE
Error	SSE	$bt-b-t-2$	MSE	
Total	TSS	$bt-2$		

EXAMPLE 19.1

Prior to spinning cotton, the cotton must be processed to remove foreign matter and moisture. The most common lint cleaner is the controlled batt saw-type lint cleaner. Although the controlled-batt saw-type lint cleaner M1 is one of the most highly effective cleaners, it is also one of the cleaners that causes the most damage to the cotton fibers. A cotton researcher designed a study to investigate four alternative methods for cleaning cotton fibers, M2, M3, M4, and M5. Methods M2 and M3 are mechanical, whereas methods M4 and M5 are a combination of mechanical and chemical procedures. The researcher wanted to take into account the impact of different growers on the process and hence obtained bales of cotton from six different cotton ranchers. The ranchers will be considered as blocks in the study. After a preliminary cleaning of the cotton, the six bales were thoroughly mixed and then an equal amount of cotton was processed by each of the five lint-cleaning methods. The losses in weight (in kg) after cleaning the cotton fibers are given in Table 19.2 for the five cleaning methods. During the processing of the cotton samples, the measurements from batch 1 processed by the M1 cleaner were lost.

TABLE 19.2
Measurements of loss (kg) during cotton fiber cleaning

	Batch						
Method	1	2	3	4	5	6	Mean
M1	*	6.75	13.05	10.26	8.01	8.42	9.300
M2	5.54	3.53	11.20	7.21	3.24	6.45	6.190
M3	7.67	4.15	9.79	8.27	6.75	5.50	7.022
M4	7.89	1.97	8.97	6.12	4.22	7.84	6.170
M5	9.27	4.39	13.44	9.13	9.20	7.13	8.760
Mean	7.593	4.158	11.290	8.198	6.280	7.068	7.426

Estimate the value for the missing observation and then perform an analysis of variance to test for differences in the mean weight loss for the five methods of cleaning cotton fibers.

Solution For this randomized block design we have that $b = 6$ and $t = 5$ with one missing value in cell (1, 1). Therefore, we need to compute the following values:

$$y_{1\cdot} = \text{sum of all measurements on method M1}$$
$$= 6.75 + 13.05 + 10.26 + 8.01 + 8.42 = 46.49$$
$$y_{\cdot 1} = \text{sum of all measurements on batch 1}$$
$$= 5.54 + 7.67 + 7.89 + 9.27 = 30.37$$
$$y_{\cdot\cdot} = \text{sum of all measurements}$$
$$= 6.75 + 13.05 + \cdots + 7.13 = 215.36$$

The estimate of the missing value, y_{11}, is given by

$$\hat{y}_{11} = \frac{ty_{1.} + by_{.1} - y_{..}}{(t-1)(b-1)} = \frac{5(46.49) + 6(30.37) - 215.36}{(5-1)(6-1)}$$
$$= \frac{199.31}{20} = 9.9655$$

Replacing the missing value with its estimate, 9.9655, we next compute the sum of squares using the formulas of Chapter 15 for a balanced randomized block design with $t = 5$ and $b = 6$. First we obtain the treatment and batch means (with the missing value replaced with 9.9655) as shown in Table 19.3.

TABLE 19.3
Method and batch means

Method Means	Batch Means
$\bar{y}_{1.} = 9.409$	$\bar{y}_{.1} = 8.067$
$\bar{y}_{2.} = 6.190$	$\bar{y}_{.2} = 4.158$
$\bar{y}_{3.} = 7.022$	$\bar{y}_{.3} = 11.290$
$\bar{y}_{4.} = 6.170$	$\bar{y}_{.4} = 8.198$
$\bar{y}_{5.} = 8.760$	$\bar{y}_{.5} = 6.280$
	$\bar{y}_{.6} = 7.068$
Overall Mean	$\bar{y}_{..} = 7.511$

Note that the means for method 1, batch 1, and the overall mean incorporate the estimated value for the missing observation. We next obtain the four sums of squares.

$$\text{TSS} = \sum_{i=1}^{t}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{..})^2$$
$$= (9.9655 - 7.511)^2 + (6.75 - 7.511)^2 + \cdots + (7.13 - 7.511)^2 = 219.887$$

$$\text{SST} = b\sum_{i=1}^{t}(\bar{y}_{i.} - \bar{y}_{..})^2$$
$$= 6\,[(9.409 - 7.511)^2 + (6.190 - 7.511)^2 + (7.022 - 7.511)^2$$
$$+ (6.170 - 7.511)^2 + (8.760 - 7.511)^2] = 53.624$$

$$\text{SSB} = t\sum_{j=1}^{b}(\bar{y}_{.j} - \bar{y}_{..})^2$$
$$= 5\,[(8.067 - 7.511)^2 + (4.158 - 7.511)^2 + (11.290 - 7.511)^2$$
$$+ (8.198 - 7.511)^2 + (6.280 - 7.511)^2 + (7.068 - 7.511)^2] = 140.032$$

$$\text{SSE} = \text{TSS} - \text{SST} - \text{SSB} = 219.887 - 53.624 - 140.032 = 26.231$$

$$\text{Bias} = \frac{(y_{.1} - (t-1)\hat{y}_{11})^2}{t(t-1)} = \frac{[30.37 - (5-1)9.9655]^2}{5(5-1)} = 4.5049$$

Corrected treatment SS = SST_C = SST − Bias = 53.624 − 4.5049 = 49.119

The AOV table for Example 19.1 is shown in Table 19.4.

TABLE 19.4
AOV table for testing the effects of treatments with one missing observation

Source	SS	df	MS	F	p-value
$\text{Blocks}_{\text{unadj}}$	140.032	5	28.01		
Treatments_C	49.119	4	12.28	7.96	.0008
Error	26.231	17	1.543		
Total	219.887	23			

19

The F test for a significant difference in the five method means is highly significant (p-value = .0008). The mean loss in cotton fiber was somewhat higher when using methods 1 and 5 in comparison to the other three methods.

comparisons among treatment means

Having seen an analysis of variance, we may wish to make certain **comparisons among the treatment means.** We'll run pairwise comparisons using Fisher's least significant difference. The least significant difference between the treatment with a missing observation and any other treatment mean is

LSD

$$\text{LSD}^* = t_{\alpha/2}\sqrt{\text{MSE}\left(\frac{2}{b} + \frac{t}{b(b-1)(t-1)}\right)}$$

For any pair of treatments with no missing value, the least significant difference is as before; namely,

$$\text{LSD} = t_{\alpha/2}\sqrt{\left(\frac{2\text{MSE}}{b}\right)}$$

EXAMPLE 19.2

In Example 19.1 we found that there was significant evidence of a difference in the mean loss in cotton fiber for the five methods. The researchers would like to determine which pairs of methods have differences. Run a pairwise comparison of the five methods using Fisher's LSD procedure.

Solution Example 19.1 involved a study in which the design was a randomized block design with $t = 5$ treatments and $b = 6$ blocks. There was a single missing observation. From Table 19.4, we have MSE = 1.543 with 17 degrees of freedom. Using $\alpha = .05$, the value of LSD for comparing the method with the missing observation, method 1, with the other four methods is computed as

$t_{.05/2,19}$

$$\text{LSD}^* = t_{.05/2,17}\sqrt{\text{MSE}\left(\frac{2}{b} + \frac{t}{b(b-1)(t-1)}\right)}$$

$$= 2.11\sqrt{1.543\left(\frac{2}{6} + \frac{5}{6(6-1)(5-1)}\right)} = 1.605$$

2.093 1.592

For comparing any pair not including method 1, the value of LSD is

$$\text{LSD} = t_{.05/2,17}\sqrt{\frac{2\text{MSE}}{b}} = 2.11\sqrt{\frac{2(1.543)}{6}} = 1.513$$

$t_{.05/2,19}$ 1.501 2.093

Using the two values of LSD we obtain the following results (Table 19.5) with the mean for method 1 computed using the estimated missing observation.

TABLE 19.5 Paired comparison of five methods

Pair Compared	Difference in Means	LSD	Conclusion
M1 & M2	9.409 − 6.190 = 3.219	1.605	Significant
M1 & M3	9.409 − 7.022 = 2.387	1.605	Significant
M1 & M4	9.409 − 6.170 = 3.239	1.605	Significant
M1 & M5	9.409 − 8.760 = .649	1.605	Not Significant
M2 & M3	6.190 − 7.022 = −.832	1.513	Not Significant
M2 & M4	6.190 − 6.170 = .020	1.513	Not Significant
M2 & M5	6.190 − 8.760 = −2.570	1.513	Significant
M3 & M4	7.022 − 6.170 = .852	1.513	Not Significant
M3 & M5	7.022 − 8.760 = −1.738	1.513	Significant
M4 & M5	6.170 − 8.760 = −2.590	1.513	Significant

	Route							
Vehicle	**R1**	**R2**	**R3**	**R4**	**R5**	**R6**	**R7**	**R8**
V1	B1 12.0	B2 11.2	B3 11.8	B4 10.0	B5 20.1	B6 18.7	B7 21.7	B8 30.2
V2	B2 10.1	B3 12.2	B4 12.1	B5 12.4	B6 18.4	B7 18.6	B8 22.3	B1 15.0
V3	B3 21.4	B4 24.2	B5 26.7	B6 23.3	B7 32.5	B8 34.1	B1 21.4	B2 27.7
V4	B4 15.4	B5 20.3	B6 17.5	B7 17.6	B8 25.3	B1 12.2	B2 12.4	B3 18.9
V5	B5 25.0	B6 24.4	B7 24.0	B8 26.5	B1 20.6	B2 19.6	B3 19.6	B4 27.3
V6	B6 18.9	B7 20.9	B8 25.2	B1 8.3	B2 15.6	B3 15.1	B4 17.4	B5 25.9
V7	B7 16.2	B8 18.2	B1 ***	B2 4.4	B3 10.2	B4 9.9	B5 12.7	B6 17.9
V8	B8 29.5	B9 21.3	B2 18.3	B3 16.1	B4 26.0	B5 26.4	B6 26.0	B7 35.0

B1

a. Estimate the amount of CO emissions for vehicle V7 while driving over route R3 using blend B1.

b. Analyze the data by replacing the missing value with the estimate obtained in part (a) and then perform an analysis of variance using the formulas for a Latin square design with no missing observations.

c. Is there a significant difference in the mean CO emissions for the different blends? Use $\alpha = .05$.

19.8 Refer to Exercise 19.7. Use the least significant difference criterion to identify which pairs of blends have significantly different mean CO emissions.

19.9 Refer to Exercise 19.7. Obtain the sum of squares for an AOV table by fitting complete and reduced models using a statistical software program. Compare your results with those in Exercise 19.7.

19.10 Refer to Exercise 19.7. Suppose upon examining the data logs from the study the researcher determined that the CO emission monitoring device was probably not functioning properly for the following two data values: vehicle V7 on route R4 using blend B2, y_{742}, and vehicle V6 on route R4 using blend B1, y_{641}. Reanalyze the data deleting these two values. Do your conclusions about the differences in the eight blends change?

19.11 Refer to Exercise 19.10.

a. Identify vehicle and route as fixed or random effects.

b. How would you test for a significant effect due to vehicle?

c. How would you test for a significant effect due to route?

Sci.

19.12 A horticulturist is interested in examining the yield potential of three new varieties of asparagus. She designed a study to evaluate the three new varieties relative to standard variety. There were 16 plots available on a large test field for the study, but the plots were not homogeneous in that there was a distinct sloping from north to south throughout the field. Also a soil analysis revealed a discernible nitrogen gradient, which ran from west to east across the field. Therefore, the horticulturists decided to assign the varieties V1, V2, V3, and V4, with V1 being the standard variety, to the plots in a Latin square arrangement. The values for marketable yield per plot (in kg/ha) are given in the following table. Note that there is a missing yield for variety V4 in row 4 and column 1. This was due to a problem that occurred during one of the harvesting periods.

	Sloping			
Nitrogen	**S1**	**S2**	**S3**	**S4**
N1	V3 1,045.38	V1 807.69	V2 967.36	V4 1,084.23
N2	V1 821.40	V2 992.56	V4 992.47	V3 1,029.53
N3	V2 1,004.02	V4 1,091.23	V3 1,062.01	V1 836.53
N4	V4 *	V3 1,090.97	V1 893.32	V2 1,053.97

Friedman's test, 978-982