CHAPTER 9

Association and Prediction: Linear Models with One Predictor Variable

9.1 INTRODUCTION

Motivation for the methods of this chapter is aided by the use of examples. For this reason, we first consider three data sets. These data are used to motivate the methods to follow. The data are also used to illustrate the methods used in Chapter 11. After the three examples are presented, we return to this introduction.

Example 9.1. Table 9.1 and Figure 9.1 contain data on mortality due to malignant melanoma of the skin of white males during the period 1950–1969 for each state in the United States as well as the District of Columbia. No mortality data are available for Alaska and Hawaii for this period. It is well known that the incidence of melanoma can be related to the amount of sunshine and, somewhat equivalently, the latitude of the area. The table contains the latitude as well as the longitude for each state. These numbers were obtained simply by estimating the center of the state and reading off the latitude as given in a standard atlas. Finally, the 1965 population and contiguity to an ocean are noted, where "1" indicates contiguity: the state borders one of the oceans.

In the next section we shall be particularly interested in the relationship between the melanoma mortality and the latitude of the states. These data are presented in Figure 9.1.

Definition 9.1. When two variables are collected for each data point, a plot is very useful. Such plots of the two values for each of the data points are called *scatter diagrams* or *scattergrams*.

Note several things about the scattergram of malignant melanoma rates vs. latitude. There appears to be a rough relationship. As the latitude increases, the melanoma rate decreases. Nevertheless, there is no one-to-one relationship between the values. There is considerable scatter in the picture. One problem is to decide whether or not the scatter could be due to chance or whether there is some relationship. It might be of interest to estimate the melanoma rate for various latitudes. In this case, how would we estimate the relationship? To convey the relationship to others, it would also be useful to have some simple way of summarizing the relationship. There are two aspects of the relationship that might be summarized. One is how the melanoma rate changes with latitude; it would also be useful to summarize the variability of the scattergram.

Table 9.1   Mortality Rate [per 10 Million (10^7)] of White Males Due to Malignant Melanoma of the Skin for the Period 1950–1959 by State and Some Related Variables

State	Mortality per 10,000,000	Latitude (deg)	Longitude (deg)	Population (millions, 1965)	Ocean State[a]
Alabama	219	33.0	87.0	3.46	1
Arizona	160	34.5	112.0	1.61	0
Arkansas	170	35.0	92.5	1.96	0
California	182	37.5	119.5	18.60	1
Colorado	149	39.0	105.5	1.97	0
Connecticut	159	41.8	72.8	2.83	1
Delaware	200	39.0	75.5	0.50	1
Washington, DC	177	39.0	77.0	0.76	0
Florida	197	28.0	82.0	5.80	1
Georgia	214	33.0	83.5	4.36	1
Idaho	116	44.5	114.0	0.69	0
Illinois	124	40.0	89.5	10.64	0
Indiana	128	40.2	86.2	4.88	0
Iowa	128	42.2	93.8	2.76	0
Kansas	166	38.5	98.5	2.23	0
Kentucky	147	37.8	85.0	3.18	0
Louisiana	190	31.2	91.8	3.53	1
Maine	117	45.2	69.0	0.99	1
Maryland	162	39.0	76.5	3.52	1
Massachusetts	143	42.2	71.8	5.35	1
Michigan	117	43.5	84.5	8.22	0
Minnesota	116	46.0	94.5	3.55	0
Mississippi	207	32.8	90.0	2.32	1
Missouri	131	38.5	92.0	4.50	0
Montana	109	47.0	110.5	0.71	0
Nebraska	122	41.5	99.5	1.48	0
Nevada	191	39.0	117.0	0.44	0
New Hampshire	129	43.8	71.5	0.67	1
New Jersey	159	40.2	74.5	6.77	1
New Mexico	141	35.0	106.0	1.03	0
New York	152	43.0	75.5	18.07	1
North Carolina	199	35.5	79.5	4.91	1
North Dakota	115	47.5	100.5	0.65	0
Ohio	131	40.2	82.8	10.24	0
Oklahoma	182	35.5	97.2	2.48	0
Oregon	136	44.0	120.5	1.90	1
Pennsylvania	132	40.8	77.8	11.52	0
Rhode Island	137	41.8	71.5	0.92	1
South Carolina	178	33.8	81.0	2.54	1
South Dakota	86	44.8	100.0	0.70	0
Tennessee	186	36.0	86.2	3.84	0
Texas	229	31.5	98.0	10.55	1
Utah	142	39.5	111.5	0.99	0
Vermont	153	44.0	72.5	0.40	1
Virginia	166	37.5	78.5	4.46	1
Washington	117	47.5	121.0	2.99	1
West Virginia	136	38.8	80.8	1.81	0
Wisconsin	110	44.5	90.2	4.14	0
Wyoming	134	43.0	107.5	0.34	0

Source: U.S. Department of Health, Education, and Welfare [1974].

[a] 1 = state borders on ocean.
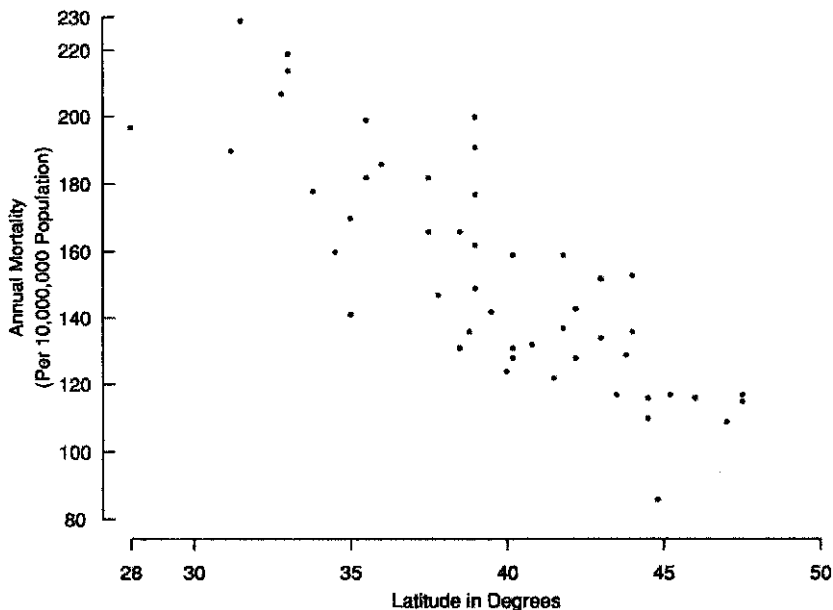
Figure 9.1 Annual mortality (per 10,000,000 population) due to malignant melanoma of the skin for white males by state and latitude of the center of the state for the period 1950–1959.

Example 9.2. To assess physical conditioning in normal subjects, it is useful to know how much energy they are capable of expending. Since the process of expending energy requires oxygen, one way to evaluate this is to look at the rate at which they use oxygen at peak physical activity. To examine the peak physical activity, tests have been designed where a person runs on a treadmill. At specified time intervals, the speed at which the treadmill moves and the grade of the treadmill both increase. The person is then run systematically to maximum physical capacity. The maximum capacity is determined by the person, who stops when unable to go further. Data from Bruce et al. [1973] are discussed.

The oxygen consumption was measured in the following way. The patient's nose was blocked off by a clip. Expired air was collected from a silicone rubber mouthpiece fitted with a very low resistance valve. The valve was connected by plastic tubes into a series of evacuated neoprene balloons. The inlet valve for each balloon was opened for 60 seconds to sample the expired air. Measurements were made of the volumes of expired air, and the oxygen content was obtained using a paramagnetic analyzer capable of measuring the oxygen. From this, the rate at which oxygen was used in mm/min was calculated. Physical conditioning, however, is relative to the size of the person involved. Smaller people need less oxygen to perform at the same speed. On the other hand, smaller people have smaller hearts, so relatively, the same level of effort may be exerted. For this reason, the maximum oxygen content is normalized by body weight; a quantity, $VO_{2\ MAX}$, is computed by looking at the volume of oxygen used per minute per kilogram of body weight. Of course, the effort expended to go further on the treadmill increases with the duration of time on the treadmill, so there should be some relationship between $VO_{2\ MAX}$ and duration on the treadmill. This relationship is presented below.

Other pertinent variables that are used in the problems and in additional chapters are recorded in Table 9.2, including the maximum heart rate during exercise, the subject's age, height, and weight. The 44 subjects listed in Table 9.2 were all healthy. They were classified as active if they usually participated at least three times per week in activities vigorous enough to raise a sweat.

Table 9.2   Exercise Data for Healthy Active Males

Case	Duration (s)	VO$_2$ MAX	Heart Rate (beats/min)	Age	Height (cm)	Weight (kg)
1	706	41.5	192	46	165	57
2	732	45.9	190	25	193	95
3	930	54.5	190	25	187	82
4	900	60.3	174	31	191	84
5	903	60.5	194	30	171	67
6	976	64.6	168	36	177	78
7	819	47.4	185	29	174	70
8	922	57.0	200	27	185	76
9	600	40.2	164	56	180	78
10	540	35.2	175	47	180	80
11	560	33.8	175	46	180	81
12	637	38.8	162	55	180	79
13	593	38.9	190	50	161	66
14	719	49.5	175	52	174	76
15	615	37.1	164	46	173	84
16	589	32.2	156	60	169	69
17	478	31.3	174	49	178	78
18	620	33.8	166	54	181	101
19	710	43.7	184	57	179	74
20	600	41.7	160	50	170	66
21	660	41.0	186	41	175	75
22	644	45.9	175	58	173	79
23	582	35.8	175	55	160	79
24	503	29.1	175	46	164	65
25	747	47.2	174	47	180	81
26	600	30.0	174	56	183	100
27	491	34.1	168	82	183	82
28	694	38.1	164	48	181	77
29	586	28.7	146	68	166	65
30	612	37.1	156	54	177	80
31	610	34.5	180	56	179	82
32	539	34.4	164	50	182	87
33	559	35.1	166	48	174	72
34	653	40.9	184	56	176	75
35	733	45.4	186	45	179	75
36	596	36.9	174	45	179	79
37	580	41.6	188	43	179	73
38	550	22.7	180	54	180	75
39	497	31.9	168	55	172	71
40	605	42.5	174	41	187	84
41	552	37.4	166	44	185	81
42	640	48.2	174	41	186	83
43	500	33.6	180	50	175	78
44	603	45.0	182	42	176	85

Source: Data from Bruce et al. [1973].

The duration of the treadmill exercise and VO$_2$ MAX data are presented in Figure 9.2. In this scattergram, we see that as the treadmill time increases, by and large, the VO$_2$ MAX increases. There is, however, some variability. The increase is not an infallible rule. There are subjects who run longer but have less oxygen consumption than someone else who has exercised for a shorter time period. Because of the expense and difficulty in collecting the expired air volumes,
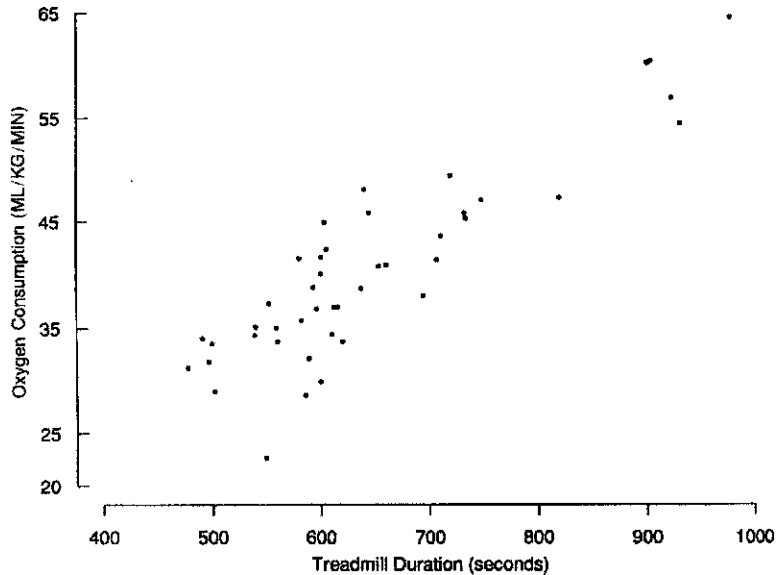
Figure 9.2   Oxygen consumption vs. treadmill duration.

it is useful to evaluate oxygen consumption and conditioning by having the subjects run on the treadmill and recording the duration. As we can see from Figure 9.2, this would not be a perfect solution to the problem. Duration would not totally determine the $VO_{2\ MAX}$ level. Nevertheless, it would give us considerable information. When we do this, how should we predict what the $VO_{2\ MAX}$ level would be from the duration? Clearly, such a predictive equation should be developed from the data at hand. When we do this, we want to characterize the accuracy of such predictions and succinctly summarize the relationship between the two variables.

Example 9.3.   Dern and Wiorkowski [1969] collected data dealing with the erythrocyte adenosine triphosphate (ATP) levels in youngest and older sons in 17 families. The purpose of the study was to determine the effect of storage of the red blood cells on the ATP level. The level is important because it determines the ability of the blood to carry energy to the cells of the body. The study found considerable variation in the ATP levels, even before storage. Some of the variation could be explained on the basis of variation by family (genetic variation). The data for the oldest and youngest sons are extracted from the more complete data set in the paper. Table 9.3 presents the data for 17 pairs of brothers along with the ages of the brothers.

Figure 9.3 is a scattergram of the values in Table 9.3. Again, there appears to be some relationship between the two values, with both brothers tending to have high or low values at the same time. Again, we would like to consider whether or not such variability might occur by chance. If chance is not the explanation, how could we summarize the pattern of variation for the pairs of numbers?

The three scattergrams have certain features in common:

1. Each scattergram refers to a situation where two quantities are associated with each experimental unit. In the first example, the melanoma rate for the state and the latitude of the state are plotted. The state is the individual unit. In the second example, for each person studied on the treadmill, $VO_{2\ MAX}$ vs. the treadmill time in seconds was plotted. In the third example, the experimental unit was the family, and the ATP values of the youngest and oldest sons were plotted.

Table 9.3   Erythrocyte Adenosine Triphosphate (ATP) Levels[a] in Youngest and Oldest Sons in 17 Families Together with Age (Before Storage)

Family	Youngest		Oldest	
	Age	ATP Level	Age	ATP Level
1	24	4.18	41	4.81
2	25	5.16	26	4.98
3	19	4.85	27	4.48
4	28	3.43	32	4.19
5	22	4.53	25	4.27
6	7	5.13	23	4.87
7	21	4.10	24	4.74
8	17	4.77	25	4.53
9	25	4.12	26	3.72
10	24	4.65	25	4.62
11	12	6.03	25	5.83
12	16	5.94	24	4.40
13	9	5.99	22	4.87
14	18	5.43	24	5.44
15	14	5.00	26	4.70
16	24	4.82	26	4.14
17	20	5.25	24	5.30

Source: Data from Dern and Wiorkowski [1969].

[a] ATP levels expressed as micromoles per gram of hemoglobin.
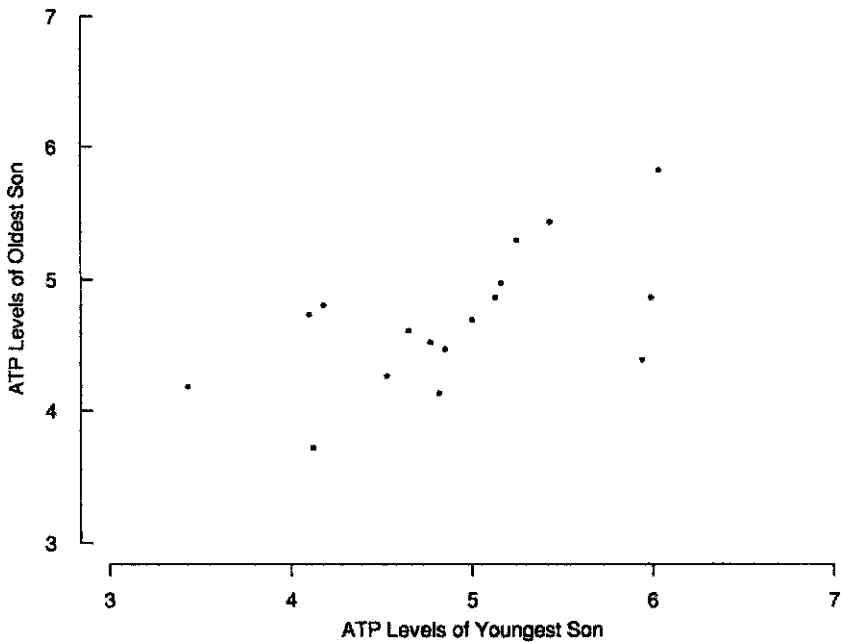
Figure 9.3   ATP levels (μmol/g of hemoglobin) of youngest and oldest sons in 17 families. (Data from Dern and Wiorkowski [1969].)

2. In each of the three diagrams, there appears to be a rough trend or association between the variables. In the melanoma rate date, as the latitude increases, the melanoma rate tends to decrease. In the treadmill data, as the duration on the treadmill increased, the VO_2 MAX also increased. In the ATP data, both brothers tended to have either a high or a low value for ATP.

3. Although increasing and decreasing trends were evident, there was not a one-to-one relationship between the two quantities. It was not true that every state with a higher latitude had a lower melanoma rate in comparison with a state at a lower latitude. It was not true that in each case when individual A ran on the treadmill a longer time than individual B that individual A had a higher VO_2 MAX value. There were some pairs of brothers for which one pair did not have the two highest values when compared to the other pair. This is in contrast to certain physical relationships. For example, if one plotted the volume of a cube as a function of the length of a side, there is the one-to-one relationship: the volume increases as the length of the side increases. In the data we are considering, there is a rough relationship, but there is still considerable variability or scatter.

4. To effectively use and summarize such scattergrams, there is a need for a method to quantitate how much of a change the trends represent. For example, if we consider two states where one has a latitude 5° south of the other, how much difference is expected in the melanoma rates? Suppose that we train a person to increase the duration of treadmill exercise by 70 seconds; how much of a change in VO_2 MAX capacity is likely to occur?

5. Suppose that we have some method of quantitating the overall relationship between the two variables in the scattergram. Since the relationship is not precisely one to one, there is a need to summarize how much of the variability the relationship explains. Another way of putting this is that we need a summary quantity which tells us how closely the two variables are related in the scattergram.

6. If we have methods of quantifying these things, we need to know whether or not any estimated relationships might occur by chance. If not, we still want to be able to quantify the uncertainty in our estimated relationships.

The remainder of this chapter deals with the issues we have just raised. In the next section we use a linear equation (a straight line) to summarize the relationship between two variables in a scattergram.

9.2 SIMPLE LINEAR REGRESSION MODEL

9.2.1 Summarizing the Data by a Linear Relationship

The three scattergrams above have a feature in common: the overall relationship is roughly linear; that is, a straight line that characterizes the relationships between the two variables could be placed through the data. In this and subsequent chapters, we look at linear relationships. A linear relationship is one expressed by a linear equation. For variables $U, V, W, \ldots$, and constants $a, b, c, \ldots$, a linear equation for Y is given by

$$Y = a + bU + cV + dW + \cdots$$

In the scattergrams for the melanoma data and the exercise data, let X denote the variable on the horizontal axis (*abscissa*) and Y be the notation for the variable on the vertical axis (*ordinate*). Let us summarize the data by fitting the straight-line equation $Y = a + bX$ to the data. In each case, let us think of the X variable as predicting a value for Y. In the first two

examples, that would mean that given the latitude of the state, we would predict a value for the melanoma rate; given the duration of the exercise test, we would predict the VO_2 MAX value for each subject.

There is terminology associated with this procedure. The variable being predicted is called the *dependent variable* or *response variable*; the variable we are using to predict is called the *independent variable*, the *predictor variable*, or the *covariate*. For a particular value, say, X_i of the predictor variable, our value predicted for Y is given by

$$\widehat{Y_i} = a + bX_i \tag{1}$$

The fit of the values predicted to the values observed (X_i, Y_i) may be summarized by the difference between the value Y_i observed and the value $\widehat{Y_i}$ predicted. This difference is called a *residual value*:

$$\text{residual value} = y_i - \widehat{y_i} = \text{value observed} - \text{value predicted} \tag{2}$$

It is reasonable to fit the line by trying to make the residual values as small as possible. The *principle of least squares* chooses a and b to minimize the sum of squares of the residual values. This is given in the following definition:

Definition 9.2.   Given data $(x_i, y_i), i = 1, 2, \ldots, n$, the *least squares fit* to the data chooses a and b to minimize

$$\sum_{i=1}^{n}(y_i - \widehat{y_i})^2$$

where $\widehat{y_i} = a + bx_i$.

The values a and b that minimize the sum of squares are described below. At this point, we introduce some notation similar to that of Section 7.3:

$$[y^2] = \sum_i (y_i - \overline{y})^2$$

$$[x^2] = \sum_i (x_i - \overline{x})^2$$

$$[xy] = \sum_i (x_i - \overline{x})(y_i - \overline{y})$$

We decided to choose values a and b so that the quantity

$$\sum_i (y_i - \widehat{y_i})^2 = \sum_i (y_i - a - bx_i)^2$$

is minimized. It can be shown that the values for a and b that minimize the quantity are given by

$$b = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2} = \frac{[xy]}{[x^2]}$$

and

$$a = \overline{y} - b\overline{x}$$

Note 9.4 gives another equivalent formula for b that emphasizes its role as a summary statistic of the slope of the X–Y relationship.

**Table 9.4   Predicted Mortality Rates by Latitude for the
Data of Table 9.1[a]**

Latitude (x)	Predicted Mortality (y)	s_1	s_2	s_3
30	209.9	19.12	6.32	20.13
35	180.0	19.12	3.85	19.50
39.5 (mean)	152.9 (mean)	19.12	2.73	19.31
40	150.1	19.12	2.74	19.31
45	120.2	19.12	4.26	19.58
50	90.3	19.12	6.83	20.30

[a]For the quantities s_2 and s_3, see Section 9.2.3.

For the melanoma data, we have the following quantities:

$$\overline{x} = 39.533, \qquad \overline{y} = 152.878$$

$$\sum_i (x_i - \overline{x})(y_i - \overline{y}) = [xy] = -6100.171$$

$$\sum_i (x_i - \overline{x})^2 = [x^2] = 1020.499$$

$$\sum_i (y_i - \overline{y})^2 = [y^2] = 53,637.265$$

The least squares slope b is

$$b = \frac{-6100.171}{1020.499} = -5.9776$$

and the least squares intercept a is

$$a = 152.878 - (-5.9776 \times 39.533) = 389.190$$

Figure 9.4 presents the melanoma data with the line of least squares fit drawn in. Because of
the method of selecting the line, the line goes through the data, of course. The least squares
line always has the property that it goes through the point in the scattergram corresponding
to the sample mean of the two variables. The sample means of the variables are located by
the intersection of dotted lines. Further, the point for Tennessee is detailed in the box in the
lower left-hand corner. The value predicted from the equation was 174, whereas the actual
melanoma rate for this state was 186. Thus, the residual value is the difference, 12. We see
that the value predicted, 174, is closer to the value observed than to the overall Y mean, which
is 152.9.

For the melanoma data, the line of least squares fit is $Y = 389.19 - 5.9776X$. For each
state's observed mortality rate, there is then a predicted mortality rate based on knowledge of
the latitude. Some predicted values are listed in Table 9.4. The farther north the state, the lower
the mortality due to malignant melanoma; but now we have quantified the change.

Note that the predicted mortality at the mean latitude (39.5°) is exactly the mean value of
the mortalities *observed*; as noted above, the regression line goes through the point $(\overline{x}, \overline{y})$.

9.2.2   Linear Regression Models

With the line of least squares fit, we shall associate a mathematical model. This *linear regression
model* takes the predictor or covariate observation as being fixed. Even if it is sampled at random,
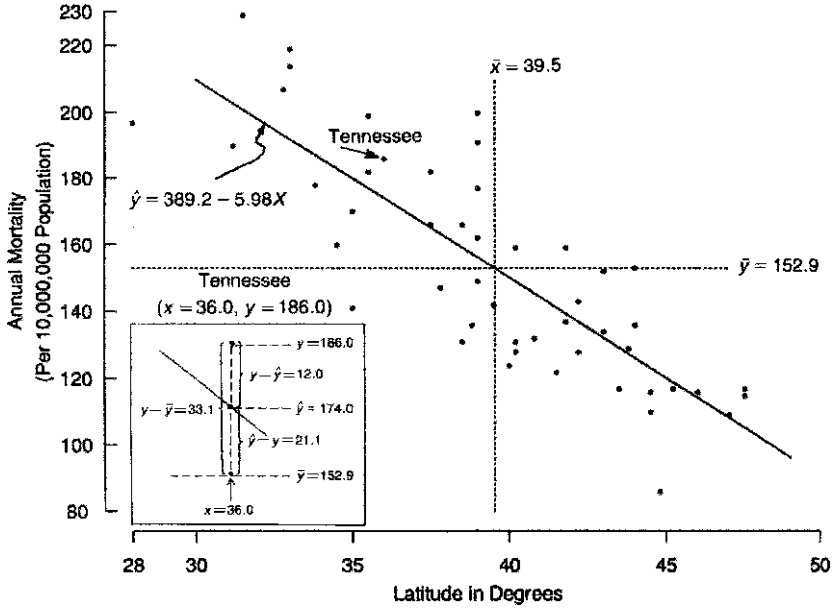
Figure 9.4   Annual mortality (per 10,000,000 population) due to malignant melanoma of the skin for white males by state and latitude of the center of the state for the period 1950–1959 (least squares regression line is given).

the analysis is conditional upon knowing the value of X. In the first example above, the latitude of each state is fixed. In the second example, the healthy people may be considered to be a representative—although not random—sample of a larger population; in this case, the duration may be considered a random quantity. In the linear regression analysis of this chapter, we know X and are interested in predicting the value of Y. The regression model assumes that for a fixed value of X, the expected value of Y is some function. In addition to this expected value, a random error term is added. It is assumed that the error has a mean value of zero. We shall restrict ourselves to situations where the expected value of Y for known X is a linear function. Thus, our linear regression model is the following:

$$\text{expected value of } Y \text{ knowing } X = E(Y|X) = \alpha + \beta X$$

$$Y = \alpha + \beta X + e, \qquad \text{where } e \text{ (error) has } E(e) = 0$$

The parameters α and β are population parameters. Given a sample of observations, the estimates a and b that we found above are estimates of the population parameters. In the mortality rates of the states, the random variability arises both because of the randomness of the rates in a given year and random factors associated with the state, other than latitude. These factors make the observations during a particular time period reasonably modeled as a random quantity. For the exercise test data, we may consider the normal subjects tested as a random sample from a population of active normal males who might have been tested.

Definition 9.3.   The line $E(Y|X) = \alpha + \beta X$ is called the *population regression line*. Here, $E(Y|X)$ is the expected value of Y at X (assumed known). The coefficients α and β are called *population regression coefficients*. The line $Y = a + bX$ is called the *estimated regression line*,

and a and b are called *estimated regression coefficients*. The term *estimated* is often dropped, and *regression line* and *regression coefficients* are used for these estimated quantities.

For each X, $E(Y|X)$ is the mean of a population of observations. On the left of Figure is shown a linear regression situation; on the right, the regression $E(Y|X)$ is not linear.

To simplify statistical inference, another assumption is often added: that the error term is normally distributed with mean zero and variance σ_1^2. As we saw with the t-test, the assumption of normality is important for testing and confidence interval estimation only in fairly small samples. In larger samples the central limit theorem replaces the need for distributional assumptions. Note that the variance of the error term is *not* the variance of the Y variable. It is the variance of the Y variable *when* the value of the X variable is known.

Given data, the variance σ_1^2 is estimated by the quantity $s_{y \cdot x}^2$, where this quantity is defined as

$$s_{y \cdot x}^2 = \sum \frac{(Y_i - \widehat{Y}_i)^2}{n - 2}$$

Recall that the usual sample variance was divided by $n - 1$. The $n - 2$ occurs because two parameters, α and β, are estimated in fitting the data rather than one parameter, the sample mean, that was estimated before.

9.2.3 Inference

We have the model

$$Y = \alpha + \beta X + e, \qquad \text{where } e \sim N(0, \sigma_1^2)$$

On the basis of n pairs of observations we presented estimates a and b of α and β, respectively. To test hypotheses regarding α and β, we need to assume the normality of the term e.

The left panel of Figure 9.5 shows a situation where these assumptions are satisfied. Note that:

1. $E(Y|X)$ is linear.
2. For each X, the normal Y-distribution has the same variance.
3. For each X, the Y-distribution is normal (less important as the sample size is large).

The right panel of Figure 9.5 shows a situation where all these assumptions don't hold.

1. $E(Y|X)$ is not a straight line; it curves.
2. The variance of Y increases as X increases.
3. The distribution becomes more highly skewed as X increases.

It can be shown, under the correct normal model or in large samples, that

$$b \sim N\left(\beta, \frac{\sigma_1^2}{[x^2]}\right) \quad \text{and} \quad a \sim N\left(\alpha, \sigma_1^2\left[\frac{1}{n} + \frac{\overline{x}^2}{[x^2]}\right]\right)$$

Recall that σ_1^2 is estimated by $s_{y \cdot x}^2 = \sum (Y_i - \widehat{Y}_i)^2/(n - 2)$. Note that the divisor is $n - 2$: the number of degrees of freedom. The reason, as just mentioned, is that now *two* parameters are estimated: α and β. Given these facts, we can now either construct confidence intervals or tests of hypotheses after constructing appropriate pivotal variables:

$$\frac{b - \beta}{\sigma_1/\sqrt{[x^2]}} \sim N(0, 1), \qquad \frac{b - \beta}{s_{y \cdot x}/\sqrt{[x^2]}} \sim t_{n-2}$$

and similar terms involving the intercept a are discussed below.
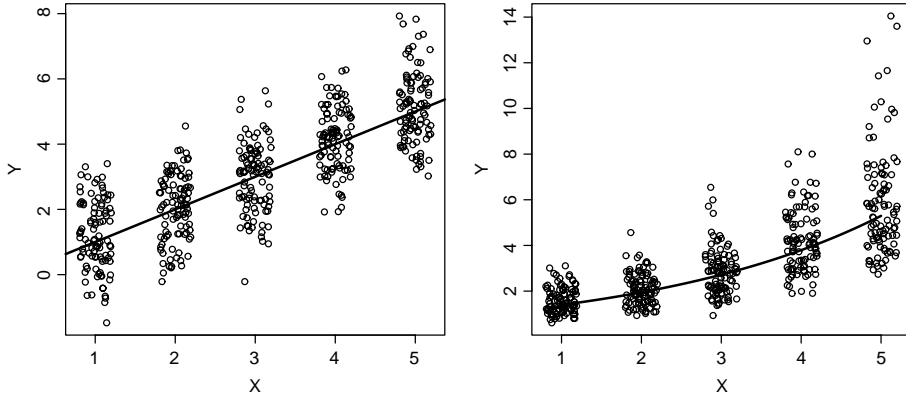
Figure 9.5   Linear regression assumptions and violations. On the left, the expected values of Y for given X values fall on a straight line. The variation about the line has the same variance at each X. On the right, the expected values fall on a curve, not a straight line. The distribution of Y is different for different X values, with variance and skewness increasing with X.

Returning to Example 9.1, the melanoma data by state, the following quantities are known or can be calculated:

$$a = 389.190, \qquad s_{y \cdot x}^2 = \sum_i \frac{(Y_i - \widehat{Y}_i)^2}{n - 2} = \frac{17{,}173.1}{47} = 365.3844$$

$$b = -5.9776, \qquad [x^2] = 1020.499, \qquad s_{y.x} = 19.1150$$

On the assumption that there is no relationship between latitude and mortality, that is, $\beta = 0$, the variable b has mean zero. A t-test yields

$$t_{47} \doteq \frac{-5.9776}{19.1150 / \sqrt{1020.499}} \doteq \frac{-5.9776}{0.59837} \doteq -9.99$$

From Table A.4, the critical value for a t-variable with 47 degrees of freedom, at the 0.0001 level (two-tailed) is approximately 4.25; hence, the hypothesis is rejected and we conclude that there is a relationship between latitude and mortality; the mortality increases about 6.0 persons per 10,000,000 for every degree farther south. This, of course, comes from the value of $b = -5.9776 \doteq -6.0$. Similarly, a 95% confidence interval for β can be constructed using the t-value of 2.01, and the standard error of the slope, $0.59837 = s_{y \cdot x} / \sqrt{[x^2]}$.

A 95% confidence interval is $-5.9776 \pm (2.01 \times 0.59837)$, producing lower and upper limits of -7.18 and -4.77, respectively. Again, the confidence interval does not include zero, and the same conclusion is reached as in the case of the hypothesis test.

The inference has been concerned with the slope β and intercept α up to now. We now want to consider two additional situations:

1. Inference about population means, $\alpha + \beta X$, for a fixed value of X
2. Inference about a future observation at a fixed value of X

To distinguish between the two cases, let $\widehat{\mu}_x$ and $\widehat{y}_x$ be the predicted mean and a new random single observation at the point x, respectively. It is important to note that for inference about a future observation the normality assumption is critical even in large samples. This is in contrast to inference about the predicted mean or about a and b, where normal distributions are required only in small samples and the central limit theorem substitutes in large samples.

First, then, inference about the population mean at a fixed X value: It is natural to estimate $\alpha + \beta X$ by $a + bx$; the predicted value of Y at the value of $X = x$. Rewrite this quantity as

$$\widehat{\mu}_x = \overline{y} + b(x - \overline{x})$$

It can be shown that $\overline{y}$ and b are statistically independent so that the variance of the quantity is

$$\text{var}[\overline{y} + b(x - \overline{x})] = \text{var}(\overline{y}) + (x - \overline{x})^2 \, \text{var}(b)$$

$$= \frac{\sigma_1^2}{n} + (x - \overline{x})^2 \frac{\sigma_1^2}{[x^2]}$$

$$= \sigma_1^2 \left[\frac{1}{n} + \frac{(x - \overline{x})^2}{[x^2]} \right] = \sigma_2^2, \qquad \text{say}$$

Tests and confidence intervals for $E(Y|X)$ at a fixed value of x may be based on the t-distribution.

The quantity σ_2^2 reduces to the variance for the intercept, a, at $X = 0$. It is useful to study this quantity carefully; there are important implications for design (see Note 9.3). The variance, σ_2^2, is not constant but depends on the value of x. The more x differs from $\overline{x}$, the greater the contribution of $(x - \overline{x})^2/[x^2]$ to the variance of $a + bx$. The contribution is zero at $x = \overline{x}$. At $x = \overline{x}$, $y = \overline{y}$ the slope is not used. Regardless of the slope the line goes through mean point $(\overline{X}, \overline{Y})$. Consider Example 9.1 again. We need the following information:

$$s_{y \cdot x} = 19.1150$$

$$n = 49$$

$$\overline{x} = 39.533$$

$$[x^2] = 1020.499$$

Let

$$s_2^2 = s_{y \cdot x}^2 \left[\frac{1}{n} + \frac{(x - \overline{x})^2}{[x^2]} \right]$$

That is, s_2^2 estimates σ_2^2. Values of s_2 as related to latitude are given in Table 9.4. Confidence interval bands for the mean, $\alpha + \beta X$ (at the 95% level), are given in Figure 9.6 by the narrower bands. The curvature is slight due to the large value of $[x^2]$ and the relatively narrow range of prediction.

We now turn to the second problem: predicting a future observation on the basis of the observed data. The variance is given by

$$s_3^2 = s_{y \cdot x}^2 \left[1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{[x^2]} \right]$$

This is reasonable in view of the following argument: At the point $\alpha + \beta X$ an observation has variance σ_2^2 (estimated by $s_{y \cdot x}^2$). In addition, there is uncertainty in the true value $\alpha + \beta X$. This adds variability to the estimate. A future observation is assumed to be independent of past observations. Hence the variance can be added and the quantity s_3^2 results when σ_1^2 is estimated by $s_{y \cdot x}^2$. Confidence interval bands for future observations (95% level) are represented by outer lines in Figure 9.6. This band means that we are 95% certain that the next observation at the fixed point x will be within the given bands. Note that the curvature is not nearly as marked.
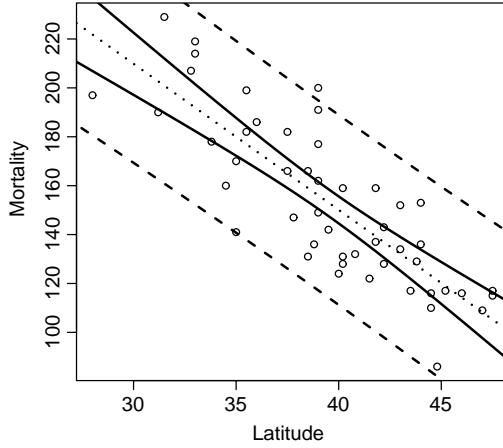
Figure 9.6 Data of Figure 9.1: 95% confidence bands for population means (solid) and 95% confidence bands for a future observation (dashed).

9.2.4 Analysis of Variance

Consider Example 9.1 and the data for Tennessee, as graphed in Figure 9.4. The basic data for this state are (omitting subscripts)

$$y = 186.0 = \text{observed mortality}$$

$$x = 36.0 = \text{latitude of center of state}$$

$$\widehat{y} = 174.0 = \text{predicted mortality using latitude of 36.0}$$

$$\overline{y} = 152.9 = \text{average mortality for United States}$$

Partition the data as follows:

$$(y - \overline{y}) = (\widehat{y} - \overline{y}) + (y - \widehat{y})$$

$$\text{total variation} = \text{attributable to regression} + \text{residual from regression}$$

$$186.0 - 152.9 = (174.0 - 152.9) \qquad + (186.0 - 174.0)$$

$$33.1 = 21.1 \qquad\qquad + 12.0$$

Note that the quantity

$$\widehat{y} - \overline{y} = a + bx - \overline{y}$$

$$= \overline{y} - b\overline{x} + bx - \overline{y}$$

$$= b(x - \overline{x})$$

The quantity is zero if $b = 0$, that is, if there is no regression relationship between Y and X. In addition, it is zero if prediction is made at the point $x = \overline{x}$.

These quantities can be calculated for each state, as indicated in abbreviated form in Table 9.5. The sums of squares of these quantities are given at the bottom of the table. The remarkable fact is that

$$\sum(y_i - \overline{y})^2 = \sum(\widehat{y}_i - \overline{y})^2 + \sum(y_i - \widehat{y}_i)^2$$

$$53{,}637.3 = 36{,}464.2 \quad + 17{,}173.1$$

Table 9.5 Deviations from Mean and Regression Based on Data of Table 9.1

				Variation				
Case	State	Observed Mortality (y)	Latitude (x)	Predicted Mortality[a]	Total $y - \overline{y}$	=	Regression $\widehat{y} - \overline{y}$	+ Residual $y - \widehat{y}$
1	Alabama	219.0	33.0	191.9	66.1	=	39.0	+ 27.1
2	Arizona	160.0	34.5	183.0	7.1	=	30.1	+ −23.0
⋮	⋮		⋮				⋮	
41	Tennessee	186.0	36.0	174.0	33.1	=	21.1	+ 12.0
⋮	⋮		⋮				⋮	
48	Wisconsin	110.0	44.5	123.2	−42.9	=	−29.7	+ −13.2
49	Wyoming	134.0	43.0	132.2	−18.9	=	−20.7	+ 1.8
Total					0	=	0	+ 0
Mean		152.9	39.5	152.9	0	=	0	+ 0
Sum of squares					53,637.3	=	36,464.2	+ 17,173.1

[a]Predicted mortality based on regression line $y = 389.19 - 5.9776x$, where x is the latitude at the center of the state.

that is, the total variation as measured by $\sum(y_i - \overline{y})^2$ has been partitioned additively into a part attributable to regression and the residual from regression. The quantity $\sum(\widehat{y}_i - \overline{y})^2 = \sum b^2(x_i - \overline{x})^2 = b^2[x^2]$. (But since $b = [xy]/[x^2]$, this becomes $\sum(\widehat{y}_i - \overline{y})^2 = [xy]^2/[x^2]$.) Associated with each sum of squares is a degree of freedom (d.f.) which can also be partitioned as follows:

$$\text{total variation} = \text{attributable to regression} + \text{residual variation}$$

$$\text{d.f.} = n - 1 = 1 + n - 2$$

$$49 = 1 + 48$$

The total variation has $n - 1$ d.f., not n, since we adjusted Y about the mean $\overline{Y}$. These quantities are commonly entered into an analysis of variance table as follows:

Source of Variation	d.f.	SS	MS	F-Ratio
Regression	1	36,464.2	36,464.2	99.80
Residual	47	17,173.1	365.384	
Total	48	53,637.3		

The quantity 365.384 is precisely $s_{y \cdot x}^2$. The F-ratio is discussed below. The mean square is the sum of squares divided by the degrees of freedom. The analysis of variance table of any set of n pairs of observations $(x_i, y_i), i = 1, \ldots, n$, is

Source of Variation	d.f.	SS	MS	F-Ratio
Regression	1	$[xy]^2/[x^2]$	$[xy]^2/[x^2]$	$\dfrac{[xy]^2/[x^2]}{s_{y \cdot x}^2}$
Residual	$n - 2$	By subtraction	$s_{y \cdot x}^2$	
Total	$n - 1$	$[y^2]$		

Several points should be noted about this table and the regression procedure:

1. Only five quantities need to be calculated from the raw data to completely determine the regression line and sums of squares: $\sum x_i, \sum y_i, \sum x_i^2, \sum y_i^2,$ and $\sum x_i y_i$. From these quantities one can calculate

$$[x^2] = \sum(x_i - \overline{x})^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

$$[y^2] = \sum(y_i - \overline{y})^2 = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}$$

$$[xy] = \sum(x_i - \overline{x})(y_i - \overline{y}) = \sum x_i y_i - \frac{\sum y_i \sum x_i}{n}.$$

2. The greater the slope, the greater the SS due to regression. That is,

$$\text{SS(regression)} = b^2 \sum(x_i - \overline{x})^2 = \frac{[xy]^2}{[x^2]}$$

If the slope is "negligible," SS(regression) will tend to be "small."

3. The proportion of the total variation attributable to regression is usually denoted by r^2; that is,

$$r^2 = \frac{\text{variation attributable to regression}}{\text{total variation}}$$

$$= \frac{[xy]^2/[x^2]}{[y^2]}$$

$$= \frac{[xy]^2}{[x^2][y^2]}$$

It is clear that $0 \leq r^2 \leq 1$ (why?). If $b = 0$, then $[xy]^2/[x^2] = 0$ and the variation attributable to regression is zero. If $[xy]^2/[x^2]$ is equal to $[y^2]$, all of the variation can be attributed to regression; to be more precise, to *linear* regression; that is, all the observations fall on the line $a + bx$. Thus, r^2 measures the degree of *linear* relationship between X and Y. The *correlation coefficient*, r, is studied in Section 9.3. For the data in Table 9.4,

$$r^2 = \frac{36{,}464.2}{53{,}637.3} = 0.67983$$

That is, approximately 68% of the variation in mortality can be attributed to variation in latitude. Equivalently, the variation in mortality can be reduced 68% knowing the latitude.

4. Now consider the ratio

$$F = \frac{[xy]^2/[x^2]}{s_{y \cdot x}^2}$$

Under the assumption of the model [i.e., $y \sim N(\alpha + \beta X, \sigma_1^2)$], the ratio F tends to be near 1 if $\beta = 0$ and tends to be larger than 1 if $\beta \neq 0$ (either positively or negatively). F has the F-distribution, as introduced in Chapter 5. In the example $F_{1,47} = 99.80$, the critical value at the 0.05 level is $F_{1,47} = 4.03$ (by interpolation). The critical value at the 0.001 level is $F_{1,47} = 12.4$ (by interpolation). Hence, we reject the hypotheses that $\beta = 0$. We tested the significance of the slope using a t-test given the value

$$t_{47} = -9.9898$$

The F-value we obtained was

$$F_{1,47} = 99.80$$

In fact,

$$(-9.9898)^2 = 99.80$$

That is,

$$t_{47}^2 = F_{1,47}$$

Recall that

$$t_v^2 = F_{1,v}$$

Thus, the t-test and the F-test for the significance of the slope are equivalent.

9.2.5 Appropriateness of the Model

In Chapter 5 we considered the appropriateness of the model $y \sim N(\mu, \sigma^2)$ for a set of data and discussed briefly some ways of verifying the appropriateness of this model. In this section we have the model

$$y \sim N(\alpha + \beta X, \sigma_1^2)$$

and want to consider its validity. At least three questions can be asked:

1. Is the relationship between Y and X linear?
2. The variance σ_1^2 is assumed to be constant for all values of X (homogeneity of variable). Is this so?
3. Does the normal model hold?

Two very simple graphical procedures, both utilizing the residuals from regression $y_i - \widehat{y_i}$, can be used to verify the assumptions above. Also, one computation on the residuals is useful. The two graphical procedures are considered first.

To Check for:	Graphical Procedure
1. Linearity of regression and homogeneity of variance	Plot $(y_i - \widehat{y_i})$ against $\widehat{y_i}$, $i = 1, \ldots, n$
2. Normality	Normal probability plot of $y_i - \widehat{y_i}, i = 1, \ldots, n$

We illustrate these with data created by Anscombe [1973]. As we noted above, just five summaries of the data specify everything about the linear regression model. Anscombe created four data sets in which these five summaries, and thus the fitted model, were identical, but where the data were very different. Only one of these sets of data is appropriate for linear regression.

Linearity of Regression and Homogeneity of Variance

Given only one predictor variable, X, the graph of Y vs. X will suggest nonlinearity or hetero-geneity of variance, see the top row of regression patterns in Figure 9.7. But if there is more than one predictor variable, as in Chapter 11, the simple two-dimensional graph is not possible. But there is a way of detecting such patterns by considering residual plots $y - \widehat{y}$ against a variety of variables. A common practice is to plot $y - \widehat{y}$ against $\widehat{y}$; this graph is usually referred to as a *residual plot*. The advantage is, of course, that no matter how many predictor variables are used,
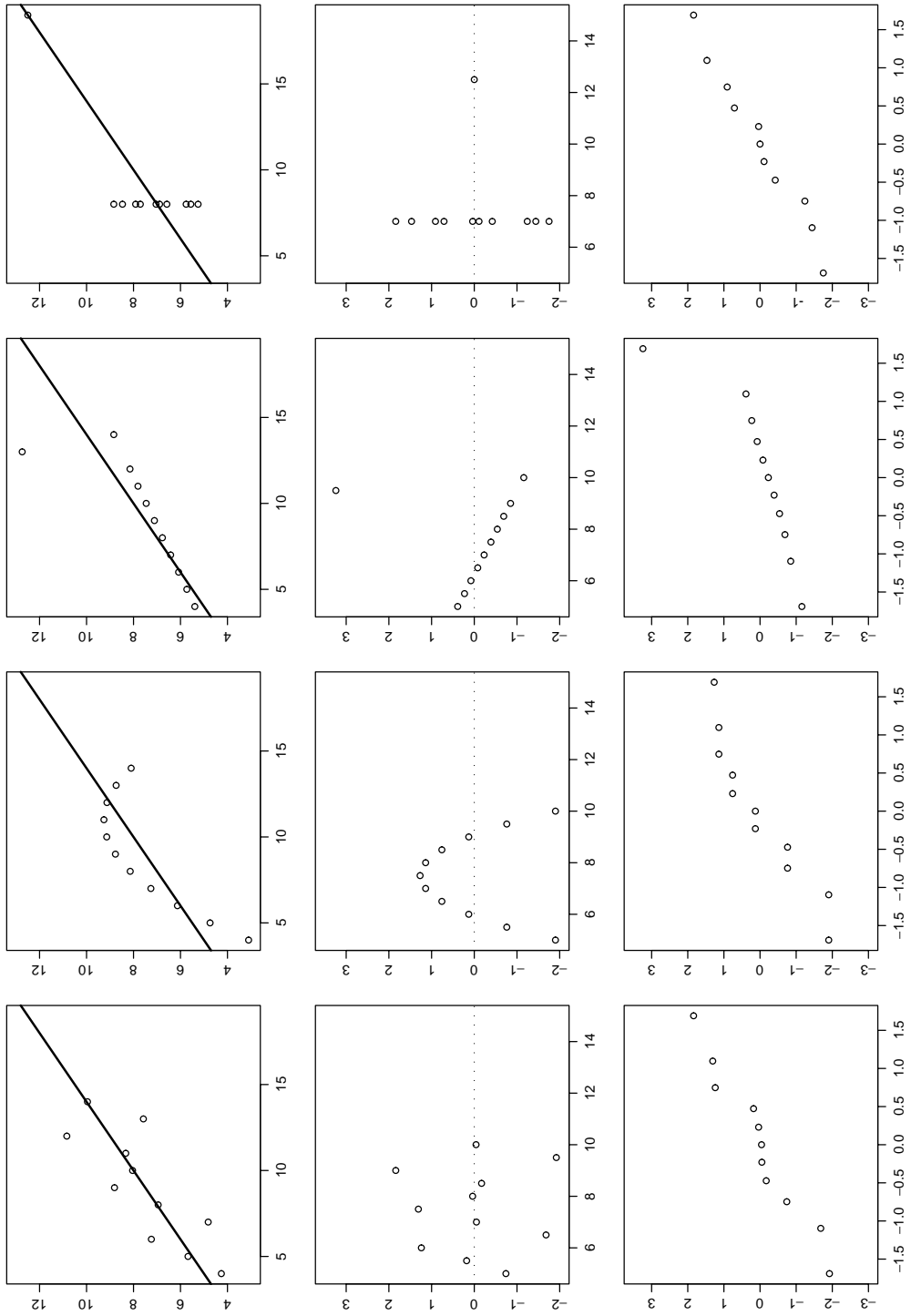
Figure 9.7 Regression patterns (Y vs. X), residuals patterns ($y - \widehat{y}$ vs. $\widehat{y}$) and normal probability plot of residuals ($y - \widehat{y}$).

it is always possible to plot $y - \widehat{y}$. The second row of graphs in Figure 9.7 indicate the residual patterns associated with the regression patterns of the top row. Pattern 1 indicates a reasonable linear trend, pattern 2 shows a very strong pattern in the residuals. Pattern 3 has a single very large residual, and in pattern 4 it is the distribution of X rather than Y that is suspicious.

Before turning to the questions of normality of the data, consider the same kind of analysis carried out on the melanoma data. The residuals are plotted in the left panel of Figure 9.8. There is no evidence that there is nonlinearity or heterogeneity of variance.

Normality

One way of detecting gross deviations from normality is to graph the residuals from regression against the expected quantiles of a normal distribution as introduced in Chapter 4. The last row of patterns in Figure 9.6 are the normal probability plots of the deviations from linear regression. The last row in Figure 9.6 indicates that a normal probability plot indicates outliers clearly but is not useful in detecting heterogeneity of variance or curvilinearity.

Of particular concern are points not fit closely by the data. The upper right and lower left points often tail in toward the center in least squares plot. Points on the top far to the right and on the bottom far to the left (as in pattern 2) are of particular concern.

The normal probability plot associated with the residuals of the melanoma are plotted in the right panel of Figure 9.8. There is no evidence against the normality assumption.

9.2.6 Two-Sample t-Test as a Regression Problem

In this section we show the usefulness of the linear model approach by illustrating how the two sample t-test can be considered a special kind of linear model. For an example, we again return to the data on mortality rates due to melanoma contained in Table 9.1. This time we consider the rates in relationship to contiguity to an ocean; there are two groups of states: those that border on an ocean and those that do not. The question is whether the average mortality rate for the first group differs from that of the second group. The t-test and analysis are contained in Table 9.6.

The mean difference, $\overline{y}_1 - \overline{y}_2 = 31.486$, has a standard error of 8.5468 so that the calculated t-value is $t = 3.684$ with 47 degrees of freedom, which exceeds the largest value in the t-table at 40 or 60 degrees of freedom and consequently, $p < 0.001$. The conclusion then is that the mortality rate due to malignant melanoma is appreciably higher in states contiguous to an ocean as compared to "inland" states, the difference being approximately 31 deaths per 10^7 population per year.
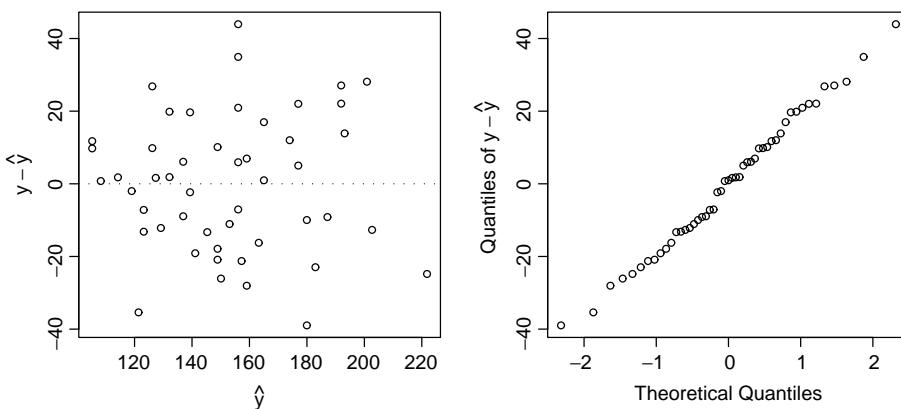
Figure 9.8 Melanoma data (left) residuals $(y - \widehat{y})$ from regression lines $Y = 389.19 - 589.8X$ plotted against $\widehat{y}$ and (right) normal quantile plot of residuals, $y - \widehat{y}$.

Table 9.6   Comparison by Two-Sample t-Test of Mortality Rates Due to Melanoma (Y) by Contiguity to Ocean

Contiguity to ocean	No $= 0$	Yes $= 1$
Number of states	$n_1 = 27$	$n_2 = 22$
Mean mortality	$\bar{y}_1 = 138.741$	$\bar{y}_2 = 170.227$
Variance[a]	$s_1^2 = 697.97$	$s_2^2 = 1117.70$
Pooled variance	$s_p^2 = 885.51$	
Standard error of difference	$s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}} = 8.5468$	
Mean difference	$\bar{y}_2 - \bar{y}_1 = 31.487$	
t-Value	$t = 3.684$	
Degrees of freedom	d.f. $= 47$	
p-Value	$p < 0.001$	

[a]Subscripts on variances denote group membership in this table.

Now consider the following (equivalent) regression problem. Let Y be the mortality rate and X the predictor variable; "$X = $ contiguity to ocean" and X takes on only two values, 0, 1. (For simplicity, we again label all the variables and parameters, Y, X, α, β, and σ_1^2, but except for Y, they obviously are different from the way they were defined in earlier sections.) The model is

$$Y \sim N(\alpha + \beta X, \sigma_1^2)$$

The data are graphed in Figure 9.9. The calculations for the regression line are as follows:

$$n = 49, \qquad b = \frac{[xy]}{[x^2]} = 31.487$$

$$[y^2] = 53637.265, \qquad a = 138.741$$

$$[xy] = 381.6939, \qquad \text{Regression line}$$

$$[x^2] = 12.12245, \qquad Y = 138.741 + 31.487X$$

$$\bar{y} = 152.8776, \qquad (n-2)s_{y \cdot x}^2 = [y^2] - \frac{[xy]^2}{[x^2]}$$

$$\bar{x} = 0.44898, \qquad = 41{,}619.0488$$

$$s_{y \cdot x}^2 = 885.51$$

The similarity to the t-test becomes obvious, the intercept $a = 138.741$ is precisely the mean mortality for the "inland" states. The "slope," $b = 31.487$, is the mean difference between the two groups of states, and $s_{y \cdot x}^2$, the residual variance, is the pooled variance. The t-test for the slope is equivalent to the t-test for the difference in the two means.

$$\text{variance of slope} = s_b^2 = \frac{s_{y \cdot x}^2}{[x^2]}$$

$$= \frac{885.51}{12.12245}$$

$$= 73.0471$$
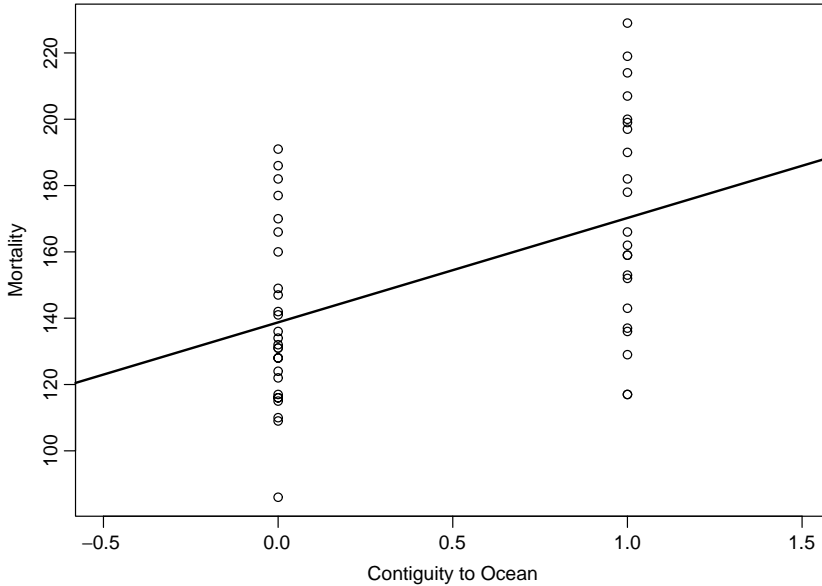
Figure 9.9 Melanoma data: regression of mortality rate on contiguity to ocean, coded 0 if not contiguous to ocean, 1 if contiguous to ocean.

$$s_b = 8.5468$$
$$t = \frac{31.487}{8.5468}$$
$$= 3.684$$

The t-test for the slope has 47 degrees of freedom, as does the two-sample t-test. Note also that s_b is the standard error of the differences in the two-sample t-test.

Finally, the regression analysis can be put into analysis of variance form as displayed in Table 9.7:

$$SS(\text{regression}) = \frac{[xy]^2}{[x^2]}$$
$$= \frac{(381.6939)^2}{12.12245}$$
$$= 12,018.22$$

$$SS(\text{residual}) = [y^2] - \frac{[xy]^2}{[x^2]}$$
$$= 53,637.26 - 12,018.22$$
$$= 41,619.04$$

We note that the proportion of variation in mortality rates attributable to "contiguity to ocean" is

$$r^2 = \frac{[xy]^2/[x^2]}{[y^2]}$$
$$= \frac{12,018.22}{53,637.06}$$
$$= 0.2241$$

Table 9.7   Regression Analysis of Mortality and Contiguity to Ocean

Source of Variation	d.f.	SS	MS	F-Ratio
Regression	1	12,018.22	12,018.22	13.57[a]
Residual	47	41,619.04	885.51	
Total	48	53,637.26		

[a]Significant at the 0.001 level.

Approximately 22% of the variation in mortality can be attributed to the predictor variable: "contiguity to ocean."

In Chapter 11 we deal with the relationships among the three variables: mortality, latitude, and contiguity to an ocean. The predictor variable "contiguity to ocean," which takes on only two values, 0 and 1 in this case, is called a *dummy variable* or *indicator variable*. In Chapter 11 more use is made of such variables.

9.3   CORRELATION AND COVARIANCE

In Section 9.2 the method of least squares was used to find a line for predicting one variable from the other. The response variable Y, or dependent variable Y, was random for given X. Even if X and Y were jointly distributed so that X was a random variable, the model only had assumptions about the distribution of Y given the value of X. There are cases, however, where both variables vary jointly, and there is a considerable amount of symmetry. In particular, there does not seem to be a reason to predict one variable from the other. Example 9.3 is of that type. As another example, we may want to characterize the length and weight relationship of newborn infants. The basic sampling unit is an infant, and two measurements are made, both of which vary. There is a certain symmetry in this situation: There is no "causal direction"—length does not cause weight, or vice versa. Both variables vary together in some way and are probably related to each other through several other underlying variables which determine (cause) length and weight. In this section we provide a quantitative measure of the strength of the relationship between the two variables and discuss some of the properties of this measure. The measure (the correlation coefficient) is a measure of the strength of the linear relationship between two variables.

9.3.1   Correlation and Covariance

We would like to develop a measure (preferable one number) that summarizes the strength of any linear relationship between two variables X and Y. Consider Example 9.2, the exercise test data. The X variable is measured in seconds and the Y variable is measured in milliliters per minute per kilogram. When totally different units are used on the two axes, one can change the units for one of the variables, and the picture seems to change. For example, if we went from seconds to minutes where 1 minute was graphed over the interval of 1 second in Figure 9.2, the data of Figure 9.2 would go almost straight up in the air. Whatever measure we use should not depend on the choice of units for the two variables. We already have one technique of adjusting for or removing the units involved: to standardize the variables. We have done this for the t-test, and we often had to do it for the construction of test statistics in earlier chapters. Further, since we are just concerned with how closely the family of points is related, if we shift our picture (i.e., change the means of the X and Y variables), the strength of the relationship between the two variables should not change. For that reason, we subtract the mean of each variable, so that the pictures will be centered about zero. In order that we have a solution that does not depend
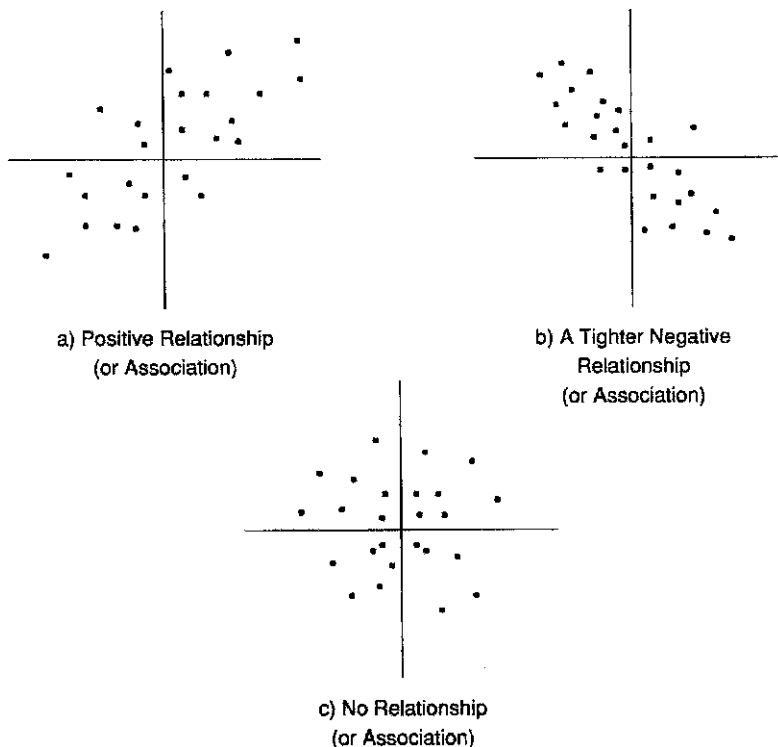
Figure 9.10 Scatter diagrams for the standardized variables.

on units, we standardize each variable by dividing by the standard deviation. Thus, we are now working with two new variables, say U and V, which are related to X and Y as follows:

$$U_i = \frac{X_i - \overline{X}}{s_x}, \qquad V_i = \frac{Y_i - \overline{Y}}{s_y}$$

where

$$s_x^2 = \sum \frac{(X_i - \overline{X})^2}{n - 1} \quad \text{and} \quad s_y^2 = \sum \frac{(Y_i - \overline{Y})^2}{n - 1}$$

Let us consider how the variables U_i and V_i vary together. In Figure 9.10 we see three possible types of association. Part (a) presents a positive relationship, or association between, the variables. As one increases, the other tends to increase. Part (b) represents a tighter, negative relationship. As one decreases, the other tends to increase, and vice versa. By the word *tighter*, we mean that the variability about a fitted regression line would not be as large. Part (c) represents little or no association, with a somewhat circular distribution of points.

One mathematical function that would capture these aspects of the data results from multiplying U_i and V_i. If the variables tend to be positive or negative together, the product will always be positive. If we add up those multiples, we would get a positive number. On the other hand, if one variable tends to be negative when the other is positive, and vice versa, when we multiply the U_i and V_i together, the product will be negative; when we add them, we will get a negative number of substantial absolute value.

On the other hand, if there is no relationship between U and V, when we multiply them, half the time the product will be positive and half the time the product will be negative; if we

sum them, the positive and negative terms will tend to cancel out and we will get something close to zero. Thus, adding the products of the standardized variables seems to be a reasonable method of characterizing the association between the variables. This gives us our definition of the correlation coefficient.

Definition 9.4. The *sample Pearson product moment correlation coefficient*, denoted by r, or r_{XY}, is defined to be

$$r = \frac{[xy]}{\sqrt{[x^2][y^2]}} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2 \sum (y_i - \overline{y})^2}} = \frac{1}{n-1} \sum u_i v_i$$

This quantity is usually called the *correlation coefficient*.

Note that the denominator looks like the product of the sample standard deviations of X and Y except for a factor of $n - 1$. If we define the sample covariance by the following equation, we could define the correlation coefficient according to the second alternative definition.

Definition 9.5. The *sample covariance*, s_{xy}, is defined by

$$s_{xy} = \sum_i \frac{(x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

Alternative Definition 9.4. The *sample Pearson product moment correlation coefficient* is defined by

$$r = \frac{[xy]}{\sqrt{[x^2][y^2]}} = \frac{s_{xy}}{s_x s_y}$$

The prefix co- is a prefix meaning "with," "together," and "in association," occurring in words derived from Latin: thus, the co-talks about the two variables varying together or in association. The term *covariance* has the same meaning as the variance of one variable: how spread out or variable things are. It is hard to interpret the value of the covariance alone because it is composed of two parts; the variability of the individual variables and their linear association. A small covariance can occur because X and/or Y has small variability. It can also occur because the two variables are not associated. Thus, in interpreting the covariance, one usually needs to have some idea of the variability in both variables. A large covariance, however, does imply that at least one of the two variables has a large variance.

The correlation coefficient is a rescaling of the covariance by the standard deviations of X and Y. The motivation for the construction of the covariance and correlation coefficient is the following: s_{xy} is the average of the product of the deviations about the means of X and Y. If X tends to be large when Y is large, both deviations will be positive and the product will be positive. Similarly, if X is small when Y is small, both deviations will be negative but their products will still be positive. Hence, the average of the products for all the cases will tend to be positive. If there is no relationship between X and Y, a positive deviation in X may be paired with a positive or negative deviation in Y and the product will either be positive or negative, and on the average will tend to center around zero. In the first case X and Y are said to be positively correlated, in the second case there is no correlation between X and Y. A third case results when large values of X tend to be associated with small values of Y, and vice versa. In this situation, the product of deviations will tend to be negative and the variables are said to be negatively correlated. The statistic r rescales the average of the product of the deviations about the means by the standard deviations of X and Y.

The statistic r has the following properties:

1. r has value between -1 and 1.
2. $r = 1$ if and only if all the observations are on a straight line with positive slope.
3. $r = -1$ if and only if all observations are on a straight line with negative slope.
4. r takes on the same value if X, or Y, changes units or has a constant added or subtracted.
5. r measures the extent of *linear* association between two variables.
6. r tends to be close to zero if there is no linear association between X and Y.

Some typical scattergrams and associated values of r are given in Figure 9.11. Figure 9.11(*a*) and (*b*) indicate perfect linear relationships between two variables. Figure 9.11(*c*) indicates no correlation. Figure 9.11(*d*) and (*e*) indicate typical patterns representing less than perfect correlation. Figure 9.11(*f*) to (*j*) portray various pathological situations. Figure 9.11(*f*) indicates that although there is an explicit relationship between X and Y, the linear relationship is zero; thus $r = 0$ does not imply that there is no relationship between X and Y. In statistical terminology, $r = 0$ does not imply that the variables are statistically independent. There is one important exception to this statement that is discussed in Section 9.3.3. Figure 9.11(*g*) indicates that except for the one extreme point there is no correlation. The coefficient of correlation is very sensitive to such outliers, and in Section 9.3.7 we discuss correlations that are not as sensitive, that is, more robust. Figure 9.11(*h*) indicates that an explicit relationship between X and Y is not identified by the correlation coefficient if the relationship is not linear. Finally, Figure 9.11(*j*)
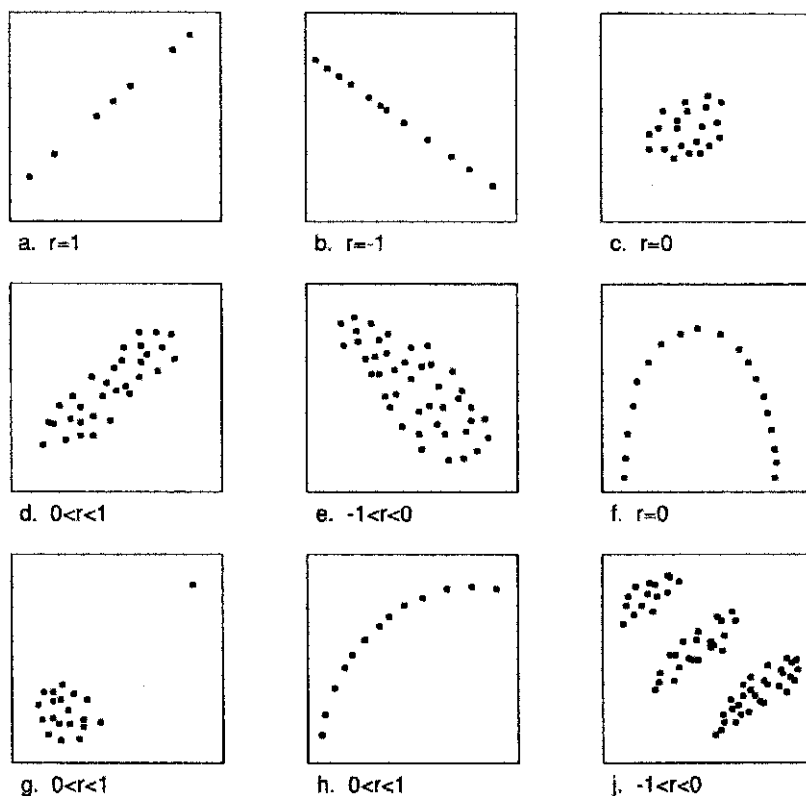

Figure 9.11   Some patterns of association.

suggests that there are three subgroups of cases; within each subgroup there is a positive cor-
relation, but the correlation is negative when the subgroups are combined. The reason is that
the subgroups have different means and care must be taken when combining data. For example,
natural subgroups defined by gender or race may differ in their means in a direction opposite
to the correlation within each subgroup.

Now consider Example 9.3. The scattergram in Figure 9.3 suggests a positive association
between the ATP level of the youngest son (X) and that of the oldest son (Y). The data for
this example produce the following summary statistics (the subscripts on the values of X and
Y have been suppressed: for example, $\sum x_i = \sum x$).

$$
\begin{aligned}
&n = 17, \\
&\sum x = 83.38, && \overline{x} = 4.90, \\
&\sum y = 79.89, && \overline{y} = 4.70, \\
&\sum x^2 = 417.1874, && \sum (x - \overline{x})^2 = 8.233024, && s_x = 0.717331 \\
&\sum y^2 = 379.6631, && \sum (y - \overline{y})^2 = 4.227094, && s_y = 0.513997 \\
&\sum xy = 395.3612, && \sum (x - \overline{x})(y - \overline{y}) = 3.524247, && s_{xy} = 0.220265
\end{aligned}
$$

$$
r = \frac{0.220265}{(0.717331)(0.513997)} = 0.597
$$

In practice, r will simply be calculated from the equivalent formula

$$
r = \frac{[xy]}{\sqrt{[x^2][y^2]}} = \frac{3.524247}{\sqrt{(8.233024)(4.227094)}} = \frac{3.524247}{5.899302} = 0.597
$$

The sample correlation coefficient and covariance estimate the population parameters. The
expected value of the covariance is

$$
\begin{aligned}
E(S_{xy}) &= E((X - \mu_x)(Y - \mu_y)) \\
&= \sigma_{xy}
\end{aligned}
$$

where

$$
\mu_x = E(X) \quad \text{and} \quad \mu_y = E(Y)
$$

The population covariance is the average of the product of X about its mean times Y about its
mean.

The sample correlation coefficient estimates the population correlation coefficient ρ, defined
as follows:

Definition 9.6. Let (X, Y) be two jointly distributed random variables. The (*population*)
correlation coefficient is

$$
\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{E((X - \mu_x)(Y - \mu_y))}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}}
$$

where σ_{xy} is the covariance of X and Y, σ_x the standard deviation of X, and σ_y the standard
deviation of Y. ρ is zero if X and Y are statistically independent variables.

There is now a question about the statistical "significance" of a value r. In practical terms,
suppose that we have sampled 17 families and calculated the correlation coefficient in ATP

levels between the youngest son and the oldest son. How much variation could we have expected relative to the value observed for this set? Could the population correlation coefficient $\rho = 0$? In the next two sections we deal with this question.

9.3.2 Relationship between Correlation and Regression

In Section 9.2.4, r^2 was presented, indicating a close connection between correlation and regression. In this section, the connection will be made explicit in several ways. Formally, one of the variables X, Y could be considered the dependent variable and the other the predictor variable and the techniques of Section 9.2 applied. It is easy to see that in most cases the slope of the regression of Y on X will not be the same as that of X on Y. To keep the distinction clear, the following notation will be used:

b_{yx} = slope of the regression of the "dependent" variable Y on the "predictor" variable X

a_y = intercept of the regression of Y on X

Similarly,

b_{xy} = slope of the regression of the "dependent" variable X on the "predictor" variable Y

a_x = intercept of the regression of X on Y

These quantities are calculated as follows:

	Regress Y on X	Regress X on Y
Slope	$b_{yx} = \dfrac{[xy]}{[x^2]}$	$b_{xy} = \dfrac{[xy]}{[y^2]}$
Intercept	$a_y = \overline{y} - b_{yx}\overline{x}$	$a_x = \overline{x} - b_{xy}\overline{y}$
Residual variance	$S^2_{y \cdot x} = \dfrac{[y^2] - [xy]^2/[x^2]}{n-2}$	$S^2_{x \cdot y} = \dfrac{[x^2] - [xy]^2/[y^2]}{n-2}$

From these quantities, the following relationships can be derived:

1. Consider the product

$$b_{yx}b_{xy} = \frac{[xy]^2}{[x^2][y^2]}$$
$$= r^2$$

Hence

$$r = \pm\sqrt{b_{yx}b_{xy}}$$

In words, r is the geometric mean of the slope of the regression of Y on X and the slope of the regression of X on Y.

2.

$$b_{yx} = r\frac{S_y}{S_x}, \qquad b_{xy} = r\frac{S_x}{S_y}$$

where S_x and S_y are the sample standard deviations of X and Y, respectively.

3. Using the relationships in (2), the regression line of Y on X,

$$\widehat{Y} = a_y + b_{yx}X$$

can be transformed to

$$\widehat{Y} = a_y + \frac{rS_y}{S_x}X$$

$$= \overline{y} + \frac{rS_y}{S_x}(X - \overline{x})$$

4. Finally, the t-test for the slope, in the regression of Y on X,

$$t_{n-2} = \frac{b_{yx}}{S_{b_{yx}}}$$

$$= \frac{b_{yx}}{S_{y \cdot x}/\sqrt{[x^2]}}$$

is algebraically equivalent to

$$r \Big/ \sqrt{\frac{1 - r^2}{n - 2}}$$

Hence, testing the significance of the slope is equivalent to testing the significance of the correlation.

Consider Example 9.3 again. The data are summarized in Table 9.8. This table indicates that the two regression lines are not the same but that the t-tests for testing the significance of the slopes produce the same observed value, and this value is identical to the test of significance of the correlation coefficient. If the corresponding analyses of variance are carried out, it will be found that the F-ratio in the two analyses are identical and give an equivalent statistical test.

9.3.3 Bivariate Normal Distribution

The statement that a random variable Y has a normal distribution with mean μ and variance σ^2 is a statement about the distribution of the values of Y and is written in a shorthand way as

$$Y \sim N(\mu, \sigma^2)$$

Such a distribution is called a *univariate distribution*.

Definition 9.7. A specification of the distribution of two (or more) variables is called a *bivariate* (or *multivariate*) *distribution*.

The definition of such a distribution will require the specification of the numerical characteristics of each of the variables separately as well as the relationships among the variables. The most common bivariate distribution is the normal distribution. The equation for the density of this distribution as well as additional properties are given in Note 9.6.

We write that (X, Y) have a bivariate normal distribution as

$$(X, Y) \sim N(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$$

Table 9.8   Regression Analyses of ATP Levels of Oldest and Youngest Sons

Dependent variable	Y^a		X^a
Predictor variable	X^b		Y^b
Slope	$b_{yx} = 0.42806$		$b_{xy} = 0.83373$
Intercept	$a_y = 2.59989$		$a_x = 0.98668$
Regression line	$\widehat{Y} = 2.600 + 0.428X$		$\widehat{X} = 0.987 + 0.834Y$
Variance about mean	$s_y^2 = 0.26419$		$s_x^2 = 0.51456$
Residual variance	$s_{y \cdot x}^2 = 0.18123$		$s_{x \cdot y}^2 = 0.35298$
Standard error of slope	$s_{b_{y \cdot x}} = 0.14837$		$s_{b_{x \cdot y}} = 0.28897$
Test of significance	$t_{15} = \dfrac{0.42806}{0.14837} = 2.885$		$t_{15} = \dfrac{0.83373}{0.28897} = 2.885$
Correlation		$r_{xy} = r_{yx} = r = 0.597401$	
Test of significance		$t_{15} = \dfrac{0.597401}{\sqrt{\dfrac{1 - (0.597401)^2}{17 - 2}}}$	
		$= \dfrac{0.597401}{0.20706}$	
		$= 2.885$	

Source: Data from Dern and Wiorkowski [1969].
[a] ATP level of oldest son.
[b] ATP level of youngest son.

Here μ_x, μ_y, σ_x^2, and σ_y^2 are the means and variances of X and Y, respectively. The quantity ρ is the (population) correlation coefficient. If we assume this model, it is this quantity, ρ, that is estimated by the sample correlation, r.

The following considerations may help to give you some feeling for the bivariate normal distribution. A continuous distribution of two variables, X and Y, may be modeled as follows. Pour 1 pound of sand on a floor (the X–Y plane). The probability that a pair (X, Y) falls into an area, say A, on the floor is the weight of the sand on the area A. For a bivariate normal distribution, the sand forms one mountain, or pile, sloping down from its peak at (μ_x, μ_y), the mean of (X, Y). Cross sections of the sand at constant heights are all ellipses. Figure 9.12 shows a bivariate normal distribution. On the left is shown a view of the sand pile; on the right, a topographical map of the terrain.

The bivariate normal distribution has the property that at every fixed value of X (or Y) the variable Y (or X) has a univariate normal distribution. In particular, write

$$Y_x = \text{random variable } Y \text{ at a fixed value of } X = x$$

It can be shown that at this fixed value of $X = x$,

$$Y_x \sim N\left(\alpha_y + \frac{\sigma_y}{\sigma_x}\rho x, \quad \sigma_y^2(1 - \rho^2)\right)$$

This is the regression model discussed previously:

$$Y_x \sim N(\alpha + \beta x, \sigma_1^2)$$

where

$$\alpha = \mu_y - \beta\mu_x, \qquad \beta = \frac{\sigma_y}{\sigma_x}\rho, \qquad \sigma_1^2 = \sigma_y^2(1 - \rho^2)$$
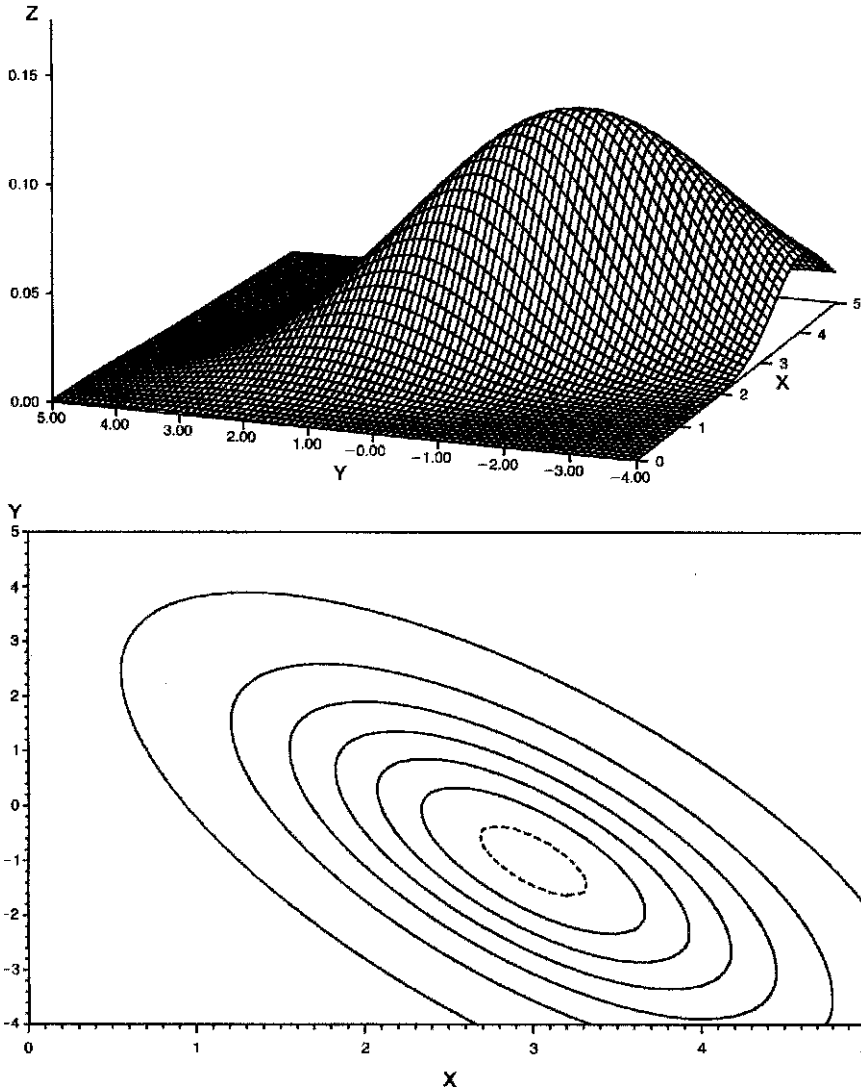
Figure 9.12 Bivariate normal distribution.

Similarly, for

$$X_y = \text{random variable } X \text{ at a fixed value of } Y = y$$

it can be shown that

$$X_y \sim N \left(\alpha_x + \frac{\sigma_x}{\sigma_y} \rho y, \sigma_x^2 (1 - \rho^2) \right)$$

The null hypothesis $\beta_{yx} = 0$ (or, $\beta_{xy} = 0$) is equivalent then to the hypothesis $\rho = 0$, and the t-test for $\beta = 0$ can be applied.

Suppose now that the null hypothesis is

$$\rho = \rho_0$$

where ρ_0 is an arbitrary but specified value. The sample correlation coefficient r does not have a normal distribution and the usual normal theory cannot be applied. However, R. A. Fisher showed that the quantity

$$Z_r = \frac{1}{2} \log_e \frac{1+r}{1-r}$$

has the following approximate normal distribution:

$$Z_r \sim N\left(\frac{1}{2} \log_e \frac{1+\rho}{1-\rho}, \frac{1}{n-3}\right)$$

where n is the number of pairs of values of X and Y from which r is computed. Not only does Z_r have approximately a normal distribution, but the variance of this normal distribution does not depend on the true value ρ; that is, $Z_r - Z_\rho$ is a pivotal quantity (5.2). This is illustrated graphically in Figure 9.13, which shows the distribution of 1000 simulated values of r and Z_r from distributions with $\rho = 0$ and $\rho = 1/\sqrt{2} \approx 0.71$. The distribution of r has a different variance and different shape for the two values of ρ, but the distribution of Z_r has the same shape and same variance, differing only in location.
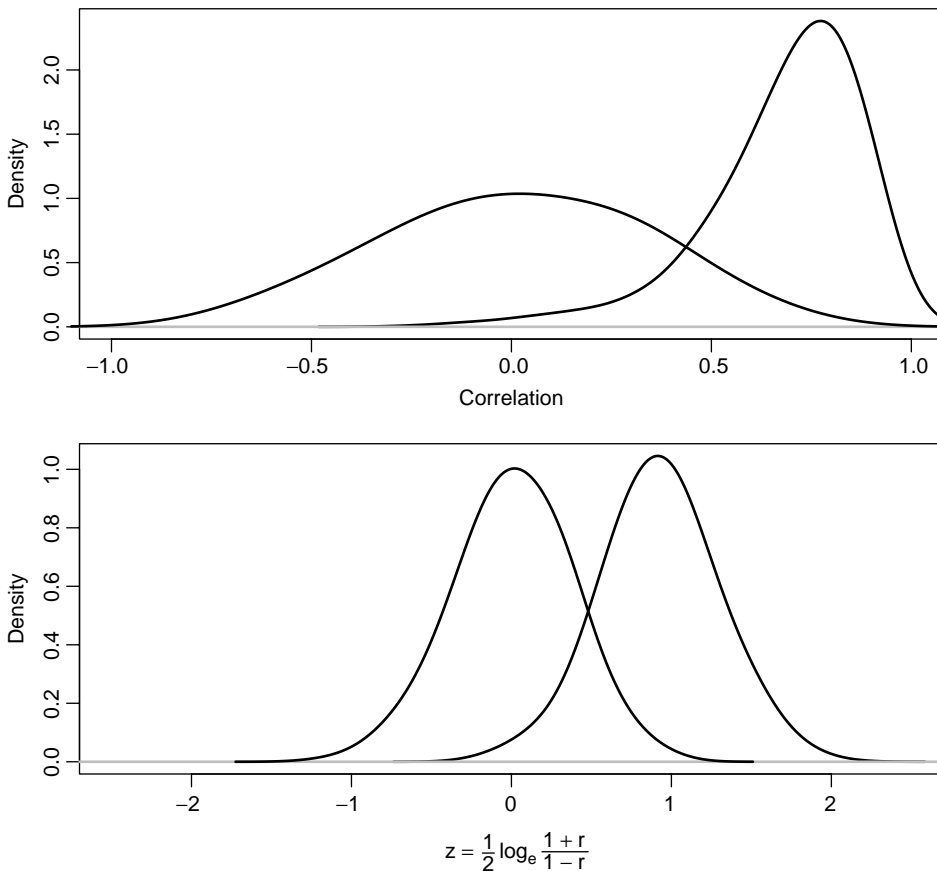
Figure 9.13 Sampling distribution of correlation coefficient, r, before and after transformation, for $\rho = 0, 1/\sqrt{2}$. Estimated from 1000 samples of size 10.

Although the approximate distribution of Z_r was derived under the assumption of a bivariate normal distribution for X and Y, it is not very sensitive to this assumption and is useful quite broadly. Z_r may be used to test hypotheses about ρ and to construct a confidence interval for ρ. This is illustrated below. The inverse, or reverse, function to r is $(e^{2Z} - 1)/(e^{2Z} + 1)$. Z_r is also the inverse of the hyperbolic tangent, tanh. To "undo" the operation, tanh is used.

Consider again Example 9.3 involving the ATP levels of youngest and oldest sons in the 17 families. The correlation coefficient was calculated to be

$$r = 0.5974$$

This value was significantly different from zero; that is, the null hypothesis $\rho = 0$ was rejected. However, the authors show in the paper that genetic theory predicts the correlation to be $\rho = 0.5$. Does the observed value differ significantly from this value? To test this hypothesis we use the Fisher Z_r transformation. Under the genetic theory, the null hypothesis stated in terms of Z_r is

$$Z_r \sim N\left(\frac{1}{2}\log_e\left(\frac{1+0.5}{1-0.5}\right), \frac{1}{17-3}\right)$$
$$\sim N(0.5493, 0.07143)$$

The value observed is

$$Z_r = \frac{1}{2}\log_e\left(\frac{1+0.5974}{1-0.5974}\right) = 0.6891$$

The corresponding standard normal deviate is

$$z = \frac{0.6891 - 0.5493}{\sqrt{0.07143}} = \frac{0.1398}{0.2673} = 0.5231$$

This value does not exceed the critical values at, say, the 0.05 level, and there is no evidence to reject this null hypothesis.

Confidence intervals for ρ may be formed by first using Z_r to find a confidence interval for $1/2\log_e[(1+\rho)/(1-\rho)]$. We then transform back to find the confidence interval for ρ. To illustrate: a $100(1-\alpha)\%$ confidence interval for $1/2\log_e[(1+\rho)/(1-\rho)]$ is given by

$$Z_r \pm z_{1-\alpha/2}\sqrt{\frac{1}{n-3}}$$

For a 90% confidence interval with these data, the interval is $(0:6891 - 1.645\sqrt{1/14}, 0.6891 + 1.645\sqrt{1/14}) = (0.249, 1.13)$. When $Z_r = 0.249, r = 0.244$, and when $Z_r = 0.811$. Thus the 90% confidence interval for ρ is $(0.244, 0.811)$. This value straddles 0.5.

9.3.4   Critical Values and Sample Size

We discussed the t-test for testing the hypothesis $\rho = 0$. The formula was

$$t_{n-2} = \frac{r}{\sqrt{(1-r^2)/(n-2)}}$$

This formula is very simple and can be used for finding critical values and for sample size calculations: Given that the number of observation pairs is specified, the critical value for t with

$n - 2$ degrees of freedom is determined, and hence the r critical value can be calculated. For simplicity, write $t_{n-1} = t$; solving the equation above for r^2 yields

$$r^2 = \frac{t^2}{t^2 + n - 2}$$

For example, suppose that $n = 20$, the corresponding t-value with 18 degrees of freedom at the 0.05 level is $t_{18} = 2.101$. Hence,

$$r^2 = \frac{(2.101)^2}{(2.101)^2 + 18} = 0.1969$$

and the corresponding value for r is $r \pm 0.444$; that is, with 20 observations the value of r must exceed 0.444 or be less than -0.444 to be significant at the 0.05 level. Table A.11 lists critical values for r, as a function of sample size.

Another approach is to determine the sample size needed to "make" an observed value of r significant. Algebraic manipulation of the formula gives

$$n = \frac{t^2}{r^2} - t^2 + 2$$

A useful approximation can be derived if it is assumed that we are interested in reasonably small values of r, say $r < 0.5$; in this case, $t \doteq 2$ at the 0.05 level and the formula becomes

$$n = \left(\frac{2}{r}\right)^2 - 2$$

For example, suppose that $r = 0.3$; the sample size needed to make this value significant is

$$n = \left(\frac{2}{0.3}\right)^2 - 2 = 44 - 2 = 42$$

A somewhat more refined calculation yields $n = 43$, so the approximation works reasonably well.

9.3.5 Using the Correlation Coefficient as a Measure of Agreement for Two Methods of Measuring the Same Quantity

We have seen that for X and Y jointly distributed random variables, the correlation coefficient ρ is a population parameter value: ρ is a measure of how closely X and Y have a linear association, ρ^2 is the proportion of the Y variance that can be explained by linear prediction from X, and vice versa.

Suppose that the regression holds and we may choose X. Figure 9.14 shows data from a regression model with three different patterns of X variables chosen. The same errors were added in each figure. The X values were spread out over larger and larger intervals. Since the spread *about* the regression line remains the same and the range of Y increases as the X range increases, the proportion of Y variability explained by X increases: 0.50 to 0.68 to 0.79. For the same random errors and population regression line, r can be anywhere between 0 and 1, depending on which X values are used! In this case the correlation coefficient depends not only on the model, but also on experimental design, where the X's are taken. For this reason some authors say that the r should never be used unless one has a *bi*variate sample: Otherwise, we do not know what r means; another experimenter with the same regression model could choose different X values and obtain a radically different result.

We discuss these ideas in the context of the exercise data of Example 9.2. Suppose that we were strong supporters of maximal treadmill stress testing and wanted to show how closely treadmill duration and $VO_{2\ MAX}$ are related. Our strategy for obtaining a large correlation coefficient will be to obtain a large spread of X values, duration. We may know that some of the largest duration and $VO_{2\ MAX}$ values were obtained by world-class cross-country skiers; so we would recruit some. For low values we might search for elderly overweight and deconditioned persons. Taking a combined group of these two types of subjects should result in a large value of r. If the same experiment is run using only very old, very overweight, and very deconditioned subjects, the small range will produce a small, statistically insignificant r value.

Since the same treadmill test procedure is associated with large and small r values, what does r mean? A preferable summary indicator is the estimate, $s_{y \cdot x}$ of the residual standard deviation σ_1. If the linear regression model holds, this would be estimated to be the same in each case.

Is it wrong to calculate or present r when a bivariate sample is not obtained? Our answer is a qualified no; that is, it is all right to present r in regression situations provided that:

1. The limitations are kept in mind and discussed. Possible comments on the situation for other sorts of X values might be appropriate.
2. The standard deviation of the residuals should be estimated and presented.

In Chapter 7, the kappa statistic was presented. This was a measure of the amount of agreement when two categorical measurements of the same objects were available. If the two measurements were continuous, the correlation coefficient r is often used as a measure of the agreement of the two techniques. Such use of r is subject to the comments above.

9.3.6   Errors in Both Variables

An assumption in the linear regression model has been that the predictor variable could be measured without error and that the variation in the dependent variable was of one kind only
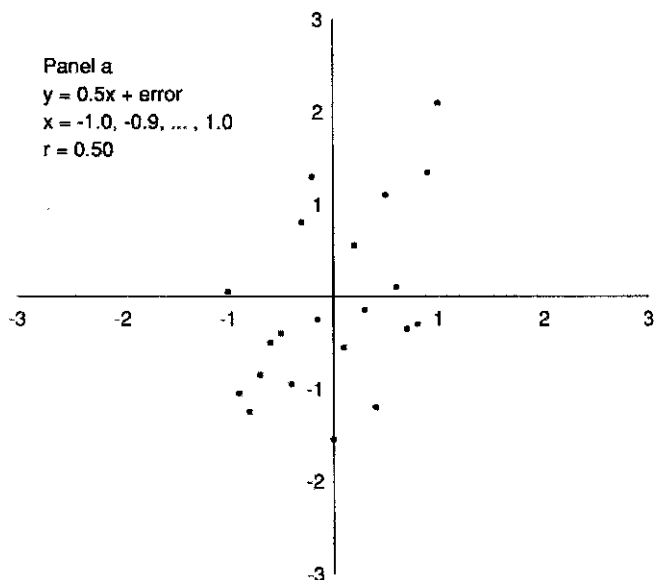
Figure 9.14   The regression model $Y = 0.5X + e$ was used. Twenty-one random $N(0, 1)$ errors were generated by computer. The same errors were used in each panel.

Panel b
y = 0.5x + error
x = -2.0, -1.8, ... , 2.0
r = 0.68

Panel c
y = 0.5x + error
x = -3.0, -2.7, ... , 3.0
r = 0.79
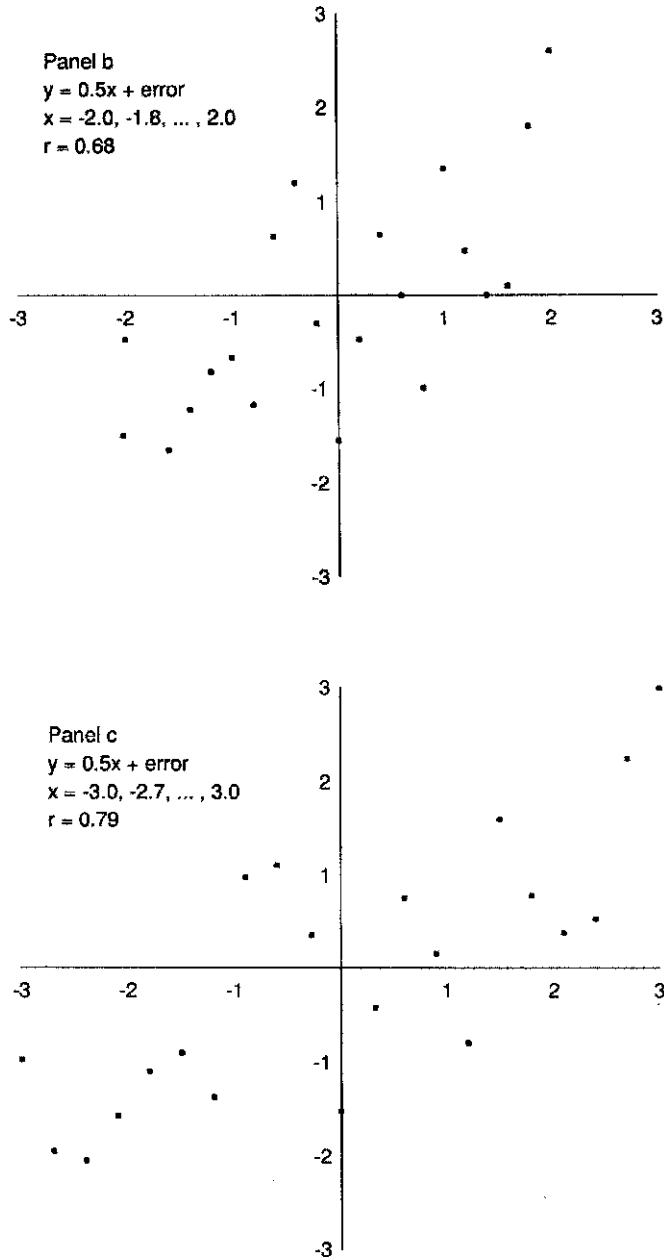
Figure 9.14   (*continued*)

and could be modeled completely if the value of the predictor variable was fixed. In almost all cases, these assumptions do not hold. For example, in measuring psychological characteristics of individuals, there is (1) variation in the characteristics from person to person; and (2) error in the measurement of these psychological characteristics. It is almost certainly true that this problem is present in all scientific work. However, it may be that the measurement error is "small" relative to the variation of the individuals, and hence the former can be neglected.

Another context where the error is unimportant is where the scientific interest is in the variable as measured, not some underlying quantity. For example, in examining how well blood pressure predicts stroke, we are interested in practical prediction, not in what might hypothetically be possible with perfect measurements.

The problem is difficult and we will not discuss it beyond the effect of errors on the correlation coefficient. For a more complete treatment, consult Acton [1984] or Kendall and Stuart [1967, Vol. 2], and for a discussion of measurement error in more complex models, see Carrol et al. [1995].

Suppose that we are interested in the correlation between two random variables X and Y which are measured with errors so that instead of X and Y, we observe that

$$W = X + d, \qquad V = Y + e$$

where d and e are errors. The sampling we have in mind is the following: a "case" is selected at random from the population of interest. The characteristics X and Y are measured but with random independent errors d and e. It is assumed that these errors have mean zero and variances σ_1^2 and σ_2^2, respectively. Another "case" is then selected and the measurement process is repeated with error. Of interest is the correlation ρ_{XY} between X and Y, but the correlation ρ_{VW} is estimated. What is the relationship between these two correlations? The correlation ρ_{XY} can be written

$$\rho_{XY} = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2 \sigma_Y^2}}$$

The reason for writing the correlation this way can be understood when the correlation between V and W is considered:

$$\rho_{VW} = \frac{\sigma_{XY}}{\sqrt{(\sigma_X^2 + \sigma_1^2)(\sigma_Y^2 + \sigma_2^2)}}$$

$$= \frac{\sigma_{XY}}{\sigma_X \sigma_Y \sqrt{\left(1 + \sigma_1^2/\sigma_X^2\right)\left(1 + \sigma_2^2/\sigma_Y^2\right)}}$$

$$= \frac{\rho_{XY}}{\sqrt{\left(1 + \sigma_1^2/\sigma_X^2\right)\left(1 + \sigma_2^2/\sigma_Y^2\right)}}$$

The last two formulas indicate that the correlation between V and W is smaller in absolute value than the correlation between X and Y by an amount determined by the ratio of the measurement errors to the variance in the population. Table 9.9 gives the effect on ρ_{XY} as related to the ratios of σ_1^2/σ_X^2 and σ_2^2/σ_Y^2.

A 10% error of measurement in the variables X and Y produces a 9% reduction in the correlation coefficient. The conclusion is that errors of measurement reduce the correlation between two variables; this phenomenon is called *attenuation*.

Table 9.9   Effect of Errors of Measurement on the Correlation between Two Random Variables

$\dfrac{\sigma_1^2}{\sigma_X^2}$	$\dfrac{\sigma_2^2}{\sigma_Y^2}$	ρ_{VW}	$\dfrac{\sigma_1^2}{\sigma_X^2}$	$\dfrac{\sigma_2^2}{\sigma_Y^2}$	ρ_{VW}
0	0	1 ρ_{XY}	0.20	0.10	$0.87\rho_{XY}$
0.05	0.05	$0.95\rho_{XY}$	0.20	0.20	$0.83\rho_{XY}$
0.10	0.10	$0.91\rho_{XY}$	0.30	0.30	$0.77\rho_{XY}$

Table 9.10 Schema for Spearman Rank Correlation

Case	X	Rank(X)	Y	Rank(Y)	$d = \text{Rank}(X) - \text{Rank}(Y)$
1	x_1	R_{x_1}	y_1	R_{y_1}	$d_1 = R_{x_1} - R_{y_1}$
2	x_2	R_{x_2}	y_2	R_{y_2}	$d_2 = R_{x_2} - R_{y_2}$
3	x_3	R_{x_3}	y_3	R_{y_3}	$d_3 = R_{x_3} - R_{y_3}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_n	R_{x_n}	y_n	R_{y_n}	$d_n = R_{x_n} - R_{y_n}$

9.3.7 Nonparametric Estimates of Correlation

As indicated earlier, the correlation coefficient is quite sensitive to outliers. There are many ways of getting estimates of correlation that are more robust; the paper by Devlin et al. [1975] contains a description of some of these methods. In this section we want to discuss two methods of testing correlations derived from the ranks of observations.

The procedure leading to the Spearman rank correlation is as follows: Given a set of n observations on the variables X, Y, the values for X are replaced by their ranks, and similarly, the values for Y. Ties are simply assigned the average of the ranks associated with the tied observations. The scheme shown in Table 9.10 illustrates the procedure.

The correlation is then calculated between R_x and R_y. In practice, the *Spearman rank correlation formula* is used:

$$r_s = r_{R_x R_y} = 1 - \frac{6 \sum d_i^2}{n^3 - n}$$

It can be shown that the usual Pearson product-moment correlation formula reduces to this formula when the calculations are made on the ranks, if there are no ties. *Note:* For one or two ties, the results are virtually the same. It is possible to correct the Spearman formula for ties, but a simpler procedure is to calculate r_s by application of the usual product-moment formula to the ranks. Table A.12 gives percentile points for testing the hypothesis that X and Y are independent.

Example 9.4. Consider again the data in Table 9.3 dealing with the ATP levels of the oldest and youngest sons. These data are reproduced in Table 9.11 together with the ranks, the ATP levels being ranked from lowest to highest.

Note that the oldest sons in families 6 and 13 had the same ATP levels; they would have been assigned ranks 12 and 13 if the values had been recorded more accurately; consequently, they are both assigned a rank of 12.5. For this example,

$$n = 17$$
$$\sum d_i^2 = 298.5$$
$$r_s = 1 - \frac{(6)(298.5)}{17^3 - 17} = 1 - 0.3658 = 0.6342$$

This value compares reasonably well with the value $r_{xy} = 0.597$ calculated on the actual data. If the usual Pearson product-moment formula is applied to the ranks, the value $r_s = 0.6340$ is obtained. The reader may verify that this is the case. The reason for the slight difference is due to the tie in values for two of the oldest sons. Table A.12 shows the statistical significance at the two-sided 0.05 level since $r_s = 0.6342 > 0.490$.

The second nonparametric correlation coefficient is the *Kendall rank correlation coefficient*. Recall our motivation for the correlation coefficient r. If there is positive association, increase in

Table 9.11 Rank Correlation Analysis of ATP Levels in Youngest and Oldest Sons in 17 Families

	Youngest		Oldest		
Family	ATP Level	Rank (X)	ATP Level	Rank (Y)	d^a
1	4.18	4	4.81	11	−7
2	5.16	12	4.98	14	−2
3	4.85	9	4.48	6	3
4	3.43	1	4.19	3	−2
5	4.53	5	4.27	4	1
6	5.13	11	4.87	12.5	−1.51
7	4.10	2	4.74	10	−8
8	4.77	7	4.53	7	0
9	4.12	3	3.72	1	2
10	4.65	6	4.62	8	−2
11	6.03	17	5.83	17	0
12	5.94	15	4.40	5	10
13	5.99	16	4.87	12.5	3.5
14	5.43	14	5.44	16	−2
15	5.00	10	4.70	9	1
16	4.82	8	4.14	2	6
17	5.25	13	5.30	15	−2
					$\sum d = 0$
					$\sum d^2 = 298.5$

^{a}Rank(X) − rank(Y).

X will tend to correspond to increase in Y. That is, given two data points (X_1, Y_1) and (X_2, Y_2), if $X_1 − X_2$ is positive, $Y_1 − Y_2$ is positive. In this case, $(X_1 − X_2)(Y_1 − Y_2)$ is usually positive. If there is negative association, $(X_1 − X_2)(Y_1 − Y_2)$ will usually be negative. If X and Y are independent, the expected value is zero. Kendall's rank correlation coefficient is based on this observation.

Definition 9.8. Consider a bivariate sample of size n, $(X_1, Y_1), \ldots, (X_n, Y_n)$. For each pair, count 1 if $(X_i − X_j)(Y_i − Y_j) > 0$. Count −1 if $(X_i − X_j)(Y_i − Y_j) < 0$. Count zero if $(X_i − X_j)(Y_i − Y_j) = 0$. Let κ be the sum of these $n(n − 1)/2$ counts. (Note that this κ is not related to the kappa of Chapter 7.) Kendall's τ is

$$\tau = \frac{\kappa}{n(n − 1)/2}$$

1. The value of τ is between −1 and 1. Under the null hypothesis of independence, τ is symmetric about zero.
2. Note that $(R_{X_i} − R_{X_j})(R_{Y_i} − R_{Y_j})$ has the same sign as $(X_i − X_j)(Y_i − Y_j)$. That is, both are positive or both are negative or both are zero. If we calculated τ from the ranks of the (X_i, Y_i), we get the same number. Thus, τ is a nonparametric quantity based on ranks; it does not depend on the distributions of X and Y.
3. The expected value of τ is

$$P[(X_i − X_j)(Y_i − Y_j) > 0] − P[(X_i − X_j)(Y_i − Y_j) < 0]$$

Table 9.12 Data for Example 9.4[a]

								j								
i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
2	1															
3	−1	1														
4	1	1	1													
5	−1	1	1	1												
6	1	1	1	1	1											
7	1	1	−1	1	−1	1										
8	−1	1	−1	1	1	1	−1									
9	1	1	1	−1	1	1	−1	1								
10	−1	1	−1	1	1	1	−1	−1	1							
11	1	1	1	1	1	1	1	1	1	1						
12	−1	−1	−1	1	1	−1	−1	−1	1	−1	1					
13	1	−1	1	1	1	0	1	1	1	1	1	1				
14	1	1	1	1	1	1	1	1	1	1	1	−1	−1			
15	−1	1	1	1	1	1	−1	1	1	1	1	−1	1	1		
16	−1	1	1	−1	−1	1	−1	−1	1	−1	1	1	1	1	1	
17	1	1	1	1	1	1	1	1	1	1	1	−1	−1	1	1	1

[a]Consider $(X_i − X_j)(Y_i − Y_j)$: the entries are 1 if this is positive, 0 if this equals 0, and −1 if this is negative.

4. For moderate to large n and no or few ties, an approximate standard normal test statistic is

$$Z = \frac{\kappa}{\sqrt{n(n-1)(2n+5)/18}}$$

More information where there are ties is given in Note 9.7.

5. If $(X_i − X_j)(Y_i − Y_j) > 0$, the pairs are said to be concordant. If $(X_i − X_j)(Y_i − Y_j) < 0$, the pairs are discordant.

Return to the ATP data of Table 9.11. $(X_1 − X_2)(Y_1 − Y_2) = (4.18 − 5.16)(4.81 − 4.98) > 0$, so we count +1. Comparing each of the $17 \times 16/2 = 136$ pairs gives the +1's, 0's and −1's in Table 9.12. Adding these numbers, $\kappa = 67$, and $\tau = 67/(17 \times 16/2) = 0.493$. The asymptotic Z-value is

$$Z = \frac{67}{\sqrt{17 \times 16 \times 39/18}} = 2.67$$

with $p = 0.0076$ (two-sided).

9.3.8 Change and Association

Consider two continuous measurements of the same quantity on the same subjects at different times or under different circumstances. The two times might be before and after some treatment. They might be for a person taking a drug and not taking a drug. If we want to see if there is a difference in the means at the two times or under the two circumstances, we have several statistical tests: the paired t-test, the signed rank test, and the sign test. Note that we have observed pairs of numbers on each subject.

We now have new methods when pairs of numbers are observed: linear regression and correlation. Which technique should be used in a given circumstance? The first set of techniques looks for *changes between the two measurements*. The second set of techniques look for association and sometimes *the ability to predict*. The two concepts are different ideas:

1. Consider two independent length measurements from the same x-rays of a sample of patients. Presumably there is a "true" length. The measurements should fluctuate about the true length. Since the true length will fluctuate from patient to patient, the two readings should be associated, hopefully highly correlated. Since both measurements are of the same quantity, there should be little or no change. This would be a case where one expects association, but no change.

2. Consider cardiac measurements on patients before and after a heart transplant. The initial measurements refer to a failing heart. After heart transplant the measurements refer to the donor heart. There will be little or no association because the measurements of output, and so on, refer to different (somewhat randomly paired) hearts.

There are situations where both change and prediction or association are relevant. After observing a change, one might like to investigate how the new changed values relate to the original values.

9.4   COMMON MISAPPLICATION OF REGRESSION AND CORRELATION METHODS

In this section we discuss some of the pitfalls of regression and correlation methods.

9.4.1   Regression to the Mean

Consider Figure 9.15, which has data points with approximately zero correlation or association, considered as measurements before and after some intervention. On the left we see that the before and after measurements have no association. The solid line indicates before = 0, and the dashed line indicates before = after. On the right we plot the change against the value before intervention. Again, the two lines are before = 0 and before = after (i.e., change = 0), and we can see how selecting based on the value of the measurement before intervention distorts the average change.

Cases with low initial values (circles on the graph) tend to have positive changes; those with high initial values (triangles) have negative changes. If we admitted to our study only the subjects with low values, it would appear that the intervention led to an increase. In fact, the change would be due to random variability and the case selection. This phenomenon is called *regression to the mean*.

As another example, consider subjects in a quantitative measurement of the amount of rash due to an allergy. Persons will have considerable variability due to biology and environment. Over time, in a random fashion, perhaps related to the season, the severity of rash will ebb and flow. Such people will naturally tend to seek medical help when things are in a particularly bad state. Following the soliciting of help, biological variability will give improvement with or without treatment. Thus, if the treatment is evaluated (using before and after values), there would be a natural drop in the amount of rash simply because medical help was solicited during particularly bad times. This phenomenon again is *regression to the mean*. The phenomenon of regression to the mean is one reason that control groups are used in clinical studies. Some approaches to addressing it are given by Yanez et al. [1998].

9.4.2   Spurious Correlation

Consider a series of population units, for example, states. Suppose that we wish to relate the occurrence of death from two distinct causes, for example, cancer at two different sites on the body. If we take all the states and plot a scatter diagram of the number of deaths from the two causes, there will be a relationship simply because states with many more people, such as
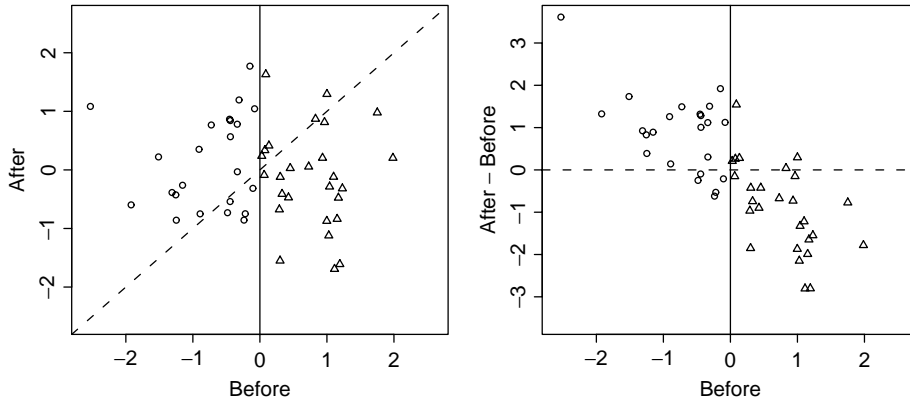
Figure 9.15 Regression to the mean in variables with no association: Before vs. after and before vs. change.

California or New York, will have a large number of deaths, compared to a smaller state such as Wyoming or New Hampshire.

It is necessary to somehow adjust for or take the population into account. The most natural thing to do is to take the death rate from certain causes, that is, to divide the number of deaths by the population of the state. This would appear to be a good solution to the problem. This introduces another problem, however. If we have two variables, X and Y, which are *not related* and we divide them by a third variable, Z, which is random, the two ratios X/Z and Y/Z *will be related*. Suppose that Z is the true denominator measured with error. The reason for the relationship is that when Z is on the low side, since we are dividing by Z, we will increase both numbers at the same time; when Z is larger than it should be and we divide X and Y by Z, we decrease both numbers. The introduction of correlation due to computing rates using the same denominator is called *spurious correlation*. For further discussion on this, see Neyman [1952] and Kronmal [1993], who gives a superb, readable review. A preferable way to adjust for population size is to use the techniques of multiple regression, which is discussed in Chapter 11.

9.4.3   Extrapolation beyond the Range of the Data

For many data sets, including the three of this chapter, the linear relationship does a reasonable job of summarizing the association between two variables. In other situations, the relationship may be reasonably well modeled as linear over a part of the range of X but not over the entire range of X. Suppose, however, that data had been collected on only a small range of X. Then a linear model might fit the accumulated data quite well. If one takes the regression line and uses it as an indication of what would happen for data values *outside the range covered by the actual data*, trouble can result. To have confidence in such extrapolation, one needs to know that indeed the linear relationship holds over a broader range than the range associated with the actual data. Sometimes this assumption is valid, but often, it is quite wrong. There is no way of knowing in general to what extent extrapolation beyond the data gives problems. Some of the possibilities are indicated graphically in Figure 9.16. Note that virtually any of these patterns of curves, when data are observed over a short range, can reasonably be approximated by a linear function. Over a wider range, a linear approximation is not adequate. But if one does not have data over the wide range, this cannot be seen.

Sometimes it is necessary to extrapolate beyond the range of the data. For example, there is substantial concern in Britain over the scale of transmission of "mad cow disease" to humans, causing variant Creutzfeld–Jakob disease (vCJD). Forecasting the number of future cases is
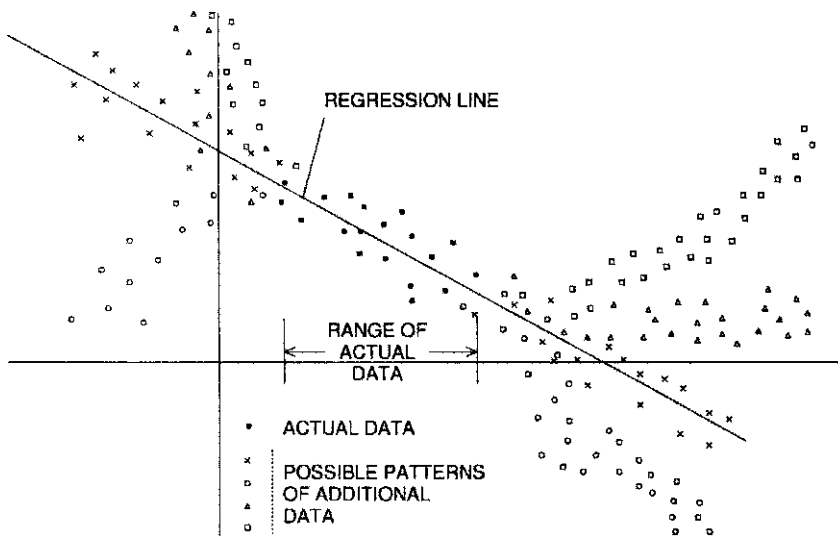
Figure 9.16   Danger of extrapolating beyond observed data.

important for public health, and intrinsically, requires extrapolation. A responsible approach to this type of problem is to consider carefully what models (linear or otherwise) are consistent with the data available and more important, with other existing knowledge. The result is a range of predictions that acknowledge both the statistical uncertainty within each model and the (often much greater) uncertainty about which model to use.

9.4.4   Inferring Causality from Correlation

Because two variables are associated does not necessarily mean that there is any causal connection between them. For example, if one recorded two numbers for each year—the numbers of hospital beds and the total attendance at major league baseball games—there would be a positive association because both of these variables have increased as the population increased. The direct connection is undoubtedly slight at best. Thus, regression and correlation approaches show observed relationships, which may or may not represent a causal relationship. In general, the strongest inference for causality comes from experimental data; in this case, factors are changed by the experimenter to observe change in a response. Regression and correlation from observational data may be very suggestive but do not definitively establish causal relationships.

9.4.5   Interpretation of the Slope of the Regression Line

During the discussion, we have noted that the regression equation implies that if the predictor or independent variable X is higher by an amount ΔX, then on the average, Y is higher by an amount $\Delta Y = b\,\Delta X$. This is sometimes interpreted to mean that if we can modify a situation such that the X variable is changed by ΔX, the Y variable will change correspondingly; this may or may not be the case. For example, if we look at a scatter diagram of adults' height and weight, it does not follow if we induce a change in a person's weight, either by dieting or by excess calories that the person's height will change correspondingly. Nevertheless, there is an association between height and weight. Thus, the particular inference depends on the science involved. Earlier in this chapter, it was noted that from the relation between $VO_{2\ MAX}$ and the duration of the exercise test that if a person is trained to have an increased duration, the $VO_{2\ MAX}$ will also increase. This particular inference is correct and has been documented by

serial studies recording both variables. It follows from other data and scientific understanding. It is *not* a logical consequence of the observed association.

9.4.6 Outlying Observations

As noted above, outlying observations can have a large effect on the actual regression line (see Figure 9.7, for example). If one examines these scattergrams or residual plots, the problem should be recognized. In many situations, however, people look at large numbers of correlations and do not have the time, the wherewithal, or possibly the knowledge to examine all of the necessary visual presentations. In such a case, an outlier can be missed and data may be interpreted inappropriately.

9.4.7 Robust Regression Models

The least squares regression coefficients result from minimizing

$$\sum_{i=1}^{n} g(Y_i - a - bX_i)$$

where the function $g(z) = z^2$. For large z (large residuals) this term is very large. In the second column of figures in Figure 9.7 we saw that one outlying value could heavily modify an otherwise nice fit.

One way to give less importance to large residuals is to choose the function g to put less weight on outlying values. Many robust regression techniques take this approach. We can choose g so that for most z, $g(z) = z^2$, as in the least squares estimates, but for very large $|z|$, $g(z)$ is less than z^2, even zero for extreme z! See Draper and Smith [1998, Chap. 25] and Huber [2003, Chap. 7]. These *resistant M–estimators* protect against outlying Y but not against outlying X, for which even more complex estimators are needed. It is also important to note that protection against outliers is not always desirable. Consider the situation of a managed care organization trying to determine if exercise reduces medical costs. A resistant regression estimator would effectively ignore information on occasional very expensive subjects, who may be precisely the most important in managing costs. See Chapter 8 and Lumley et al. [2002] for more discussion of these issues.

NOTES

9.1 Origin of the Term Regression

Sir Francis Galton first used the term in 1885. He studied the heights of parents and offspring. He found (on the average) that children of tall parents were closer to the average height (were shorter); children of short parents were taller and closer to the average height. The children's height *regressed* to the average.

9.2 Maximum Likelihood Estimation of Regression and Correlation Parameters

For a data set from a continuous probability density, the probability of observing the data is proportional to the probability density function. It makes sense to *estimate the parameters by choosing parameters to make the probability of the observed data as large as possible*. Such estimates are called *maximum likelihood estimates* (MLEs). Knowing $X_1, \ldots, X_n$ in the regression

problem, the likelihood function for the observed $Y_1, \ldots, Y_n$ is (assuming normality)

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2\sigma^2}[Y_i - (\alpha + \beta X_i)]^2 \right\}$$

The maximum likelihood estimates of α and β are the least squares estimates a and b. For the bivariate normal distribution, the MLE of ρ is r.

9.3 Notes on the Variance of a, Variance of $a + bx$, and Choice of x for Small Variance (Experimental Design)

1. The variance of a in the regression equation $y = a + bx$ can be derived as follows: $a = \overline{y} + b\overline{x}$; it is true that $\overline{y}$ and b are statistically independent; hence,

$$\begin{aligned}
\text{var}(a) &= \text{var}(\overline{y} + b\overline{x}) \\
&= \text{var}(\overline{y}) + \overline{x}^2 \text{var}(b) \\
&= \frac{\sigma_1^2}{n} + \overline{x}^2 \frac{\sigma_1^2}{[x^2]} \\
&= \sigma_1^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{[x^2]} \right)
\end{aligned}$$

2. Consider the variance of the estimate of the mean of y at some arbitrary fixed point X:

$$\sigma_1^2 \left(\frac{1}{n} + \frac{(x - \overline{x})^2}{[x^2]} \right)$$

   a. Given a choice of x, the quantity is minimized at $x = \overline{x}$.
   b. For values of x close to $\overline{x}$ the contribution to the variance is minimal.
   c. The contribution increase as the *square* of the distance the predictor variable x is from $\overline{x}$.
   d. If there was a choice in the selection of the predictor variables, the quantity $[x^2] = \sum(x_i - \overline{x})^2$ is maximized if the predictor variables are spaced as far apart as possible. If X can have a range of values, say, X_{min} to X_{max}, the quantity $[x^2]$ is maximized if half the observations are placed at X_{min} and the other half at X_{max}. The quantity $(x - \overline{x})^2/[x^2]$ will then be as small as possible. Of course, a price is paid for this design: it is not possible to check the linearity of the relationship between Y and X.

9.4 Average-Slope Formula for b

An alternative formula for the slope estimate b emphasizes the interpretation as an average difference in Y for each unit difference in X. Suppose that we had just two points (X_1, Y_1) and (X_2, Y_2). The obvious estimate of the slope comes from simply joining the points with a line:

$$b_{21} = \frac{Y_2 - Y_1}{X_2 - X_1}$$

With more than two points we could calculate all the pairwise slope estimates

$$b_{ij} = \frac{Y_i - Y_j}{X_i - X_j}$$

and then take some summary of these as the overall slope. More weight should be give to estimates b_{ij} where $X_i - X_j$ is larger, as the expected difference in Y, $\beta(X_i - X_j)$ is larger relative to the residual error in Y_i and Y_j. If we assign weights $w_{ij} = (X_i - X_j)^2$, a little algebra shows that an alternative formula for the least squares estimate b is

$$b = \frac{\sum_{i,j} w_{ij} b_{ij}}{\sum_{i,j} w_{ij}}$$

a weighted average of the pairwise slopes.

This formulation makes it clear that b estimates the average slope of Y with respect to X under essentially no assumptions. Of course, if the relationship is not at least roughly linear, the average slope may be of little practical interest, and in any case some further assumptions are needed for statistical inference.

9.5   Regression Lines through the Origin

Suppose that we want to fit the model $Y \sim N(\beta X, \sigma^2)$, that is, the line goes through the origin. In many situations this is an appropriate model (e.g., in relating body weight to height, it is reasonable to assume that the regression line must go through the origin). However, the regression relationship may not be linear over the entire range, and often, the interval of interest is quite far removed from the origin.

Given n pairs of observation (x_i, y_i), $i = 1, \dots, n$, and a regression line through the origin is desired, it can be shown that the least squares estimate, b, of β is

$$b = \frac{\sum x_i y_i}{\sum x_i^2}$$

The residual sum of squares is based on the quantity

$$\sum (y_i - \widehat{y}_i)^2 = \sum (y_i - b x_i)^2$$

and has associated with it, $n - 1$ degrees of freedom, since only one parameter, β, is estimated.

9.6   Bivariate Normal Density Function

The formula for the density of the bivariate normal distribution is

$$f_{X,Y}(x, y) = \frac{1}{2\pi \sigma_X \sigma_Y \sqrt{1 - p^2}} \exp\left[-\frac{1}{2(1 - \rho^2)} (Z_X^2 - 2\rho Z_X Z_Y + Z_Y^2) \right]$$

where

$$Z_X = \frac{x - \mu_X}{\sigma_X} \quad \text{and} \quad Z_Y = \frac{y - \mu_Y}{\sigma_Y}$$

The quantities $\mu_X, \mu_Y, \sigma_X,$ and σ_Y are, as usual, the means and standard deviations of X and Y, respectively. Several characteristics of this distribution can be deduced from this formula:

1. If $\rho = 0$, the equation becomes

$$f_{X,Y}(x, y) = \frac{1}{2\pi \sigma_X \sigma_Y} \exp\left[-\frac{1}{2}(Z_X^2 + Z_Y^2) \right]$$

and can be written as

$$= \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left(-\frac{1}{2}Z_X^2\right) \frac{1}{\sqrt{2\pi}\sigma_Y} \exp\left(-\frac{1}{2}Z_Y^2\right)$$

$$= f_X(x)f_Y(y)$$

Thus in the case of the bivariate normal distribution, $\rho = 0$ (i.e., the correlation is zero), implies that the random variables X and Y are statistically independent.

2. Suppose that $f_{X,Y}(x, y)$ is fixed at some specified value; this implies that the expression in the exponent of the density $f_{X,Y}(x, y)$ has a fixed value, say, K:

$$K = \frac{-1}{2(1-\rho^2)} \left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\frac{x-\mu_X}{\sigma_X}\frac{y-\mu_Y}{\sigma_Y} + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]$$

This is the equation of an ellipse centered at (μ_X, μ_Y).

9.7 Ties in Kendall's Tau

When there are ties in the X_i and/or Y_i values for Kendall's tau, the variability is reduced. The asymptotic formula needs to be adjusted accordingly [Hollander and Wolfe, 1999]. Let the X_i values have g distinct values with ties with t_j tied observations at the jth tied value. Let the Y_i values have h distinct tied values with u_k tied observations at the kth tied value. Under the null hypothesis of independence between the X and Y values, the variance of K is

$$
\begin{aligned}
\text{var}(K) = {} & \frac{n(n-1)(2n+5)}{18} \\
& - \sum_{j=1}^{g} \frac{t_j(t_j-1)(2t_j+5)}{18} \\
& - \sum_{j=1}^{h} \frac{u_k(u_k-1)(2u_k+5)}{18} \\
& + \frac{\left[\sum_{j=1}^{g} t_j(t_j-1)(t_j-2)\right]\left[\sum_{k=1}^{h} u_k(u_k-1)(u_k-2)\right]}{9n(n-1)(n-2)} \\
& + \frac{\left[\sum_{j=1}^{g} t_j(t_j-1)\right]\left[\sum_{k=1}^{h} u_k(u_k-1)\right]}{2n(n-1)}
\end{aligned}
$$

The asymptotic normal Z value is

$$Z = \frac{K}{\sqrt{\text{var}(K)}}$$

Note that the null hypothesis is independence, not $\tau = 0$. If the data are not independent but nevertheless have $\tau = 0$ (e.g., a U-shaped relationship), the test will be incorrect.

9.8 Weighted Regression Analysis

In certain cases the assumption of homogeneity of variance of the dependent variable, Y, at all levels of X is not tenable. Suppose that the precision of value $Y = y$ is proportional to a value

W, the weight. Usually, the precision is the reciprocal of the variance at X_i. The data can then be modeled as follows:

Case	X	Y	W
1	x_1	y_1	w_1
2	x_2	y_2	w_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	x_i	y_i	w_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n	w_n

Define $\sum w_i(x_i - \overline{x}_i)^2 = [wx^2]$, $\sum w(x_i - \overline{x})(y_i - \overline{y}) = [wxy]$. It can be shown that the *weighted* least squares line has slope and intercept,

$$b = \frac{[wxy]}{[wx^2]} \quad \text{and} \quad a = \overline{y} - b\overline{x}$$

where

$$\overline{y} = \frac{\sum w_i y_i}{\sum w_i} \quad \text{and} \quad \overline{x} = \frac{\sum w_i x_i}{\sum w_i}$$

It is a weighted least squares solution in that the quantity $\sum w_i(y_i - \widehat{y}_i)^2$ is minimized. If all the weights are the same, say equal to 1, the ordinary least squares solutions are obtained.

9.9   Model-Robust Standard Error Estimates

We showed that Student's t-test can be formulated as a regression problem. This raises the question of whether we can also find a regression formulation of the Z-test or the unequal-variance approximate t-test of Note 5.2. The answer is in the affirmative. Standard error estimates are available that remove subsidiary assumptions such as equality of variance for a wide range of statistical estimators. These model-robust or "sandwich" standard errors were discovered independently in different fields of statistics and are typically attributed to Huber in biostatistics and to White in econometrics. The Huber–White standard error estimates are available for linear models in SAS and for nearly all regression models in State. In the case of linear regression with a binary X variable, they are equivalent to the unequal-variance t-test except that there is not complete agreement on whether n or $n - 1$ should be used as a denominator in computing variances. See Huber [2003] for further discussion.

PROBLEMS

In most of the problems below, you are asked to perform some subset of the following tasks:

(a) Plot the scatter diagram for the data.
(b) Compute for $\overline{X}$, $\overline{Y}$, $[x^2]$, $[y^2]$, and $[xy]$ those quantities not given.

(c) Find the regression coefficients a and b.

(d) Place the regression line on the scatter diagram.

(e) Give $s_{y \cdot x}^2$ and $s_{y \cdot x}$.

(f) Compute the missing predicted values, residuals, and normal deviates for the given portion of the table.

(g) Plot the residual plot.

(h) Interpret the residual plot.

(i) Plot the residual normal probability plot.

(j) Interpret the residual normal probability plot.

(k)   **i.** Construct the 90% confidence interval for β.

    ii. Construct the 95% confidence interval for β.

    iii. Construct the 99% confidence interval for β.

    iv. Compute the t-statistic for testing $\beta = 0$. What can you say about its p-value?

(l)   **i.** Construct the 90% confidence interval for α.

    ii. Construct the 95% confidence interval for α.

    iii. Construct the 99% confidence interval for α.

(m) Construct the ANOVA table and use Table A.7 to give information about the p-value.

(n) Construct the 95% confidence interval for $\alpha + \beta X$ at the X value(s) specified.

(o) Construct the interval such that one is 95% certain that a new observation at the specified X value(s) will fall into the interval.

(p) Compute the correlation coefficient r.

(q)   **i.** Construct the 90% confidence interval for ρ.

    ii. Construct the 95% confidence interval for ρ.

    iii. Construct the 99% confidence interval for ρ.

(r) Test the independence of X and Y using Spearman's rank correlation coefficient. Compute the coefficient.

(s) Test the independence of X and Y using Kendall's rank correlation coefficient. Compute the value of the coefficient.

(t) Compute Student's paired t-test for the data, if not given; in any case, interpret.

(u) Compute the signed rank statistic, if not given; in any case, interpret.

The first set of problems, 9.1 to 9.4, come from the exercise data in Example 9.2.

9.1 Suppose that we use duration, X, to predict $VO_{2\ MAX}$, Y. The scatter diagram is shown in Figure 9.2. $\overline{X} = 647.4$, $\overline{Y} = 40.57$, $[x^2] = 673,496.4$, $[y^2] = 3506.2$, and $[xy] = 43,352.5$. Do tasks (c), (e), (f), (h), (k-ii), (k-iv), (l-ii), (m), (n) at $x = 650$, (p), and (q-ii) (the residual plot is Figure 9.17). All the data are listed in Table 9.13. What proportion of the Y variance is explained by X? (In practice, duration is used as a reasonable approximation to $VO_{2\ MAX}$.)
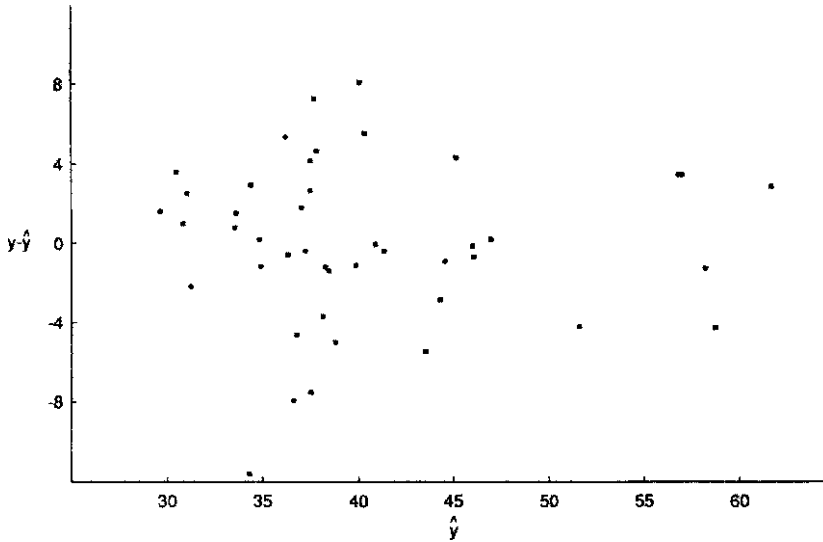
Figure 9.17 Residual plot for the data of Example 9.2; $VO_{2 \ MAX}$ predicted from duration.

Table 9.13 Oxygen Data for Problem 9.1

X	Y	$\widehat{Y}$	$Y - \widehat{Y}$	Normal Deviate
706	41.5	44.5	−3.0	−0.80
732	45.9	46.13	−0.23	−0.06
930	54.5	?	?	?
900	60.3	?	3.59	0.96
903	60.5	56.90	3.60	0.97
976	64.6	61.50	3.10	0.83
819	47.4	?	−4.21	−1.13
922	57.0	58.10	−1.10	−0.29
600	40.2	37.82	?	0.64
540	35.2	?	1.16	0.31
560	33.8	35.30	−1.50	?
637	38.8	40.15	−1.35	−0.36
593	38.9	?	1.52	0.41
719	49.5	45.31	?	1.23
615	37.1	38.77	−1.67	−0.45
589	32.2	37.13	?	−1.32
478	31.3	30.14	1.16	0.31
620	33.8	39.08	−5.28	?
710	43.7	44.75	−1.05	−0.28
600	41.7	37.82	3.88	1.04
660	41.0	41.60	−0.60	−0.16

9.2 One expects exercise performance to reduce with age. In this problem, X = age and Y = duration. $\overline{X} = 47.2, \overline{Y} = 647.4, [x^2] = 4303.2, [y^2] = 673,496.4$, and $[xy] = -36,538.5$. Do tasks (c), (e), (k-i), (l-i), (p), and (q-i).

9.3 To see if maximum heart rate changes with age, the following numbers are found where X = age and Y = maximum heart rate. $\overline{X} = 47.2, \overline{Y} = 174.8, [x^2] = 4303.2, [y^2] = 5608.5$, and $[xy] = -2915.4$. Do tasks (c), (e), (k-iii), (k-iv), (m), (p), and (p-iii).

9.4 The relationship between height and weight was examined in these active healthy males. X = height, Y = weight, $\overline{X} = 177.7, \overline{Y} = 77.8, [x^2] = 1985.2, [y^2] = 3154.5$, and $[xy] = 1845.6$. Do tasks (c), (e), (m), (p), and (q-i). How do the p-values for the F-test [in part (m)] and for the transformed Z for r compare? There were two normal deviates of values 3.44 and 2.95. If these two people were removed from the calculation, $\overline{X} = 177.5, \overline{Y} = 76.7, [x^2] = 1944.5, [y^2] = 2076.12$, and $[xy] = 1642.5$. How much do the regression coefficients a and b, and correlation coefficient r, change?

Problems 9.5 to 9.8 also refer to the Bruce et al. [1973] paper, as did Example 9.2 and Problems 9.1 to 9.4. The data for 43 active females are given in Table 9.14.

9.5 The duration and VO$_2$ MAX relationship for the active females is studied in this problem. $\overline{X} = 514.9, \overline{Y} = 29.1, [x^2] = 251,260.4, [y^2] = 1028.7$, and $[xy] = 12,636.5$. Do tasks (c), (e), (f), (g), (h), (i), (j), (k-iv), (m), (p), and (q-ii). Table 9.15 contains the residuals. If the data are rerun with the sixth case omitted, the values of $\overline{X}, \overline{Y}, [x^2], [y^2]$, and $[xy]$ are changed to 512.9, 29.2, 243,843.1, 1001.5, and 13,085.6, respectively. Find the new estimates a, b, and r. By what percent are they changed?

9.6 With X = age and Y = duration, $\overline{X} = 45.1, \overline{Y} = 514.9, [x^2] = 4399.2, [y^2] = 251,260.4$, and $[xy] = -22,911.3$. For each 10-year increase in age, how much does duration tend to change? What proportion of the variability in VO$_2$ MAX is accounted for by age? Do tasks (m) and (q-ii).

9.7 With X = age and Y = maximum heart rate, $\overline{X} = 45.1, \overline{Y} = 180.6, [x^2] = 4399.2, [y^2] = 5474.6$, and $[xy] = -2017.3$. Do tasks (c), (e), (k-i), (k-iv), (l-i), (m), (n) at $X = 30$ and $X = 50$, (o) at $X = 45$, (p), and (q-ii).

9.8 X = height and Y = weight, $\overline{X} = 164.7, \overline{Y} = 61.3, [x^2] = 1667.1, [y^2] = 2607.4$, and $[xy] = 1006.2$. Do tasks (c), (e), (h), (k-iv), (m), and (p). Check that $t^2 = F$. The residual plot is shown in Figure 9.18.

For Problems 9.9 to 9.12, additional Bruce et al. [1973] data are used. Table 9.16 presents the data for 94 sedentary males.

9.9 The duration, X, and VO$_2$ MAX, Y, give $\overline{X} = 577.1, \overline{Y} = 35.6, [x^2] = 1,425,990.9, [y^2] = 5245.3$, and $[xy] = 78,280.1$. Do tasks (c), (e), (j), (k-i), (k-iv), (l-i), (m), and (p). The normal probability plot is shown in Figure 9.19.

9.10 X = age is related to Y = duration. $\overline{X} = 49.8, \overline{Y} = 577.1, [x^2] = 11,395.7, [y^2] = 1,425,990.9$, and $[xy] = -87,611.9$. Do tasks (c), (e), (m), (p), and (q-ii).

9.11 The prediction of age by maximal heart rate for sedentary males is considered here. $\overline{X} = 49.8, \overline{Y} = 18.6, [x^2] = 11,395.7, [y^2] = 32,146.4$, and $[xy] = -12,064.1$. Do tasks (c), (m), and (p). Verify (to accuracy given) that $(\overline{X}, \overline{Y})$ lies on the regression line.

9.12 The height and weight data give $\overline{X} = 177.3, \overline{Y} = 79.0, [x^2] = 4030.1, [y^2] = 7060.0$, and $[xy] = 2857.0$. Do tasks (c), (e), (k-iv), (n) at $X = 160, 170$, and 180, and (p).

Table 9.14   Exercise Data for Healthy Active Females

Duration	VO$_2$ MAX	Heart Rate	Age	Height	Weight
660	38.1	184	23	177	83
628	38.4	183	21	163	52
637	41.7	200	21	174	61
575	33.5	170	42	160	50
590	28.6	188	34	170	68
600	23.9	190	43	171	68
562	29.6	190	30	172	63
495	27.3	180	49	157	53
540	33.2	184	30	178	63
470	26.6	162	57	161	63
408	23.6	188	58	159	54
387	23.1	170	51	162	55
564	36.6	184	32	165	57
603	35.8	175	42	170	53
420	28.0	180	51	158	47
573	33.8	200	46	161	60
602	33.6	190	37	173	56
430	21.0	170	50	161	62
508	31.2	158	65	165	58
565	31.2	186	40	154	69
464	23.7	166	52	166	67
495	24.5	170	40	160	58
461	30.5	188	52	162	64
540	25.9	190	47	161	72
588	32.7	194	43	164	56
498	26.9	190	48	176	82
483	24.6	190	43	165	61
554	28.8	188	45	166	62
521	25.9	184	52	167	62
436	24.4	170	52	168	62
398	26.3	168	56	162	66
366	23.2	175	56	159	56
439	24.6	156	51	161	61
549	28.8	184	44	154	56
360	19.6	180	56	167	79
566	31.4	184	40	165	56
407	26.6	156	53	157	52
602	30.6	194	52	161	65
488	27.5	190	40	178	64
526	30.9	188	55	162	61
524	33.9	164	39	166	59
562	32.3	185	57	168	68
496	26.9	178	46	156	53

Source: Data from Bruce et al. [1973].

Mehta et al. [1981] studied the effect of the drug dipyridamole on blood platelet function in eight patients with at least 50% narrowing of one or more coronary arteries. Active platelets are sequestered in the coronary arteries, giving reduced platelet function in the coronary venous blood, that is, in blood leaving the heart muscle after delivering oxygen and nutrients. More active platelets in the coronary arteries can lead to thrombosis, blood clots, and a heart attack. Drugs lessening the chance of thrombosis may be useful in treatment.

Table 9.15   Data for Problem 9.5

X	Y	$\widehat{Y}$	Residual	Normal Deviate
660	38.1	36.35	1.75	0.56
628	38.4	34.74	3.66	1.18
637	41.7	35.19	6.51	2.10
575	33.5	32.08	1.42	0.46
590	28.6	32.83	−4.23	−1.37
600	23.9	?	?	?
562	29.6	31.42	−1.82	−0.59
495	27.3	28.05	−0.75	−0.24
540	33.2	?	2.88	0.93
470	26.6	26.80	−0.20	−0.06
408	23.6	23.68	−0.07	−0.02
387	23.1	22.62	0.48	0.15
564	36.6	31.52	5.08	1.64
603	35.8	33.49	2.21	0.75
420	28.0	24.28	3.72	1.20
573	33.8	?	?	0.59
602	33.6	33.43	0.17	0.05
430	21.0	24.78	−3.78	?
508	31.2	28.71	2.49	?
565	31.2	31.57	−0.37	−0.12
464	23.7	26.49	−2.79	−0.90
495	24.5	28.05	−3.55	−1.10
461	30.5	26.34	4.16	1.34
540	25.9	30.32	−4.42	−1.43
588	32.7	?	−0.03	−0.00
498	26.9	?	−1.30	−0.42
483	24.6	27.45	−2.85	−0.92
554	28.8	31.02	−2.22	−0.72
521	25.9	29.36	−3.46	−1.12
436	24.4	25.09	−0.69	−0.22
398	26.3	23.18	3.12	1.01
366	23.2	21.57	1.63	0.53
439	24.6	25.24	−0.64	−0.21
549	28.8	30.77	−1.97	−0.64
360	19.6	21.26	−1.66	−0.54
566	31.4	31.62	−0.22	−0.07
407	26.6	23.63	2.97	0.96
602	30.6	33.43	−2.83	−0.92
488	27.5	27.70	−0.20	−0.06
526	30.9	29.61	1.29	0.42
524	33.9	29.51	4.39	1.42
562	32.3	31.42	0.88	0.28
496	26.9	28.10	−1.20	−0.39

Platelet aggregation measures the extent to which platelets aggregate or cluster together in the presence of a chemical that stimulates clustering or aggregation. The measure used was the percent increase in light transmission after an aggregating agent was added to plasma. (The clustering of the cells make more "holes" in the plasma to let light through.) Two aggregating agents, adenosine diphosphate (ADP) and epinephrine (EPI), were used in this experiment. A second measure taken from the blood count was the count of platelets.
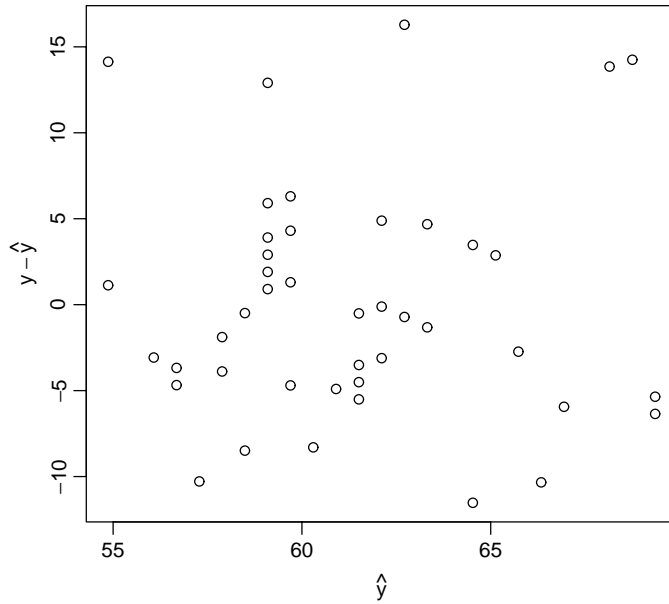
Figure 9.18 Residual plot for Problem 9.8.
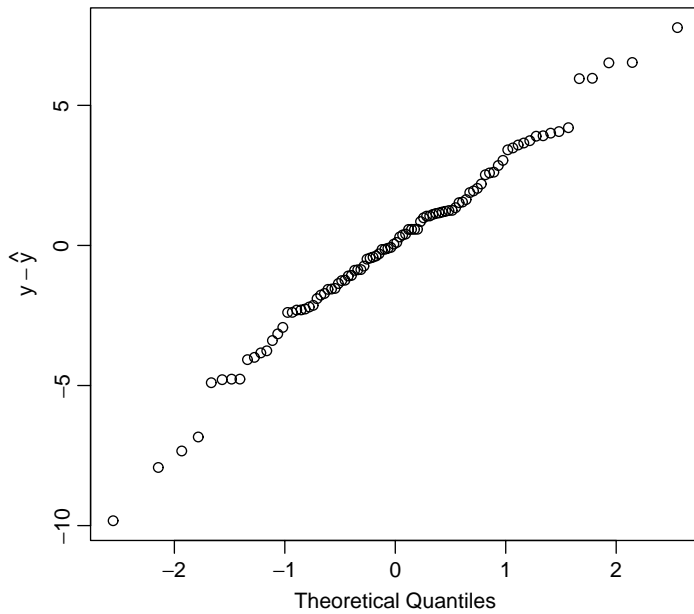
Figure 9.19 Normal probability plot for Problem 9.9.

Blood was sampled from two sites, the aorta (blood being pumped from the heart) and the coronary sinus (blood returning from nourishing the heart muscle). Control samples as well as samples after intravenous infusion of 100 mg of dipyridamole were taken. The data are given in Table 9.17 and 9.18. Problems 9.13 to 9.22 refer to these data.

Table 9.16   Exercise Data for Sedentary Males

Duration	VO_2 MAX	Heart Rate	Age	Height	Weight
360	24.7	168	40	175	96
770	46.8	190	25	168	68
663	41.2	175	41	187	82
679	31.4	190	37	176	82
780	45.7	200	26	179	73
727	47.6	210	28	185	84
647	38.6	208	26	177	77
675	43.2	200	42	162	72
735	48.2	196	30	188	85
827	50.9	184	21	178	69
760	47.2	184	33	182	87
814	41.8	208	31	182	82
778	42.9	184	29	174	73
590	35.1	174	42	188	93
567	37.6	176	40	184	86
648	47.3	200	40	168	80
730	44.4	204	44	183	78
660	46.7	190	44	176	81
663	41.6	184	40	174	78
589	40.2	200	43	193	92
600	35.8	190	41	176	68
480	30.2	174	44	172	84
630	38.4	164	39	181	72
646	41.3	190	39	187	90
630	31.2	190	42	173	69
630	42.6	190	53	181	53
624	39.4	172	57	172	57
572	35.4	164	58	181	58
622	35.9	190	61	168	61
209	16.0	104	74	171	74
536	29.3	175	57	181	57
602	36.7	175	49	175	49
727	43.0	168	53	172	53
260	15.3	112	75	170	75
622	42.3	175	47	185	47
705	43.7	174	51	169	51
669	40.3	174	65	170	65
425	28.5	170	56	167	56
645	38.0	175	50	177	50
576	30.8	184	48	188	48
605	40.2	156	46	187	46
458	29.5	148	61	185	61
551	32.3	188	49	182	49
607	35.5	179	53	179	53
599	35.3	166	55	182	55
453	32.3	160	69	182	69
337	23.8	204	68	176	68
663	41.4	182	47	171	47
603	39.0	180	48	180	48
610	38.6	190	55	180	55
472	31.5	175	53	192	85

Table 9.16   (*continued*)

Duration	VO$_2$ $_{MAX}$	Heart Rate	Age	Height	Weight
458	25.7	166	58	178	81
446	24.6	160	50	178	77
532	30.0	160	51	175	82
656	42.0	186	52	176	73
583	34.4	175	52	172	77
595	34.9	180	48	179	78
552	35.5	156	45	167	89
675	38.7	162	58	183	85
622	38.4	186	45	175	76
591	32.4	170	62	175	79
582	33.6	156	63	171	69
518	30.0	166	57	174	75
444	28.9	170	48	180	105
473	29.5	175	52	177	77
490	30.4	168	59	173	74
596	34.4	192	46	190	92
529	37.0	175	54	168	82
652	43.4	156	54	180	85
714	46.0	175	46	174	77
646	43.0	184	45	178	80
551	29.3	160	54	172	86
601	36.8	184	48	169	82
579	35.0	170	54	180	80
325	21.9	140	61	175	76
392	25.4	168	60	180	89
659	40.7	178	45	181	81
631	33.8	184	48	173	74
405	28.8	170	63	168	79
560	35.8	180	60	181	82
615	40.3	190	47	178	78
580	33.4	180	66	173	68
530	39.0	174	47	169	64
495	23.2	145	69	171	84
330	20.5	138	60	185	87
600	36.4	200	50	182	81
443	23.5	166	50	175	84
508	29.7	188	61	188	80
596	43.2	168	57	174	66
461	30.4	170	47	171	65
583	34.7	164	46	187	83
620	37.1	174	61	165	71
620	41.4	190	45	171	79
180	19.8	125	71	185	80

Source: Data from Bruce et al. [1973]

9.13 Relate the control platelet counts in the aorta, X, and coronary sinus, Y. Do tasks (a), (b), (c), (d), (e), compute the (X, Y, $\widehat{Y}$, residual, normal deviate) table, (g), (h), (i), (j), (k-i), (k-iv), (l), (m), (p), (r), and (s).

9.14 Look at the association between the platelet counts in the aorta, X, and coronary sinus, Y, when being treated with dipyridamole. Do tasks (a), (b), (c), (d), (m), (r), and (s).

Table 9.17   Platelet Aggregation Data for Problem 9.12

	Platelet Aggregation (%)							
	Control				Dipyridamole			
	Aorta		Coronary Sinus		Aorta		Coronary Sinus	
Case	EPI	ADP	EPI	ADP	EPI	ADP	EPI	ADP
1	87	75	89	23	89	75	89	35
2	70	23	42	14	45	16	47	18
3	96	75	96	31	96	83	96	84
4	65	51	70	33	70	55	70	57
5	85	16	79	4	69	13	53	22
6	98	83	98	80	83	70	94	88
7	77	14	97	13	84	35	73	67
8	98	50	99	40	85	50	91	48
Mean	85	48	84	30	78	50	77	52
±SEM	5	10	7	8	6	9	7	9

Source: Data from Mehta et al. [1981].

9.15 Examine the control platelet aggregation percent for EPI, X, and ADP, Y, in the aorta. Do tasks (a), (b), (c), (d), (e), and (m).

9.16 Examine the association between the EPI, X, and ADP, Y, in the control situation at the coronary sinus. Do tasks (a), (b), (c), (d), (e), (m), (p), (r), and (s).

9.17 Interpret at the 5% significance level. Look at the platelet aggregation % for epinephrine in the aorta and coronary sinus under the control data. Do tasks (m), (p) and (t), (u). Explain in words how there can be association but no (statistical) difference between the values at the two locations.

9.18 Under dipyridamole treatment, study the platelet aggregation percent for EPI in the aorta, X, and coronary sinus, Y. Do tasks (a), (b), (c), (d), (e), (g), (h), (m), (p), (r), (s), (t), and (u).

9.19 The control aggregation percent for ADP is compared in the aorta, X, and coronary sinus, Y, in this problem. Do tasks (a), (b), (c), (d), (e), (f), (g), (h), (i), (j), (m), (p), and (q-ii).

9.20 Under dipyridamole, the aggregation percent for ADP in the aorta, X, and coronary sinus, Y, is studied here. Do tasks (b), (c), (e), (k-ii), (k-iv), (l-ii), (m), (p), (q-ii), (r), and (s).

9.21 The aortic platelet counts under the control, X, and dipyridamole, Y, are compared in this problem. Do tasks (b), (c), (e), (m), (p), (q-ii), (t), and (u). Do the platelet counts differ under the two treatments? (Use $\alpha = 0.05$.) Are the platelet counts associated under the two treatments? ($\alpha = 0.05$.)

9.22 The coronary sinus ADP aggregation percent was studied during the control period, the X variable, and on dipyridamole, the Y variable. Do tasks (b), (c), (d), (e), (m), and (t). At the 5% significance level, is there a change between the treatment and control periods? Can you show association between the two values? How do you reconcile these findings?

Table 9.18   Platelet Count Data for Problem 9.12

	Platelet Counts $(\times 1000/mm^3)^a$			
	Control		Dipyridamole	
Case	Aorta	Coronary Sinus	Aorta	Coronary Sinus
1	390	355	455	445
2	240	190	248	205
3	135	125	150	145
4	305	268	285	290
5	255	195	230	220
6	283	307	291	312
7	435	350	457	374
8	290	250	301	284
Mean	292	255	302	284
±SEM	32	29	38	34

Source: Data from Mehta et al. [1981].

Problems 9.23 to 9.29 deal with the data in Tables 9.19 and 9.20. Jensen et al. [1980] studied 19 patients with coronary artery disease. Thirteen had a prior myocardial infarction (heart attack); three had coronary bypass surgery. The patients were evaluated before and after three months or more on a structured supervised training program.

The cardiac performance was evaluated using radionuclide studies while the patients were at rest and also exercising with bicycle pedals (while lying supine). Variables measured included (1) ejection fraction (EF), the fraction of the blood in the left ventricle ejected during a heart beat, (2) heart rate (HR) at maximum exercise in beats per minute, (3) systolic blood pressure (SBP) in millimeters of mercury, (4) the rate pressure product (RPP) maximum heart rate times the maximum systolic blood pressure divided by 100, and (5) the estimated maximum oxygen consumption in cubic centimeters of oxygen per kilogram of body weight per minute.

9.23 The resting ejection fraction is measured before, X, and after, Y, training. $\overline{X} = 0.574, \overline{Y} = 0.553, [x^2] = 0.29886, [y^2] = 0.32541, [xy] = 0.23385$, and paired $t = -0.984$. Do tasks (c), (e), (k-iv), (m), and (p). Is there a change in resting ejection fraction demonstrated with six months of exercise training? Are the two ejection fractions associated?

9.24 The ejection fraction at maximal exercise was measured before, X, and after, Y, training. $\overline{X} = 0.556, \overline{Y} = 0.564, [x^2] = 0.30284, [y^2] = 0.46706,$ and $[xy] = 0.2809$. Is there association ($\alpha = 0.05$) between the two ejection fractions? If yes, do tasks (c), (k-iii), (l-iii), (p), and (q-ii). Is there a change ($\alpha = 0.05$) between the two ejection fractions? If yes, find a 95% confidence interval for the average difference.

9.25 The maximum systolic blood pressure was measured before, X, and after, Y, training. $\overline{X} = 173.8, \overline{Y} = 184.2, [x^2] = 11,488.5, [y^2] = 10,458.5, [xy] = 7419.5$, and paired $t = 2.263$. Do tasks (a), (b), (c), (d), (e), (m), (p), and (t). Does the exercise training produce a change? How much? Can we predict individually the maximum SBP after training from that before? How much of the variability in maximum SBP after exercise is accounted for by knowing the value before exercise?

9.26 The before, X, and after, Y, rate pressure product give $\overline{X} = 223.0, \overline{Y} = 245.7, [x^2] = 58,476, [y^2] = 85,038, [xy] = 54,465$, and paired $t = 2.256$ (Table 9.21). Do tasks (c), (e), (f), (g), (h), and (m). Find the large-sample p-value for Kendall's tau for association.

Table 9.19   Resting and Maximal Ejection Fraction Measured by Radionuclide Ventriculography, and Maximal Heart Rate

Case	Resting EF Pre	Resting EF Post	Maximal EF Pre	Maximal EF Post	Maximal HR Pre	Maximal HR Post
1	0.39	0.48	0.46	0.48	110	119
2	0.57	0.49	0.51	0.57	120	125
3	0.77	0.63	0.70	0.82	108	105
4	0.48	0.50	0.51	0.51	85	88
5	0.55	0.46	0.45	0.55	107	103
6	0.60	0.50	0.52	0.54	125	115
7	0.63	0.61	0.75	0.68	170	166
8	0.73	0.61	0.53	0.71	160	142
9	0.70	0.68	0.80	0.79	125	114
10	0.66	0.68	0.54	0.43	131	150
11	0.40	0.31	0.42	0.30	135	174
12	0.48	0.46	0.48	0.30	97	94
13	0.63	0.78	0.60	0.75	135	132
14	0.41	0.37	0.41	0.44	127	162
15	0.75	0.54	0.76	0.57	126	148
16	0.58	0.64	0.62	0.72	102	112
17	0.50	0.58	0.54	0.65	145	140
18	0.71	0.81	0.65	0.60	152	145
19	0.37	0.38	0.32	0.31	155	170
Mean	0.57	0.55	0.56	0.56	127	132
±SD	0.13	0.13	0.13	0.16	23	26

Table 9.20   Systolic Blood Pressure, Rate Pressure Product and Estimate VO$_2$ $_{MAX}$ before (Pre) and after (Post) Training

Case	Maximal SBP Pre	Maximal SBP Post	Maximal RPP Pre	Maximal RPP Post	Est. VO$_2$ $_{MAX}$ (cm^3/kg · min) Pre	Est. VO$_2$ $_{MAX}$ (cm^3/kg · min) Post
1	148	156	163	186	24	30
2	180	196	216	245	28	44
3	185	200	200	210	28	28
4	150	148	128	130	34	38
5	150	156	161	161	20	28
6	164	172	205	198	30	36
7	180	210	306	349	64	54
8	182	176	291	250	44	40
9	186	170	233	194	30	28
10	220	230	288	345	30	30
11	188	205	254	357	28	44
12	120	165	116	155	22	20
13	175	160	236	211	20	36
14	190	180	241	292	36	38
15	140	170	176	252	36	44
16	200	230	204	258	28	36
17	215	185	312	259	44	44
18	165	190	251	276	28	34
19	165	200	256	340	44	52
Mean	174	184	223	246	31	37
±SD	25	24	57	69	8	9

Table 9.21   Blood Pressure Data for Problem 9.26

| Maximal SBP | | | | |
| Pre | Post | | | |
X	Y	$\widehat{Y}$	$Y - \widehat{Y}$	Normal Deviate
163	186	189.90	−3.80	−0.08
216	245	239.16	?	?
200	210	224.26	−14.26	−0.32
128	130	157.20	−27.20	−0.61
161	161	?	−26.94	?
205	198	228.92	−30.92	−0.69
306	349	322.99	26.01	?
291	250	309.02	−59.02	−1.31
233	194	255.00	−61.00	−1.36
288	345	306.22	38.77	0.86
254	357	?	?	?
116	155	146.02	8.98	0.20
236	211	257.79	−46.79	−1.04
241	292	262.45	29.55	0.66
176	252	201.91	50.09	1.12
204	258	227.99	30.01	0.67
312	259	328.58	−69.58	−1.55
251	276	271.76	4.24	0.09
256	340	276.42	63.58	1.42

9.27 The maximum oxygen consumption, $VO_{2\ MAX}$, is measured before, X, and after, Y. Here $\overline{X} = 32.53, \overline{Y} = 37.05, [x^2] = 2030.7, [y^2] = 1362.9, [xy] = 54465$, and paired $t = 2.811$. Do tasks (c), (k-ii), (m), (n), at $x = 30$, 35, and 40, (p), (q-ii), and (t).

9.28 The ejection fractions at rest, X, and at maximum exercise, Y, before training is used in this problem. $\overline{X} = 0.574, \overline{Y} = 0.556, [x^2] = 0.29886, [y^2] = 0.30284, [xy] = 0.24379$, and paired $t = -0.980$. Analyze these data, including a scatter diagram, and write a short paragraph describing the change and/or association seen.

9.29 The ejection fractions at rest, X, and after exercises, Y, for the subjects after training: (1) are associated, (2) do not change on the average, (3) explain about 52% of the variability in each other. Justify statements (1)–(3). $\overline{X} = 0.553, \overline{Y} = 0.564, [x^2] = 0.32541, [y^2] = 0.4671, [xy] = 0.28014$, and paired $t = 0.424$.

Problems 9.30 to 9.33 refer to the following study. Boucher et al. [1981] studied patients before and after surgery for isolated aortic regurgitation and isolated mitral regurgitation. The aortic valve is in the heart valve between the left ventricle, where blood is pumped from the heart, and the aorta, the large artery beginning the arterial system. When the valve is not functioning and closing properly, some of the blood pumped from the heart returns (or regurgitates) as the heart relaxes before its next pumping action. To compensate for this, the heart volume increases to pump more blood out (since some of it returns). To correct for this, open heart surgery is performed and an artificial valve is sewn into the heart. Data on 20 patients with aortic regurgitation and corrective surgery are given in Tables 9.22 and 9.23.

"NYHA Class" measures the amount of impairment in daily activities that the patient suffers: I is least impairment, II is mild impairment, III is moderate impairment, and IV is severe impairment; HR, heart rate; SBP, the systolic (pumping or maximum) blood pressure; EF, the ejection fraction, the fraction of blood in the left ventricle pumped out during a beat; EDVI,

Table 9.22   Preoperative Data for 20 Patients with Aortic Regurgitation

Case	Age (yr) and Gender	NYHA Class	HR (beats/min)	SBP (mmHG)	EF	EDVI (mL/m^2)	SVI (mL/m^2)	ESVI (mL/m^2)
1	33M	I	75	150	0.54	225	121	104
2	36M	I	110	150	0.64	82	52	30
3	37M	I	75	140	0.50	267	134	134
4	38M	I	70	150	0.41	225	92	133
5	38M	I	68	215	0.53	186	99	87
6	54M	I	76	160	0.56	116	65	51
7	56F	I	60	140	0.81	79	64	15
8	70M	I	70	160	0.67	85	37	28
9	22M	II	68	140	0.57	132	95	57
10	28F	II	75	180	0.58	141	82	59
11	40M	II	65	110	0.62	190	118	72
12	48F	II	70	120	0.36	232	84	148
13	42F	III	70	120	0.64	142	91	51
14	57M	III	85	150	0.60	179	107	30
15	61M	III	66	140	0.56	214	120	94
16	64M	III	54	150	0.60	145	87	58
17	61M	IV	110	126	0.55	83	46	37
18	62M	IV	75	132	0.56	119	67	52
19	64M	IV	80	120	0.39	226	88	138
20	65M	IV	80	110	0.29	195	57	138
Mean	49		75	143	0.55	162	85	77
±SD	14		14	25	0.12	60	26	43

Table 9.23   Postoperative Data for 20 Patients with Aortic Regurgitation

Case	Age (yr) and Gender	NYHA Class	HR (beats/min)	SBP (mmHG)	EF	EDVI (mL/m^2)	SVI (mL/m^2)	ESVI (mL/m^2)
1	33M	I	80	115	0.38	113	43	43
2	36M	I	100	125	0.58	56	32	24
3	37M	I	100	130	0.27	93	25	68
4	38M	I	85	110	0.17	160	27	133
5	38M	I	94	130	0.47	111	52	59
6	54M	I	74	110	0.50	83	42	42
7	56F	I	85	120	0.56	59	33	26
8	70M	I	85	130	0.59	68	40	28
9	22M	II	120	136	0.33	119	39	80
10	28F	II	92	160	0.32	71	23	48
11	40M	II	85	110	0.47	70	33	37
12	48F	II	84	120	0.24	149	36	113
13	42F	III	84	100	0.63	55	35	20
14	57M	III	86	135	0.33	91	72	61
15	61M	III	100	138	0.34	92	31	61
16	64M	III	60	130	0.30	118	35	83
17	61M	IV	88	130	0.62	63	39	24
18	62M	IV	75	126	0.29	100	29	71
19	64M	IV	78	110	0.26	198	52	147
20	65M	IV	75	90	0.26	176	46	130
Mean	49		87	123	0.40	102	38	65
±SD	14		13	15	0.14	41	11	39

Table 9.24   Preoperative Data for 20 Patients with Mitral Regurgitation

| Case	Age (yr) and Gender	NYHA Class	HR (beats/min)	SBP (mmHG)	EF	EDVI (mL/m^2)	SVI (mL/m^2)	ESVI (mL/m^2)
1	23M	II	75	95	0.69	71	49	22
2	31M	II	70	150	0.77	184	142	42
3	40F	II	86	90	0.68	84	57	30
4	47M	II	120	150	0.51	135	67	66
5	54F	II	85	120	0.73	127	93	34
6	57M	II	80	130	0.74	149	110	39
7	61M	II	55	120	0.67	196	131	65
8	37M	III	72	120	0.70	214	150	64
9	52M	III	108	105	0.66	126	83	43
10	52F	III	80	115	0.52	167	70	97
11	52M	III	80	105	0.76	130	99	31
12	56M	III	80	115	0.60	136	82	54
13	58F	III	65	110	0.62	146	91	56
14	59M	III	102	90	0.63	82	52	30
15	66M	III	60	100	0.62	76	47	29
16	67F	III	75	140	0.71	94	67	27
17	71F	III	88	140	0.65	111	72	39
18	55M	IV	80	125	0.66	136	90	46
19	59F	IV	115	130	0.72	96	69	27
20	60M	IV	64	140	0.60	161	97	64
Mean	53		81	121	0.66	131	86	45
±SD	12		17	17	0.09	40	30	19

Table 9.25   Postoperative Data for 20 Patients with Mitral Regurgitation

Case	Age (yr) and Gender	NYHA Class	HR (beats/min)	SBP (mmHG)	EF	EDVI (mL/m^2)	SVI (mL/m^2)	ESVI (mL/m^2)
1	23M	II	90	100	0.60	67	40	27
2	31M	II	95	110	0.64	64	41	23
3	40F	II	80	110	0.77	59	45	14
4	47M	II	90	120	0.36	96	35	61
5	54F	II	100	110	0.41	59	24	35
6	57M	II	75	115	0.54	71	38	33
7	61M	II	140	120	0.41	165	68	97
8	37M	III	95	120	0.25	84	21	63
9	52M	III	100	125	0.43	67	29	38
10	52F	III	90	90	0.44	124	55	69
11	52M	III	98	116	0.55	68	37	31
12	56M	III	61	108	0.56	112	63	49
13	58F	III	88	120	0.50	76	38	38
14	59M	III	100	100	0.48	40	19	21
15	66M	III	85	124	0.51	31	16	15
16	67F	III	84	120	0.39	81	32	49
17	71F	III	100	100	0.44	76	33	43
18	55M	IV	108	124	0.43	63	27	36
19	59F	IV	100	110	0.49	62	30	32
20	60M	IV	90	110	0.36	93	34	60
Mean	53		93	113	0.48	78	36	42
±SD	12		15	9	0.11	30	14	21

the volume of the left ventricle after the heart relaxes (adjusted for physical size, to divide by an estimate of the patient's body surface area (BSA); SVI, the volume of the left ventricle after the blood is pumped out, adjusted for BSA; ESVI, the volume of the left ventricle pumped out during one cycle, adjusted for BSA; ESVI = EDVI − SVI. These values were measured before and after valve replacement surgery. The patients in this study were selected to have left ventricular volume overload; that is, expanded EDVI.

Another group of 20 patients with mitral valve disease and left ventricular volume overload were studied. The mitral valve is the valve allowing oxygenated blood from the lungs into the left ventricle for pumping to the body. Mitral regurgitation allows blood to be pumped "backward" and to be mixed with "new" blood coming from the lungs. The data for these patients are given in Tables 9.24 and 9.25.

9.30 **(a)** The preoperative, X, and postoperative, Y, ejection fraction in the patients with aortic valve replacement gave $\overline{X} = 0.549, \overline{Y} = 0.396, [x^2] = 0.26158, [y^2] = 0.39170, [xy] = 0.21981$, and paired $t = -6.474$. Do tasks (a), (c), (d), (e), (m), (p), and (t). Is there a change? Are ejection fractions before and after surgery related?

(b) The mitral valve cases had $\overline{X} = 0.662, \overline{Y} = 0.478, [x^2] = 0.09592, [y^2] = 0.24812, [xy] = 0.04458$, and paired $t = -7.105$. Perform the same tasks as in part (a).

(c) When the emphasis is on the change, rather than possible association and predictive value, a figure like Figure 9.20 may be preferred to a scatter diagram. Plot the scatter diagram for the aortic regurgitation data and comment on the relative merits of the two graphics.
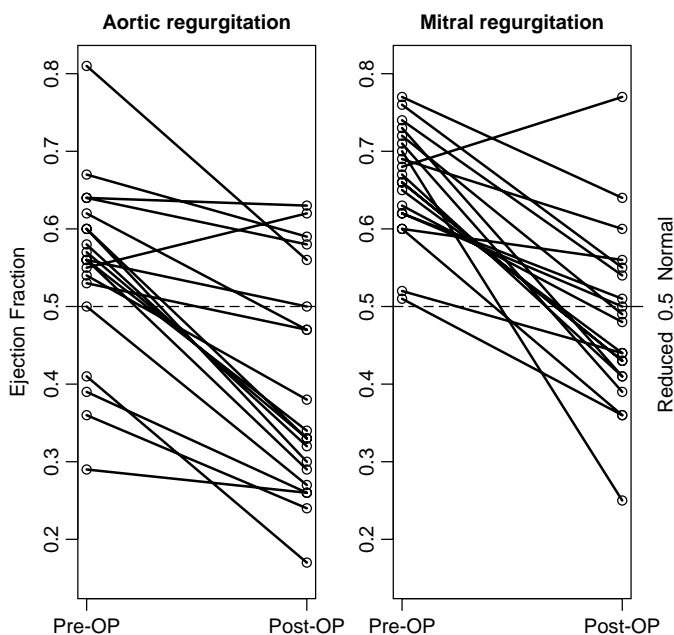
Figure 9.20 Figure for Problem 9.30(c). Individual values for ejection fraction before (pre-OP) and early after (post-OP) surgery are plotted; preoperatively, only four patients with aortic regurgitation had an ejection fraction below normal. After operation, 13 patients with aortic regurgitation and 9 with mitral regurgitation had an ejection fraction below normal. The lower limit of normal (0.50) is represented by a dashed line. (From Boucher et al. [1981].).

Table 9.26 Data for Problem 9.31

| X	Y	$\widehat{Y}$	Residuals	Normal Deviate
22	67	51.26	15.74	0.75
42	64	74.18	−10.18	−0.48
30	59	60.42	−1.42	−0.06
66	96	101.68	−5.68	−0.27
34	59	65.01	−6.01	−0.28
39	71	70.74	0.26	0.01
65	165	?	?	?
64	84	99.39	15.29	−0.73
43	67	75.32	?	−0.39
97	124	137.20	−13.20	?
31	68	61.57	?	?
54	112	87.93	24.07	1.14
56	76	?	?	−0.67
30	40	?	−20.42	−0.97
29	31	?	?	?
27	81	56.99	24.01	1.14
39	76	70.74	5.26	0.25
46	63	78.76	−15.76	−0.75
27	62	56.99	5.01	0.24
64	93	99.39	−6.39	−0.30

9.31 **(a)** For the mitral valve cases, we use the end systolic volume index (ESVI) before surgery to try to predict the end diastolic volume index (EDVI) after surgery. $\overline{X} = 45.25, \overline{Y} = 77.9, [x^2] = 6753.8, [y^2] = 16, 885.5$, and $[xy] = 7739.5$. Do tasks (c), (d), (e), (f), (h), (j), (k-iv), (m), and (p). Data are given in Table 9.26. The residual plot and normal probability plot are given in Figures 9.21 and 9.22.

(b) If subject 7 is omitted, $\overline{X} = 44.2, \overline{Y} = 73.3, [x^2] = 6343.2, [y^2] = 8900.1$, and $[xy] = 5928.7$. Do tasks (c), (m), and (p). What are the changes in tasks (a), (b), and (r) from part (a)?

(c) For the aortic cases; $\overline{X} = 75.8, \overline{Y} = 102.3, [x^2] = 35,307.2, [y^2] = 32,513.8, [xy] = 27, 076$. Do tasks (c), (k-iv), (p), and (q-ii).

9.32 We want to investigate the predictive value of the preoperative ESVI to predict the postoperative ejection fraction, EF. For each part, do tasks (a), (c), (d), (k-i), (k-iv), (m), and (p).

(a) The aortic cases have $\overline{X} = 75.8, \overline{Y} = 0.396, [x^2] = 35307.2, [y^2] = 0.39170$, and $[xy] = 84.338$.

(b) The mitral cases have $\overline{X} = 45.3, \overline{Y} = 0.478, [x^2] = 6753.8, [y^2] = 0.24812$, and $[xy] = -18.610$.

9.33 Investigate the relationship between the preoperative heart rate and the postoperative heart rate. If there are outliers, eliminate (their) effect. Specifically address these questions: (1) Is there an overall change from preop to postop HR? (2) Are the preop and postop HRs associated? If there is an association, summarize it (Tables 9.27 and 9.28).

(a) For the aortic cases, $\sum X = 1502, \sum Y = 17.30, \sum X^2 = 116, 446, \sum Y^2 = 152, 662$, and $\sum XY = 130, 556$. Data are given in Table 9.27.

(b) For the mitral cases: $\sum X = 1640, \sum Y = 1869, \sum X^2 = 140, 338, \sum Y^2 = 179, 089$, and $\sum XY = 152, 860$. Data are given in Table 9.28.
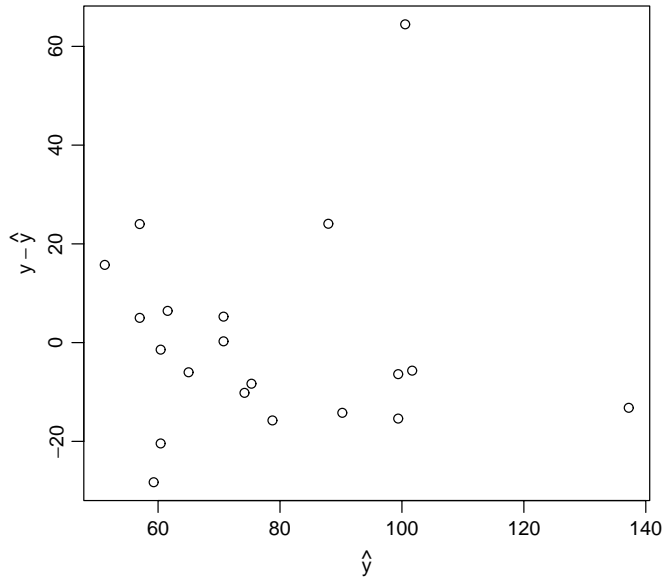
Figure 9.21   Residual plot for Problem 9.31(a).
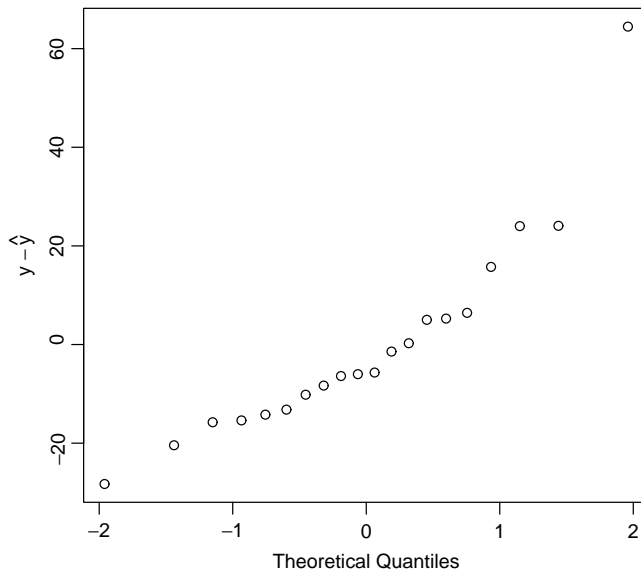
Figure 9.22   Normal probability plot for Problem 9.31(a).

9.34 The Web appendix to this chapter contains county-by-county electoral data for the state
of Florida for the 2000 elections for president and for governor of Florida. The major
Democratic and Republican parties each had a candidate for both positions, and there
were two minor party candidates for president and one for governor. In Palm Beach
County a poorly designed ballot was used, and it was suggested that this led to some
voters who intended to vote for Gore in fact voting for Buchanan.

Table 9.27   Data for Problem 9.33(a)

X	Y	$\widehat{Y}$	Residuals	Normal Deviate
75	80	86.48	−6.48	−0.51
110	100	92.56	7.44	0.59
75	100	86.48	13.52	1.06
70	85	85.61	0.61	−0.04
68	94	85.27	8.73	0.69
76	74	86.66	−12.66	−1.00
60	85	83.88	1.12	0.08
70	85	85.61	0.61	−0.04
68	120	85.27	34.73	2.73
75	92	86.48	5.52	0.43
65	85	84.75	0.25	0.02
70	84	85.61	−1.61	−0.13
70	84	85.61	−1.61	−0.13
85	86	88.22	−2.22	−0.17
66	100	84.92	15.08	1.19
54	60	82.84	−22.84	−1.80
110	88	92.56	−4.56	0.36
75	75	86.48	−11.48	−0.90
80	78	87.35	−9.35	−0.74
80	75	87.35	−12.35	−0.97

Table 9.28   Data for Problem 9.33(b)

X	Y	$\widehat{Y}$	Residuals	Normal Deviate
75	90	93.93	−3.93	−0.25
70	95	94.27	0.73	0.04
86	80	93.18	−13.18	−0.84
120	90	90.87	−0.87	−0.05
85	100	93.25	6.75	0.43
80	75	93.59	−18.59	−1.19
55	140	95.28	44.72	2.86
72	95	94.13	0.87	0.05
108	100	91.68	8.32	0.53
80	90	93.59	−3.59	−0.23
80	98	93.59	4.41	0.28
80	61	93.95	−32.59	−2.08
65	88	94.61	−6.61	0.42
102	100	92.09	7.91	0.51
60	85	94.94	−9.94	−0.64
75	84	93.93	−9.93	−0.63
88	100	93.04	6.96	0.44
80	108	93.59	14.41	0.92
115	100	91.21	8.79	0.56
64	90	94.67	−4.67	−0.30

(a) Using simple linear regression and graphs, examine whether the data support this claim.

(b) Read the analyses linked from the Web appendix and critically evaluate their claims.

REFERENCES

Acton, F. S. [1984]. *Analysis of Straight-Line Data*. Dover Publications, New York.

Anscombe, F. J. [1973]. Graphs in statistical analysis. *American Statistician*, **27**: 17–21.

Boucher, C. A., Bingham, J. B., Osbakken, M. D., Okada, R. D., Strauss, H. W., Block, P. C., Levine, F. H., Phillips, H. R., and Phost, G. M. [1981]. Early changes in left ventricular volume overload. *American Journal of Cardiology*, **47**: 991–1004.

Bruce, R. A., Kusumi, F., and Hosmer, D. [1973]. Maximal oxygen intake and nomographic assessment of functional aerobic impairment in cardiovascular disease. *American Heart Journal*, **65**: 546–562.

Carroll, R. J., Ruppert, D., and Stefanski, L. A. [1995]. *Measurement Error in Nonlinear Models*. Chapman & Hall, London.

Dern, R. J., and Wiorkowski, J. J. [1969]. Studies on the preservation of human blood: IV. The hereditary component of pre- and post storage erythrocyte adenosine triphosphate levels. *Journal of Laboratory and Clinical Medicine*, **73**: 1019–1029.

Devlin, S. J., Gnanadesikan, R., and Kettenring, J. R. [1975]. Robust estimation and outlier detection with correlation coefficients. *Biometrika*, **62**: 531–545.

Draper, N. R., and Smith, H. [1998]. *Applied Regression Analysis*, 3rd ed. Wiley, New York.

Hollander, M., and Wolfe, D. A. [1999]. *Nonparametric Statistical Methods*. 2nd ed. Wiley, New York.

Huber, P. J. [2003]. *Robust Statistics*. Wiley, New York.

Jensen, D., Atwood, J. E., Frolicher, V., McKirnan, M. D., Battler, A., Ashburn, W., and Ross, J., Jr., [1980]. Improvement in ventricular function during exercise studied with radionuclide ventriculography after cardiac rehabilitation. *American Journal of Cardiology*, **46**: 770–777.

Kendall, M. G., and Stuart, A. [1967]. *The Advanced Theory of Statistics*, Vol. 2, *Inference and Relationships*, 2nd ed. Hafner, New York.

Kronmal, R. A. [1993]. Spurious correlation and the fallacy of the ratio standard revisited. *Journal of the Royal Statistical Society, Series A*, **60**: 489–498.

Lumley, T., Diehr, P., Emerson, S., and Chen, L. [2002]. The importance of the normality assumption in large public health data sets. *Annual Review of Public Health*, **23**: 151–169.

Mehta, J., Mehta, P., Pepine, C. J., and Conti, C. R. [1981]. Platelet function studies in coronary artery disease: X. Effects of dipyridamole. *American Journal of Cardiology*, **47**: 1111–1114.

Neyman, J. [1952]. On a most powerful method of discovering statistical regularities. *Lectures and Conferences on Mathematical Statistics and Probability*. U.S. Department of Agriculture, Washington, DC, pp. 143–154.

U.S. Department of Health, Education, and Welfare [1974].

U.S. Cancer Mortality by County: 1950–59. DHEW Publication (NIH) 74–615. U.S. Government Printing Office, Washington, DC.

Yanez, N. D., Kronmal, R. A., and Shemanski, L. R. [1998]. The effects of measurement error in response variables and test of association of explanatory variables in change models. *Statistics in Medicine* **17**(22): 2597–2606.

C H A P T E R 10

Analysis of Variance

10.1 INTRODUCTION

The phrase *analysis of variance* was coined by Fisher [1950], who defined it as "the separation of variance ascribable to one group of causes from the variance ascribable to other groups." Another way of stating this is to consider it as a partitioning of total variance into component parts. One illustration of this procedure is contained in Chapter 9, where the total variability of the dependent variable was partitioned into two components: one associated with regression and the other associated with (residual) variation about the regression line. Analysis of variance models are a special class of linear models.

Definition 10.1. An *analysis of variance model* is a linear regression model in which the predictor variables are classification variables. The categories of a variable are called the *levels* of the variable.

The meaning of this definition will become clearer as you read this chapter.

The topics of analysis of variance and design of experiments are closely related, which has been evident in earlier chapters. For example, use of a paired t-test implies that the data are paired and thus may indicate a certain type of experiment. Similarly, a partitioning of total variation in a regression situation implies that two variables measured are linearly related. A general principle is involved: The analysis of a set of data should be appropriate for the design. We indicate the close relationship between design and analysis throughout this chapter.

The chapter begins with the one-way analysis of variance. Total variability is partitioned into a variance between groups and a variance within groups. The groups could consist of different treatments or different classifications. In Section 10.2 we develop the construction of an analysis of variance from group means and standard deviations, and consider the analysis of variance using ranks. In Section 10.3 we discuss the two-way analysis of variance: A special two-way analysis involving randomized blocks and the corresponding rank analysis are discussed, and then two kinds of classification variables (random and fixed) are covered. Special but common designs are presented in Sections 10.4 and 10.5. Finally, in Section 10.6 we discuss the testing of the assumptions of the analysis of variance, including ways of transforming the data to make the assumptions valid. Notes and specialized topics conclude our discussion.

A few comments about notation and computations: The formulas for the analysis of variance look formidable but follow a logical pattern. The following rules are followed or held (we remind you on occasion):

1. Indices for groups follow a mnemonic pattern. For example, the subscript i runs from $1, \ldots, I$; the subscript j from $1, \ldots, J$; k from $1, \ldots, K$, and so on.

2. Sums of values of the random variables are indicated by replacing the subscript by a dot. For example,

$$Y_{i\cdot} = \sum_{j=1}^{J} Y_{ij}, \qquad Y_{\cdot jk} = \sum_{i=1}^{I} Y_{ijk}, \qquad Y_{\cdot j\cdot} = \sum_{i=1}^{I} \sum_{k=1}^{K} Y_{ijk}$$

3. It is expensive to print subscripts and superscripts on $\sum$ signs. A very simple rule is that summations are always over the given subscripts. For example,

$$\sum Y_i = \sum_{i=1}^{I} Y_i, \qquad \sum Y_{ijk} = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} Y_{ijk}$$

We may write expressions initially with the subscripts and superscripts, but after the patterns have been established, we omit them. See Table 10.6 for an example.

4. The symbol n_{ij} denotes the number of Y_{ijk} observations, and so on. The total sample size is denoted by n rather than $n_{\ldots}$; it will be obvious from the context that the total sample size is meant.

5. The means are indicated by $\overline{Y}_{ij\cdot}, \overline{Y}_{\cdot j\cdot}$, and so on. The number of observations associated with a mean is always n with the same subscript (e.g., $\overline{Y}_{ij\cdot} = Y_{ij\cdot}/n_{ij}$ or $\overline{Y}_{\cdot j\cdot} = Y_{\cdot j\cdot}/n_{\cdot j}$).

6. The analysis of variance is an analysis of variability associated with a single observation. This implies that sums of squares of subtotals or totals must always be divided by the number of observations making up the total; for example, $\sum Y_{i\cdot}^2/n_i$ if $Y_{i\cdot}$ is the sum of n_i observations. The rule is then that the divisor is always the number of observations represented by the dotted subscripts. Another example: $Y_{\cdot\cdot}^2/n_{\cdot\cdot}$, since $Y_{\cdot\cdot}$ is the sum of $n_{\cdot\cdot}$ observations.

7. Similar to rules 5 and 6, a sum of squares involving means always have as weighting factor the number of observations on which the mean is based. For example,

$$\sum_{i=1}^{I} n_i (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2$$

because the mean $\overline{Y}_{i\cdot}$ is based on n_i observations.

8. The ANOVA models are best expressed in terms of means and deviations from means. The computations are best carried out in terms of totals to avoid unnecessary calculations and prevent rounding error. (This is similar to the definition and calculation of the sample standard deviation.) For example,

$$\sum n_i (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2 = \sum \frac{Y_{i\cdot}^2}{n_i} - \frac{Y_{\cdot\cdot}^2}{n_{\cdot\cdot}}$$

See Problem 10.25.

10.2   ONE-WAY ANALYSIS OF VARIANCE

10.2.1   Motivating Example

Example 10.1.   To motivate the one-way analysis of variance, we return to the data of Zelazo et al. [1972], which deal with the age at which children first walked (see Chapter 5). The experiment involved reinforcement of the walking and placing reflexes in newborns. The walking and placing reflexes disappear by about 8 weeks of age. In this experiment, newborn children were randomly assigned to one of four treatment groups: active exercise; passive exercise; no exercise; or an 8-week control group. Infants in the active-exercise group received walking and placing stimulation four times a day for eight weeks, infants in the passive-exercise group received an equal amount of gross motor stimulation, infants in the no-exercise group were tested along with the first two groups at weekly intervals, and the eight-week control group consisted of infants observed only at 8 weeks of age to control for possible effects of repeated examination. The response variable was age (in months) at which the infant first walked. The data are presented in Table 10.1. For purposes of this example we have added the mean of the fourth group to that group to make the sample sizes equal; this will not change the mean of the fourth group. Equal sample sizes are not required for the one-way analysis of variance.

Assume that the age at which an infant first walks alone is normally distributed with variance σ^2. For the four treatment groups, let the means be $\mu_1, \mu_2, \mu_3,$ and μ_4. Since σ^2 is unknown, we could calculate the sample variance for each of the four groups and come up with a pooled estimate, s_p^2, of σ^2. For this example, since the sample sizes per group are assumed to be equal, this is

$$s_p^2 = \frac{1}{4}(2.0938 + 3.5938 + 2.3104 + 0.7400) = 2.1845$$

But we have one more estimate of σ^2. If the four treatments do not differ ($H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu$), the sample means are normally distributed with variance $\sigma^2/6$. The quantity $\sigma^2/6$ can be estimated by $s_{\bar{Y}}^2$, the variance of the sample means. For this example it is

$$s_{\bar{Y}}^2 = 0.87439$$

Table 10.1   Distribution of Ages (in Months) at which Infants First Walked Alone

|	Active Group	Passive Group	No-Exercise Group	Eight-Week Control Group
	9.00	11.00	11.50	13.25
	9.50	10.00	12.00	11.50
	9.75	10.00	9.00	12.00
	10.00	11.75	11.50	13.50
	13.00	10.50	13.25	11.50
	9.50	15.00	13.00	12.35[a]
Mean	10.125	11.375	11.708	12.350
Variance	2.0938	3.5938	2.3104	0.7400
$Y_i.$	60.75	68.25	70.25	74.10

Source: Data from Zelazo et al. [1972].

[a]This observation is missing from the original data set. For purposes of this illustration, it is estimated by the sample mean. See the text for further discussion.

Hence, $6s_{\bar{Y}}^2 = 5.2463$ is also an estimate of σ^2. Under the null hypothesis, $6s_{\bar{Y}}^2/s_p^2$ will follow an F-distribution. How many degrees of freedom are involved? The quantity $s_{\bar{Y}}^2$ has three degrees of freedom associated with it (since it is a variance based on four observations). The quantity s_p^2 has 20 degrees of freedom (since each of its four component variances has five degrees of freedom). So the quantity $6s_{\bar{Y}}^2/s_p^2$ under the null hypothesis has an F-distribution with 3 and 20 degrees of freedom. What if the null hypothesis is not true (i.e., the μ_1, μ_2, μ_3, and μ_4 are not all equal)? It can be shown that $6s_{\bar{Y}}^2$ then estimates $\sigma^2 +$ *positive constant*, so that the ratio $6s_{\bar{Y}}^2/s_p^2$ tends to be larger than 1. The usual hypothesis-testing approach is to reject the null hypothesis if the ratio is "too large," with the critical value selected from an F-table. The analysis is summarized in an *analysis of variance table* (ANOVA), as in Table 10.2.

The variances $6s_{\bar{Y}}^2/s_p^2$ and s_p^2 are called *mean squares* for reasons to be explained later. It is clear that the first variance measures the variability between groups, and the second measures the variability within groups. The F-ratio of 2.40 is referred to an F-table. The critical value at the 0.05 level is $F_{3,20,0.95} = 3.10$, the observed value 2.40 is smaller, and we do not reject the null hypothesis at the 0.05 level. The data are displayed in Figure 10.1. From the graph it can be seen that the active group had the lowest mean value. The nonsignificance of the F-test suggests that the active group mean is not significantly lower than that of the other three groups.

Table 10.2   Simplified ANOVA Table of Data of Table 10.1

Source of Variation	d.f.	MS	F-Ratio
Between groups	3	$6s_{\bar{Y}}^2 = 5.2463$	$\dfrac{6s_{\bar{Y}}^2}{s_p^2} = \dfrac{5.2463}{2.1845} = 2.40$
Within groups	20	$s_p^2 = 2.1845$	
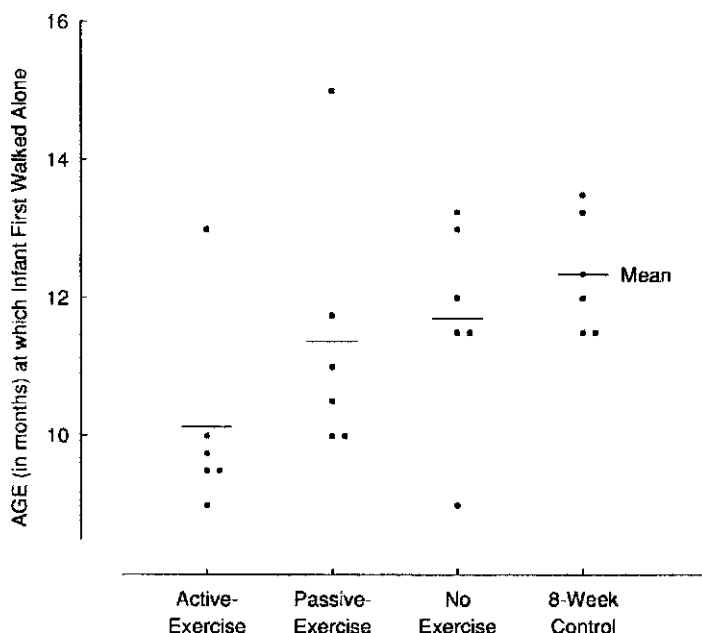
Figure 10.1   Distribution of ages at which infants first walked alone. (Data from Zelazo et al. [1972]; see Table 10.1.)

10.2.2 Using the Normal Distribution Model

Basic Approach

The one-way analysis of variance is a generalization of the t-test. As in the motivating example above, it can be used to examine the age at which groups of infants first walk alone, each group receiving a different treatment; or we may compare patient costs (in dollars per day) in a sample of hospitals from a metropolitan area. (There is a subtle distinction between the two examples; see Section 10.3.4 for a further discussion.)

Definition 10.2. An analysis of variance of observations, each of which belongs to one of I disjoint groups, is a *one-way analysis of variance of I groups*.

Suppose that samples are taken from I normal populations that differ at most in their means; the observations can be modeled by

$$Y_{ij} = \mu_i + \epsilon_{ij}, \qquad i = 1, \dots, I, \quad j = 1, \dots, n_i \qquad (1)$$

The mean for normal population i is μ_i; we assume that there are n_i observations from this population. Also, by assumption, the ϵ_{ij} are independent $N(0, \sigma^2)$ variables. In words: Y_{ij} denotes the jth sample from a population with mean μ_i and variance σ^2. If $I = 2$, you can see that this is precisely the model for the two-sample t-test.

The only difference between the situation now and that of Section 10.2.1 is that we allow the number of observations to vary from group to group. The within-group estimate of the variance σ^2 now becomes a weighted sum of sample variances. Let s_i^2 be the sample variance from group i, where $i = 1, \dots, I$. The within-group estimate σ^2 is

$$\frac{\sum (n_i - 1)s_i^2}{\sum (n_i - 1)} = \frac{\sum (n_i - 1)s_i^2}{n - I}$$

where $n = n_1 + n_2 + \cdots + n_I$ is the total number of observations.

Under the null hypothesis $H_0 : \mu_1 = \mu_2 = \cdots = \mu_I = \mu$, the variability among the group of sample means also estimates σ^2. We will show below that the proper expression is

$$\frac{\sum n_i (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2}{I - 1}$$

where

$$\overline{Y}_{i\cdot} = \sum_{j=1}^{n_i} \frac{Y_{ij}}{n_i}$$

is the sample mean for group i, and

$$\overline{Y}_{\cdot\cdot} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} \frac{Y_{ij}}{n} = \sum \frac{n_i \overline{Y}_{i\cdot}}{n}$$

is the grand mean. These quantities can again be arranged in an ANOVA table, as displayed in Table 10.3. Under the null hypothesis, $H_0 : \mu_1 = \mu_2 = \cdots = \mu_I = \mu$, the quantity A/B in Table 10.3 follows an F-distribution with $(I - 1)$ and $(n - I)$ degrees of freedom.

We now reanalyze our first example in Section 10.2.1, deleting the sixth observation, 12.35, in the eight-week control group. The means and variances for the four groups are now:

Table 10.3 One-Way ANOVA Table for I Groups and n_i Observations per Group ($i = 1, \ldots, I$)

Source of Variation	d.f.	MS	F-Ratio
Between groups	$I - 1$	$A = \dfrac{\sum n_i (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2}{I - 1}$	A/B
Within groups	$n - I$	$B = \sum \dfrac{(n_i - 1)s_i^2}{n - I}$	

Table 10.4 ANOVA of Data from Example 10.1, Omitting the Last Observation

Source of Variation	d.f.	MS	F-Ratio
Between groups	3	4.9253	2.14
Within groups	19	2.2994	

	Active	**Passive**	**No Exercise**	**Control**	**Overall**
Mean ($\overline{Y}_{i\cdot}$)	10.125	11.375	11.708	12.350	11.348
Variance (s_i^2)	2.0938	3.5938	2.3104	0.925	—
n_i	6	6	6	5	23

Therefore,

$$
\begin{aligned}
\sum n_i (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2 &= 6(10.125 - 11.348)^2 + 6(11.375 - 11.348)^2 \\
&\quad + 6(11.708 - 11.348)^2 + 5(12.350 - 11.348)^2 \\
&= 14.776
\end{aligned}
$$

The between-group mean square is $14.776/(4 - 1) = 4.9253$. The within-group mean square is

$$
\frac{1}{23 - 4}[5(2.0938) + 5(3.5938) + 5(2.3104) + 4(0.925)] = 2.2994
$$

The ANOVA table is displayed in Table 10.4.

The critical value $F_{3,19,0.95} = 3.13$, so again, the four groups do not differ significantly.

Linear Model Approach

In this section we approach the analysis of variance using linear models. The model $Y_{ij} = \mu_i + \epsilon_{ij}$ is usually written as

$$
Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \qquad i = 1, \ldots, I, \quad j = 1, \ldots, n_i \tag{2}
$$

The quantity μ is defined as

$$
\mu = \sum_{i=1}^{I} \sum_{j=1}^{n_i} \frac{\mu_i}{n}
$$

where $n = \sum n_i$ (the total number of observations). The quantity α_i is defined as $\alpha_i = \mu - \mu_i$. This implies that

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i}\alpha_i = \sum n_i\alpha_i = 0 \tag{3}$$

Definition 10.3. The quantity $\alpha_i = \mu - \mu_i$ is the *main effect* of the ith population.

Comments:

1. The symbol α with a subscript will denote an element of the analysis of variance model, not the type I error. The context will make it clear which meaning is intended.
2. The equation $\sum n_i\alpha_i = 0$ is a constraint. It implies that fixing any $(I - 1)$ of the main effects determines the remaining value.

If we hypothesize that the I populations have the same means,

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_I = \mu$$

then an equivalent statement is

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0 \quad \text{or} \quad H_0 : \alpha_i = 0, \qquad i = 1, \dots, I$$

How are the quantities $\mu_i, i = 1, \dots, I$ and σ^2 to be estimated from the data? (Or, equivalently, $\mu, \alpha_i, i = 1, \dots, I$ and σ^2.) Basically, we follow the same strategy as in Section 10.2.1. The variances within the I groups are pooled to provide an estimate of σ^2, and the variability between groups provides a second estimate under the null hypothesis. The data can be displayed as shown in Table 10.5. For this set of data, a partitioning can be set up that mimics the model defined by equation (2):

$$\left.\begin{array}{ll}\text{Model}: & Y_{ij} = \mu + \alpha_i + \epsilon_{ij}\\[4pt] \text{Data}: & Y_{ij} = \overline{Y}.. + a_i + e_{ij}\end{array}\right\} i = 1, \dots, I, \quad j = 1, \dots, n_i \tag{4}$$

where $a_i = \overline{Y}_{i.} - \overline{Y}..$ and $e_{ij} = Y_{ij} - \overline{Y}_{i.}$ for $i = 1, \dots, I$ and $j = 1, \dots, n_i$. It is easy to verify that the condition $\sum n_i\alpha_i = 0$ is mimicked by $\sum n_i a_i = 0$. Each data point is partitioned into three component estimates:

$$Y_{ij} = \overline{Y}.. + (\overline{Y}_i - \overline{Y}..) + (Y_{ij} - \overline{Y}_{i.}) = \text{mean} + i\text{th main effect} + \text{error}$$

Table 10.5   Pooled Variances of I Groups

	Sample				
	1	2	3	$\cdots$	I
	Y_{11}	Y_{21}	Y_{31}	$\cdots$	Y_{I1}
	Y_{12}	Y_{22}	Y_{32}	$\cdots$	Y_{I2}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	Y_{1n_1}	Y_{2n_2}	Y_{3n_3}	$\cdots$	Y_{In_I}
Observations	n_1	n_2	n_3	$\cdots$	n_I
Means	$\overline{Y}_1.$	$\overline{Y}_2.$	$\overline{Y}_3.$	$\cdots$	$\overline{Y}_I.$
Totals	$Y_1.$	$Y_2.$	$Y_3.$	$\cdots$	$Y_I.$

The expression on the right side of Y_{ij} is an algebraic identity. It is a remarkable property of this partitioning that the sum of squares of the Y_{ij} is equal to the sum of the three sums of squares of the elements on the right side:

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i} Y_{ij}^2 = \sum_{i=1}^{I}\sum_{j=1}^{n_i} \overline{Y}_{..}^2 + \sum_{i=1}^{I}\sum_{j=1}^{n_i}(\overline{Y}_{i\cdot} - \overline{Y}_{..})^2 + \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\cdot})^2$$

$$= n\overline{Y}_{..}^2 + \sum_{i=1}^{I} n_i(\overline{Y}_{i\cdot} - \overline{Y}_{..})^2 + \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\cdot})^2 \tag{5}$$

and the degrees of freedom can also be partitioned: $n = 1 + (I-1) + (n-I)$. You will recognize the terms on the right side as the ingredients needed for setting up the analysis of variance table as discussed in the preceding section. It should also be noted that the quantities on the right side are random variables (since they are based on statistics). It can be shown that their expected values are

$$E\left(\sum n_i(\overline{Y}_{i\cdot} - \overline{Y}_{..})^2\right) = \sum n_i \alpha_i^2 + (I-1)\sigma^2 \tag{6}$$

and

$$E\left(\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\cdot})^2\right) = (n-I)\sigma^2 \tag{7}$$

If the null hypothesis $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0$ is true (i.e., $\mu_1 = \mu_2 = \cdots = \mu_I = \mu$), then $\sum n_i \alpha_i^2 = 0$, and both of the terms above provide an estimate of σ^2 [after division by $(I-1)$ and $(n-I)$, respectively]. This layout and analysis is summarized in Table 10.6.

The quantities making up the component parts of equation (5) are called *sums of squares* (SS). "Grand mean" is usually omitted; it is used to test the null hypothesis that $\mu = 0$. This is rarely of very much interest, particularly if the null hypothesis $H_0 : \mu_1 = \mu_2 = \cdots = \mu_I$ is rejected (but see Example 10.7). "Between groups" is used to test the latter null hypothesis, or the equivalent hypothesis, $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0$.

Before returning to Example 10.1, we give a few computational notes.

Computational Notes

As in the case of calculating standard deviations, the computations usually are not based on the means but rather, on the group totals. Only three quantities have to be calculated for the one-way ANOVA. Let

$$Y_{i\cdot} = \sum_{j=1}^{n_i} Y_{ij} = \text{total in the } i\text{th treatment group} \tag{8}$$

and

$$Y_{..} = \sum Y_{i\cdot} = \text{grand total} \tag{9}$$

The three quantities that have to be calculated are

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i} Y_{ij}^2 = \sum\sum Y_{ij}^2, \qquad \sum_{i=1}^{I}\frac{Y_{i\cdot}^2}{n_i} = \sum\frac{Y_{i\cdot}^2}{n_i}, \qquad \frac{Y_{..}^2}{n}$$

Table 10.6 Layout for the One-Way ANOVA

Source of Variation	d.f.	SS[a]	MS	F-Ratio	d.f. of F-Ratio	E(MS)	Hypothesis Tested
Grand mean	1	$SS_\mu = n\overline{Y}_{..}^2$	$MS_\mu = SS_\mu$	$\dfrac{MS_\mu}{MS_\epsilon}$	$(1, n-1)$	$n\mu^2 + \sigma^2$	$\mu = 0$
Between groups (main effects)	$I-1$	$SS_\alpha = \sum n_i(\overline{Y}_{i\cdot} - \overline{Y}_{..})^2$	$MS_\alpha = \dfrac{SS_\alpha}{I-1}$	$\dfrac{MS_\alpha}{MS_\epsilon}$	$(I-1, n-I)$	$\dfrac{\sum n_i\alpha_i^2}{I-1} + \sigma^2$	$\alpha_1 = \cdots = \alpha_I$ or $\mu_1 = \cdots = \mu_I$
Within groups (residuals)	$n-I$	$SS_\epsilon = \sum\sum(Y_{ij} - \overline{Y}_{i\cdot})^2$	$MS_\epsilon = \dfrac{SS_\epsilon}{n-I}$	—	—	σ^2	σ^2
Total	n	$\sum\sum Y_{ij}^2$					

[a]Summation is over all displayed subscripts.

$$\text{Model}: \quad Y_{ij} = \mu_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim \text{iid } N(0, \sigma^2) \quad i = 1, \ldots, I, \quad j = 1, \ldots, n_i$$

$$= \mu + \alpha_i + \epsilon_{ij},$$

$$\text{Data}: \quad Y_{ij} = \overline{Y}_{..} + (\overline{Y}_{i\cdot} - \overline{Y}_{..}) + (Y_{ij} - \overline{Y}_{i\cdot})$$

(iid = independent and identically distributed). An equivalent model is

$$Y_{ij} \sim N(\mu_i, \sigma^2), \quad \text{where } Y_{ij}\text{'s are independent}$$

where $n = \sum n_i = $ total observations. It is easy to establish the following relationships:

$$SS_\mu = \frac{Y_{..}^2}{n} \tag{10}$$

$$SS_\alpha = \sum \frac{Y_{i.}^2}{n_i} - \frac{Y_{..}^2}{n} \tag{11}$$

$$SS_\epsilon = \sum \sum Y_{ij}^2 - \sum \frac{Y_{i.}^2}{n_i} \tag{12}$$

The subscripts are omitted.

We have an algebraic identity in $\sum \sum Y_{ij}^2 = SS_\mu + SS_\alpha + SS_\epsilon$. Defining SS_{TOTAL} as $SS_{\text{TOTAL}} = \sum \sum Y_{ij}^2 - SS_\mu$, we get $SS_{\text{TOTAL}} = SS_\alpha + SS_\epsilon$ and degrees of freedom $(n-1) = (i-1) + (n-I)$.

This formulation is a simplified version of equation (5). Note that the original data are needed only for $\sum \sum Y_{ij}^2$; all other sums of squares can be calculated from group or overall totals.

Continuing Example 10.1, omitting again the last observation (12.35):

$$\sum \sum Y_{ij}^2 = 9.00^2 + 9.50^2 + \cdots + 11.50^2 = 3020.2500$$

$$\sum \frac{Y_{i.}^2}{n_i} = \frac{60.75^2}{6} + \frac{68.25^2}{6} + \frac{70.25^2}{6} + \frac{61.75^2}{5} = 2976.5604$$

$$\frac{Y_{..}^2}{n} = \frac{261.00^2}{23} = 2961.7826$$

The ANOVA table omitting rows for SS_μ and SS_{TOTAL} becomes

Source of Variation	d.f.	SS	MS	F-Ratio
Between groups	3	14.7778	4.9259	2.14
Within groups	19	43.6896	2.2995	

The numbers in this table are not subject to rounding error and differ slightly from those in Table 10.4.

Estimates of the components of the expected mean squares of Table 10.6 can now be obtained. The estimate of σ^2 is $\widehat{\sigma}^2 = 2.2995$, and the estimate of $\sum n_i \alpha_i^2 / (I-1)$ is

$$\frac{\sum n_i \widehat{\alpha}_i^2}{I-1} = 4.9259 - 2.2995 = 2.6264$$

How is this quantity to be interpreted in view of the nonrejection of the null hypothesis? Theoretically, the quantity can never be less than zero (all the terms are positive). The best interpretation looks back to MS_α, which is a random variable which (under the null hypothesis) estimates σ^2. Under the null hypothesis, MS_α and MS_ϵ both estimate σ^2, and $\sum n_i \alpha_i^2 / (I-1)$ is zero.

10.2.3 One-Way ANOVA from Group Means and Standard Deviation

In many research papers, the raw data are not presented but rather, the means and standard deviations (or variances) for each of the, say, I treatment groups under consideration. It is instructive to construct an analysis of variance from these data and see how the assumption

of the equality of the population variances for each of the groups enters in. Advantages of constructing the ANOVA table are:

1. Pooling the sample standard deviations (variances) of the groups produces a more precise estimate of the population standard deviation. This becomes very important if the sample sizes are small.
2. A simultaneous comparison of all group means can be made by means of the F-test rather than by a series of two-sample t-tests. The analysis can be modeled on the layout in Table 10.3.

Suppose that for each of I groups the following quantities are available:

Group	Sample Size	Sample Mean	Sample Variance
i	n_i	$\overline{Y}_i.$	s_i^2

The quantities $n = \sum n_i$, $Y_i. = n_i\overline{Y}_i.$, and $Y.. = \sum Y_i.$ can be calculated. The "within groups" SS is the quantity B in Table 10.3 times $n - I$, and the "between groups" SS can be calculated as

$$SS_\alpha = \sum \frac{Y_i^2.}{n_i} - \frac{Y_{..}^2}{n}$$

Example 10.2. Barboriak et al. [1972] studied risk factors in patients undergoing coronary bypass surgery for coronary artery disease. The authors looked for an association between cholesterol level (a putative risk factor) and the number of diseased blood vessels. The data are:

Diseased Vessels (i)	Sample Size (n_i)	Mean Cholesterol Level ($\overline{Y}_i.$)	Standard Deviation (s_i)
1	29	260	56.0
2	49	289	87.5
3	76	295	72.4

Using equations (8)–(12), we get $n = 29 + 49 + 76 = 154$,

$$Y_1. = n_1\overline{Y}_1. = 29(260) = 7540, \qquad Y_3. = n_3\overline{Y}_3. = 76(295) = 22{,}420$$

$$Y_2. = n_2\overline{Y}_2. = 49(289) = 14{,}161, \qquad Y.. = \sum n_i\overline{Y}_i. = \sum Y_i. = 44{,}121$$

$$SS_\alpha = \frac{7540^2}{29} + \frac{14{,}161^2}{49} + \frac{22{,}420^2}{76} - \frac{44{,}121^2}{154}$$

$$= 12{,}666{,}829.0 - 12{,}640{,}666.5 = 26{,}162.5$$

$$SS_\epsilon = \sum (n_i - 1)s_i^2 = 28 \times 56.0^2 + 48 \times 87.5^2 + 75 \times 72.4^2 = 848{,}440$$

The ANOVA table (Table 10.7) can now be constructed. (There is no need to calculate the total SS.)

The critical value for F at the 0.05 level with 2 and 120 degrees of freedom is 3.07; the observed F-value does not exceed this critical value, and the conclusion is that the average cholesterol levels do not differ significantly.

Table 10.7 ANOVA of Data of Example 10.2

Source	d.f.	SS	MS	F-Ratio
Main effects (disease status)	2	26,162.50	13,081.2	2.33
Residual (error)	151	848,440.0	5,618.5	—

10.2.4 One-Way ANOVA Using Ranks

In this section the rank procedures discussed in Chapter 8 are extended to the one-way analysis of variance. For three or more groups, Kruskal and Wallis [1952] have given a one-way ANOVA based on ranks. The model is

$$Y_{ij} = \mu_i + \epsilon_{ij}, \qquad i = 1, \ldots, I, \quad j = 1, \ldots, n_i$$

The only assumption about the ϵ_{ij} is that they are independently and identically distributed, not necessarily normal. It is assumed that there are no ties among the observations. For a small number of ties in the data, the average of the ranks for the tied observations is usually assigned (see Note 10.1). The test procedure will be conservative in the presence of ties (i.e., the p-value will be smaller when adjustment for ties is made).

The null hypothesis of interest is

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_I = \mu$$

The procedure for obtaining the ranks is similar to that for the two-sample Wilcoxon rank-sum procedure: The $n_1 + n_2 + \cdots + n_I = n$ observations are ranked without regard to which group they belong. Let $R_{ij} = $ rank of observation j in group i.

$$T_{KW} = \frac{12 \sum n_i (\overline{R}_{i\cdot} - \overline{R}_{\cdot\cdot})^2}{n(n+1)} \tag{13}$$

where $\overline{R}_{i\cdot}$ is the average of the ranks of the observations in group i:

$$\overline{R}_{i\cdot} = \sum_{j=1}^{n_i} \frac{R_{ij}}{n_i}$$

and $\overline{R}_{\cdot\cdot}$ is the grand mean of the ranks. The value of the mean ($\overline{R}_{\cdot\cdot}$) must be $(n+1)/2$ (why?) and this provides a partial check on the arithmetic. Large values of T_{KW} imply that the average ranks for the group differ, so that the null hypothesis is rejected for large values of this statistic. If the null hypothesis is true and all the n_i become large, the distribution of the statistic T_{KW} approaches a χ^2-distribution with $I - 1$ degrees of freedom. Thus, for large sample sizes, critical values for T_{KW} can be read from a χ^2-table. For small values of n_i, say, in the range 2 to 5, exact critical values have been tabulated (see, e.g., CRC Table X.9 [Beyer, 1968]). Such tables are available for three or four groups.

An equivalent formula for T_{KW} as defined by equation (13) is

$$T_{KW} = \frac{12 \sum R_{i\cdot}^2 / n_i}{n(n+1)} - 3(n+1) \tag{14}$$

where $R_{i\cdot}$ is the total of the ranks for the ith group.

Example 10.3. Chikos et al. [1977] studied errors in the reading of chest x-rays. The opinion of 10 radiologists about the status of the left ventricle of the heart ("normal" vs. "abnormal") was compared to data obtained by ventriculography (which consists of the insertion of a catheter into the left ventricle, injection of a radiopague fluid, and the taking of a series of x-rays). The ventriculography data were used to classify a subject's left ventricle as "normal" or "abnormal." Using this gold standard, the percentage of errors for each radiologist was computed. The authors were interested in the effect of experience, and for this purpose the radiologists were classified into one of three groups: senior staff, junior staff, and residents. The data for these three groups are shown in Table 10.8.

To compute the Kruskal–Wallis statistic T_{KW}, the data are ranked disregarding groups:

Observation	7.3	7.4	10.6	13.3	14.7	15.0	20.7	22.7	23.0	26.6
Rank	1	2	3	4	5	6	7	8	9	10
Group	1	1	2	2	3	2	2	3	3	3

The sums and means of the ranks for each group are calculated to be

$$R_{1.} = 1 + 2 = 3, \qquad \overline{R}_{1.} = 1.5$$
$$R_{2.} = 3 + 4 + 6 + 7 = 20, \qquad \overline{R}_{2.} = 5.0$$
$$R_{3.} = 5 + 8 + 9 + 10 = 32, \qquad \overline{R}_{3.} = 8.0$$

[The sum of the ranks is $R_1 + R_2 + R_3 = 55 = (10 \times 11)/2$, providing a partial check of the ranking procedure.]

Using equation (14), the T_{KW} statistic has a value of

$$T_{KW} = \frac{12(3^2/2 + 20^2/4 + 32^2/4)}{10(10 + 1)} - 3(10 + 1) = 6.33$$

This value can be referred to as a χ^2-table with two degrees of freedom. The p-value is $0.025 < p < 0.05$. The exact p-value can be obtained from, for example, Table X.9 of the CRC tables [Beyer, 1968]. (This table does not list the critical values of T_{KW} for $n_1 = 2$, $n_2 = 4$, $n_3 = 4$; however, the order in which the groups are labeled does not matter, so that the values $n_1 = 4, n_2 = 4$, and $n_3 = 2$ may be used.) From this table it is seen that $0.011 < p < 0.046$, indicating that the chi-square approximation is satisfactory even for these small sample sizes. The conclusion from both analyses is that among staff levels there are significant differences in the accuracy of reading left ventricular abnormality from a chest x-ray.

Table 10.8   Data for Three Radiologist Groups

	Senior Staff	Junior Staff	Residents
i	1	2	3
n_i	2	4	4
Y_{ij}	7.3	13.3	14.7
	7.4	10.6	23.0
(Percent error)		15.0	22.7
		20.7	26.6

10.3 TWO-WAY ANALYSIS OF VARIANCE

10.3.1 Using the Normal Distribution Model

In this section we consider data that arise when a response variable can be classified in two ways. For example, the response variable may be blood pressure and the classification variables type of drug treatment and gender of the subject. Another example arises from classifying people by type of health insurance and race; the response variable could be number of physician contacts per year.

Definition 10.4. An analysis of variance of observations, each of which can be classified in two ways is called a *two-way analysis of variance*.

The data are usually displayed in "cells," with the row categories the values of one classification variable and the columns representing values of the second classification variable.

A completely general two-way ANOVA model with each cell mean any value could be

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk} \tag{15}$$

where $i = 1, \ldots, I, j = 1, \ldots, J$, and $k = 1, \ldots, n_{ij}$. By assumption, the ϵ_{ijk} are iid $N(0, \sigma^2)$: independently and identically distributed $N(0, \sigma^2)$. This model could be treated as a one-way ANOVA with IJ groups with a test of the hypothesis that all μ_{ij} are the same, implying that the classification variables are not related to the response variable. However, if there is a significant difference among the IJ group means, we want to know whether these differences can be attributed to:

1. One of the classification variables,
2. Both of the classification variables acting separately (no interaction), or
3. Both of the classification variables acting separately and jointly (interaction).

In many situations involving classification variables, the mean μ_{ij} may be modeled as the sum of two terms, an effect of variable 1 plus an effect of variable 2:

$$\mu_{ij} = u_i + v_j, \qquad i = 1, \ldots, I, \quad j = 1, \ldots, J \tag{16}$$

Here μ_{ij} depends, in an additive fashion, on the ith level of the first variable and the jth level of the second variable. One problem is that u_i and v_j are not defined uniquely; for any constant C, if $\mu_i^* = u_i + C$ and $v_j^* = v_j - C$, then $\mu_{ij} = u_i^* + v_j^*$. Thus, the values of u_i and v_j can be pinned down to within a constant. The constant is specified by convention and is associated with the experimental setup. Suppose that there are n_{ij} observations at the ith level of variable 1 and the jth level of variable 2. The frequencies of observations can be laid out in a contingency table as shown in Table 10.9.

The experiment has a total of $n..$ observations. The notation is identical to that used in a two-way contingency table layout. (A major difference is that all the frequencies are usually chosen by the experimenter; we shall return to this point when talking about a balanced ANOVA design.) Using the model of equation (16), the value of μ_{ij} is defined as

$$\mu_{ij} = \mu + \alpha_i + \beta_j \tag{17}$$

where $\mu = \sum \sum n_{ij}\mu_{ij}/n.., \sum n_i.\alpha_i = 0$, and $\sum n_{.j}\beta_j = 0$. This is similar to the constraints put on the one-way ANOVA model; see equations (2) and (10.3), and Problem 10.25(d).

Example 10.4. An experimental setup involves two explanatory variables, each at three levels. There are 24 observations distributed as shown in Table 10.10. The effects of the first

Table 10.9   Contingency Table for Variables

Levels of Variable 1	Levels of Variable 2						Total
	1	2	$\cdots$	j	$\cdots$	J	
1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1j}	$n_{1\cdot}$
2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2J}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{iJ}	$n_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
I	n_{I1}	n_{I2}	$\cdots$	n_{Ij}	$\cdots$	n_{IJ}	$n_{I\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot j}$	$\cdots$	$n_{\cdot J}$	$n_{\cdot\cdot}$

Table 10.10   Observation Data

Levels of Variable 1	Levels of Variable 2			Total
	1	2	3	
1	2	2	2	6
2	3	3	3	9
3	3	3	3	9
Total	8	8	8	24

Table 10.11   Data for Variable Effects

Effects of the First Variable	Effects of the Second Variable			Total
	$\beta_1 = 1$	$\beta_2 = -3$	$\beta_3 = 2$	
$\alpha_1 = 3$	$\mu_{11} = 24$	$\mu_{12} = 20$	$\mu_{13} = 25$	$\mu_{1\cdot} = 23$
$\alpha_2 = 6$	$\mu_{21} = 27$	$\mu_{22} = 23$	$\mu_{23} = 28$	$\mu_{2\cdot} = 26$
$\alpha_3 = -8$	$\mu_{31} = 13$	$\mu_{32} = 9$	$\mu_{33} = 14$	$\mu_{3\cdot} = 12$
Total	$\mu_{\cdot 1} = 21$	$\mu_{\cdot 2} = 17$	$\mu_{\cdot 3} = 22$	$\mu = 20$

variable are assumed to be $\alpha_1 = 3, \alpha_2 = 6$, and $\alpha_3 = -8$; the effects of the second variable are $\beta_1 = 1, \beta_2 = -3$, and $\beta_3 = 2$. The overall level is $\mu = 20$. If the model defined by equation (17) holds, the cell means μ_{ij} are specified completely as shown in Table 10.11.

For example, $\mu_{11} = 20 + 3 + 1 = 24$ and $\mu_{33} = 20 - 8 + 2 = 14$. Note that $\sum n_i \cdot \alpha_i = 6.3 + 9.6 + 9(-8) = 0$ and, similarly, $\sum n_{\cdot j} \beta_j = 0$. Note also that $\mu_{1\cdot} = \sum n_{1j}\mu_{1j} / \sum n_{ij} = \mu + \alpha_1 = 20 + 3 = 23$; that is, a marginal mean is just the overall mean plus the effect of the variable associated with that margin. The means are graphed in Figure 10.2. The points have been joined by dashed lines to make the pattern clear; there need not be any continuity between the levels. A similar graph could be made with the level of the second variable plotted on the abscissa and the lines indexed by the levels of the first variable.

Definition 10.5.   A two-way ANOVA model satisfying equation (17) is called an *additive model*.
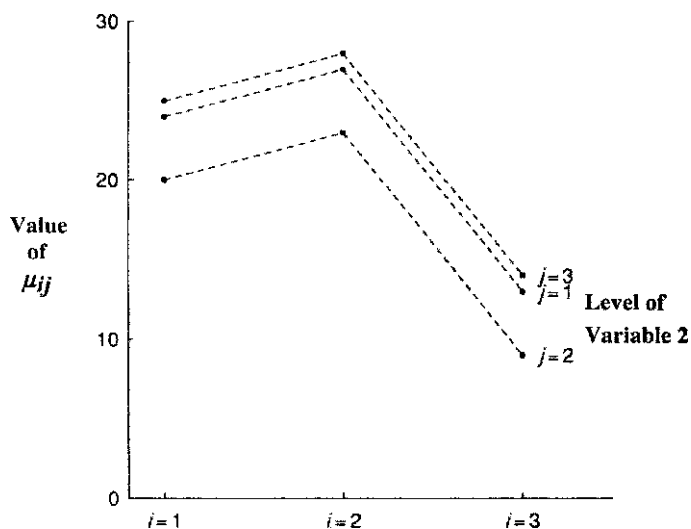
Figure 10.2   Graph of additive ANOVA model (see Example 10.4).

Some implications of this model are discussed. You will find it helpful to refer to Example 10.4 and Figure 10.2 in understanding the following:

1. The statement of equation (17) is equivalent to saying that "changing the level of variable 1 while the level of the second variable remains fixed changes the value of the mean by the same amount regardless of the (fixed) level of the second variable."
2. Statement 1 holds with variables 1 and 2 interchanged.
3. If the values of $\mu_{ij}(i = 1, \dots, I)$ are plotted for the various levels of the second variable, the curves are parallel (see Figure 10.2).
4. Statement 3 holds with the roles of variables 1 and 2 interchanged.
5. The model defined by equation (17) imposes $1 + (I - 1) + (J - 1)$ constraints on the *IJ* means μ_{ij}, leaving $(I - 1)(J - 1)$ degrees of freedom.

We now want to define a nonadditive model, but before doing so, we must introduce one other concept.

Definition 10.6.   A two-way ANOVA has a *balanced (orthogonal) design* if for every i and j,

$$n_{ij} = \frac{n_i. n_{\cdot j}}{n_{..}}$$

That is, the cell frequencies are functions of the product of the marginal totals. The reason this characteristic is needed is that only for balanced designs can the total variability be partitioned in an additive fashion. In Section 10.5 we introduce a discussion of unbalanced or nonorthogonal designs; the topic is treated in terms of multiple regression models in Chapter 11.

Definition 10.7.   A *balanced two-way* ANOVA *model with interaction* (a nonadditive model) is defined by

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}, \qquad \begin{matrix} i = 1, \dots, I \\ j = 1, \dots, J \\ k = 1, \dots, n_{ij} \end{matrix} \qquad (18)$$

subject to the following conditions:

1. $n_{ij} = n_i.n_{\cdot j}/n_{\cdot\cdot}$ for every i and j.
2. $\sum n_i.\alpha_i = \sum n_{\cdot j}\beta_j = 0$.
3. $\sum n_i.\gamma_{ij} = 0$ for all $j = 1, \ldots, J$, $\sum n_{\cdot j}\gamma_{ij} = 0$ for all $i = 1, \ldots, I$.
4. The ϵ_{ijk} are iid $N(0, \sigma^2)$. This assumption implies homogeneity of variances among the IJ cells.

If the γ_{ij} are zero, the model is equivalent to the one defined by equation (17), there is no interaction, and the model is additive.

As in Section 10.2, equations (4) and (5), a set of data as defined at the beginning of this section can be partitioned into parts, each of which estimates the component part of the model:

$$Y_{ijk} = \overline{Y}... + a_i + b_j + g_{ij} + e_{ijk} \tag{19}$$

where

$$\overline{Y}... = \text{grand mean}$$

$$a_i = \overline{Y}_{i}.. - \overline{Y}... = \text{main effect of } i\text{th level of variable 1}$$

$$b_j = \overline{Y}_{\cdot j}. - \overline{Y}... = \text{main effect of } j\text{th level of variable 2}$$

$$g_{ij} = \overline{Y}_{ij}. - \overline{Y}_{i}.. - \overline{Y}_{\cdot j}. + \overline{Y}... = \text{interaction of } i\text{th and } j\text{th levels of variables 1 and 2}$$

$$e_{ijk} = \overline{Y}_{ijk} - \overline{Y}_{ij}. = \text{residual effect (error)}$$

The quantities $\overline{Y}_{i}..$ and $\overline{Y}_{\cdot j}.$ are the means of the ith level of variable 1 and the jth level of variable 2. In symbols,

$$\overline{Y}_{i}.. = \sum_{j=1}^{J} \sum_{k=1}^{n_{ij}} \frac{Y_{ijk}}{n_i.} \quad \text{and} \quad \overline{Y}_{\cdot j}. = \sum_{i=1}^{I} \sum_{k=1}^{n_{ij}} \frac{Y_{ijk}}{n_{\cdot j}}$$

The interaction term, g_{ij}, can be rewritten as

$$g_{ij} = (\overline{Y}_{ij}. - \overline{Y}...) - (\overline{Y}_{i}.. - \overline{Y}...) - (\overline{Y}_{\cdot j}. - \overline{Y}...)$$

which is the overall deviation of the mean of the ijth cell from the grand mean minus the main effects of variables 1 and 2. If the data can be fully explained by main effects, the term g_{ij} will be zero. Hence, g_{ij} measures the extent to which the data deviate from an additive model.

For a balanced design the total sum of squares, $\text{SS}_{\text{TOTAL}} = \sum\sum\sum(Y_{ijk} - \overline{Y}...)^2$ and degrees of freedom can be partitioned additively into four parts:

$$\text{SS}_{\text{TOTAL}} = \text{SS}_\alpha + \text{SS}_\beta + \text{SS}_\gamma + \text{SS}_\epsilon$$

$$n_{\cdot\cdot} - 1 = (I - 1) + (J - 1) + (I - 1)(J - 1) + (n_{\cdot\cdot} - IJ) \tag{20}$$

Let

$$Y_{ij}. = \sum_{k=1}^{n_{ij}} Y_{ijk} = \text{total for cell } ij$$

$$Y_{i\cdot\cdot} = \sum_{j=1}^{J} Y_{ij\cdot} = \text{total for row } i$$

$$Y_{\cdot j\cdot} = \sum_{i=1}^{I} Y_{ij\cdot} = \text{total for column } j$$

Then the equations for the sums of squares together with computationally simpler formulas are

$$\text{SS}_\alpha = \sum n_{i\cdot}(\overline{Y}_{i\cdot\cdot} - \overline{Y}...)^2 = \sum \frac{Y_{i\cdot\cdot}^2}{n_{i\cdot}} - \frac{Y_{...}^2}{n_{\cdot\cdot}}$$

$$\text{SS}_\beta = \sum n_{\cdot j}(\overline{Y}_{\cdot j\cdot} - \overline{Y}...)^2 = \sum \frac{Y_{\cdot j\cdot}^2}{n_{\cdot j}} - \frac{Y_{...}^2}{n_{\cdot\cdot}} \tag{21}$$

$$\text{SS}_\gamma = \sum\sum n_{ij}(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}...)^2 = \sum\sum \frac{Y_{ij\cdot}^2}{n_{ij}} - \frac{Y_{...}^2}{n} - \text{SS}_\alpha - \text{SS}_\beta$$

$$\text{SS}_\epsilon = \sum\sum\sum(Y_{ijk} - \overline{Y}_{ij\cdot})^2 = \sum\sum\sum Y_{ijk}^2 - \sum\sum \frac{Y_{ij\cdot}^2}{n_{ij}}$$

The partition of the sum of squares, the mean squares, and the expected mean squares are given in Table 10.12.

A series of F-tests can be carried out to test the significance of the components of the model specified by equation (18). The first test carried out is usually the test for interaction: $\text{MS}_\gamma/\text{MS}_\epsilon$. Under the null hypothesis $H_0 : \gamma_{ij} = 0$ for all i and j, this ratio has an F-distribution with $(I-1)(J-1)$ and $n - IJ$ degrees of freedom. The null hypothesis is rejected for large values of this ratio. Interaction is indicated by nonparallelism of the treatment effects. In Figure 10.3, some possible patterns are indicated. The expected results of F-tests are given at the top of each graph. For example, pattern 1 shows NO–YES–NO, implying that the test for the main effect of variable 1 was not significant, the test for main effect of variable 2 was significant, and the test for interaction was not significant. It now becomes clear that if interaction is present, main effects are going to be difficult to interpret. For example, pattern 4 in Figure 10.3 indicates significant interaction but no significant main effects. But the significant interaction implies that at level 1 of variable 1 there is a significant difference in the main effect of variable 2. What is happening is that the effect of variable 2 is in the opposite direction at the second level of variable 1. This pattern is extreme. A more common pattern is that of pattern 6. How is this pattern to be interpreted? First, there is interaction; second, above the interaction there are significant main effects.

There are substantial practical problems associated with significant interaction patterns. For example, suppose that the two variables represent two drugs for pain relief administered simultaneously to a patient. With pattern 2, the inference would be that the two drugs together are more effective than either one acting singly. In pattern 4 (and pattern 3), the drugs are said to act *antagonistically*. In pattern 6, the drugs are said to act *synergistically*; the effect of both drugs combined is greater than the sum of each acting alone. (For some subtle problems associated with these patterns, see the discussion of transformations in Section 10.6.)

If interaction is not present, the main effects can be tested by means of the F-tests $\text{MS}_\alpha/\text{MS}_\epsilon$ and $\text{MS}_\beta/\text{MS}_\epsilon$ with $(I-1, n-IJ)$ and $(J-1, n-IJ)$ degrees of freedom, respectively. If a main effect is significant, the question arises: Which levels of the main effect differ significantly? At this point, a visual inspection of the levels may be sufficient to establish the pattern; in Chapter 12 we establish a more formal approach.

As usual, the test $\text{MS}_\mu/\text{MS}_\epsilon$ is of little interest, and this line is frequently omitted in an analysis of variance table.

Table 10.12 Layout for the Two-Way ANOVA[a]

Source of Variation	d.f.	SS[b]	MS	F-Ratio	d.f. of F-Ratio	E(MS)	Hypothesis Being Tested
Grand mean	1	$SS_\mu = n\overline{Y}_{...}^2$	$MS_\mu = SS_\mu$	$\dfrac{MS_\mu}{MS_\epsilon}$	$(1, n-IJ)$	$\sigma^2 + n\mu^2$	$\mu = 0$
Row main effects	$I-1$	$SS_\alpha = \sum n_{i\cdot}(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{...})^2$	$MS_\alpha = \dfrac{SS_\alpha}{I-1}$	$\dfrac{MS_\alpha}{MS_\epsilon}$	$(I-1, n-IJ)$	$\sigma^2 + \dfrac{\sum n_{i\cdot}\alpha_i^2}{I-1}$	$\alpha_i = 0$ for all i
Column main effects	$J-1$	$SS_\beta = \sum n_{\cdot j}(\overline{Y}_{\cdot j\cdot} - \overline{Y}_{...})^2$	$MS_\beta = \dfrac{SS_\beta}{J-1}$	$\dfrac{MS_\beta}{MS_\epsilon}$	$(J-1, n-IJ)$	$\sigma^2 + \dfrac{\sum n_{\cdot j}\beta_j^2}{J-1}$	$\beta_j = 0$ for all j
Row × column interaction	$(I-1)(J-1)$	$SS_\gamma = \sum n_{ij}(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot}\,\overline{Y}_{...})^2$	$MS_\gamma = \dfrac{SS_\gamma}{(I-1)(J-1)}$	$\dfrac{MS_\gamma}{MS_\epsilon}$	$((I-1)(J-1), n-IJ)$	$\sigma^2 + \dfrac{\sum n_{ij}\gamma_{ij}^2}{(I-1)(J-1)}$	$\gamma_{ij} = 0$ for all i and j, or $\mu_{ij} = u_i + v_j$
Residual	$n-IJ$	$SS_\epsilon = \sum(Y_{ijk} - \overline{Y}_{ij\cdot})^2$	$MS_\epsilon = \dfrac{SS_\epsilon}{n-IJ}$	—	—	σ^2	
Total	n	$\sum Y_{ijk}^2$	—				

[a] Model: $Y_{ijk} = \mu_{ij} + \epsilon_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$ [where $\epsilon_{ijk} \sim$ iid $N(0, \sigma^2)$].
Data: $Y_{ijk} = \overline{Y}_{...} + (\overline{Y}_{i\cdot\cdot} - \overline{Y}_{...}) + (\overline{Y}_{\cdot j\cdot} - \overline{Y}_{...}) + (\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} - \overline{Y}_{...}) + (Y_{ijk} - \overline{Y}_{ij\cdot})$.
Equivalent model: $Y_{ijk} \sim N(\mu_{ij}, \sigma^2)$, where the Y_{ijk}'s are independent.
[b] Summation is over all subscripts displayed.
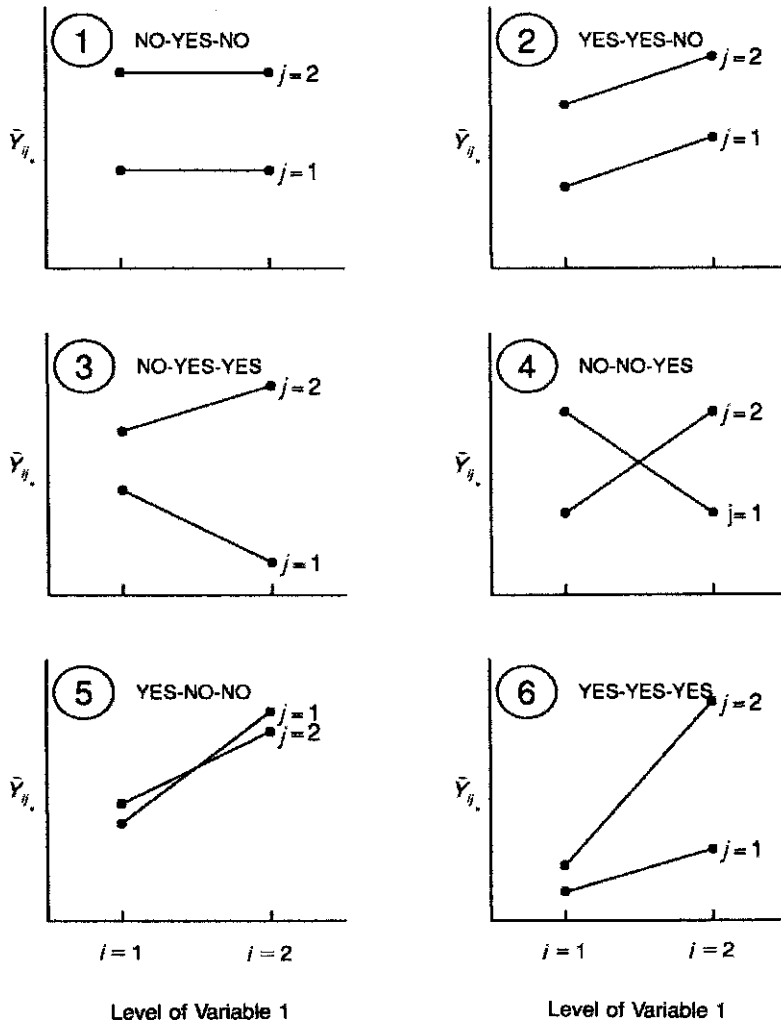
Figure 10.3 Some possible patterns for observed cell means in two-way ANOVA with two levels for each variable. Results of F-tests for main effects variable 1, variable 2, and interaction are indicated by YES or NO. See the text for a discussion.

Example 10.5. Nitrogen dioxide (NO_2) is an automobile emission pollutant, but less is known about its effects than those of other pollutants, such as particulate matter. Several animal models have been studied to gain an understanding of the effects of NO_2. Sherwin and Layfield [1976] studied protein leakage in the lungs of mice exposed to 0.5 part per million (ppm) NO_2 for 10, 12, and 14 days. Half of a total group of 44 animals was exposed to the NO_2; the other half served as controls. Control and experimental animals were matched on the basis of weight, but this aspect will be ignored in the analysis since the matching did not appear to influence the results. Thirty-eight animals were available for analysis; the raw data and some basic statistics are listed in Table 10.13.

The response is the percent of serum fluorescence. High serum fluorescence values indicate a greater protein leakage and some kind of insult to the lung tissue. The authors carried out t-tests and state that with regard to serum fluorescence, "no significant differences" were found.

Table 10.13 **Serum Fluorescence Readings of Mice Exposed to Nitrogen Dioxide (NO₂) for 10, 12, and 14 Days Compared with Control Animals**

	Serum Fluorescence		
	10 Days ($j = 1$)	12 Days ($j = 2$)	14 Days ($j = 3$)
Control ($i = 1$)	143	179	76
	169	160	40
	95	87	119
	111	115	72
	132	171	163
	150	146	78
	141	—	—
Exposed ($i = 2$)	152	141	119
	83	132	104
	91	201	125
	86	242	147
	150	209	200
	108	114	178
	75	—	—

n_{ij}

	j		
i	1	2	3
1	7	6	6
2	7	6	6

$Y_{ij\cdot}$

	j		
i	1	2	3
1	941	858	548
2	745	1039	873

$\overline{Y}_{ij\cdot}$

	j		
i	1	2	3
1	134.4	143.0	91.3
2	106.4	173.2	145.5

s_{ij}

	j		
i	1	2	3
1	24.7	35.5	43.2
2	32.1	51.0	37.1

The standard deviations are very similar, suggesting that the homogeneity of variance assumption is probably valid. It is a good idea again to graph the results to get some "feel" for the data, and this is done in Figure 10.4. We can see from this figure that there are no outlying observations that would invalidate the normality assumption of the two-way ANOVA model.
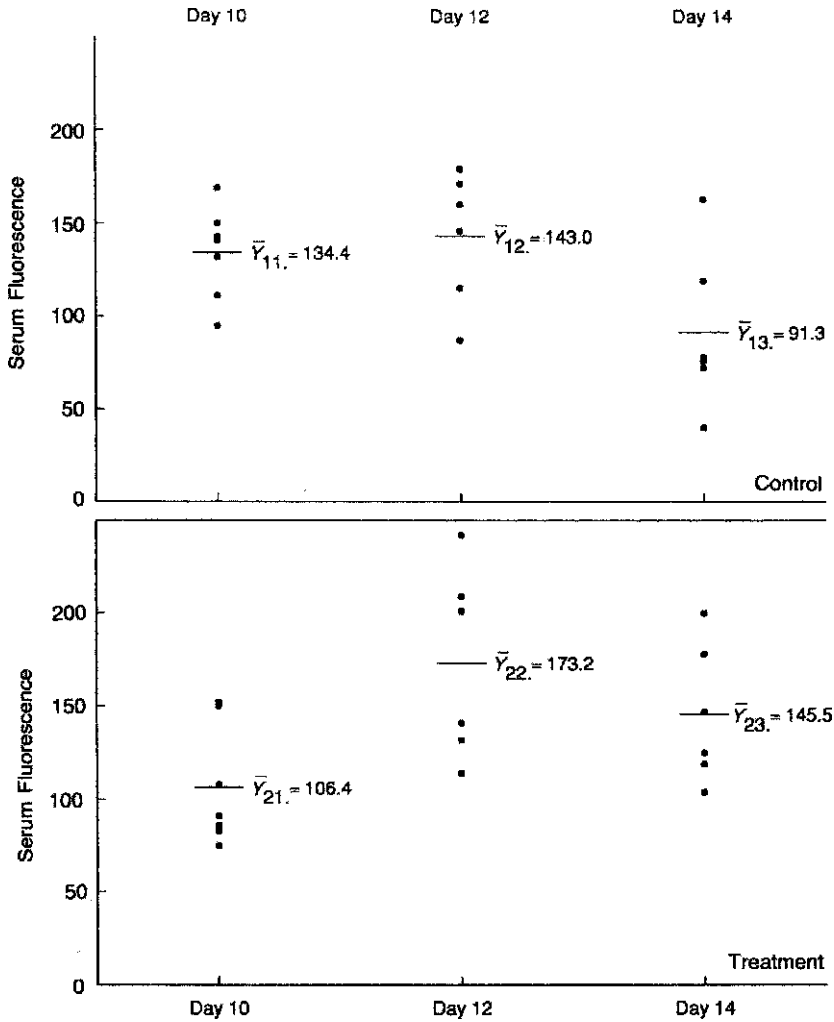
To obtain the entries for the two-way ANOVA table, we basically need six quantities:

$$n, \quad Y..., \quad \sum Y_{ijk}^2, \quad \sum \frac{Y_{i\cdot\cdot}^2}{n_{i\cdot}}, \quad \sum \frac{Y_{\cdot j\cdot}^2}{n_{\cdot j}}, \quad \sum \frac{Y_{ij\cdot}^2}{n_{ij}}$$

With these quantities, and using equations (20) and (21), the entire table can be computed. The values are as follows:

$$n = 38, \qquad Y... = 5004, \qquad \sum Y_{ijk}^2 = 730{,}828$$

$$\sum \frac{Y_{i\cdot\cdot}^2}{n_{i\cdot}} = 661{,}476.74, \qquad \sum \frac{Y_{\cdot j\cdot}^2}{n_{\cdot j}} = 671{,}196.74, \qquad \sum \frac{Y_{ij\cdot}^2}{n_{ij}} = 685{,}472.90$$

Figure 10.4 Serum fluorescence of mice exposed to nitrogen dioxide. (Data from Sherwin and Layfield [1976]; see Example 10.5.)

Sums of squares can now be calculated:

$$SS_\alpha = SS_{\text{TREATMENT}} = 661,476.74 - \frac{5004^2}{38} = 2528.95$$

$$SS_\beta = SS_{\text{DAYS}} = 671196.74 - \frac{5004^2}{38} = 12,248.95$$

$$SS_\gamma = SS_{\text{TREATMENT} \times \text{DAYS}} = 685,472.90 - \frac{5004^2}{38} - 2528.95 - 12,248.95 = 11,747.21$$

$$SS_\epsilon = SS_{\text{RESIDUAL}} = 730,828 - 685,472.90 = 45,355.10$$

(It can be shown that $SS_\epsilon = \sum (n_{ij} - 1)s_{ij}^2$. You can verify this for these data.) The ANOVA table is presented in Table 10.14.

Table 10.14   ANOVA **of Serum Fluorescence Levels of Mice Exposed to Nitrogen Dioxide (NO$_2$)**

Source of Variation	d.f.	SS	MS	F-Ratio	p-Value
Treatment	1	2,528.95	2528.95	1.78	> 0.10
Days	2	12,248.95	6124.48	4.32	< 0.05
Treatment × days	2	11,747.21	5873.60	4.14	< 0.05
Residual	32	45,355.10	1417.35	—	—
Total	37	71,880.21	—	—	—

Source: Data from Sherwin and Layfield [1976].

The MS for interaction is significant at the 0.05 level ($F_{2,32} = 4.14$, $p < 0.05$). How is this to be interpreted? The means $\overline{Y}_{ij}$. are graphed in Figure 10.5. There clearly is nonparallelism, and the model is not an additive one. But more should be said in order to interpret the results, particularly regarding the role of the control animals. Clearly, control animals were used to provide a measurement of background variation. The differences in mean fluorescence levels among the control animals indicate that the baseline response level changed from day 10 to day 14. If we consider the response of the animals exposed to nitrogen dioxide standardized by the control level, a different picture emerges. In Figure 10.5, the differences in means between exposed and unexposed animals is plotted as a dashed line with scale on the right-hand side of the graph. This line indicates that there is an increasing effect of exposure with time. The interpretation of the significant interaction effect then is, possibly, that exposure did induce increased protein leakage, with greater leakage attributable to longer exposure. This contradicts the authors' analysis of the data using t-tests. If the matching by weight was retained, it would
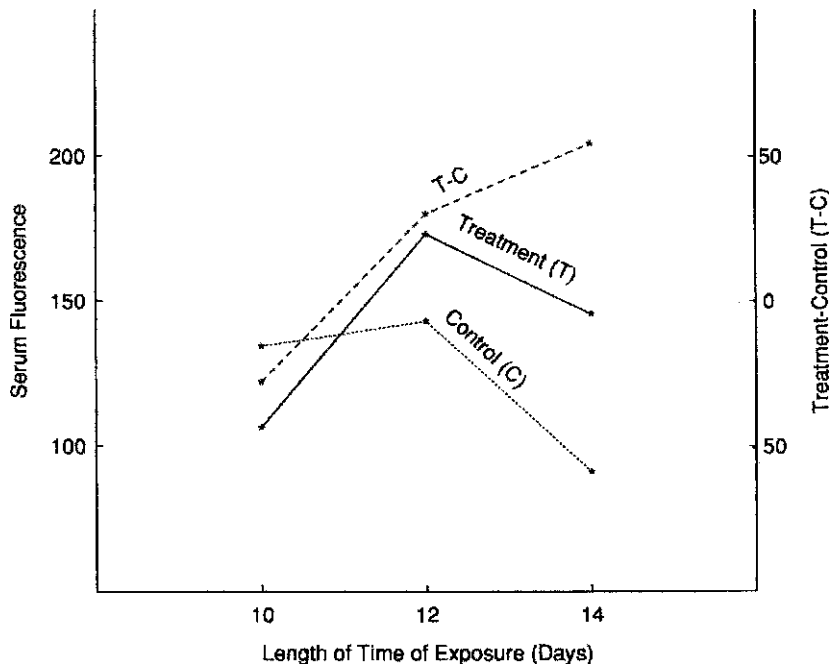
Figure 10.5   Mean serum fluorescence level of mice exposed to nitrogen dioxide, treatment vs. control. The difference (treatment − control) is given by the dashed line. (Data from Sherwin and Layfield [1976]; see Example 10.5.)

have been possible to consider the differences between exposed and control animals and carry out a one-way ANOVA on the differences. See Problem 10.5.

Two-Way ANOVA from Means and Standard Deviations

As in the one-way ANOVA, a two-way ANOVA can be reconstructed from means and standard deviations. Let $\overline{Y}_{ij\cdot}$ be the mean, s_{ij} the standard deviation, and n_{ij} the sample size associated with cell $ij (i = 1, \dots , I, j = 1, \dots , J)$, assuming a balanced design. Then

$$Y_{\cdots} = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \overline{Y}_{ij\cdot}, \qquad Y_{i\cdot\cdot} = \sum_{j=1}^{J} n_{ij} \overline{Y}_{ij\cdot}, \qquad Y_{\cdot j\cdot} = \sum_{i=1}^{I} n_{ij} \overline{Y}_{ij\cdot}$$

Using equation (21), SS_α and SS_β can now be calculated. The term $\sum Y_{ij\cdot}^2 / n_{ij}$ in SS_γ is equivalent to

$$\sum \frac{Y_{ij\cdot}^2}{n_{ij}} = \sum n_{ij} \overline{Y}_{ij\cdot}^2.$$

Finally, SS_ϵ can be calculated from

$$SS_\epsilon = \sum (n_{ij} - 1) s_{ij}^2 \tag{22}$$

Problems 10.22 and 10.23 deal with data presented in terms of means and standard deviations. There will be some round-off error in the two-way analysis constructed in this way, but it will not affect the conclusion.

It is easy to write a computer subroutine that produces such a table upon input of means, standard deviations, and sample sizes.

10.3.2   Randomized Block Design

In Chapter 2 we discussed the statistical concept of blocking. A block consists of a subset of homogeneous experimental units. The background variability among blocks is usually much greater than within blocks, and the experimental strategy is to assign all treatments randomly to the units of a block. A simple example of blocking is illustrated by the paired t-test. Suppose that two antiepileptic agents are to be compared. One possible (valid) design is to assign randomly half of a group of patients to one agent and half to the other. By this randomization procedure, the variability among patients is "turned" into error. Appropriate analyses are the two-sample t-test, the one-way analysis of variance, or a two-sample nonparametric test. However, if possible, a better design would be to test both drugs on the same patient; this would eliminate patient-to-patient variability, and comparisons are made within patients. The patients in this case act as *blocks*. A paired t-test or analogous nonparametric test is now appropriate. For this design to work, we would want to assign the drugs randomly within a patient. This would eliminate a possible additive sequence effect; hence, the term *randomized block design*. In addition, we would want to have a reasonably large time interval between drugs to eliminate possible carryover effects; that is, we cannot permit a treatment $\times$ period interaction. Other examples of naturally occurring blocks are animal litters, families, and classrooms. Constructed blocks could be made up of sets of subjects matched on age, race, and gender.

Blocking is done for two purposes:

1. To obtain smaller residual variability
2. To examine treatments under a wide range of conditions

A basic design principle is to partition a population of study units in such a way that background variability between blocks is maximized, and consequently, background variability within blocks is minimized.

Definition 10.8. In a *randomized block design*, each treatment is given once and only once in each block. Within a block, the treatments are assigned randomly to the experimental units.

Note that a randomized block design, by definition, is a balanced design: This is somewhat restrictive. For example, in animal experiments it would require litters to be of the same size.

The statistical model associated with the randomized block design is

$$Y_{ij} = \mu + \beta_i + \tau_j + \epsilon_{ij}, \qquad i = 1, \dots, I, \quad j = 1, \dots, J \tag{23}$$

and (**1**) $\sum \beta_i = \sum \tau_j = 0$ and (**2**) ϵ are iid $N(0, \sigma^2)$. In this model, β_i is the effect of block i and τ_j the effect of treatment j. In this model, as indicated, we assume no interaction between blocks and treatments (i.e., if there is a difference between treatments, the magnitude of this effect does not vary from block to block except for random variation). In Section 10.6 we discuss a partial check on the validity of the assumption of no interaction.

The analysis of variance table for this design is a simplified version of Table 10.12: The number of observations is the same in each block and for each treatment. In addition, there is no SS for interaction; another way of looking at this is that the SS for interaction is the error SS. The calculations are laid out in Table 10.15.

Tests of significance proceed in the usual way. The expected mean squares can be derived from Table 10.12, making use of the simpler design.

The computations for the randomized block design are very simple. You can verify that

$$\mathrm{SS}_\mu = \frac{Y_{..}^2}{n}, \qquad \mathrm{SS}_\beta = \frac{\sum Y_{i\cdot}^2}{J} - \frac{Y_{..}^2}{n}, \qquad \mathrm{SS}_\tau = \frac{\sum Y_{\cdot j}^2}{I} - \frac{Y_{..}^2}{n} \tag{24}$$

$$\mathrm{SS}_\epsilon = \sum Y_{ij}^2 - \frac{Y_{..}^2}{n} - \mathrm{SS}_\beta - \mathrm{SS}_\tau$$

Example 10.6. The pancreas, a large gland, secretes digestive enzymes into the intestine. Lack of this fluid results in bowel absorption problems (steatorrhea); this can be diagnosed by excess fat in feces. Commercial pancreatic enzyme supplements are available in three forms: capsule, tablets, and enteric-coated tablets. The enteric-coated tablets have a protective shell to prevent gastrointestinal reaction. Graham [1977] investigated the effectiveness of these three formulations in six patients with steatorrhea; the three randomly assigned treatments were preceded by a control period. For purposes of this example, we will consider the control period as a treatment, even though it was not randomized. The data are displayed in Table 10.16.

To use equation 4, we will need the quantities

$$Y_{..} = 618.6, \qquad \frac{\sum Y_{i\cdot}^2}{4} = 21{,}532.80, \qquad \frac{\sum Y_{\cdot j}^2}{6} = 17{,}953.02, \qquad \sum Y_{ij}^2 = 25{,}146.8$$

The analysis of variance table, omitting SS_μ, is displayed in Table 10.17.

The treatment effects are highly significant. A visual inspection of Table 10.16 suggests that capsules and tablets are the most effective, enteric-coated tablets less effective. The author points out that the "normal" amount of fecal fat is less than 6 g per day, suggesting that, at best, the treatments are palliative. The F-test for patients is also highly significant, indicating that the levels among patients varied considerably: Patient 4 had the lowest average level at 6.1 g in 24 hours; patient 5 had the highest level, with 47.1 g in 24 hours.

Table 10.15 Layout for the Randomized Block Design[a]

Source of Variation	d.f.	SS[b]	MS	F-Ratio	d.f. of F-Ratio	E(MS)	Hypothesis Being Tested
Grand mean	1	$SS_\mu = n\overline{Y}^2_{..}$	$MS_\mu = SS_\mu$	$\dfrac{MS_\mu}{MS_\epsilon}$	$(1, (I-1)(J-1))$	$\sigma^2 + ij\mu^2$	$\mu = 0$
Blocks	$I-1$	$SS_\beta = J\sum(\overline{Y}_{i\cdot} - \overline{Y}_{..})^2$	$MS_\beta = \dfrac{SS_\beta}{I-1}$	$\dfrac{MS_\beta}{MS_\epsilon}$	$(I-1, (I-1)(J-1))$	$\sigma^2 + \dfrac{J\sum\beta_i^2}{I-1}$	$\beta_i = 0$ for all i
Treatments	$J-1$	$SS_\tau = I\sum(\overline{Y}_{\cdot j} - \overline{Y}_{..})^2$	$MS_\tau = \dfrac{SS_\tau}{J-1}$	$\dfrac{MS_\tau}{MS_\epsilon}$	$(J-1, (I-1)(J-1))$	$\sigma^2 + \dfrac{I\sum\tau_j^2}{J-1}$	$\tau_j = 0$ for all j
Residual	$(I-1)(J-1)$	$SS_\epsilon = \sum(Y_{ij} - \overline{Y}_{i\cdot} - \overline{Y}_{\cdot j} + \overline{Y}_{..})^2$	$MS_\epsilon = \dfrac{SS_\epsilon}{(I-1)(J-1)}$	—	—	σ^2	
Total	IJ	$\sum Y_{ij}^2$	—				

[a]Model: $Y_{ij} = \mu + \beta_i + \tau_j + \epsilon_{ij}$ [where $\epsilon_{ij} \sim$ iid $N(0, \sigma^2)$].
Data : $Y_{ij} = \overline{Y}_{..} + (\overline{Y}_{i\cdot} - \overline{Y}_{..}) + (\overline{Y}_{\cdot j} - \overline{Y}_{..}) + (\overline{Y}_{ij} - \overline{Y}_{i\cdot} - \overline{Y}_{\cdot j} + \overline{Y}_{..})$.
Equivalent model : $Y_{ij} \sim N(\mu_{ij}, \sigma^2)$, where the Y_{ij}'s are independent.
[b]Summation is over all displayed subscripts.

Table 10.16 Effectiveness of Pancreatic Supplements on Fat Absorption in Patients with Steatorrhea (Grams/Day)

Case	None (Control)	Tablet	Capsule	Enteric-Coated Tablet	$Y_i.$	$\overline{Y}_i.$
1	44.5	7.3	3.4	12.4	67.6	16.9
2	33.0	21.0	23.1	25.4	102.5	25.6
3	19.1	5.0	11.8	22.0	57.9	14.5
4	9.4	4.6	4.6	5.8	24.4	6.1
5	71.3	23.3	25.6	68.2	188.4	47.1
6	51.2	38.0	36.0	52.6	177.8	44.4
$Y._j$	228.5	99.2	104.5	186.4	618.6	—
$\overline{Y}._j$	38.1	16.5	17.4	31.1	$\overline{Y}.. = 25.8$	

Source: Data from Graham [1977].

Table 10.17 Randomized Block Analysis of Fecal Fat Excretion of Patients with Steatorrhea

Source of Variation	d.f.	SS	MS	F-Ratio	p-Value
Patients (blocks)	5	5588.38	1117.68	10.44	<0.001
Treatments	3	2008.60	669.53	6.26	<0.01
Residual	15	1605.40	107.03	—	—
Total	23	9202.38	—	—	—

Source: Data from Graham [1977].

10.3.3 Analyses of Randomized Block Designs Using Ranks

A nonparametric analysis of randomized block data using only the ranks was developed by Friedman [1937]. The model is that of equation (23), but the ϵ_{ij} are no longer required to be normally distributed. We assume that there are no ties in the data; for a small number of ties the average ranks may be used. The idea of the test is simple: If there are no treatment effects ($\tau_j = 0$ for all j), the ranks of the observations within a block are randomly distributed. For block i, let

$$R_{ij} = \text{rank of } Y_{ij} \text{ among } Y_{i1}, Y_{i2}, \ldots, Y_{iJ}$$

The Friedman statistic for testing the null hypothesis $H_0 : \tau_j = 0$ (where $j = 1, \ldots, J$) is

$$T_{\text{FR}} = 12I \sum_{j=1}^{J} \frac{(\overline{R}._j - \overline{R}..)^2}{J(J+1)} \tag{25}$$

Computationally, the following formula is easier:

$$T_{\text{FR}} = \frac{12}{IJ(J+1)} \sum_{j=1}^{J} R._j^2 - 3(I)(J+1) \tag{26}$$

The null hypothesis is rejected for large values of T_{FR}. For small randomized block designs, the critical values of T_{FR} are tabulated; see, for example, Table 39 in Odeh et al. [1977], which goes up to $I = J = 6$. As the number of blocks becomes very large, the distribution of T_{FR}

approaches that of a χ^2-distribution with $(J-1)$ degrees of freedom. See also Notes 10.1 and 10.2.

Example 10.6. (*continued*)   Replacing the observations for each *individual* by their ranks produces Table 10.18. For individual 4, the two tied observations are replaced by the average of the two ranks. [As a check, the total $R_{..}$ of ranks must be $R_{..} = IJ(J+1)/2$. (Why?) For this example $I = 6$, $J = 4$, $IJ(J+1)/2 = (6 \cdot 4 \cdot 5)/2 = 60$, and $R_{..} = 22 + 8.5 + 9.5 + 20 = 60$.] The Friedman statistic, using equation (26), has the value

$$\begin{aligned} T_{\text{FR}} &= \frac{12}{6 \times 4 \times 5}(22^2 + 8.5^2 + 9.5^2 + 20^2) - (3 \times 6 \times 5) \\ &= 104.65 - 90 = 14.65 \end{aligned}$$

This quantity is compared to a χ^2 distribution with 3 d.f. $(14.65/3 = 4.88)$; the *p*-value is $p = 0.0021$. From exact tables such as Odeh et al. [1977], the exact *p*-value is $p < 0.001$. The conclusion is the same as that of the analysis of variance in Section 10.3.2. Note also that the ranking of treatments in terms of the total ranks is the same as in Table 10.11. For an alternative rank analysis of these data, see Problem 10.20.

10.3.4   Types of ANOVA Models

In Section 10.2.2, two examples were mentioned of one-way analyses of variance. The first dealt with the age at which children begin to walk as a function of various training procedures; the second example dealt with patient hospitalization costs, based on an examination of some hospitals (treatments) selected randomly from all the hospitals in a large metropolitan area (from each hospital selected, a specified number of patient records are selected for cost analysis). The experimental design associated with the first example differs from the second: In a repetition of the first study, the same set of treatments could be used; in the second study, a new set of hospitals could presumably be selected; that is, the "treatment levels" are randomly selected from a larger set of treatment levels.

Definition 10.9.   If the levels of a classification variable in an ANOVA situation are selected at random from a population, the variable is said to be a *random factor* or *random effect*. Factors with the levels fixed by those conducting the study or which are fixed classifications (e.g., gender) are called *fixed factors* or *fixed effects*.

Table 10.18   Rank Values for Supplement Use

	Treatment			
Case	Control	Tablet	Capsule	Enteric-Coated Tablet
1	4	2	1	3
2	4	1	2	3
3	3	1	2	4
4	4	1.5	1.5	3
5	4	1	2	3
6	3	2	1	4
$R_{.j}$	22	8.5	9.5	20

Definition 10.10. ANOVA situations with all classification variables fixed are called *fixed effects models* (model I). If all the classification variables are random effects, the design is a *random effects model* (model II). If both random and fixed effects are present, the design is a *mixed effects model*.

Historically, no distinction was made between model I and II designs, in part due to identical analyses in simple situations and similar analyses in more complicated situations. Eisenhart [1947] was the first to describe systematically the differences between the two models. Some other examples of random effects models are:

1. A manufacturer of spectrophotometers randomly selects five instruments from its production line and obtains a series of replicated readings on each machine.

2. To estimate the maximal exercise performance in a healthy adult population, 20 subjects are selected randomly and 10 independent estimates of maximal exercise performance for each person are obtained.

3. To determine knowledge about the effect of drugs among sixth graders, a researcher randomly selects five sixth-grade classes from among the 100 sixth-grade classes in a large school district. Each child selected fills out a questionnaire.

How can we determine whether a design is model I or model II? The basic criterion deals with the population to which inferences are to be made. Another way of looking at this is to consider the number of times randomness is introduced (ideally). In Example 10.2 there are two sources of randomness: subjects and observations within subjects. If more than one "layer of randomness" has to be passed through in order to reach the population of interest, we have a random effects model.

An example of a mixed model is example 2 above with a further partitioning of subjects into male and female. The factor, gender, is fixed.

Sometimes a set of data can be modeled by either a fixed or random effects model. Consider example 1 again. Suppose that a cancer research center has bought the five instruments and is now running standardization experiments. For the purpose of the research center, the effects of machines are fixed effects.

To distinguish a random effects model from a fixed effects model, the components of the model are written as random variables. The two-way random effects ANOVA model with interaction is written as

$$Y_{ijk} = \mu + A_i + B_j + G_{ij} + e_{ijk}, \qquad i = 1, \ldots, I, \quad j = 1, \ldots, J, \quad k = 1, \ldots, n_{ij} \quad (27)$$

The assumptions are:

1. e_{ijk} are iid $N(0, \sigma^2)$, as before.
2. A_i are iid $N(0, \sigma_\alpha^2)$.
3. B_j are iid $N(0, \sigma_\beta^2)$.
4. G_{ij} are iid $N(0, \sigma_\gamma^2)$.

The total variance can now be partitioned into several components (hence another term for these models: *components of variance models*). Assume that the experiment is balanced with $n_{ij} = m$ for all i and j. The difference between the fixed effect and random effect model is in the expected mean squares. Table 10.19 compares the EMS for both models, taking the EMS for the fixed effect model from Table 10.12.

The test for interaction is the same in both models. However, if interaction is present, to be valid the test for main effects in the random effects model must use MS_γ in the denominator rather than MS_ϵ.

Table 10.19 **Comparison of Expected Mean Squares in the Two-Way ANOVA, Fixed Effect vs. Random Effect Models**[a]

Source of Variation	d.f.	EMS	
		Fixed Effect	Random Effect
Row main effects	$I - 1$	$\sigma^2 + \dfrac{Jm \sum \alpha_i^2}{I - 1}$	$\sigma^2 + m\sigma_\gamma^2 + mJ\sigma_\alpha^2$
Column main effects	$J - 1$	$\sigma^2 + \dfrac{Im \sum \beta_j^2}{J - 1}$	$\sigma + m\sigma_\gamma^2 + mI\sigma_\beta^2$
Row × column interaction	$(I - 1)(J - 1)$	$\sigma^2 + \dfrac{IJm \sum \gamma_{ij}^2}{(I - 1)(J - 1)}$	$\sigma^2 + m\sigma_\gamma^2$
Residual	$n.. - IJ$	σ^2	σ^2

[a]There are **m** observations in each cell.

The null hypothesis

$$H_0 : \gamma_{ij} = 0 \qquad \text{all } i \text{ and } j$$

in the fixed effect model has as its counterpart,

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

in the random effect model. In both cases the test is carried out using the ratio MS_γ / MS_ϵ with $(I - 1)(J - 1)$ and $n - IJ$ degrees of freedom. If interaction is not present, the tests for main effects are the same in both models. However, if H_0 is not rejected, the tests for main effects are different in the two models. In the random effects model the expected mean square for main effects now contains a term involving σ_γ^2. Hence the appropriate F-test involves MS_γ in the denominator rather than MS_ϵ; the degrees of freedom are changed accordingly.

Several comments can be made:

1. Frequently, the degrees of freedom associated with MS_γ are fewer than those of MS_ϵ, so that there is a loss of precision if MS_γ has to be used to test main effects.

2. From a design point of view, if m, I, and J can be chosen, it may pay to choose m small and I, J relatively large if a random effects model is appropriate. A minimum of two replicates per treatment combination is needed to obtain an estimate of σ^2. If possible, the rest of the observations should be allocated to the levels of the variables. This may not always be possible, due to costs or other considerations. If the total cost of the experiment is fixed, an algorithm can be developed for choosing the values of m, I, and J.

3. The difference between the fixed and random effects models for the two-way ANOVA designs is not as crucial as it seems. We have indicated caution in proceeding to the tests of main effects if interaction is present in the fixed model (see Figure 10.3 and associated discussion). In the random effects model, the same precaution holds. It is perhaps too strong to say that main effects should not be tested when interaction is present, but you should certainly be able to explain what information you hope to obtain from such tests after a full interpretation of the (significant) interaction.

4. Expected mean squares for an unbalanced random effects model are not derivable or are very complicated. A more useful approach is that of multiple regression, discussed in Chapter 11. See also Section 10.5.

5. For the randomized block design the MS_ϵ can be considered the mean square for interaction. Hence, in this case the F-tests are appropriate for both models. (Does this contradict the

statement made in comment 3?) Note also that there is little interest in the test of block effects, except as a verification that the blocking was effective.

Good discussions about inference in the case of random effects models can be found in Snedecor and Cochran [1988] and Winer [1991].

10.4   REPEATED MEASURES DESIGNS AND OTHER DESIGNS

10.4.1   Repeated Measures Designs

Consider a situation in which blood pressures of two populations are to be compared. One person is selected at random from each population. The blood pressure of each of the two subjects is measured 100 times. How would you react to data analysis that used the two-sample *t*-test with two samples of size 100 and showed that the blood pressures differed in the two populations? The idea is ridiculous, but in one form or another appears frequently in the research literature. Where does the fallacy lie? There are two sources of variability: within individuals and among individuals. The variability within individuals is assumed incorrectly to represent the variability among individuals. Another way of saying this is that the 100 readings are not independent samples from the population of interest. They are repeated measurements on the same experimental unit. The repeated measures may be useful in this context in pinning down more accurately the blood pressure of the two people, but they do not make up for the small sample size. Another feature we want to consider is that the sequence of observations within the person cannot be randomized, for example, a sequence of measurements of growth. Thus, typically, we do not have a randomized block design.

Definition 10.11.   In a *repeated measures design*, multiple (two or more) measurements are made sequentially on the same observational unit.

A repeated measures design usually is an example of a mixed model with the observational unit a random effect (e.g., persons or animals, and the treatments on the observational units fixed effects). Frequently, data from repeated measure designs are somewhat unbalanced and this makes the analysis more difficult. One approach is to summarize the repeated measures in some meaningful way by single measures and then analyze the single measures in the usual way. This is the way many computer programs analyze such data. We motivate this approach by an example. See Chapter 18 for further discussion.

Example 10.7.   Hillel and Patten [1990] were interested in the effect of accessory nerve injury as result of neck surgery in cancer. The surgery frequently decreases the strength of the arm on the affected side. To assess the potential recovery, the unaffected arm was to be used as a control. But there is a question of the comparability of arms due to dominance, age, gender, and other factors. To assess this effect, 33 normal volunteers were examined by several measurements. The one discussed here is that of torque, or the ability to abduct (move or pull) the shoulder using a standard machine built for that purpose. The subjects were tested under three consecutive conditions (in order of increasing strenuousness): $90°, 60°$, and $30°$ per second. The data presented in Table 10.20 are the best of three trials under each condition. For completeness, the age and height of each of the subjects are also presented. The researchers wanted answers to at least five questions, all dealing with differences between dominant and nondominant sides:

1. Is there a difference between the dominant and nondominant arms?
2. Does the difference vary between men and women?

Table 10.20   Peak Torque for 33 Subjects by Gender, Dominant Arm, and Age Group under Three Conditions

Subject		Age	Height (in.)	Weight (lb)	90°		60°		30°	
					DM[a]	ND[a]	DM	ND	DM	ND
Female	1	20	64	107	17	13	20	17	23	22
	2	23	68	140	25	25	28	29	31	31
	3	23	67	135	27	28	30	31	32	33
	4	23	67	155	23	28	27	29	27	32
	5	25	65	115	15	11	15	13	17	17
	6	26	68	147	27	17	25	21	32	27
	7	31	62	147	25	17	25	21	29	24
	8	31	66	137	19	15	17	17	21	19
	9	33	66	160	28	26	31	27	31	31
	10	36	66	118	23	23	26	27	27	25
	11	56	67	210	23	31	37	44	49	53
	12	59	67	130	15	17	17	19	20	20
	13	60	63	132	17	15	19	21	24	28
	14	60	64	180	15	15	17	19	19	21
	15	67	62	135	13	5	15	8	15	14
	16	73	62	124	11	9	13	13	19	17
Male	1	26	69	140	43	43	44	43	49	41
	2	28	71	175	45	43	48	45	53	52
	3	28	70	125	25	29	29	37	39	41
	4	28	70	175	39	41	49	47	55	44
	5	29	72	150	38	33	40	33	44	37
	6	30	68	145	53	41	51	40	59	44
	7	31	74	240	60	49	71	54	68	53
	8	32	67	168	32	31	37	31	39	30
	9	40	69	174	47	37	43	47	49	53
	10	41	72	190	33	25	29	25	39	27
	11	41	68	184	39	24	43	25	39	33
	12	56	70	200	21	11	23	12	33	24
	13	58	72	168	41	35	45	37	49	39
	14	59	73	170	31	32	31	31	35	38
	15	60	73	225	39	41	47	45	55	49
	16	68	67	140	31	23	33	27	37	33
	17	72	69	125	13	17	17	19	17	25

Source: Data from Hillel and Patten [1990].

[a]DM, dominant arm; ND, nondominant arm.

3. Does the difference depend on age, height, or weight?

4. Does the difference depend on treatment condition?

5. Is there interaction between any of the factors or variables mentioned in questions 1 to 4?

For purposes of this example, we only address questions 1, 2, 4, and 5, leaving question 3 for the discussion of analysis of covariance in Chapter 11.

The second to fourth columns in Table 10.21 contain the differences between the dominant and nondominant arms; the fifth to seventh columns are reexpressions of the three differences as follows. Let d90, d60, and d30 be the differences between the dominant and nondominant

Table 10.21 Differences in Torque under Three Conditions and Associated Orthogonal Contrasts[a]

		DM–ND			Orthogonal Contrasts		
		90°	60°	30°	Constant	Linear	Quadratic
Female	1	4	3	1	4.6	2.1	−0.4
	2	0	−1	0	−0.6	0.0	0.8
	3	−1	−1	−1	−1.7	0.0	0.0
	4	−5	−2	−5	−6.9	0.0	−2.4
	5	4	2	0	3.5	2.8	0.0
	6	10	4	5	11.0	3.5	2.9
	7	8	4	5	9.8	2.1	2.0
	8	4	0	2	3.5	1.4	2.4
	9	2	4	0	3.5	1.4	−2.4
	10	0	−1	2	0.6	−1.4	1.6
	11	−8	−7	−4	−11.0	−2.8	0.8
	12	−2	−2	0	−2.3	−1.4	0.8
	13	2	−2	−4	−2.3	4.2	0.8
	14	0	−2	−2	−2.3	1.4	0.8
	15	8	7	1	9.2	4.9	−2.0
	16	2	0	2	2.3	0.0	1.6
Male	1	0	1	8	5.2	−5.7	2.4
	2	2	3	1	3.5	0.7	−1.2
	3	−4	−8	−2	−8.1	−1.4	4.1
	4	−2	2	11	6.4	−9.2	2.0
	5	5	7	7	11.0	−1.4	−0.8
	6	12	11	15	21.9	−2.1	2.0
	7	11	17	15	24.8	−2.8	−3.3
	8	1	6	9	9.2	−5.7	−0.8
	9	10	−4	−4	1.2	9.9	5.7
	10	8	4	12	13.9	−2.8	4.9
	11	15	18	6	22.5	6.4	−6.1
	12	10	11	9	17.3	0.7	−1.2
	13	6	8	10	13.9	−2.8	0.0
	14	−1	0	−3	−2.3	1.4	−1.6
	15	−2	2	6	3.5	−5.7	0.0
	16	8	6	4	10.4	2.8	0.0
	17	−4	−2	−8	−8.1	2.8	−3.3

Source: Data from Hillel and Patten [1990].
[a] See Table 10.20 for notation.

arms under each of the three conditions. Then we define

$$\text{constant} = \frac{d90 + d60 + d30}{\sqrt{3}}$$

$$\text{linear} = \frac{d90 - d30}{\sqrt{2}}$$

$$\text{quadratic} = \frac{d90 - 2 \cdot d60 + d30}{\sqrt{6}}$$

For example, for the first female subject, rounding off to one decimal place yields

$$\frac{4 + 3 + 1}{\sqrt{3}} = 4.6$$

$$\frac{4 - 1}{\sqrt{2}} = 9.9$$

$$\frac{4 - 2 \times (3) + 1}{\sqrt{6}} = -0.4$$

The first component clearly represents an average difference of dominance over the three conditions. The divisor is chosen to make the variance of this term equal to the variance of a single difference. The second term represents a slope within an individual. If the three conditions were considered as values of a predictor variable with values -1 (for $30°$), 0 (for $60°$), and 1 (for $90°$), the slope would be expressed as in the second, or linear, term. The linear term assesses a possible trend in the differences over the three conditions within an individual. The last term, the quadratic term, fits a quadratic curve through the data assessing possible curvature or nonlinearity within an individual. This partitioning of the observations within an individual has the property that sums of squares are maintained. For example, for the first female subject,

$$4^2 + 3^2 + 1^2 = 26 = (4.6)^2 + (2.1)^2 + (-0.4)^2$$

except for rounding. (If you were to calculate these terms to more decimal places, you would find that the right side is identical to the left side.) In words, the variability in response within an individual has been partitioned into a constant component, a linear component, and a quadratic component. The questions posed can now be answered unambiguously since the three components have been constructed to be *orthogonal*, or uncorrelated. An analysis of variance is carried out on the three terms; unlike the usual analysis of variance, a term for the mean is included; results are summarized in Table 10.22. We start by discussing the analysis of the quadratic component. The analysis indicates that there are no significant differences between males and females in terms of the quadratic or nonlinear component. Nor is there an overall effect. Next, conclusions are similar for the linear effect. We conclude that there is no linear trend for abductions at $90°, 60°$, and $30°$. This leaves the constant term, which indicates (1)

Table 10.22   ANOVA and Means of the Data in Table 10.21

Source of Variation		d.f.	SS	MS	F-Ratio
		Analysis of Variance			
Constant	Mean	1	900.7	900.7	13.3
	Gender	1	438.5	438.5	6.48
	Error 1	31	2099.2	67.72	
Linear	Mean	1	0.33	0.33	0.02
	Gender	1	33.43	33.43	2.43
	Error 2	31	426.0	13.74	
Quadratic	Mean	1	3.09	3.09	0.50
	Gender	1	0.70	0.70	0.11
	Error 3	31	191.2	6.17	
		Means			

		Constant	Linear	Quadratic
Female ($n = 16$)	Mean	1.306	1.138	0.456
	Standard deviation	5.920	2.121	1.609
Male ($n = 17$)	Mean	8.600	-0.876	0.165
	Standard deviation	9.917	4.734	3.085

that there is a significant gender effect of dominance ($F_{1,31} = 6.48$, $p < 0.05$) and an overall dominance effect. The average of the constant term for females is 1.31, for males is 8.6. One question that can be raised is whether the difference between female and male is a true gender difference or can be attributed to differences is size. An analysis of covariance can answer this question (see Problem 11.38).

Data from a repeated measures design often look like those of a randomized block design. The major difference is the way the data are generated. In the randomized block, the treatments are allocated randomly to a block. In the repeated measures design, this is not the case; not being possible, as in the case of observations over time, or because of experimental constraints, as in the example above. If the data are analyzed as a randomized block, care must be taken that the assumptions of the randomized block design are satisfied. The key assumption is that of *compound symmetry*: The sample correlations among treatments over subjects must all estimate the same population correlation. The randomization ensures this in the randomized block design. For example, for the data in Table 10.16, the correlations are as follows:

	Control	Tablet	Capsule
Tablet	0.658		
Capsule	0.599	0.960	
Coated tablet	0.852	0.784	0.833

These correlations are reasonably comparable. If the correlations are not assumed equal, a conservative F-test can be carried out by referring the observed value of F for treatments to an F-table with 1 and $(I - 1)$ [rather than $(J - 1)$ and $(I - 1)(J - 1)$] degrees of freedom). Alternatives to the foregoing two approaches include multivariate analyses. There is a huge literature on repeated measures analysis. The psychometric literature contains many papers on this topic. To explore this area, consult recent issues of journals such as *American Statistician*. One example is a paper by Looney and Stanley [1989]. See also Chapter 18.

10.4.2 Factorial Designs

An experimental layout that is very common in agricultural and nutritional studies is the balanced factorial design. It is less common in medical research, due to the ever-present risk of missing observations and ethical constraints.

Definition 10.12. In a *factorial design* each level of a factor occurs with every level of every other factor. Experimental units are assigned randomly to treatment combinations.

Suppose that there are three factors with levels $I = 3$, $J = 2$, and $K = 4$. Then there are $3 \times 2 \times 4 = 24$ treatment combinations. If there are three observations per combination, 72 experimental units are needed. Factorial designs, if feasible, are very economical and permit assessment of joint effects of treatments that are not possible with experiments dealing with one treatment at a time. The two-way analysis of variance can be thought of as dealing with a two-factor experiment. The generalization to three or more factors does not require new concepts or strategies, just increased computational complexity.

10.4.3 Hierarchical or Nested Designs

A hierarchical or nested design is illustrated by the following example. As part of a program to standardize measurement of the blood level of phenytoin, an antiepileptic drug, samples with known amounts of active ingredients are sent to four commercial laboratories for analysis. Each

laboratory employs a number of technicians who make one or more determinations of the blood level. A possible layout is the following:

Laboratory	1		2			3		4	
Technician	1	2	3	4	5	6	7	8	9
Assay	∧	∧	∧	∧	∧	∧	∧	∧	∧

In this example, laboratory 2 employs three technicians who routinely do this assay; all other laboratories use two technicians. In laboratory 3, each technician runs three assays; in the other laboratories each technician runs two assays. There are three factors: laboratories, technicians, and assays; the arrangement is *not* factorial: there is no reason to match technician 1 with any technician from another laboratory.

Definition 10.13. In a *hierarchical or nested design* levels of one or more factors are subsampled within one or more other factors. In other words, the levels of one or more factors are not crossed with one or more other factors.

In the example above, the factors, "technicians" and "assay," are not "crossed" with the first factor but rather nested within that factor. For the factor "technician" to be "crossed," its levels would have to repeat within each level of "laboratory." That is why we deliberately labeled the levels of "technician" consecutively and introduced some imbalance. Determining whether a design is factorial or hierarchical is not always easy. If the first of the two technicians within a laboratory was the senior technician and the second (or second and third) a junior technician, then "technician" could perhaps be thought of as having two levels, "senior" and "junior," which could then be crossed with "laboratory." A second reason is that designs are sometimes mixed, having both factorial and hierarchical components. In the example above, if "technician" occurred at two levels, "technician" and "laboratory" could be crossed or factorial, but "assay" would continue to be nested within "technician."

10.4.4   Split-Plot Designs

A related experimental design is the split-plot design. We illustrate it with an example. We want to test the effect of physiotherapy in conjunction with drug therapy on the mobility of patients with arthritis. Patients are randomly assigned to physiotherapy, and each patient is given a standard drug and a placebo in random order. The experimental layout is as follows:

		Physiotherapy				
		$i = 1$ (Yes)		$i = 2$ (No)		
k	**Patient**	1	$2 \cdots J$	1	$2 \cdots J$
1	Drug	Y_{111}	—$\cdots$—	Y_{211}	—$\cdots$—
2	Placebo	Y_{112}	—$\cdots$—	Y_{212}	—$\cdots$—

The patients form the "whole plots" and the drug administration, the "split plot." These designs are characterized by almost separate analyses of specified effects. To illustrate in this example, let

$$D_{ij} = Y_{ij1} - Y_{ij2} \quad \text{and} \quad T_{ij} = Y_{ij1} + Y_{ij2}, \qquad i = 1, 2, \quad j = 1, \ldots, J$$

In words, D_{ij} is the difference between drug and placebo for patient j receiving physiotherapy level i; T_{ij} is the sum of readings for drug and placebo. Now carry out an analysis of variance (or two-sample t-test) on each of these variables; see Table 10.23.

Table 10.23 ANOVA **Table for Split-Plot Design**

| One-Way ANOVA | d.f. | Interpretation of Split-Plot Analyses | |
		Differences	Sums
Mean	1	Mean differences	Mean sums
Between groups	1	Differences × physiotherapy	Sums × physiotherapy
Within groups	$2(J-1)$	Differences within physiotherapy	Sums within physiotherapy
Total	$2J$	"Total"	"Total"

An analysis of variance of the sums is, in effect, an assessment of physiotherapy (averaged or summed over drug and placebo), that is, a comparison of $\overline{T}_1.$ and $\overline{T}_2..$

The analysis of differences is very interesting. The assessment of the significance of "between groups" is a comparison of the average differences between drug and placebo with physiotherapy and without physiotherapy; that is, $\overline{D}_1. - \overline{D}_2.$ is a test for interaction. Additionally, the "mean differences" term can be used to test the hypothesis that $\overline{D}..$ comes from a population with mean zero, that is, it is a comparison of drug and placebo. This test makes sense only if the null hypothesis of no interaction is not rejected.

These remarks are intended to give you an appreciation for these designs. For more details, consult a text on design of experiments, such as Winer [1971].

10.5 UNBALANCED OR NONORTHOGONAL DESIGNS

In previous sections we have discussed balanced designs. The balanced design is necessary to obtain an additive partition of the sum of squares. If the design is not balanced, there are basically three strategies available; the first is to try to restore balance. If only one or two observations are "missing," this is a possible strategy, but if more than two or three are missing, a second or third alternative will have to be used. The second alternative is to use an unweighted means analysis. The third strategy is to use a multiple regression approach; this is discussed in detail in Section 11.10.

10.5.1 Causes of Imbalance

Perhaps the most important thing you can do in the case of unbalanced data is to reflect on the reason(s) for the imbalance. If the imbalance is due to some random mechanism unrelated to the factors under study, the procedures discussed below are appropriate. If the imbalance is due to a specific reason, perhaps related to the treatment, it will be profitable to think very carefully about the implications. Usually, such imbalance suggests a bias in the treatment effects. For example, if a drug has major side effects which cause patients to drop out of a study, the effect of the drug may be estimated inappropriately if only the remaining patients are used in the analysis; if one does the analysis only on patients for whom "all data are available," biased estimates may result.

10.5.2 Restoring Balance

Missing Data in the Randomized Block Design

Suppose that the ijth observation is missing in a randomized block design consisting of I blocks and J treatments. The usual procedure is to:

1. Estimate the missing data point by least squares using the formula

$$\widehat{Y}_{ij} = \frac{IY_i. + JY._j - Y..}{(I-1)(J-1)} \tag{28}$$

where the row, column, and grand totals are those for the values present.

2. Carry out the usual analysis of variance on this augmented data set.

3. Reduce the degrees of freedom for MS_ϵ by 1.

If more than one observation is missing, say two or three, values are guessed for all but one, the latter is estimated by equation (28), a second missing value is deleted, and the process is repeated until convergence. The degrees of freedom for MS_ϵ are now reduced by the number of observations that are missing.

Example 10.6. (*continued*)   We return to Table 10.11. Suppose that observation $Y_{31} = 19.1$ is missing and we want to estimate it. For this example, $I = 6$, $J = 4$, $Y_3. = 38.8$, $Y_{\cdot 1} = 209.4$, and $Y.. = 599.5$. We estimate Y_{31} by

$$\widehat{Y}_{31} = \frac{6(38.8) + 4(209.4) - 599.5}{(6-1)(4-1)} = 31.4$$

This value appears to be drastically different from 19.1; it is. It also indicates that there is no substitute for real data. The analysis of variance is not altered a great deal (see Table 10.24).

The F-ratios have not changed much from those in Table 10.12. So in this case, the conclusions are unchanged. Note that the degrees of freedom for residual are reduced by 1. This means that the critical values of the F-statistics are increased slightly. Therefore, this experiment has less power than the one without missing data.

Missing Data in Two-Way and Factorial Designs

If a cell in a two-way design has a missing observation, it is possible to replace the missing point by the mean for that cell, carry out the analysis as before, and subtract one degree of freedom for MS_ϵ. A second approach is to carry out an unweighted means analysis. We illustrate both procedures by means of an example.

Example 10.8.   These data are part of data used in Wallace et al. [1977]. The observations are from a patient with prostatic carcinoma. The question of interest is whether the immune system of such a patient differs from that of noncarcinoma subjects. One way of assessing this is to stimulate in vitro the patient's lymphocytes with phytohemagglutinin (PHA). This causes blastic transformation. Of interest is the amount of blastogenic generation as measured by DNA incorporation of a radioactive compound. The data observed are the mean radioactive counts per minute both when stimulated with PHA and when not stimulated by PHA. As a control, the amount of PHA stimulation in a pooled sera of normal blood donors was used. To examine the response of a subject's lymphocytes, the quantity

$$\frac{\dfrac{\text{subject's mean count/minute stimulated with PHA}}{\text{subject's mean count/minute without PHA}}}{\dfrac{\text{normal sera mean count/minute stimulated with PHA}}{\text{normal sera mean count/minute without PHA}}} = \frac{X_{11}/X_{12}}{X_{21}/X_{22}} \qquad (29)$$

Table 10.24   ANOVA for Example 10.6

Source of Variable	d.f.	SS	MS	F-Ratio
Patients (blocks)	5	5341.93	1068.39	9.90
Treatments	3	2330.30	776.77	7.20
Residual	14	1510.94	107.92	—
Total	22	9183.17	—	—

Table 10.25   DNA Incorporation of Sera of Patient with Prostatic Carcinoma Compared to Sera from Normal Blood Donors[a]

Subject	Radioactivity (counts/min)	
	With PHA	Without PHA
Patient sera	129,594 (11.772)	301 (5.707)
	143,687 (11.875)	333 (5.808)
	115,953 (11.661)	295 (5.687)
	103,098 (11.543)	285 (5.652)
	98,125 (11.494)	
Blood donor sera	43,125 (10.672)	247 (5.509)
	46,324 (10.743)	298 (5.697)
	42,117 (10.648)	387 (5.958)
	45,482 (10.725)	
	31,192 (10.348)	

[a] $\log_e$ of counts in parentheses.

was used. If the lymphocytes responded in the same way to the subject's sera and the pooled sera, the ratio should be approximately equal to 1. The data are displayed in Table 10.25.

There is a great deal of variability in the counts/minute values as related to level. In Section 10.6.3 we suggest that logarithms are appropriate for stabilization of the variability. There is a bonus involved in this case. Under the null hypothesis of no difference in patient and blood donor sera, the ratio in equation (28) is 1; that is,

$$H_0 : \frac{E(X_{11})/E(X_{12})}{E(X_{21})/E(X_{22})} = 1$$

This is equivalent to

$$H_0 : \log_e \frac{E(X_{11})/E(X_{12})}{E(X_{21})/E(X_{22})} = \log_e 1 = 0$$

or

$$\log_e E(X_{11}) - \log_e E(X_{12}) - \log_e E(X_{21}) + \log_e E(X_{22}) = 0 \tag{30}$$

Now define

$$Y_{ijk} = \log_e X_{ijk}, \qquad i = 1, 2, \quad j = 1, 2, \quad k = 1, \dots, n_{ij}$$

It can be shown that equation (30) is zero only if the true interaction term is zero. Thus, the hypothesis that the patient's immune system does not differ from that of noncarcinoma subjects is translated into a null hypothesis about interaction involving the logarithms of the radioactive counts.

We finally get to the "missing data" problem. The data are not balanced: $n_{ij} \neq n_i.n._j/n..$ [we could delete one observation from the (1,2) cell, but considering the small numbers, we want to retain as much information as possible]. One strategy is to add an observation to cell (2,2) equal to the mean for that cell and adjust the degrees of freedom for interaction. The mean $\overline{Y}_{22}.$ is 5.721. The analysis of variance becomes as shown in Table 10.26.

Note that the MS for error has 13 degrees of freedom, not 14. The MS for error will be the correct estimate using this procedure, but the MS for interaction (and main effects) will not be the same as the one obtained by techniques of Chapter 11. However, it should be close.

Table 10.26   ANOVA for the Missing Data Problem

Source	d.f.	SS	MS	F-Ratio	p-Value
Subject	1	1.4893	1.4893	—	—
PHA	1	131.0722	131.0722	—	—
PHA × subject	1	1.2247	1.2247	50.0	<0.001
Error	13	0.3184	0.02449	—	—
Total	16	—	—	—	—

10.5.3   Unweighted Means Analysis

The second approach is that of unweighted mean analysis. Again, assuming that the unequal cell frequencies are not due to treatment effects, the cell means are used and an average sample size calculated for each cell. The appropriate average sample size is given by the harmonic mean. In the context of our example, the harmonic mean is defined to be

$$\widetilde{n} = \frac{IJ}{1/n_{11} + 1/n_{12} + 1/n_{21} + 1/n_{22}}$$

where n_{ij} is the number of observations in cell (i, j). The harmonic mean is used because the standard error of the mean of cell (i, j) is proportional to $1/n_{ij}$. All calculations for row and column effects are now based on cell means and the harmonic mean of the cell sample sizes. Write the cell means and marginal means as follows:

$$\begin{array}{cc|c} \overline{Y}_{11.} & \overline{Y}_{12.} & \widehat{M}_{1.} \\ \overline{Y}_{21.} & \overline{Y}_{22.} & \widehat{M}_{2.} \\ \hline \widehat{M}_{.1} & \widehat{M}_{.2} & \widehat{M}_{..} \end{array}$$

The marginal and overall means are just the arithmetic average of the cell means, that is, the unweighted average (hence the name *unweighted mean analysis*). The row and column sums of squares are calculated as follows:

$$SS_\alpha = \widetilde{n}J \sum (\overline{M}_{i.} - \overline{M}_{..})^2$$

$$SS_\beta = \widetilde{n}I \sum (\overline{M}_{.j} - \overline{M}_{..})^2$$

$$SS_\gamma = \widetilde{n} \sum (\overline{Y}_{ij.} - \overline{M}_{i.} - \overline{M}_{.j} + \overline{M}_{..})^2$$

SS_ϵ is calculated in the usual way: $SS_\epsilon = \sum (Y_{ijk} - \overline{Y}_{ij.})^2$. For the example, the calculations are

Means

$$\begin{array}{cc|c} 11.669000 & 5.713500 & 8.691250 \\ 10.627200 & 5.721333 & 8.174266 \\ \hline 11.148100 & 5.717416 & 8.432758 \end{array}$$

The harmonic mean $\widetilde{n}$ is

$$\widetilde{n} = \frac{(2)(2)}{1/5 + 1/4 + 1/5 + 1/3} = 4.067797$$

$$SS_\mu = (4.067797)(2) \left[(8.691250 - 8.432758)^2 + (8.174266 - 8.432758)^2 \right] = 1.0872$$

Table 10.27 ANOVA **Table for Unweighted Means**

Source	d.f.	SS	MS	F-Ratio	p-Value
Subject	1	1.0872	1.0872		
PHA	1	119.688	119.6888	1	
PHA $\times$ subject	1	1.1204	1.1204	45.7	<0.001
Error	13	0.3184	0.02449		
Total	16				

$$SS_\beta = (4.067797)(2)\left[(11.148100 - 8.432758)^2 + (5.717416 - 8.432758)^2\right] = 119.6888$$

$$SS_\gamma = (4.067797)\left[(4)(0.262408)^2\right] = 1.1204$$

making use of the fact that all the interaction deviations are equal in absolute value:

$$\overline{Y}_{11\cdot} - \overline{M}_{1\cdot} - \overline{M}_{\cdot 1} + \overline{M}_{\cdot\cdot} = 0.262408$$
$$\overline{Y}_{12\cdot} - \overline{M}_{1\cdot} - \overline{M}_{\cdot 2} + \overline{M}_{\cdot\cdot} = -0.262408, \ldots$$

The ANOVA table based on the unweighted means is shown in Table 10.27.

The conclusion remains unchanged. It turns out in this case that the test for interaction is identical to the multiple regression procedure of Chapter 11.

10.6 VALIDITY OF ANOVA MODELS

10.6.1 Assumptions in ANOVA Models

All the models considered in this chapter have assumed at least the following:

1. Homogeneity of variance
2. Normality of the residual error
3. Statistical independence of the residual errors
4. Linearity of the model

For example, consider again the model associated with the one-way analysis of variance (omitting the subscripts):

$$Y = \mu + \alpha + \epsilon$$

We assumed that (1) the error term ϵ had constant variance for all values of μ and α, and was normally distributed; (2) values of ϵ were randomly (independently) selected; and (3) the response Y was related linearly to μ, α, and ϵ.

In addition, the random effects and repeated measures models made assumptions about the covariances of the random factors and the residual error; other models assumed zero interaction (additivity).

If one or more of the assumptions does not hold, one of the following approaches is frequently used:

1. The data are analyzed by a method that makes fewer assumptions: for example, nonparametric analysis.

2. Part of the data is eliminated or not used, for example, extreme values (i.e., outliers) are deleted or replaced by less extreme values. Deletion usually induces bias.

3. The measurement variables are replaced by categorical variables and some kind of analysis of frequencies is carried out; for example, "age at first pregnancy" is replaced by "teenage mother: yes–no," and the number of observations in various categories is now the outcome variable.

4. A weighted analysis is done; for example, if the variance is not constant at all levels of response, the responses are weighted by the inverse of the variances. The log-linear models of Chapter 7 are an example of a weighting procedure.

5. The data are "transformed" to make the assumptions valid. Typical transformations are: logarithmic, square root, reciprocal, and arcsin $\sqrt{\quad}$. These transformations are nonlinear. Linear transformations do not alter the analysis of variance tests.

6. Finally, appeal is made to the "robustness" of the ANOVA and the analysis is carried out anyway. This is a little bit like riding a bicycle without holding onto the handle bars; it takes experience and courage. If you arrive safely, everyone is impressed, if not, they told you so.

The most common approach is to transform the data. There are advantages and disadvantages to transformations. A brief discussion is presented in the next section. In the other sections we present specific tests of the assumptions of the ANOVA model.

10.6.2 Transformations

Some statisticians recommend routine transformations of data before any analysis is carried out. We recommend the contrary approach; do not carry out transformations unless necessary, and then be very careful, particularly in estimation. We discuss this more fully below, but first we present some common transformations. Table 10.28 lists seven of the most commonly used transformations and one somewhat more specialized one. Each row in the table lists some of the characteristics of the transformation and its uses. A large number of these transformations are variance stabilizing. For example, if the variance of Y is $\lambda^2 \mu_Y$, where λ is a constant and μ_Y is the expected value of Y, then $\sqrt{Y}$ tends to have a variance that is constant and equal to $\lambda^2/4$. Hence, this transformation is frequently associated with a Poisson random variable: in this case $\lambda = 1$, so that $\sqrt{Y}$ tends to have a variance of 1/4 regardless of the value of μ_Y. This result is approximate in that it holds for large values of μ_Y. However, the transformation works remarkably well even for small μ_Y, say, equal to 10. Freeman and Tukey [1950] have proposed a modification of the square root transformation which stabilizes the variance for even smaller values of μ_Y. Variance stabilizing transformations tend to be normalizing as well and can be derived explicitly as a function of the variance of the original variable.

The logarithmic transformation is used to stabilize the variance and/or change a multiplicative model into an linear model. When the standard deviation of Y is proportional to μ_Y the logarithmic transformation tends to stabilize the variance. The reciprocal transformation (one per observation) is used when the variance is proportional to μ_Y^4. These first three transformations deal with a progression in the dependence of the variance of Y on μ_Y: from μ_Y to μ_Y^4. The transformations consist of raising Y to an exponent from $Y^{1/2}$ to Y^{-1}. If we define the limit of Y^b to be $\log_e Y$ as b approaches 0, these transformations represent a gradation in exponents. A further logical step is to let the data determine the value of b. This transformation, Y^b, is an example of a power transformation. (*Power* here does not imply "powerful" but simply that Y is raised to the bth power.) See Note 10.4 for additional comments.

The next two transformations are used with proportions or rates. The first one of these is the ubiquitous logistic transformation, which is not variance stabilizing but does frequently induce linearity (cf. Section 7.5). The angle transformation is variance stabilizing but has a finite range; it is not used much anymore because computational power is now available to use the more complex but richer logistic transformation.

Table 10.28 Characteristics of Some Common Transformations of a Random Variable Y

$W = g(Y)$	Range of Y	Variance[a] Y	Variance[a] W	Normalizing	Stabilizing	Linearity	Comments	Uses
$\sqrt{Y}$	$0 \le Y \le \infty$	$\lambda^2 \mu_Y$	$\lambda^2/4$	U	Y	—	for $\mu_Y < 10$ use : $W = 1/2(\sqrt{Y} + \sqrt{Y+1})$ (Freeman Tukey transformation)	Poisson
$\log_e Y$	$0 \le Y \le \infty$	$\lambda^2 \mu_Y^2$	λ^2	U	Y	C	Use $\log_e(Y+1)$ if zeroes occur	Wide range of Y, e.g., 1–1000
$\dfrac{1}{Y}$	$0 \le Y \le \infty$	$\lambda^2 \mu_Y^4$	λ^2	U	Y	C	Use $1/(Y+1)$ if zeroes occur	Survival time, response time
Y^b	$0 \le Y \le \infty$	—	1	Y	Y	C	Box–Cox transformation, Power transformation	Generalized transformation
$\log_e \dfrac{Y}{1-Y}$	$0 < Y < 1$	$\lambda^2 \mu_Y(1-\mu_Y)$	$\dfrac{\lambda^2}{\mu_Y(1-\mu_Y)}$	U	N	C	Logit transformation	Logistic regression, binomial
$\arcsin\sqrt{Y}$	$0 \le Y \le 1$	$\lambda^2 \mu_Y(1-\mu_Y)$	$\lambda^2/4$	U	Y	C	"Angle" transformation	Binomial
$1/2\log_e \dfrac{1+Y}{1-Y}$	$-1 \le Y \le 1$	$\lambda^2(1-\mu_Y^2)^2$	λ^2	Y	Y	C	R. A. Fisher's Z-transformation	Normalize correlation coefficient
$\Phi^{-1}\left(\dfrac{\text{rank } Y}{n}\right)$	$-\infty \le Y \le \infty$	—	1	Y	Y	—	Normal scores transformation $(\text{Rank}(Y) - 1/2)/n$ is sometimes used	Nonparametric analysis

[a] C, could; N, no; U, usually; Y, yes.

399

The Fisher Z-transformation is used to transform responses whose range is between -1 and $+1$. It was developed specifically for the Pearson product-moment correlation coefficient and discussed in Chapter 9. Finally, we mention one transformation via ranks, the normal scores transformation. This transformation is used extensively in nonparametric analyses and discussed in Chapter 8.

There are benefits to the use of transformations. It is well to state them explicitly since we also have some critical comments. The benefits include the following:

1. Methods using the normal distribution can be used.

2. Tables, procedures, and computer programs are available.

3. A transformation derived for one purpose tends to achieve some other purposes as well—but not always.

4. Inferences (especially relating to hypothesis testing) can be made more easily.

5. Confidence intervals in the transformed scale can be "transformed back" (but estimates of standard errors cannot).

Transformations are more useful for testing purposes than for estimation. The following drawbacks of transformations should be kept in mind:

1. The order of statistics may not be preserved. Consider the following two sets of data: sample 1 : 1, 10; sample 2 : 5, 5. The arithmetic means are 5.5 and 5.0, respectively. The geometric means (i.e., the antilogarithms of the arithmetic mean of the logarithms of the observations) are 3.2 and 5.0, respectively. Hence, the ordering of the *means* is not preserved by the transformation (the ordering of the *raw* data is preserved).

2. Contrary to some, we think that there may be a "natural scale" of measurement. Some examples of variables with a natural scale of measurement are "life expectancy" measured in years, days, or months; cost of medical care in dollars; number of accidents attributable to alcoholism. Administrators or legislators may not be impressed with, or willing to think about, the cost of medical care in terms of "square root of millions of dollars expended."

3. Closely related is the problem of bias. An obvious approach to the criticism in our discussion of drawback 2 is to do the analysis in the transformed units and then transform back to the original scale. Unfortunately, this introduces bias as mentioned in our discussion of drawback 1. Formally, if Y is the variable of interest and $W = g(Y)$ its transform, then it is usually the case that

$$E(W) \neq g(E(Y))$$

There are ways of assessing this bias and eliminating it but such methods are cumbersome and require an additional layer of computations, something the transformation was often designed to reduce!

4. Finally, many of the virtues of transformations are asymptotic virtues; they are approached as the sample size becomes very large. This should be kept in mind when analyzing relatively small data sets.

10.6.3 Testing of Homogeneity of Variance

It is often the case that the variance or standard deviation is proportional to the mean level of response. There are two common situations where this occurs. First, where the range of response varies over two or more orders of magnitude; second, in situations where the range of response is bounded, on the left, the right or both. Examples of the former are Poisson random variables; examples of the latter, responses such as proportions, rates, or random variables that cannot be negative.
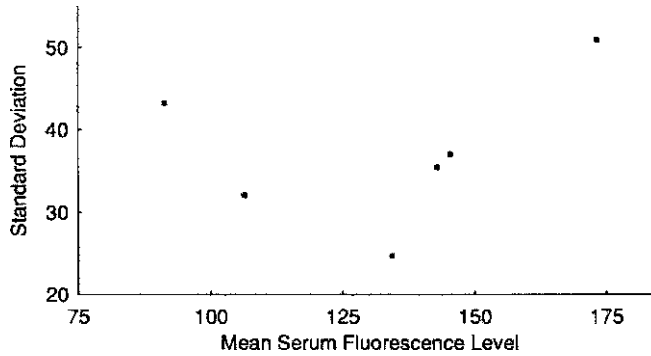
Figure 10.6 Mean serum fluorescence level and standard deviation. (Data from Sherwin and Layfield [1976]; see Example 10.5.)

The simplest verification of homogeneity of variance is provided by a graph, plotting the variance or standard deviation vs. the level of response.

Example 10.5. (*continued*)   In Table 10.8, the means and standard deviations of serum fluorescence readings of mice exposed to nitrogen dioxide are given. In Figure 10.6 the standard deviations are plotted against the means of the various treatment combinations. This example does not demonstrate any pattern between the standard deviation and the cell means. It would not be expected because the range of the cell means is fairly small.

Example 10.9.   A more interesting example is the data of Quesenberry et al. [1976] discussed in Problem 3.14. Samples of peanut kernels were analyzed for aflatoxin levels. Each sample was divided into 15 or 16 subsamples. There was considerable variability in mean levels and corresponding standard deviations.

A plot of means vs. standard deviations displays an increasing pattern, suggesting a logarithmic transformation to stabilize the variance. This transformation as well as two other transformations ($\sqrt{Y}, Y^{1/4}$) are summarized in Table 10.29. Means vs. standard deviations are

Table 10.29   Aflatoxin Levels in Peanut Kernels: Means and Standard Deviations for 11 Samples Using Transformations

		Mean and Standard Deviation of Aflatoxin Level								
		Y		$W = Y^{1/4}$		$W = \sqrt{Y}$		$W = \log Y$		
Sample	n	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	16	110	25.6	3.2	0.192	10.4	1.24	4.7	0.240
2	16	79	20.6	3.0	0.204	8.8	1.19	4.3	0.281
3	16	21	3.9	2.1	0.109	4.5	0.45	3.0	0.213
4	16	33	12.2	2.4	0.192	5.7	0.96	3.4	0.311
5	15	32	10.6	2.4	0.194	5.6	0.92	3.4	0.328
6	16	15	2.7	2.0	0.089	3.8	0.35	2.7	0.183
7	15	33	6.2	2.4	0.111	5.8	0.54	3.5	0.183
8	16	31	2.8	2.4	0.054	5.6	0.26	3.4	0.092
9	16	17	4.2	2.0	0.129	4.1	0.51	2.8	0.261
10	16	8	3.1	1.7	0.143	2.9	0.49	2.1	0.339
11	15	84	17.7	3.0	0.164	9.1	0.98	4.4	0.221

Source: Data from Quesenberry et al. [1976].
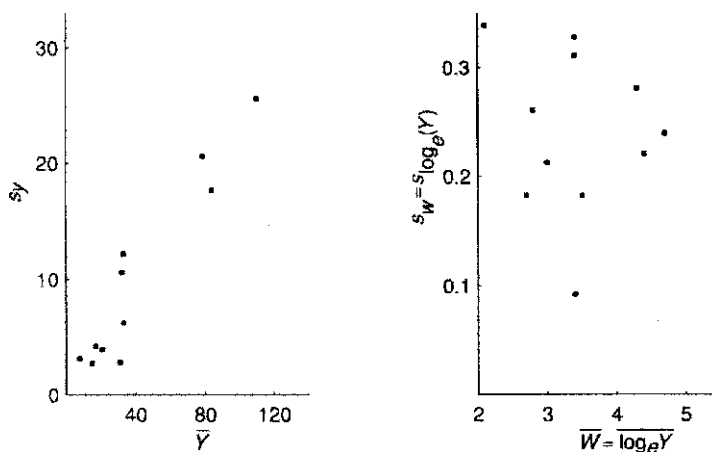
Figure 10.7 Means vs. standard deviation, arithmetic and logarithmic scales. (Data from Wallace et al. [1977]; see Example 10.8.)

plotted in Figure 10.7. The first pattern clearly indicates a linear trend; the plot for the data expressed as logarithms suggests very little pattern. This does not prove that the lognormal model is appropriate. Quesenberry et al. [1976], in fact, considered two classes of models: the 11 samples are from normal distributions with means and variances $\mu_i, \sigma_i^2, i = 1, \ldots, 11$; the second class of models assumes that the logarithms of the aflatoxin levels for the 11 samples come from normal distributions with means and variances $\gamma_i, \theta^2, i = 1, \ldots, 11$.

On the basis of their analysis, they conclude that the normal models are more appropriate. The cost is, of course, that 10 more parameters have to be estimated. Graphs of means vs. standard deviation for the $\sqrt{Y}$ and $Y^{1/4}$ scale still suggest a relationship.

The tests of homogeneity of variance developed here are graphical. There are more formal tests. All of the tests assume normality and are sensitive to departure from normality. In view of the robustness of the analysis of variance to heterogeneity of variance, Box [1953] remarked that "... to make the preliminary tests on variances is rather like putting to sea in a rowing boat to find out whether conditions are sufficiently calm for an ocean liner to leave port." There are four common tests of homogeneity of variance, associated with the names of Hartley, Cochran, Bartlett, and Scheffé. Only the first two are described here, they will be adequate for most purposes. For a description of the other tests see, for example, Winer [1971]. Suppose that there are k samples with sample size n_i and sample variance $s_i^2, i = 1, \ldots, k$. For the moment, assume that all n_i are equal to n. Hartley's test calculates

$$F_{\text{MAX}} = \frac{s_{\text{maximum}}^2}{s_{\text{minimum}}^2}$$

Cochran's test calculates

$$C = \frac{s_{\text{maximum}}^2}{\sum S_i^2}$$

In the absence of software for computing critical values, both statistics can be referred to appropriate tables in the Web appendix. If the sample sizes are not equal, the tables can be entered with the minimum sample size to give a conservative test and with the maximum

Table 10.30   Calculations for Example 10.9

Scale	$F_{\max}$	C
Y	$\left(\dfrac{25.6}{2.7}\right)^2 = 89.9$	$\dfrac{(25.7)^2}{1758.1} = 0.38$
$\sqrt{Y}$	$\left(\dfrac{1.24}{0.26}\right)^2 = 22.7$	$\dfrac{(1.24)^2}{9.787} = 0.16$
$Y^{1/4}$	$\left(\dfrac{0.204}{0.054}\right)^2 = 14.1$	$\dfrac{(0.204)^2}{0.252} = 0.16$
$\log_e Y$	$\left(\dfrac{0.339}{0.092}\right)^2 = 13.6$	$\dfrac{(0.339)^2}{0.694} = 0.17$
Critical value at 0.05 level	5.8	0.15

sample size to give a "liberal" test (i.e., the null hypothesis is rejected more frequently than the nominal significance level).

Example 10.9. (*continued*)   For the transformations considered, the F_{MAX} test and C test statistics are as shown in Table 10.30.

The critical values have been obtained by interpolation. The F_{MAX} test indicates that none of the transformations achieve satisfactory homogeneity of variance, validating one of Quesenberry et al.'s conclusions. The Cochran test suggests that there is little to choose between the three transformations.

A question remains: How valid is the analysis of variance under heterogeneity of variance? Box [1953] indicates that for three treatments a ratio of 3 in the maximum-to-minimum *population* variance does not alter the significance level of the test appreciably (one-way ANOVA model with $n_i = 5$, $I = 3$). The analysis of variance is therefore reasonably robust with respect to deviation from homogeneity of variance.

10.6.4   Testing of Normality in ANOVA

Tests of normality are not as common or well developed as tests of homogeneity of variance. There are at least two reasons: first, they are not as crucial because even if the underlying distribution of the data is not normal, appeal can be made to the central limit theorem. Second, it turns out that fairly large sample sizes are needed (say, $n > 50$) to discriminate between distributions. Again, most tests are graphical.

Consider for simplicity the one-way analysis of variance model

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \qquad i = 1, \ldots, I, \quad j = 1, \ldots, n_i$$

By assumption the ϵ_{ij} are iid $N(0, \sigma^2)$. The ϵ_{ij} are estimated by

$$\epsilon_{ij} = Y_{ij} - \overline{Y}_i.$$

The e_{ij} are normally distributed with population mean zero; $\sum e_{ij}^2/(n - I)$ is an unbiased estimate of σ^2 but the e_{ij} are not statistically independent. They can be made statistically independent, but it is not worthwhile for testing the normality. Some kind of normal probability plot is usually made and a decision made based on a visual inspection. Frequently, such a plot is used to identify outliers. Before giving an example, we give a simple procedure which is based on the use of order statistics.

Definition 10.14. Given a sample of n observations, $Y_1, Y_2, \ldots, Y_n$, the *order statistics* $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$ are the values ranked from lowest to highest.

Now suppose that we generate samples of size n from an $N(0, 1)$ distribution and average the values of the order statistics.

Definition 10.15. *Rankits* are the expected values of the order statistics of a sample of size n from an $N(0, 1)$ distribution. That is, let $Z_{(1)}, \ldots, Z_{(n)}$ be the order statistics from an $N(0, 1)$ population; then the rankits are $E(Z_{(1)}), E(Z_{(2)}), \ldots, E(Z_{(n)})$.

Rankits have been tabulated in Table A.13. A plot of the order statistics of the residuals against the rankits is equivalent to a normal probability plot. A reasonable approximation for the ith rankit is given by the formula

$$E(Z_{(i)}) \doteq 4.91[p^{0.14} - (1 - p)^{0.14}] \tag{31}$$

where

$$p = \frac{i - 3/8}{n + 1/4}$$

For a discussion, see Joiner and Rosenblatt [1971]. To illustrate its use we return to Example 10.1. A one-way analysis of variance was constructed for these data and we now want to test the normality assumption.

Example 10.1. (*continued*)   The distribution of ages at which infants first walked [discussed in Section 10.2.1 (see Table 10.1)] is now analyzed for normality. The residuals $Y_{ij} - \overline{Y}_i$. for the 23 observations are:

−1.125	−0.375	−0.208	0.900
−0.625	−1.375	0.292	−0.850
−0.375	−1.375	−2.708	−0.350
−0.125	0.375	−0.208	1.150
2.875	−0.875	1.542	−0.850
−0.625	3.625	1.292	

Note that the last observation has been omitted again so that we are working with the 23 observations given in the paper. These observations are now ranked from smallest to largest to be plotted on probability paper. To illustrate the use of rankits, we will calculate the expected values of the 23 normal $(0,1)$ order statistics using equation (31). The 23 order statistics for e_{ij}, $e_{(ij)}$, and the corresponding rankits are presented in Table 10.31.

For example, the largest deviation is -2.708; the expected value of $Z_{(1)}$ associated with this deviation is calculated as follows:

$$p = \frac{1 - 3/8}{23 + 1/4} = 0.02688$$

$$E(Z_{(1)}) = 4.91[(0.02688)^{0.14} - (1 - 0.02688)^{0.14}]$$

$$= -1.93$$

The rankits and the ordered residuals are plotted in Figure 10.8. What do we do with this graph? Is there evidence of nonnormality?

Table 10.31   Order Statistics for Example 10.1

$e_{(ij)}$	$E(Z_{(ij)})$	$e_{(ij)}$	$E(Z_{(ij)})$	$e_{(ij)}$	$E(Z_{(ij)})$
−2.708	−1.93	−0.625	−0.33	0.375	0.57
−1.375	−1.48	−0.375	−0.22	0.900	0.70
−1.375	−1.21	−0.375	−0.11	1.150	0.84
−1.125	−1.01	−0.350	0.0	1.292	1.01
−0.875	−0.84	−0.208	0.11	1.542	1.21
−0.850	−0.70	−0.208	0.22	2.875	1.48
−0.850	−0.57	−0.125	0.33	3.625	1.93
−0.625	−0.44	−0.292	0.44		
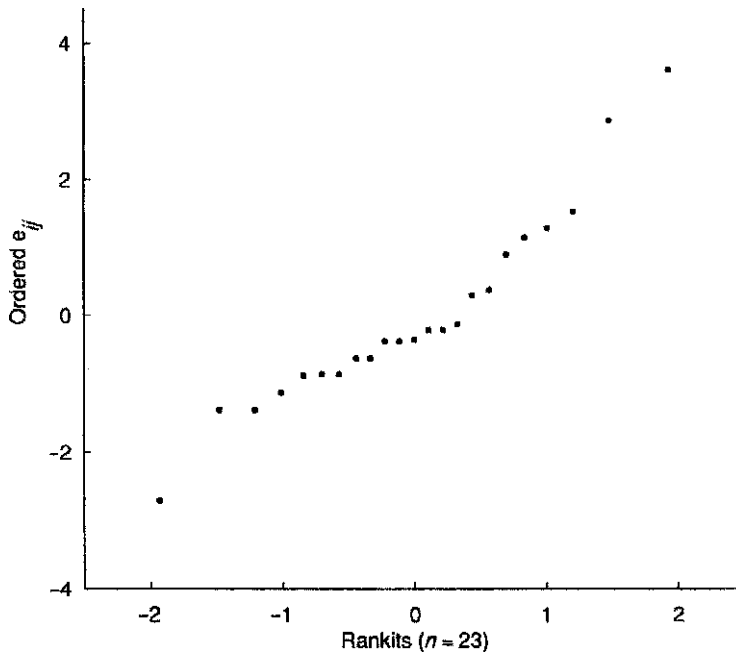
Figure 10.8   Normal probability plot of residuals from linear model. (Data from Zelazo et al. [1972]; see Example 10.1.)

There does seem to be some excessive deviation in the tails. The question is: How important is it? One way to judge this would be to generate many plots for normal and nonnormal data and compare the plots to develop a visual "feel" for the data. This has been done by Daniel and Wood [1999] and Daniel [1976]. Comparison of this plot with the plots in Daniel and Wood suggests that these data deviate moderately from normality. For a further discussion, see Section 11.8.1.

More formal tests of normality can be carried out using the Kolmogorov–Smirnov test of Chapter 8. A good test is based on the Pearson product-moment correlation of the order statistics and corresponding rankits. If the residuals are normally distributed, there should be a very high correlation between the order statistics and the rankits. The (null) hypothesis of normality is rejected when the correlation is *not large enough*. Weisberg and Bingham [1975] show that this is a very effective procedure. The critical values for the correlation have been tabulated; see, for example, Ryan et al. [1980]. For $n \geq 15$, the critical value is on the order of 0.95 or more. This

is a simple number to remember. For Example 10.1, the correlation between the order statistics of the residuals, $e_{(ij)}$ and the rankits $E(Z_{(ij)})$ is $r = 0.9128$ for $n = 23$. This is somewhat lower than the critical value of 0.95 again, suggesting that the residuals are "not quite" normally distributed.

10.6.5 Independence

One of the most difficult assumptions to verify is that of statistical independence of the residuals. There are two problems. First, tests of independence for continuous variables are difficult to implement. Frequently, such tests are, in fact, tests of no correlation among the residuals, so that if the errors are normally distributed and uncorrelated, they are independent. Second, the observed residuals in the analysis of variance have a built-in dependence due to the constraints on the linear model. For example, in the one-way analysis of variance with I treatments and, say, $n_i. = m$ observations per treatment, there are mI residuals but only $(m - 1)I$ degrees of freedom; this induces some correlation among the residuals. This is not an important dependence and can be taken care of.

Tests for dependence usually are tests for serial correlation (i.e., correlation among adjacent values). This assumes that the observations can be ordered in space or time. The most common test statistic for serial correlation is the Durbin–Watson statistic. See, for example, Draper and Smith [1998]. Computer packages frequently will print this statistic assuming that the observations are entered in the same sequence in which they were obtained. This, of course, is rarely the case and the statistic and its value should not be used. Such "free information" is sometimes hard to ignore; the motto for computer output is *caveat lector* (let the reader beware).

10.6.6 Linearity in ANOVA

Like independence, linearity is difficult to verify. In Example 10.7 we illustrated a multiplicative model. The model was transformed to a linear (nonadditive) model by considering the logarithm of the original observations. Other types of nonlinear models are discussed in Chapters 11 to 15. Evidence for a nonlinear model may consist of heterogeneity of variance or interaction. However, this need not always be the case. Scheffé [1959] gives the following example. Suppose that there are $I + J + 1$ independent Poisson variables defined as follows: $U_1, U_2, \ldots, U_I$ have means $\alpha_1, \alpha_2, \ldots, \alpha_I$; $V_1, V_2, \ldots, V_J$ have means $\beta_1, \beta_2, \ldots, \beta_J$; and W has mean γ. Let $Y_{ij} = W + U_i + V_j$; then $E(Y_{ij}) = \gamma + \alpha_i + \beta_j$; that is, we have an additive, linear model. But $\text{var}(Y_{ij}) = \gamma + \alpha_i + \beta_j$, so that there is heterogeneity of variance (unless all the α_i are equal and all the β_j are equal). The square root transformation destroys the linearity and the additivity. Scheffé [1959] states: "It is not obvious whether Y or $\sqrt{Y}$ is more nearly normal … but in the present context it hardly matters." A linear model is frequently assumed to be appropriate for a set of data without any theoretical basis. It may be a reasonable first-order approximation to the "state of nature" but should be recognized as such.

Sometimes a nonlinear model can be derived from theoretical principles. The form of the model may then suggest a transformation to linearity. But as the example above illustrates, the transformation need not induce other required properties of ANOVA models, or may even destroy them.

Another strategy for testing linearity is to add some nonlinear terms to the model and then test their significance. In Sections 11.7 and 11.8 we elaborate on this strategy.

10.6.7 Additivity

The term *additivity* is used somewhat ambiguously in the statistical literature. It is sometimes used to describe the transformation of a multiplicative model to a linear model. The effects of the treatment variables become "additive" rather than multiplicative. We have called such a transformation a *linearizing transformation*. It is not always possible to find such a transformation

(see Section 11.10.5). We have reserved the term *additivity* for the additive model illustrated by the two-way analysis of variance model (see Definition 10.4). A test for additivity then becomes a test for "no interaction." Scheffé [1959] proves that transformations to additivity exists for a very broad class of models.

The problem is that the existence of interaction may be of key concern. Consider Example 10.8. The existence of interaction in this example is taken as evidence that the immune system of a patient with prostatic carcinoma differed from that of normal blood donors. This finding has important implications for a theory of carcinogenesis. These data are an example of the importance of expressing observations in an appropriate scale. Of course, what evidence is there that the logarithms of the radioactive count is the appropriate scale? There is some arbitrariness, but the original model was stated in terms of percentage changes, and this implies constant changes on a logarithmic scale.

So the problem has been pushed back one step: Why state the original problem in terms of percentage changes? The answer must be found in the experimental situation and the nature of the data. Ultimately, the researcher will have to provide justification for the initial model used.

This discussion has been rather philosophical. One other situation will be considered: the randomized block design. There is no test for interaction because there is only one observation per cell. Tukey [1949] suggested a procedure that is an example of a general class of procedures. The validity of a model is evaluated by considering an enlarged model and testing the significance of the terms in the enlarged model. To be specific, consider the randomized block design model of equation (23):

$$Y_{ij} = \mu + \beta_i + \tau_j + \varepsilon_{ij}, \qquad i = 1, \dots, I, \quad j = 1, \dots, J$$

Tukey [1949] embedded this model in the "richer" model

$$Y_{ij} = \mu + \beta_i + \tau_j + \lambda \beta_i \tau_j + \varepsilon_{ij}, \qquad i = 1, \dots, I, \quad j = 1, \dots, J \tag{32}$$

He then proposed to test the null hypothesis,

$$H_0 : \lambda = 0$$

as a test for nonadditivity. Why this form? It is the simplest nonlinear effect involving both blocks and treatments. The term λ is estimated and tested as follows. Let the model without interaction be estimated by

$$Y_{ij} = \overline{Y}.. + b_i + t_j + e_{ij}$$

where
$$b_i = \overline{Y}_{i.} - \overline{Y}.., t_j = \overline{Y}._j - \overline{Y}.. \quad \text{and} \quad e_{ij} = Y_{ij} - \overline{Y}.. - b_i - t_j$$

We have the usual constraints,
$$\sum b_i = \sum t_j = 0$$

and
$$\sum_i e_{ij} = \sum_j e_{ij} = 0 \qquad \text{for all } i \text{ and } j$$

Now define
$$X_{ij} = b_i t_j, \qquad i = 1, \dots, I, \quad j = 1, \dots, J \tag{33}$$

It can be shown that the least squares estimate, $\widehat{\lambda}$, of λ is

$$\widehat{\lambda} = \frac{\sum X_{ij} Y_{ij}}{\sum X_{ij}^2} \tag{34}$$

Since $\overline{X} = 0$ (why?), the quantity $\widehat{\lambda}$ is precisely the regression of Y_{ij} on X_{ij}. The sum of squares for regression is the sum of squares for nonadditivity in the ANOVA table:

$$SS_\lambda = SS_{\text{nonadditivity}} = \frac{\left(\sum X_{ij} Y_{ij}\right)^2}{\sum X_{ij}^2} \tag{35}$$

The ANOVA table for the randomized block design including the test for nonadditivity is displayed in Table 10.32. As expected, the SS_λ has one degree of freedom since we are estimating a slope. But who "pays" for the one degree of freedom? A little reflection indicates that it must come out of the error term; the number of constraints on the block and treatment effects remain the same. A graph of Y_{ij} vs. X_{ij} (or equivalently, e_{ij} vs. X_{ij}) will indicate whether there is any pattern.

The idea of testing models within larger models as a way of testing the validity of the model is discussed further in Section 11.8.2.

Example 10.6. (*continued*)   We now apply the Tukey test for additivity to the experiment assessing the effect of pancreatic supplements on fat absorption in patients with steatorrhea, discussed in Section 10.3.2. We need to calculate SS_λ from equation (35) and this involves the regression of Y_{ij} on X_{ij}, where X_{ij} is defined by equation (33). To save space we calculate only a few of the X_{ij}. For example,

$$
\begin{aligned}
X_{11} &= \left(\overline{Y}_{1\cdot} - \overline{Y}_{\cdot\cdot}\right)\left(\overline{Y}_{\cdot 1} - \overline{Y}_{\cdot\cdot}\right) \\
&= (16.9 - 25.775)(38.083 - 25.775) \\
&= -109.2
\end{aligned}
$$

and

$$
\begin{aligned}
X_{23} &= \left(\overline{Y}_{2\cdot} - \overline{Y}_{\cdot\cdot}\right)\left(\overline{Y}_{\cdot 3} - \overline{Y}_{\cdot\cdot}\right) \\
&= (25.625 - 25.775)(17.417 - 25.775) \\
&= 1.3
\end{aligned}
$$

(Note that a few more decimal places for the means are used here as compared to Table 10.15.) A graph of Y_{ij} vs. X_{ij} is presented in Figure 10.9. The estimate of the slope is

$$
\begin{aligned}
\widehat{\lambda} &= \frac{\sum X_{ij} Y_{ij}}{\sum X_{ij}^2} \\
&= \frac{(-109.2)(44.5) + (82.0)(7.3) + \cdots + (98.8)(52.6)}{(-109.2)^2 + (82.0)^2 + \cdots + (98.8)^2} \\
&= \frac{13{,}807}{467{,}702} \\
&= 0.029521
\end{aligned}
$$

SS_λ is

$$SS_\lambda = \frac{(13{,}807)^2}{467{,}702} = 407.60$$

The analysis of variance is tabulated in Table 10.33.

Table 10.32 ANOVA of Randomized Block Design Incorporating Tukey Test for Additivity[a]

Source of Variation	d.f.	SS[b]	MS	F-Ratio	d.f.	E(MS)	Hypothesis Tested
Grand mean	1	$SS_\mu = n\bar{Y}^2_{..}$	$MS_\mu = SS_\mu$	$\dfrac{MS_\mu}{MS_\epsilon}$	$(1, IJ-I-J)$	$n\mu^2 + \sigma^2$	$\mu = 0$
Blocks	$I-1$	$SS_\beta = J\sum(\bar{Y}_{i\cdot} - \bar{Y}_{..})^2$	$MS_\beta = \dfrac{SS_\beta}{I-1}$	$\dfrac{MS_\beta}{MS_\epsilon}$	$(I-1, IJ-I-J)$	$\dfrac{J\sum\beta_i^2}{I-1} + \sigma^2$	$\beta_i = 0$ all i
Treatments	$J-1$	$SS_\tau = I\sum(\bar{Y}_{\cdot j} - \bar{Y}_{..})^2$	$MS_\tau = \dfrac{SS_\tau}{J-1}$	$\dfrac{MS_\tau}{MS_\epsilon}$	$(J-1, IJ-I-J)$	$\dfrac{I\sum\tau_j^2}{J-1} + \sigma^2$	$\tau_j = 0$ all j
Nonadditivity	1	$SS_\lambda^c = \dfrac{(\sum X_{ij}Y_{ij})^2}{\sum X_{ij}^2}$	$MS_\lambda = SS_\lambda$	$\dfrac{MS_\lambda}{MS_\epsilon}$	$(1, IJ-I-J)$	$\lambda^2 C^d + \sigma^2$	$\lambda = 0$
Residual	$IJ-I-J$	$SS_\epsilon =$ by subtraction	$MS_\epsilon = \dfrac{SS_\epsilon}{IJ-I-J}$				
Total	IJ	$\sum Y_{ij}^2$					

[a] Model: $Y_{ij} = \mu + \beta_i + \tau_j + \lambda\beta_i\tau_j + \epsilon_{ij}$ [$\epsilon_{ij} \sim$ iid $N(0,\sigma^2)$]. Data: $Y_{ij} = \bar{Y}_{..} + (\bar{Y}_{i\cdot} - \bar{Y}_{..}) + (\bar{Y}_{\cdot j} - \bar{Y}_{..}) + \hat{\lambda}X_{ij} + \tilde{e}_{ij}$ (residual obtained by subtraction).

[b] Summation is over all subscripts displayed.

[c] $X_{ij} = (\bar{Y}_{i\cdot} - \bar{Y}_{..})(\bar{Y}_{\cdot j} - \bar{Y}_{..})$.

[d] C is a constant that depends on the cell means; it is zero if the additive model holds.
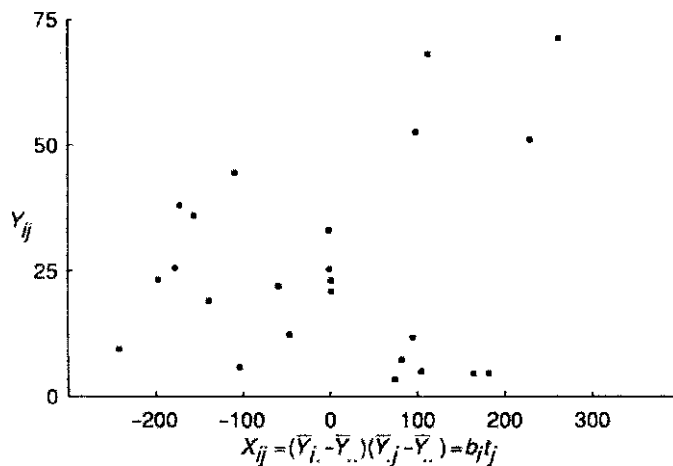
Figure 10.9   Plot of the Tukey test for additivity. See the text for an explanation.

Table 10.33   Randomized Block Analysis with Tukey Test for Additivity of Fecal Fat Excretion of Patients with Steatorrhea

Source of Variation	d.f.	SS	MS	F-Ratio	p-Value
Patients	5	5588.38	1117.68	13.1	<0.001
Treatments	3	2008.60	669.53	7.83	<0.01
Additivity	1	407.60	407.60	4.76	$0.025 < p < 0.05$
Residual	14	1197.80	85.557		
Total	23	9202.38			

Source: Data from Graham [1977].

The test for additivity indicates significance at the 0.05 level ($p = 0.047$); thus there is some evidence that the data cannot be represented by an additive model. Tukey [1949] related the constant a in Y^a (power transformation) to the degree of nonadditivity by the following formula:

$$\widehat{a} = 1 - \widehat{\lambda}\overline{Y}...$$

The quantity $\widehat{a}$ is a statistic and hence a random variable. For a particular set of data, the confidence interval on $\widehat{a}$ will tend to be fairly wide; hence, a "nice" value of "a" is usually chosen. For the example, $\widehat{a} = 1 - (0.029521)(25.775) = 0.239$. A "nice" value for "$a$" is thus 0.25, or even 0.20.

10.6.8   Strategy for Analysis of Variance

It is useful to have a checklist in carrying out an ANOVA. Not every item on the list needs to be considered, nor necessarily in the order given, but you will find it useful to be reminded of these items:

1. Describe how the data were generated: from what population? To what population will inferences be made? State explicitly at what steps in the data generation randomness entered.

2. Specify the ANOVA null hypotheses, alternative hypotheses; whether the model is fixed, random, or mixed.

3. Graph the data to get some idea of treatment effects, variability, and possible outliers.

4. If necessary, test the homogeneity of variance and the normality.

5. If ANOVA is inappropriate on the data as currently expressed, consider alternatives. If transformations are used, repeat steps 2 and 4.

6. Carry out the ANOVA. Calculate F-ratios. Watch out for F-ratios much less than 1; they usually indicate an inappropriate model.

7. State conclusions and limitations.

8. If null hypotheses are not rejected, consider the power of the study and of the analysis.

9. For more detailed analyses and estimation procedures, see Chapter 12.

NOTES

10.1 *Ties in Nonparametric Analysis of Variance (One-Way and Randomized Block)*

As indicated, both the Kruskal–Wallis and the Friedman tests are conservative in the presence of ties. The adjustment procedure is similar to those used in Chapter 8, equation (4). For the Kruskal–Wallis situation, let

$$C_{\mathrm{KW}} = \frac{\sum_{l=1}^{L} \left(t_l^3 - t_l\right)}{n^3 - n}$$

where L is the number of groups of tied ranks and t_l is the number of ties in group l, $l = 1, \ldots, L$. Then the statistic T_{KW} [equation (13)] is adjusted to $T_{\mathrm{ADJ}} = T_{\mathrm{KW}}/(1 - C_{\mathrm{KW}})$. Since $0 \leq C_{\mathrm{KW}} \leq 1$, $T_{\mathrm{ADJ}} \geq T_{\mathrm{KW}}$. Hence, if the null hypothesis is rejected with T_{KW}, it will certainly be rejected with T_{ADJ} since the degrees of freedom remain unchanged. Usually, C_{KW} will be fairly small: Suppose that there are 10 tied observations in an ANOVA of 20 observations; in this case $C_{\mathrm{KW}}(10^3 - 10)/(20^3 - 20) = 0.1241$, so that $T_{\mathrm{ADJ}} = T_{\mathrm{KW}}/(1 - 0.1241) = 1.14 T_{\mathrm{KW}}$. The adjusted value is only 14% larger than the value of T_{KW} even in this extreme situation. (If the 10 ties are made up of five groups of two ties each, the adjustment is less than 0.5%).

A similar adjustment is made for the Friedman statistic, given by equations (25) and (26). In this case,

$$C_{\mathrm{FR}} = \frac{\sum_{i=1}^{I} \sum_{l=1}^{L_i} \left(t_{il}^3 - t_{il}\right)}{I(J^3 - J)}$$

where t_{il} is the number of ties in group l within block i and untied values within a block are counted as a group of size 1. (Hence $\sum_{l=1}^{L_i} t_{il} = J$ for every i.) The adjusted Friedman statistic, T_{ADJ}, is $T_{\mathrm{ADJ}} = T_{\mathrm{FR}}/(1 - C_{\mathrm{FR}})$. Again, unless there are very many ties, the adjustment factor, C_{FR} will be relatively small.

10.2 *Nonparametric Analyses with Ordered Alternatives*

All the tests considered in this chapter have been "omnibus tests"; that is, the alternative hypotheses have been general. In the one-way ANOVA, the null hypothesis is $H_0 : \mu_1 = \mu_2 = \cdots = \mu_I = \mu$, the alternative hypothesis $H_1 : \mu_i \neq \mu_i'$ for at least one i and i'. Since the power of a test is determined by the alternative hypothesis, we should be able to "do better" using more specific alternative hypotheses. One such hypothesis involves ordered alternatives. For the one-way ANOVA (see Section 10.2), let $H_1 : \mu_1 \leq \mu_2 \leq \cdots \leq \mu_I$ with at least one strict

inequality. A regression-type parametric analysis could be carried out by coding the categories $X = 1, X = 2, \ldots, X = I$.

A nonparametric test of H_0 against an ordered alternative H_1 was developed by Terpstra and Jonckheere (see, e.g., Hollander and Wolfe [1999]). The test is based on the Mann–Whitney statistic (see Section 8.6). The Terpstra–Jonckheere statistic is

$$T_{\text{TJ}} = \sum_{i=1}^{I-1} \sum_{k=i+1}^{I} M_{ik} = \sum_{i<k} M_{ik}$$

where M_{ik} is the number of pairs with the observation in group i less than that of group $k (i < k)$ among the $n_i n_k$ pairs.

Under the null hypothesis $H_0 : \mu_1 = \mu_2 = \cdots = \mu_I = \mu$, the statistic T_{TJ} has a distribution that approaches a normal distribution as n becomes large, with mean and variance given by

$$E\,[T_{\text{TJ}}] = \frac{n^2 - \sum n_i^2}{4}$$

and

$$\text{var}[T_{\text{TJ}}] = \frac{[n^2(2n + 3) - \sum n_i^2(2n_i + 3)]}{72}$$

where $n = n_1 + n_2 + \cdots + n_I$. See Problems 10.3 and 10.11 for an application.

In Section 10.3.3, a nonparametric analysis of randomized block design was presented to test the null hypothesis $H_0 : \tau_1 = \tau_2 = \cdots = \tau_J = 0$. Again, we consider an ordered alternative, $H_1 : \tau_1 \leq \tau_2 \leq \cdots \leq \tau_J$ with at least one strict inequality. Using the notation of Section 10.3.3, let $R_{\cdot j} =$ sum of ranks for treatment j. Page [1963] developed a nonparametric test of H_0 against H_1. The statistic $T_{\text{PAGE}} = \sum_{j=1}^{J} j R_{\cdot j}$ under the null hypothesis approaches a normal distribution (as I become large) with mean and variance

$$E\,[T_{\text{PAGE}}] = \frac{I J^2(J + 1)}{4}$$

and

$$\text{var}\,[T_{\text{PAGE}}] = \frac{I(J^3 - J)^2}{144(J - 1)}$$

10.3 *Alternative Rank Analyses*

Conover and Iman [1981] in a series of papers have advocated a very simple rank analysis: Replace observations by their ranks and then carry out the usual parametric analysis. These procedures must be viewed with caution when models are nonadditive [Akritas, 1990] and discussion in Chapter 8. Hettmansperger and McKean [1978] provide an illustration of another class of rank-based analytical procedures that can be developed. There are three steps in this type of approach:

1. Define a robust or nonparametric estimate of dispersion.
2. State an appropriate statistical model for the data.
3. Given a set of data, estimate the values of the parameters of the model to minimize the robust estimate of dispersion.

A drawback of such procedures is that estimates cannot be written explicitly, and more important, the estimation procedure is nonlinear, requiring a computer to carry it out. However, with the increasing availability of microcomputers, it will only be a matter of time until software will be developed, making such procedures widely accessible.

It is possible to run a parametric analysis of the raw data routinely and compare it with some alternative rank analysis. If the two analyses do not agree, the data should be examined more carefully to determine the cause of the discrepant results; usually, it will be due to the *nonnormality* of the data. The researcher then has two choices: if the nonnormality is thought to be a characteristic of the biological system from which the data came, the rank analysis would be preferred. On the other hand, if the nonnormality is due to outliers (see Chapter 8), there are other options available, all the way from redoing the experiment (more carefully this time), to removing the outliers, to sticking with the analysis of the ranks. Clearly, there are financial, ethical, and professional costs and risks. What should *not* be done in the case of disagreement is to pick the analysis that conforms, in some sense, to the researcher's preconceptions or desires.

10.4 Power Transformation

Let Y^δ be a transformation of Y. The assumption is that Y^δ is normally distributed with mean μ (which will depend on the experimental model) and variance σ^2. The SS_ε will now be a function of δ. It can be shown that the appropriate quantity to be minimized is

$$L(\delta) = \frac{n}{2}SS_\varepsilon - \sum \ln(\delta y^\delta)$$

and defined to be

$$= \frac{n}{2}SS_\epsilon - \sum \ln y$$

for $\delta = 0$ (corresponding to the logarithmic transformation). Typically, this equation is solved by trial and error. With a computer this can be done quickly. Usually, there will be a range of values of δ over which the values of $L(\delta)$ will be close to the minimum; it is customary then to pick a value of δ that is simple. For example, if the minimum of $L(\delta)$ occurs at $\delta = 0.49$, the value chosen will be $\delta = 0.50$ to correspond to the square root transformation. For an example, see Weisberg [1985]. Empirical evidence suggests that the value of δ derived from the data is frequently close to some "natural" rescaling of the data. (This may just be a case of perfect 20/20 hindsight.)

PROBLEMS

For Problems 10.1 to 10.23, carry out one or more of the following tasks. Additional tasks are indicated at each problem.

(a) State an appropriate ANOVA model, null hypotheses, and alternative hypotheses. State whether the model is fixed, random, or mixed. Define the population to which inferences are to be made.

(b) Test the assumption of homogeneity of variance.

(c) Test the assumption of normality using a probability plot.

(d) Test the assumption of normality correlating residuals and rankits.

(e) Graph the data. Locate the cell means on the graph.

(f) Transform the data. Give a rationale for transformation.

(g) Carry out the analysis of variance. State conclusions and reservations. Compare with the conclusions of the author(s) of the paper. If possible, estimate the power if the results are not significant.

(h) Carry out a nonparametric analysis of variance. Compare with conclusions of parametric analysis.

(i) Partition each observation into its component parts [see, e.g., equations (4) and (19)] and verify that the sum of squares of each component is equal to one of the sums of squares in the ANOVA table.

(j) Construct the ANOVA table from means and standard deviations (or standard errors). Do relevant parts of (g).

10.1 Olsen et al. [1975] studied "morphine and phenytoin binding to plasma proteins in renal and hepatic failure." Twenty-eight subjects with uremia were classified into four groups. The percentage of morphine that was bound is the endpoint.

Chronic ($n_1 = 18$) : 31.5, 35.1, 32.1, 34.2, 26.7, 31.9, 30.8,

27.3, 27.3, 29.0, 30.0, 36.4, 39.8, 32.0, 35.9, 29.9, 32.2, 31.8

Acute ($n_2 = 2$) : 31.6, 28.5

Dialysis ($n_3 = 3$) : 29.3, 32.1, 26.9

Anephric ($n_4 = 5$) : 26.5, 22.7, 27.5, 24.9, 23.4

(a) Do tasks (a) to (e) and (g) to (i).

(b) In view of the nature of the response variable (percent of morphine bound), explain why, strictly speaking, the assumption of homogeneity of variance cannot hold.

10.2 Graham [1977] assayed 16 commercially available pancreatic extracts for six types of enzyme activity. See also Example 10.6. Data for one of these enzymes, proteolytic activity, are presented here. The 16 products were classified by packaging form: capsule, tablet, and enteric-coated tablets. The following data were obtained:

	Proteolytic Activity (U/unit)					
Tablet ($n = 5$)	6640	4440	240	990	410	
Capsule ($n = 4$)	6090	5840	110	195		
Coated tablet ($n = 7$)	1800	1420	980	1088	2200	870	690

(a) Do tasks (a) to (e) and (g) to (i).

(b) Is there a transformation that would make the variance more homogeneous? Why is this unlikely to be the case? What is peculiar about the values for the coated tablets?

10.3 The following data from Rifkind et al. [1976] consist of antipyrine clearance of males suffering from β-thalassemia, a chronic type of anemia. In this disease, abnormally thin red blood cells are produced. The treatment of the disease has undesirable side effects, including liver damage. Antipyrine is a drug used to assess liver function with a high clearance rate, indicating satisfactory liver function. These data deal with the antipyrine clearance rate of 10 male subjects classified according to pubertal stage.

The question is whether there is any significant difference in clearance rate among the pubertal stages (I = infant; V = adult).

Pubertal Stage	Clearance Rate (Half-Life in Hours)				
I	7.4	5.6	3.7	6.6	6.0
IV	10.9	12.2			
V	11.3	10.0	13.3		

  (a) Do tasks (a) to (e) and (g) to (i).

 *(b) Assuming that the antipyrine clearance rate increases with age, carry out a non-parametric test for trend (see Note 10.2). What is the alternative hypothesis in this case?

10.4 It is known that organisms react to stress. A more recent discovery is that the immune system's response is altered as a function of stress. In a paper by Keller et al. [1981], the immune response of rats as measured by the number of circulating lymphocytes (cells per milliliter $\times 10^{-6}$) was related to the level of stress. The following data are taken from this paper:

Group	Number of Rats	Mean Number of Lymphocytes	SE
Home-cage control	12	6.64	0.80
Apparatus control	12	4.84	0.70
Low shock	12	3.98	1.13
High shock	12	2.92	0.42

  (a) Do tasks (a), (b), (e), and (j).

  (b) The authors state: "a significant lymphocytopenia [$F(3, 44) = 3.86, p < 0.02$] was induced by the stressful conditions." Does your F-ratio agree with theirs?

  (c) Sharpen the analysis by considering a trend in the response levels as a function of increasing stress level.

10.5 This problem deals with the data in Table 10.8. The authors of the paper state that the animals were matched on the basis of weight but that there was no correlation with weight. Assume that the data are presented in the order in which the animals were matched, that is, $Y_{111} = 143$ is matched with $Y_{211} = 152$; in general, Y_{1jk} is matched with Y_{2jk}.

  (a) Construct a table of differences $D_{jk} = Y_{2jk} - Y_{1jk}$.

  (b) Carry out a one-way ANOVA on the differences; include SS_μ in your table.

  (c) Interpret SS_μ for these data.

  (d) State your conclusions and compare them with the conclusions of Example 10.5.

  (e) Relate the MS(between groups) in the one-way ANOVA to one of the MS terms in Table 10.14. Can you identify the connection and the reason for it?

 *(f) We want to correlate the Y_{1jk} observations with the Y_{2jk} observations, but the problem is that the response level changes from day to day, which would induce a correlation. So we will use the following "trick." Calculate $Y_{ijk}^* = Y_{ijk} - \overline{Y}_{ij\cdot}$;

and correlate Y^*_{1jk} with Y^*_{2jk}. Test this correlation using a t-test with $16 - 1 = 15$ degrees of freedom. Why $16 - 1$? There are $7 - 1 = 6$ independent pairs for day 10, 5 each for day 12, and day 14, for a total of 16. Since the observations sum to zero already, we subtract one more degree of freedom for estimating the correlation. If matching was not effective, this correlation should be zero.

10.6 Ross and Bras [1975] investigated the relationship between diet and length of life in 121 randomly bred rats. After 21 days of age, each rat was given a choice of several diets *ad libitum* for the rest of its life. The daily food intake (g/day) was categorized into one of six intervals, so that an equal number of rats (except for the last interval) appeared in each interval. The response variable was life span in days. The following data were obtained:

Mean food intake (g/day)	18.3	19.8	20.7	21.6	22.4	24.1
Food intake category	1	2	3	4	5	6
Number of rats	20	20	20	20	20	21
Mean life span (days)	733	653	630	612	600	556
Standard error	117	126	111	115	113	106

  (a) Carry out tasks (a), (b), (e), and (j).

  (b) Can this be thought of as a regression problem? How would the residual MS from regression be related to the MS error of the analysis of variance?

  *(c) Can you relate in detail the ANOVA procedure and the regression analysis; particularly an assessment of a nonlinear trend?

10.7 The following data from Florey et al. [1977] are the fasting serum insulin levels for adult Jamaican females after an overnight fast:

	Fasting Serum Insulin Level (μU/mL)			
Age	25–34	35–44	45–54	55–64
Number	73	97	74	53
Mean	22.9	26.2	22.0	23.8
SD	10.3	13.0	7.4	10.0

  (a) Do tasks (a), (b), (e), and (j).

  (b) Why did the authors partition the ages of the subjects into intervals? Are there other ways of summarizing and analyzing the data? What advantages or disadvantages are there to your alternatives?

10.8 The assay of insulin was one of the earliest topics in bioassay. A variety of methods have been developed over the years. In the mid-1960s an assay was developed based on the fact that insulin stimulates glycogen synthesis in mouse diaphragm tissue, in vitro. A paper by Wardlaw and van Belle [1964] describes the statistical aspects of this assay. The data in this problem deal with a qualitative test for insulin activity. A pool of 36 hemidiaphragms was collected for each day's work and the tissues incubated in tubes containing medium with or without insulin. Each tube contained three randomly selected diaphragms. For parts of this problem we ignore tube effects and assume that each treatment was based on six hemidiaphragms. Four unknown samples were

Table 10.34   Glycogen Content Data

Medium Only		Standard Insulin (0.5 mU/mL)		Test Preparation							
				A		B		C		D	
280	290	460	465	470	480	430	300	510	505	310	290
240	275	400	460	440	390	385	505	610	570	350	330
225	350	470	470	425	445	380	485	520	570	250	300

Source: Data adapted from Wardlaw and van Belle [1964].

assayed. Since the diaphragms synthesize glycogen in medium, a control preparation of medium only was added as well as a standard insulin preparation. The glycogen content (optical density in anthrone TEST × 1000) data are given in Table 10.34.

(a) Carry out tasks (a) to (e) and (g) to (i). (To simplify the arithmetic if you are using a calculator, divide the observations by 100.)

(b) Each column in the data layout represents one tube in which the three hemidiaphragms were incubated so that the design of the experiment is actually hierarchical. To assess the effect of tubes, we partition the SS_ε (with 30 d.f.) into two parts: SS(between tubes within preparations) = $SS_{BT(WP)}$ with six degrees of freedom (why?) and SS(within tubes) = SS_{WT} with 24 degrees of freedom (why?). The latter SS can be calculated by considering each tube as a treatment. The former can then be calculated as $SS_{BT(WP)} = SS_\varepsilon - SS_{WT}$. Carry out this analysis and test the null hypothesis that the variability between tubes within preparations is the same as the within-tube variability.

10.9   Schizophrenia is one of the psychiatric illnesses that is thought to have a definite physiological basis. Lake et al. [1980] assayed the concentration of norepinephrine in the cerebrospinal fluid of patients (NE in CSF) with one of three types of schizophrenia and controls. They reported the following means and *standard errors*:

NE in CSF (pg/mL)	Control Group	Schizophrenic Group		
		Paranoid	**Undifferentiated**	**Schizoaffective**
N	29	14	10	11
Mean	91	144	101	122
Standard error	6	20	11	21

Carry out tasks (a), (b), (e), and (j).

10.10   Corvilain et al. [1971] studied the role of the kidney in the catabolism (conversion) of insulin by comparing the metabolic clearance rate in seven control subjects, eight patients with chronic renal failure, and seven anephric (without kidneys) patients. The data for this problem consist of the plasma insulin concentrations (ng/mL) at 45 and 90 min after the start of continuous infusion of labeled insulin. A low plasma concentration is associated with a high metabolic clearance rate, as shown in Table 10.35.

(a) Consider the plasma insulin concentration at 45 minutes. Carry out tasks (a) to (e) and (g) to (i).

Table 10.35 Plasma Concentration Data (ng/mL)

Control			Renal Failure			Anephric		
Patient	45	90	Patient	45	90	Patient	45	90
1	3.7	3.8	1	3.0	4.2	1	6.7	9.6
2	3.4	4.2	2	3.1	3.9	2	2.6	3.4
3	2.4	3.1	3	4.4	6.1	3	3.4	—[a]
4	3.3	4.4	4	5.1	7.0	4	4.0	5.1
5	2.4	2.9	5	1.9	3.5	5	3.1	4.2
6	4.8	5.4	6	3.4	5.7	6	2.7	3.8
7	3.2	4.1	7	2.9	4.3	7	5.3	6.6
			8	3.8	4.8			

[a] Missing observation.

(b) Consider the plasma insulin concentration at 90 minutes. Carry out tasks (a) to (e) and (g) to (i).

(c) Calculate the difference in concentrations between 90 and 45 minutes for each patient. Carry out tasks (a) to (e) and (g) to (i). Omit Patient 3 in the anephric group.

(d) Graph the means for the three groups at 45 and 90 minutes on the same graph. What is the overall conclusion that you draw from the three analyses? Were all three analyses necessary? Would two of three have sufficed? Why or why not?

10.11 We return to the data of Zelazo et al. [1972] one more time. Carry out the Terpstra–Jonckheere test for ordered alternatives as discussed in Note 10.2. Justify the use of an ordered alternative hypothesis. Discuss in terms of power the reason that this analysis does indicate a treatment effect, in contrast to previous analyses.

10.12 One of the problems in the study of SIDS is the lack of a good animal model. Baak and Huber [1974] studied the guinea pig as a possible model observing the effect of lethal histamine shock on the guinea pig thymus. The purpose was to determine if changes in the thymus of the guinea pig correspond to pathological changes observed in SIDS victims. In the experiment 40 animals (20 male, 20 female) were randomly assigned either to "control" or "histamine shock." On the basis of a Wilcoxon two-sample test—which ignored possible gender differences—the authors concluded that the variable medullary blood vessel surface (mm^2/mm^3) did not differ significantly between "control" and "histamine shock." The data below have been arranged to keep track of gender differences.

	Control					Histamine Shock				
Female	6.4	6.2	6.9	6.9	5.4	8.4	10.2	6.2	5.4	5.5
	7.5	6.1	7.3	5.9	6.8	7.3	5.2	5.1	5.7	9.8
Male	4.3	7.5	5.2	4.9	5.7	7.5	6.7	5.7	4.9	6.8
	4.3	6.4	6.2	5.0	5.0	6.6	6.9	11.8	6.7	9.0

(a) Do tasks (a) to (e), (g), and (i).

(b) Replace the observations by their ranks and repeat the analysis of variance. Compare your conclusions with those of part (a).

10.13 In tumor metastasis, tumor cells spread from the original site to other organs. Usually, a particular tumor will spread preferentially to specific organs. There are two possibilities as to how this may occur: The tumor cells gradually adapt to the organ to which they have spread, or tumor cells that grow well at this organ are selected preferentially. Nicolson and Custead [1982] studied this problem by comparing the metastatic potential of melanoma tumor cells mechanically lodged in the lungs of mice or injected intravenously and allowed to metastasize to the lung. Each of these cell lines was then harvested and injected subcutaneously. The numbers of pulmonary tumor colonies were recorded for each of three treatments: original line (control), mechanical placement (adaptation), and selection. The data in Table 10.36 were obtained in three experiments involving 84 mice.

Table 10.36   Experimental Data for Three Treatments

Experiment	Control			Adaption				Selection			
1	0	4	20	32	0	3	20	7	92	141	
		0	9	22		0	6	24	64	96	149
		1	11	31		2	14	29	79	100	151
2	0	8	31	41	0	10	13	0	101	132	
		3	8	32		0	11	14	52	109	136
		6	22	39		5	12	14	89	110	140
3	0	4	36	49	0	11	21	30	79	111	
		0	18	39		0	13	27	46	89	114
		2	29	42		3	13	28	51	100	114

Number of Pulmonary Tumor Colonies (spanning header over Control, Adaption, Selection)

(a) Carry out tasks (a) to (g). You may want to try several transformations: for example, $\sqrt{\ }$, $Y^{1/4}$. An appropriate transformation is logarithmic. To avoid problems with zero values, use $\log(Y + 1)$.

(b) How would you interpret a significant "experiment × treatment" interaction?

10.14 A paper by Gruber [1976] evaluated interactions between two analgesic agents: fenoprofen and propoxyphene. The design of the study was factorial with respect to drug combinations. Propoxyphene (P) was administered in doses of 0, 5, 100, and 150 mg.; fenoprofen (F) in doses of 0, 200, 400, and 600 mg. Each combination of the two factors was studied. In addition, postepisiotomy postpartum patients were categorized into one of four pain classes: "little," "some," "lot," and "terrible" pain; for each of the 16 medication combinations, 8, 10, 10, and 2 patients in the four pain classes were used. The layout of the number of patients could be constructed as shown in Table 10.37.

(a) One response variable was "analgesic score" for a medication combination. Table 10.38 is a partial ANOVA table for this variable. Fill in the lines in the table, completing the table.

(b) The total analgesic score for the 16 sets of 30 patients classified by the two drug levels is given in Table 10.39. Carry out a "randomized block analysis" on these total scores dividing the sums of squares by 30 to return the analysis to a single reading status. Link this analysis with the table in part (a). You have, in effect, partitioned the SS for medications in that table into three parts. Test the significance of the *three* mean squares.

(c) Graph the mean analgesia score (per patient) by plotting the dose on the x-axis for fenoprofen, indicating the levels of the propoxyphene dose in the graph. State your conclusions.

Table 10.37   Design of Medication Combinations

Pain Level	Medication Combination						
	$(0P, 0F)$	$(0P, 200F)$	$\cdots$	$(0P, 600F)$	$(50P, 0F)$	$\cdots$	$(150P, 600F)$
"Little"	8	8	$\cdots$	8	8	$\cdots$	8
"Some"	10	10	$\cdots$	10	10	$\cdots$	10
"Lot"	10	10	$\cdots$	10	10	$\cdots$	10
"Terrible"	2	2	$\cdots$	2	2	$\cdots$	2

Table 10.38   ANOVA Table for Analgesic Score

Source	d.f.	SS	MS	F-Ratio	P-Value
Pain class	—	3,704	—	—	—
Medications	—	9,076	—	—	—
Interaction	—	3,408	—	—	—
Residual	—	—	—		
Total	479	41,910			

Table 10.39   Total Analgesia Score

Propoxyphene Dose (mg)	Fenoprofen Calcium Dose (mg)			
	0	200	400	600
0	409	673	634	756
50	383	605	654	785
100	496	773	760	755
150	496	723	773	755

10.15   Although the prescription, "Take two aspirins, drink lots of fluids, and go to bed," is usually good advice, it is known that aspirin induces "microbleeding" in the gastrointestinal system, as evidenced by minute amounts of blood in the stool. Hence, there is constant research to develop other anti-inflammatory and antipyretic (fever-combating) agents. Arsenault et al. [1976] reported on a new agent, R-803, studying its effect in a Latin square design, comparing it to placebo and aspirin (900 mg, q.i.d). For purposes of this exercise the data are extracted in the form of a randomized block design. Each subject received each of three treatments for a week. We will assume that the order was random. The variable measured is the amount of blood lost in mL/day as measured over a week.

Subject	Mean Blood Loss (ml/day)								
	1	2	3	4	5	6	7	8	9
Placebo	0.45	0.54	0.69	0.53	3.03	0.78	0.14	0.82	0.96
R-803	0.82	0.39	0.67	1.19	1.18	1.07	0.49	0.14	0.80
Aspirin	18.00	6.46	6.19	6.52	7.18	9.39	6.93	1.57	4.03

   (a)   Do tasks (a) to (e) and (g) to (i).
   (b)   Carry out the Tukey test for additivity. What are your conclusions?

Table 10.40 COHb Data for Problem 10.16

Subject	No. Hours Since Beginning of Exposure				
	0	2	4	6	8
1	4.4	4.9	5.2	5.7	5.7
2	3.3	5.3	6.9	7.0	8.8
3	5.0	6.4	7.2	7.7	9.3
4	5.3	5.3	7.4	7.0	8.3
5	4.1	6.8	9.6	11.5	12.0
6	5.0	6.0	6.8	8.3	8.1
7	4.6	5.2	6.6	7.4	7.1

10.16 Occupational exposures to toxic substances are being investigated more and more carefully. Ratney et al. [1974] studied the effect of daily exposure of 180 to 200 ppm of methylene chloride on carboxyhemoglobin (COHb) measured during the workday. The COHb data (% COHb) for seven subjects measured five times during the day is given in Table 10.40.

(a) Carry out tasks (a), (c) to (e), and (g) to (i).

(b) Suppose that the observation for subject 3 at time 6 ($Y_{34} = 7.7$) is missing. Estimate its value and redo the ANOVA.

(c) Carry out the Tukey test for additivity.

(d) Carry out the Page test for trend (see Note 10.2).

(e) Why do the data not form a randomized block design?

(f) Could this problem be treated by regression methods, where $X = $ hours since exposure and $Y = $ % COHb? Why or why not?

(g) Calculate all 10 pairwise correlations between the treatment combinations. Do they look "reasonably close"?

10.17 Wang et al. [1976] studied the effects on sleep of four hypnotic agents and a placebo. The preparations were: lorazepam 2 and 4 mg, and flurazepam 15 and 30 mg. Each of 15 subjects received all five treatments in a random order in five-night units. The analysis of variance of length of sleep is presented here.

Source	d.f.	SS	MS	F-Ratio	p-Value
Treatments	—	—	12.0	—	—
Patients	—	—	14.8	—	—
Residual	—	—	2.2		
Total	74	—			

(a) Do task (a).

(b) Fill in the missing values in the ANOVA table.

(c) State your conclusions.

(d) The article does not present any raw data or means. How satisfactory is this in terms of clinical significance?

10.18 High blood pressure is a known risk factor for cardiovascular disease, and many drugs are now on the market that provide symptomatic as well as therapeutic relief. One of

**Table 10.41   Blood Pressure Data (mmHg) for
Problem 10.18**

Patient	Recumbent		Upright	
	Placebo	Propranolol	Placebo	Propranolol
N.F.	96	71	73	87
A.C.	96	85	104	76
P.D.	92	89	83	90
J.L.	97	110	101	85
G.P.	104	85	112	94
A.H.	100	73	101	93
C.L.	93	81	88	85

these drugs is propranolol. Hamet et al. [1973] investigated the effect of propranolol in labile hypertension. Among the variables studied was mean blood pressure measured in mmHg (diastolic $+1/3$ pulse pressure). A placebo treatment was included in a double-blind fashion. The blood pressure was measured in the recumbent and upright positions. The blood pressure data is given in Table 10.41.

(a) Assuming that the treatments are just four treatments, carry out tasks (a) to (e) and (g) to (i) (i.e., assume a randomized block design).

(b) The sum of squares for treatments (3 d.f.) can be additively partitioned into three parts: SS_{DRUG}, $SS_{POSITION}$, and $SS_{DRUG \times POSITION}$, each with one degree of freedom. To do this, construct an "interaction table" of treatment totals.

$$SS_{DRUGS} = \frac{1340^2}{14} + \frac{1204^2}{14} - \frac{2544^2}{28} = 660.57$$

$$SS_{POSITION} = \frac{1272^2}{14} + \frac{1272^2}{14} - \frac{2544^2}{28} = 0[sic]$$

$$SS_{DRUGS \times POSITION} = \frac{678^2}{7} + \frac{594^2}{4} + \frac{662^2}{7} + \frac{610^2}{7} - \frac{2544^2}{28}$$
$$- SS_{DRUGS} - SS_{POSITION} = 36.57$$

Expand the ANOVA table to include these terms. (The $SS_{POSITION} = 0$ is most unusual; the raw data are as reported in the table.)

(c) This analysis could have been carried out as a series of three paired t-tests as follows: for each subject, calculate the following three quantities " $+ + - -$," " $+ - + -$," and " $+ - - +$." For example, for subject N.F. " $+ + - -$ " $= 96 + 71 - 73 - 87 = 7$, " $+ - + -$ " $= 96 - 71 + 73 - 87 = 11$, and " $+ - - +$ " $= 96 - 71 - 73 + 87 = 39$. These quantities represent effects of position, drug treatment, and interaction, respectively, and are called *contrasts* (see Chapter 12 for more details). Each contrast can be tested against zero by means of a one-sample t-test. Carry out these t-tests. Compare the variances for each contrast; one assumption in the analysis of variance is that these contrast variances all estimate the same variance. How is the sum of the contrast variances related to the SS_ε in the ANOVA?

(d) Let d_1 be the sum of the observations associated with the pattern $+ + - -$, d_2 the sum of the observations associated with the pattern $+ - + -$, and d_3 the sum

of the observations associated with the pattern $+ - -+$. How is $(d_1^2 + d_2^2 + d_3^2)$ related to $SS_{TREATMENT}$?

10.19 Consider the data in Example 10.5. Rank all 38 observations from lowest to highest and carry out the usual analysis of variance on these ranks. Compare your p-values with the p-values of Table 10.14. In view of Note 10.3, does this analysis give you some concern?

10.20 Consider the data of Table 10.16 dealing with the effectiveness of pancreatic supplements on fat absorption. Rank all of the observations from 1 to 24 (i.e., ignoring both treatment and block categories).

   (a) Carry out an analysis of variance on the ranks obtained above.

   (b) Compare your analysis with the analysis using the Friedman statistic. What is a potential drawback in the analysis of part (a)?

   (c) Return to the Friedman ranks in Section 10.3.3 and carry out an analysis of variance on them. How is the Friedman statistic related to SS_τ of the ANOVA of the Friedman ranks?

10.21 These data are from the same source as those in Problem 10.3. We add data for females to generate the two-way layout shown in Table 10.42.

Table 10.42   Two-Way Layout for Problem 10.21

	Antipyrine Clearance (Half-Life in Hours)					
	Stage I		Stage IV		Stage V	
Males	7.4	5.6	3.7	10.9	11.3	13.3
	6.6	6.0		12.2	10.0	
Females	9.1	6.3	7.1	11.0	8.3	
	11.3	9.4	7.9		4.3	

   (a) Do tasks (a) to (d).

   (b) Graph the data. Is there any suggestion of interaction? Of main effects?

   (c) Carry out a weighted means analysis.

   (d) Partition each observation into its component parts and verify that the sums of squares are *not* additive.

10.22 Fuertes-de la Haba et al. [1976] measured intelligence in offspring of oral and nonoral contraceptive users in Puerto Rico. In the early 1960s, subjects were randomly assigned to oral conceptive use or other methods of birth control. Subsequently, mothers with voluntary pregnancies were identified and offspring between ages 5 and 7 were administered a Spanish–Puerto Rican version of the Wechsler Intelligence Scale for Children (WISC). Table 10.43 lists the data for boys only, taken from the article.

   (a) Carry out tasks (a), (b), and (e).

   (b) Do an unweighted means analysis. Interpret your findings.

   (c) The age categories have obviously been "collapsed." What effect could such a collapsing have on the analysis? (Why introduce age as a variable since IQ is standardized for age?)

   (d) Suppose that we carried out a contingency table analysis on the cell frequencies. What could such an analysis show?

Table 10.43   Data for Problem 10.22

	Age Groups (Years)		
	5	6	7–8
Oral contraceptive WISC score			
n	9	18	14
Mean	81.44	88.50	76.00
SD	9.42	11.63	9.29
Other birth control WISC score			
n	11	28	21
Mean	82.91	87.75	83.24
SD	10.11	10.85	9.60

Table 10.44   Data for Problem 10.23

	Gender	
	Boys	Girls
Oral contraceptive WISC score		
n	41	55
Mean	82.68	86.87
SD	11.78	14.66
Other birth control WISC score		
n	60	54
Mean	85.28	85.83
SD	10.55	12.22

10.23 The data in Table 10.44 are also from the article by Fuertes-de la Haba [1976] but have been "collapsed over age" and are presented by treatment (type of contraceptive) by gender. The response variable is, again, Wechsler IQ score.

   (a) Carry out tasks (a), (b), and (e).
   (b) Do an unweighted means analysis.
   (c) Compare your conclusions with those of Problem 10.22.

10.24 This problem considers some implications of the additive model for the two-way ANOVA as defined by equation (18) and illustrated in Example 10.4.

   (a) Graph the means of Example 10.4 by using the level of the second variable for the abscissa. Interpret the difference in the patterns.
   (b) How many degrees of freedom are left for the means assuming that the model defined by equation (18) holds?
   (c) We now want to define a nonadditive model retaining the values of the α's, β's, and μ, equivalently, retaining the same marginal and overall means. You are free to vary any of the cell means subject to the constraints above. Verify that you can manipulate only four cell means. After changing the cell means, calculate for each cell ij the quantity $Y_{ij} = \mu - \alpha_i - \beta_j$. What are some characteristics of these quantities?
   (d) Graph the means derived in part (c) and compare the pattern obtained with that of Figure 10.2.

*10.25 This problem is designed to give you some experience with algebraic manipulation. It is not designed to teach you algebra but to provide additional insight into the mathematical structure of analysis of variance models. You will want to take this medicine in small doses.

   (a) Show that equation (5) follows from the model defined by equation (4).

   (b) Prove equations (6) and (7).

   (c) Prove equations (10) to (12) starting with the components of equation (5).

   (d) Consider equation (17). Let $\mu_i = \sum n_{ij}\mu_{ij}/n_i.$, and so on. Relate α_i and β_j to $\mu_i.$ and $\mu._j$.

   (e) For the two-way ANOVA model as defined by equation (21), show that $SS_\varepsilon = SS_{ERROR} = \sum (n_{ij} - 1)s_{ij}^2$, where s_{ij}^2 is the variance of the observations in cell (i, j).

   (f) Derive the expected mean squares for MS_α and MS_γ in the fixed and random effects models, as given in Table 10.19.

REFERENCES

Akritas, M. G. [1990]. The rank transform in some two-factor designs. *Journal of the American Statistical Association*, **85**: 73–78.

Arsenault, A., Le Bel, E., and Lussier, E. [1976]. Gastrointestinal microbleeding in normal subjects receiving acetylsalicylic acid, placebo and R-803, a new antiinflammatory agent, in a design balanced for residual effects. *Journal of Clinical Pharmacology*, **16**: 473–480. Used with permission from J.B. Lippincott Company.

Baak, J. P. A., and Huber, J. [1974]. Effects of lethal histamine shock in the guinea pig thymus. In *SIDS 1974, Proceedings of the Francis E. Camps International Symposium of Sudden and Unexpected Deaths in Infancy*, R. R. Robertson (ed.). Canadian Foundation for the Study of Infant Death, Toronto, Ontario, Canada.

Barboriak, J. J., Rimm, A., Tristani, F. E., Walker, J. R., and Lepley, D., Jr. [1972]. Risk factors in patients undergoing aorta-coronary bypass surgery. *Journal of Thoracic and Cardiovascular Surgery*, **64**: 92–97.

Beyer, W. H. (ed.) [1968]. *CRC Handbook of Tables for Probability and Statistics*. CRC Press, Cleveland, OH.

Box, G. E. P. [1953]. Non-normality and tests on variances. *Biometrika*, **40**: 318–335. Used with permission of the Biometrika Trustees.

Chikos, P. M., Figley, M. M., and Fisher, L. D. [1977]. Visual assessment of total heart volume and chamber size from standard chest radiographs. *American Journal of Roentgenology*, **128**: 375–380. Copyright © 1977 by the American Roentgenology Society.

Conover, W. J., and Iman, R. L. [1981]. Rank transformations as a bridge between parametric and non-parametric statistics. *American Statistician*, **35**: 124–133.

Corvilain, J., Brauman, H., Delcroix, C., Toussaint, C., Vereerstraeten, P., and Franckson, J. R. M. [1971]. Labeled insulin catabolism in chronic renal failure and the anephric state. *Diabetes*, **20**: 467–475.

Daniel, C. [1976]. *Applications of Statistics to Industrial Experiments*. Wiley, New York.

Daniel, C., and Wood, F. [1999]. *Fitting Equations to Data*, 2nd ed. Wiley, New York.

Draper, N. R., and Smith, H. [1998]. *Applied Regression Analysis*, 3rd ed. Wiley, New York.

Eisenhart, C. [1947]. The assumptions underlying the analysis of variance. *Biometrics*, **3**: 1–21.

Fisher, R. A. [1950]. *Statistical Methods for Research Workers*, 11th ed. Oliver & Boyd, London.

Florey, C. du V., Milner, R. D. G., and Miall, W. I. [1977]. Serum insulin and blood sugar levels in a rural population of Jamaican adults. *Journal of Chronic Diseases*, **30**: 49–60. Used with permission of Pergamon Press, Inc.

Freeman, M. F., and Tukey, J. W. [1950]. Transformations related to the angular and the square root. *Annals of Mathematical Statistics*, **21**: 607–611.

Friedman, M. [1937]. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, **32**: 675–701.

Fuertes-de la Haba, A., Santiago, G., and Bangdiwala, I. S. [1976]. Measured intelligence in offspring of oral and non-oral contraceptive users. *American Journal of Obstetrics and Gynecology*, **7**: 980–982.

Graham, D. Y. [1977]. Enzyme replacement therapy of exocrine pancreatic insufficiency in man. *New England Journal of Medicine*, **296**: 1314–1317.

Gruber, C. M., Jr. [1976]. Evaluating interactions between fenoprofen and propoxyphene: analgesic and adverse reports by postepisiotomy patients. *Journal of Clinical Pharmacology*, **16**: 407–417. Used with permission from J.B. Lippincott Company.

Hamet, P., Kuchel, O., Cuche, J. L., Boucher, R., and Genest, J. [1973]. Effect of propranolol on cyclic AMP excretion and plasma renin activity in labile essential hypertension. *Canadian Medical Association Journal*, **1**: 1099–1103.

Hettmansperger, T. P., and McKean, J. W. [1978]. Statistical inference based on ranks. *Psychometrika*, **43**: 69–79.

Hillel, A., and Patten, C. [1990]. Effects of age and gender on dominance for lateral abduction of the shoulder. Unpublished data; used by permission.

Hollander, M., and Wolfe, D. A. [1999]. *Nonparametrical Statistical Methods*, 2nd ed. Wiley, New York.

Joiner, B. L., and Rosenblatt, J. R. [1971]. Some properties of the range in samples from Tukey's symmetric lambda distributions. *Journal of the American Statistical Association*, **66**: 394–399.

Keller, S. E., Weiss, J. W., Schleifer, S. J., Miller, N. E., and Stein, M. [1981]. Suppression of immunity by stress. *Science*, **213**: 1397–1400. Copyright © 1981 by the AAAS.

Kruskal, W. H., and Wallis, W. A. [1952]. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, **47**: 583–621.

Lake, C. R., Sternberg, D. E., van Kammen, D. P., Ballenger, J. C., Ziegler, M. G., Post, R. M., Kopin, I. J., and Bunney, W. E. [1980]. Schizophrenia: elevated cerebrospinal fluid norepinephrine. *Science*, **207**: 331–333. Copyright © 1980 by the AAAS.

Looney, S. W., and Stanley, W. B. [1989]. Exploratory repeated measures analysis for two or more groups. *American Statistician*, **43**: 220–225.

Nicolson, G. L., and Custead, S. E. [1982]. Tumor metastasis is not due to adaptation of cells to a new organ environment. *Science*, **215**: 176–178. Copyright © 1982 by the AAAS.

Odeh, R. E., Owen, D. B., Birnbaum, Z. W., and Fisher, L. D. [1977]. *Pocket Book of Statistical Tables*. Marcel Dekker, New York.

Olsen, G. D., Bennett, W. M., and Porter, G. A. [1975]. Morphine and phenytoin binding to plasma proteins in renal and hepatic failure. *Clinical Pharmaceuticals and Therapeutics*, **17**: 677–681.

Page, E. B. [1963]. Ordered hypotheses for multiple treatments: a significance test for linear ranks. *Journal of the American Statistical Association*, **58**: 216–230.

Quesenberry, P. D., Whitaker, T. B., and Dickens, J. W. [1976]. On testing normality using several samples: an analysis of peanut aflatoxin data. *Biometrics*, **32**: 753–759.

Ratney, R. S., Wegman, D. H., and Elkins, H. B. [1974]. In vivo conversion of methylene chloride to carbon monoxide. *Archives of Environmental Health*, **28**: 223–226. Reprinted with permission of the Helen Dwight Reid Educational Foundation. Published by Heldref Publications, 4000 Albemarle Street, N.W., Washington DC 20016. Copyright © 1974.

Rifkind, A. B., Canale, V., and New, M. I. [1976]. Antipyrine clearance in homozygous beta-thalassemia. *Clinical Pharmaceuticals and Therapeutics*, **20**: 476–483.

Ross, M. H., and Bras, G. [1975]. Food preference and length of life. *Science*, **190**: 165–167. Copyright © 1975 by the AAAS.

Ryan, T. A., Jr., Joiner, B. L., and Ryan, B. F. [1980]. *Minitab Reference Manual*, Release 1/10/80. Statistics Department, Pennsylvania State University, University Park, PA.

Scheffé, H. [1959]. *The Analysis of Variance*. Wiley, New York.

Sherwin, R. P., and Layfield, L. J. [1976]. Protein leakage in lungs of mice exposed to 0.5 ppm nitrogen dioxide: a fluorescence assay for protein. *Archives of Environmental Health*, **31**: 116–118.

Snedecor, G. W., and Cochran, W. G. [1988]. *Statistical Methods*, 8th ed. Iowa State University Press, Ames, IA.

Tukey, J. W. [1949]. One degree of freedom for additivity. *Biometrics*, **5**: 232–242.

Wallace S. S, Fisher, L. D., and Tremann, J. A. [1977]. Unpublished manuscript.

Wang, R. I. H., Stockdale, S. L., and Hieb, E. [1976]. Hypnotic efficacy of lorazepam and flurazepam. *Clinical Pharmaceuticals and Therapeutics*, **19**: 191–195.

Wardlaw, A. C., and van Belle, G. [1964]. Statistical aspects of the mouse diaphragm test for insulin. *Diabetes*, **13**: 622–634.

Weisberg, S. [1985]. *Applied Linear Regression*, 2nd ed. Wiley, New York.

Weisberg, S., and Bingham, C. [1975]. Approximate analysis of variance test for non-normality suitable for machine calculation. *Technometrics*, **17**: 133–134.

Winer, B. J. [1991]. *Statistical Principles in Experimental Design*, 3rd ed. McGraw-Hill, New York.

Zelazo, P. R., Zelazo, N. A., and Kalb, S. [1972]. "Walking" in the newborn. *Science*, **176**: 314–315.