7 Linear Regression and Correlation

Simple linear regression analysis is a statistical technique that defines the functional relationship between two variables, X and Y, by the "best-fitting" straight line. A straight line is described by the equation, $Y = A + BX$, *where* Y *is the dependent variable* (ordinate), X *is the independent variable* (abscissa), and A *and* B *are the* Y *intercept* and *slope of the line*, respectively (Fig. 7.1).* Applications of regression analysis in pharmaceutical experimentation are numerous. This procedure is commonly used

1. to describe the relationship between variables where the functional relationship is known to be linear, such as in Beer's law plots, where optical density is plotted against drug concentration;
2. when the functional form of a response is unknown, but where we wish to represent a trend or rate as characterized by the slope (e.g., as may occur when following a pharmacological response over time);
3. when we wish to describe a process by a relatively simple equation that will relate the response, Y, to a fixed value of X, such as in stability prediction (concentration of drug versus time).

In addition to the specific applications noted above, regression analysis is used to define and characterize dose–response relationships, for fitting linear portions of pharmacokinetic data, and in obtaining the best fit to linear physical–chemical relationships.

Correlation is a procedure commonly used to characterize quantitatively the relationship between variables. Correlation is related to linear regression, but its application and interpretation are different. This topic is introduced at the end of this chapter.

7.1 INTRODUCTION

Straight lines are constructed from sets of data pairs, X and Y. Two such pairs (i.e., two points) uniquely define a straight line. As noted previously, a straight line is defined by the equation

$$Y = A + BX, \tag{7.1}$$

where A is the Y intercept (the value of Y when $X = 0$) and B is the slope ($\Delta Y/\Delta X$). $\Delta Y/\Delta X$ is $(Y_2 - Y_1)/(X_2 - X_1)$ for any two points on the line (Fig. 7.1). The slope and intercept define the line; once A and B are given, the line is specified. In the elementary example of only two points, a statistical approach to define the line is clearly unnecessary.

In general, with more than two X, y points,† a plot of y versus X will not *exactly* describe a straight line, even when the relationship is known to be linear. The failure of experimental data derived from truly linear relationships to lie exactly on a straight line is due to errors of observation (experimental variability). Figure 7.2 shows the results of four assays of drug samples of different, but known potency. The assay results are plotted against the known amount of drug. If the assays are performed without error, the plot results in a 45° line (slope = 1) which, if extended, passes through the origin; that is, the Y intercept, A, is 0 [Fig. 7.2(A)].

* The notation $Y = A + BX$ is standard in statistics. We apologize for any confusion that may result from the reader's familiarity with the equivalent, $Y = mX + b$, used frequently in analytical geometry.

† In the rest of this chapter, y denotes the experimentally observed point, and Y denotes the corresponding point on the least squares "fitted" line (or the true value of Y, according to context).

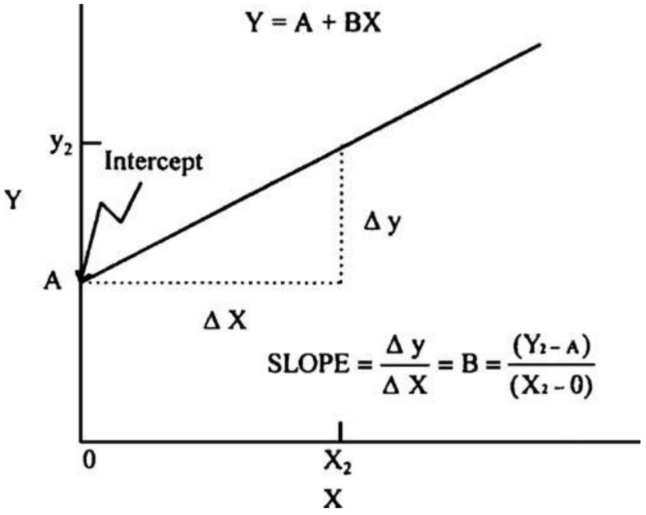

Y = A + BX
y_2
Intercept
Y
A
Δ y
Δ X
SLOPE = $\frac{\Delta y}{\Delta X}$ = B = $\frac{(Y_2 - A)}{(X_2 - 0)}$
0
X_2
X

Figure 7.1 Straight-line plot.

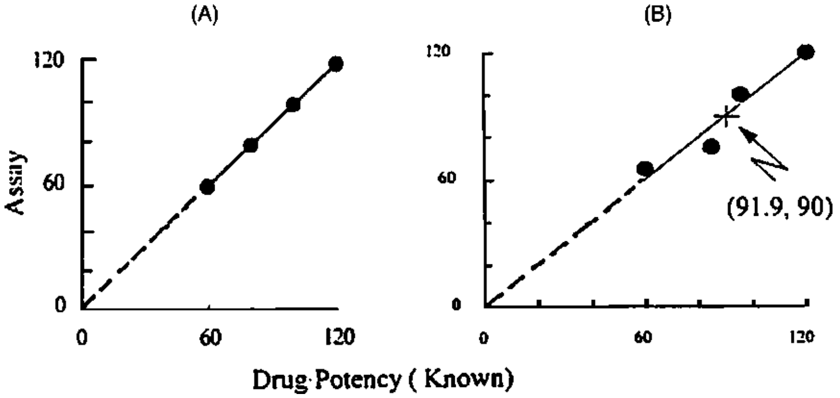


Figure 7.2 Plot of assay recovery versus known amount: theoretical and actual data.

In this example, the equation of the line $Y = A + BX$ is $Y = 0 + 1(X)$, or $Y = X$. Since there is no error in this experiment, the line passes exactly through the four X, Y points.

Real experiments are not error free, and a plot of X, y data rarely exactly fits a straight line, as shown in Figure 7.2(B). We will examine the problem of obtaining a line to fit data that are not error free. In these cases, the line does not go exactly through all of the points. A "good" line, however, should come "close" to the experimental points. When the variability is small, a line drawn by eye will probably be very close to that constructed more exactly by a statistical approach [Fig. 7.3(A)]. With large variability, the "best" line is not obvious. What single line would you draw to best fit the data plotted in Figure 7.3(B)? Certainly, lines drawn through any two arbitrarily selected points will not give the best (or a unique) line to fit the totality of data.

Given N pairs of variables, X, Y, we can define the best straight line describing the relationship of X and y as the line that minimizes the sum of squares of the vertical distances of each point from the fitted line. The definition of "sum of squares of the vertical distances of each point from the fitted line" (Fig. 7.4) is written mathematically as $\sum(y - Y)^2$, where y represents the experimental points and Y represents the corresponding points on the fitted line. The line constructed according to this definition is called the *least squares* line. Applying techniques of

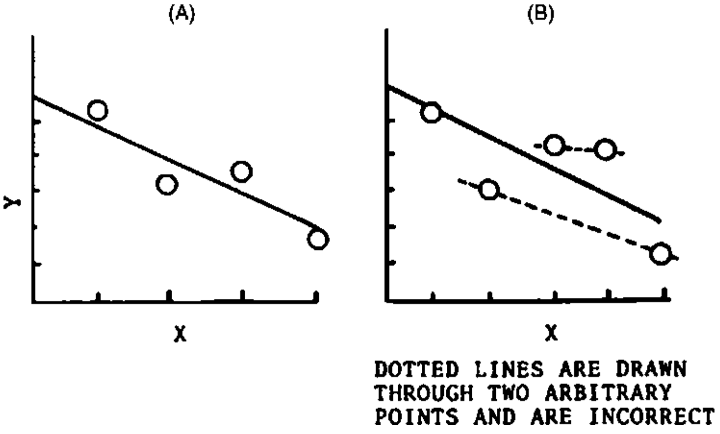


Figure 7.3 Fit of line with variable data.

calculus, the slope and intercept of the least squares line can be calculated from the sample data as follows:

$$\text{Slope} = b = \frac{\sum(X - \overline{X})(y - \overline{y})}{\sum(X - \overline{X})^2} \tag{7.2}$$

$$\text{Intercept} = a = \overline{y} - b\overline{X} \tag{7.3}$$

Remember that the slope and intercept uniquely define the line.

There is a shortcut computing formula for the slope, similar to that described previously for the standard deviation

$$b = \frac{N\sum Xy - (\sum X)(\sum y)}{N\sum X^2 - (\sum X)^2}, \tag{7.4}$$

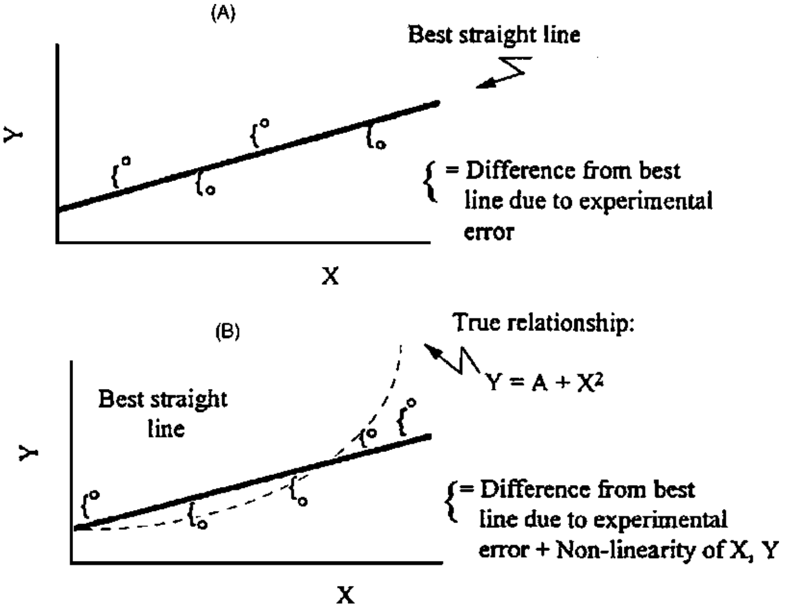


Figure 7.4 Lack of fit due to (A) experimental error and (B) nonlinearity.

Table 7.1 Raw Data from Figure 7.2(A) to Calculate the Least Squares Line

Drug potency, *X*	Assay, *y*	*Xy*
60	60	3600
80	80	6400
100	100	10,000
120	120	14,400
$\sum X = 360$	$\sum y = 360$	$\sum Xy = 34{,}400$
$\sum X^2 = 34{,}400$		

Table 7.2 Raw Data from Figure 7.2(B) Used to Calculate the Least Squares Line

Drug potency, *X*	Assay, *y*	*Xy*
60	63	3780
80	75	6000
100	99	9900
120	116	13,920
$\sum X = 360$	$\sum y = 353$	$\sum Xy = 33{,}600$
$\sum X^2 = 34{,}400$	$\sum y^2 = 32{,}851$	

where N is the number of X, y pairs. The calculation of the slope and intercept is relatively simple, and can usually be quickly computed using a computer (e.g., EXCEL) or with a hand calculator. Some calculators have a built-in program for calculating the regression parameter estimates, a and b.[‡]

For the example shown in Figure 7.2(A), the line that exactly passes through the four data points has a slope of 1 and an intercept of 0. The line, $Y = X$, is clearly the best line for these data, an exact fit. The least squares line, in this case, is exactly the same line, $Y = X$. The calculation of the intercept and slope using the least squares formulas, Eqs. (7.3) and (7.4), is illustrated below. Table 7.1 shows the raw data used to construct the line in Figure 7.2(A).

According to Eq. (7.4) ($N = 4$, $\sum X^2 = 34{,}400$, $\sum Xy = 34{,}400$, $\sum X = \sum y = 360$),

$$b = \frac{(4)(3600 + 6400 + 10{,}000 + 14{,}000) - (360)(360)}{4(34{,}400) - (360)^2} = 1$$

a is computed from Eq. (7.3); $a = \overline{y} - b\overline{X}(\overline{y} = \overline{X} = 90,\ b = 1)$. $a = 90 - 1(90) = 0$. This represents a situation where the assay results exactly equal the known drug potency (i.e., there is no error).

The actual experimental data depicted in Figure 7.2(B) are shown in Table 7.2. The slope b and the intercept a are calculated from Eqs. (7.4) and (7.3). According to Eq. (7.4),

$$b = \frac{(4)(33{,}600) - (360)(353)}{4(34{,}400) - (360)^2} = 0.915.$$

According to Eq. (7.3),

$$a = \frac{353}{4} - 0.915(90) = 5.9.$$

A perfect assay (no error) has a slope of 1 and an intercept of 0, as shown above. The actual data exhibit a slope close to 1, but the intercept appears to be too far from 0 to be attributed to random error. Exercise Problem 2 addresses the interpretation of these results as they relate to assay method characteristics.

[‡] a and b are the sample estimates of the true parameters, A and B.

This example suggests several questions and problems regarding linear regression analysis. The line that best fits the experimental data is an estimate of some true relationship between X and Y. In most circumstances, we will fit a straight line to such data only if we believe that the true relationship between X and Y is linear. The experimental observations will not fall exactly on a straight line because of variability (e.g., error associated with the assay). This situation (true linearity associated with experimental error) is different from the case where the underlying true relationship between X and Y is not linear. In the latter case, the lack of fit of the data to the least squares line is due to a combination of experimental error and the lack of linearity of the X, Y relationship (Fig. 7.4). Elementary techniques of simple linear regression will not differentiate these two situations: (a) experimental error with true linearity and (b) experimental error and nonlinearity. (A design to estimate variability due to both nonlinearity and experimental error is given in App. II.)

We will discuss some examples relevant to pharmaceutical research that make use of least squares linear regression procedures. The discussion will demonstrate how variability is estimated and used to construct estimates and tests of the line parameters A and B.

7.2 ANALYSIS OF STANDARD CURVES IN DRUG ANALYSIS: APPLICATION OF LINEAR REGRESSION

The assay data discussed previously can be considered as an example of the construction of a *standard curve* in drug analysis. Known amounts of drug are subjected to an assay procedure, and a plot of percentage recovered (or amount recovered) versus amount added is constructed. Theoretically, the relationship is usually a straight line. A knowledge of the line parameters A and B can be used to predict the amount of drug in an unknown sample based on the assay results. In most practical situations, A and B are unknown. The least squares estimates a and b of these parameters are used to compute drug potency (X) based on the assay response (y). For example, the least squares line for the data in Figure 7.2(B) and Table 7.2 is

$$\textit{Assay result} = 5.9 + 0.915\ (\textit{potency}). \tag{7.5}$$

Rearranging Eq. (7.5), an unknown sample that has an assay value of 90 can be predicted to have a true potency of

$$\textit{Potency} = X = \frac{y - 5.9}{0.915}$$

$$\textit{Potency} = \frac{90 - 5.9}{0.915} = 91.9.$$

This point (91.9, 90) is indicated in Figure 7.2 by a cross.

7.2.1 Line Through the Origin

Many calibration curves (lines) are known to pass through the origin; that is, the assay response must be zero if the concentration of drug is zero. The calculation of the slope is simplified if the line is forced to go through the point (0,0). In our example, if the intercept is *known* to be zero, the slope is (Table 7.2)

$$\begin{aligned} b &= \frac{\sum Xy}{\sum X^2} \\ &= \frac{33{,}600}{60^2 + 80^2 + 100^2 + 120^2} = 0.977. \end{aligned} \tag{7.6}$$

The least squares line fitted with the zero intercept is shown in Figure 7.5. If this line were to be used to predict actual concentrations based on assay results, we would obtain answers that are different from those predicted from the line drawn in Figure 7.2(B). However, both lines have been constructed from the same raw data. "Is one of the lines correct?" or "Is one line better than the other?" Although one cannot say with certainty which is the better line, a thorough knowledge of the analytical method will be important in making a choice. For example, a nonzero intercept suggests either nonlinearity over the range of assays or the

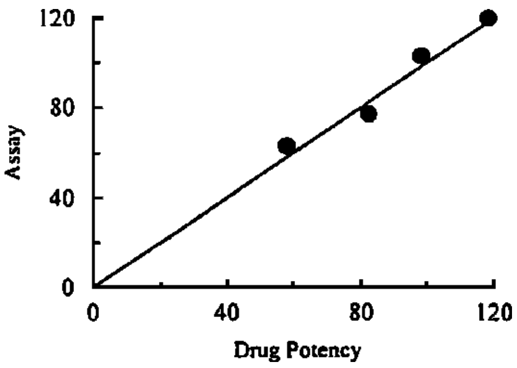


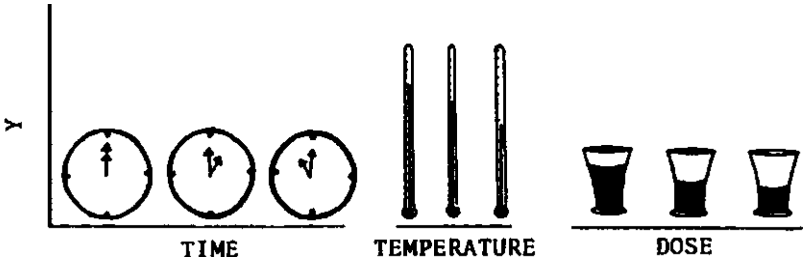
Figure 7.5 Plot of data in Table 7.2 with known (0, 0) intercept.

presence of an interfering substance in the sample being analyzed. The decision of which line to use can also be made on a statistical basis. A statistical test of the intercept can be performed under the null hypothesis that the intercept is 0 (H_0: $A = 0$, sect. 7.4.1). Rejection of the hypothesis would be strong evidence that the line with the positive intercept best represents the data.

7.3 ASSUMPTIONS IN TESTS OF HYPOTHESES IN LINEAR REGRESSION

Although there are no prerequisites for fitting a least squares line, the testing of statistical hypotheses in linear regression depends on the validity of several assumptions.

1. *The X variable is measured without error.* Although not always *exactly* true, X is often measured with relatively little error and, under these conditions this assumption can be considered to be satisfied. In the present example, X is the potency of drug in the "known" sample. If the drug is weighed on a sensitive balance, the error in drug potency will be very small. Another example of an X variable that is often used, which can be precisely and accurately measured, is "time."



2. *For each X, y is independent and normally distributed.* We will often use the notation $Y.x$ to show that the value of Y is a function of X.
3. *The variance of y is assumed to be the same at each X.* If the variance of y is not constant, but is either known or related to X in some way, other methods (see sect. 7.7) are available to estimate the intercept and slope of the line [1].
4. *A linear relationship exists between X and Y.* $Y = A + BX$, where A and B are the true parameters. Based on theory or experience, we have reason to believe that X and Y are linearly related.

These assumptions are depicted in Figure 7.6. Except for location (mean), the distribution of y is the same at every value of X; that is, y has the same variance at every value of X. In the example in Figure 7.6, the mean of the distribution of y's decreases as X increases (the slope is negative).

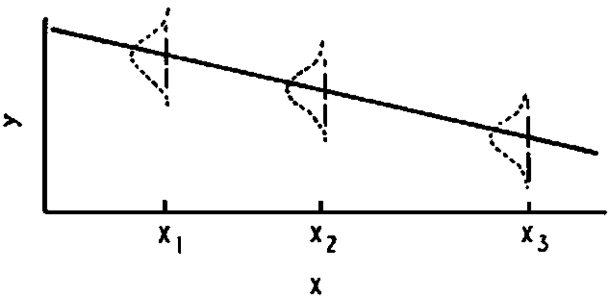


Figure 7.6 Normality and variance assumptions in linear regression.

7.4 ESTIMATE OF THE VARIANCE: VARIANCE OF SAMPLE ESTIMATES OF THE PARAMETERS

If the assumptions noted in section 7.3 hold, the distributions of *sample estimates* of the slope and intercept, b and a, are normal with means equal to B and A, respectively.[§] Because of this important result, statistical tests of the parameters A and B can be performed using normal distribution theory. Also, one can show that the sample estimates are unbiased estimates of the true parameters (similar to the sample average, $\overline{X}$, being an unbiased estimate of the true mean, μ). The variances of the estimates, a and b, are calculated as follows:

$$\sigma_a^2 = \sigma_{Y,x}^2 \left[\frac{1}{N} + \frac{\overline{X}^2}{\sum (X - \overline{X})^2} \right] \tag{7.7}$$

$$\sigma_b^2 = \frac{\sigma_{Y,x}^2}{\sum (X - \overline{X})^2}. \tag{7.8}$$

$\sigma_{Y,x}^2$ is the variance of the response variable, y. An estimate of $\sigma_{Y,x}^2$ can be obtained from the closeness of the data to the least squares line. If the experimental points are far from the least squares line, the estimated variability is larger than that in the case where the experimental points are close to the least squares line. This concept is illustrated in Figure 7.7. If the data exactly fit a straight line, the experiment shows no variability. In real experiments the chance of an exact fit with more than two X, y pairs is very small. An unbiased estimate of $\sigma_{Y,x}^2$ is obtained from the sum of squares of deviations of the observed points from the fitted line as follows:

$$S_{Y,x}^2 = \frac{\sum (y - Y)^2}{N - 2} = \frac{\sum (y - \overline{y})^2 - b^2[\sum (X - \overline{X})^2]}{N - 2}, \tag{7.9}$$

where y is the observed value and Y is the predicted value of Y from the least squares line ($Y = a + bX$) (Fig. 7.7). The variance estimate, $S_{Y,x}^2$, has $N - 2$ rather than $(N - 1)$ d.f. because two parameters are being estimated from the data (i.e., the slope and intercept).

When $\sigma_{Y,x}^2$ is unknown, the variances of a and b can be estimated, substituting $S_{Y,x}^2$ for $\sigma_{y,x}^2$ in the formulas for the variances [Eqs. (7.7) and (7.8)]. Equations (7.10) and (7.11) are used as the variance estimates, S_a^2 and S_b^2, when testing hypotheses concerning the parameters A and B. This procedure is analogous to using the sample estimate of the variance in the t test to compare sample means.

$$S_a^2 = S_{Y,x}^2 \times \left[\frac{1}{N} + \frac{\overline{X}^2}{\sum (X - \overline{X})^2} \right] \tag{7.10}$$

$$S_b^2 = \frac{S_{Y,x}^2}{\sum (X - \overline{X})^2} \tag{7.11}$$

[§] a and b are calculated as linear combinations of the normally distributed response variable, y, and thus can be shown to be also normally distributed.

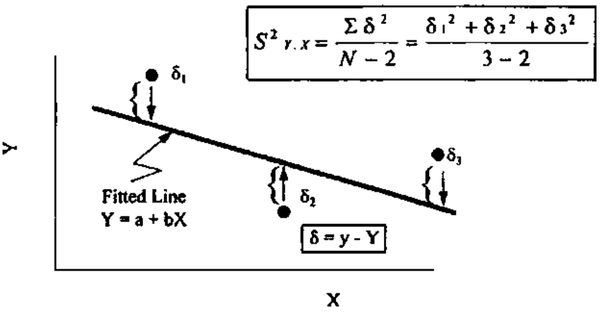

$S^2_{Y,X} = \frac{\Sigma\delta^2}{N-2} = \frac{\delta_1^2 + \delta_2^2 + \delta_3^2}{3-2}$
δ_1
δ_2
δ_3
Fitted Line
Y = a + bX
δ = y - Y
Y
X

Figure 7.7 Variance calculation from least squares line.

7.4.1 Test of the Intercept, *A*

The background and formulas introduced previously are prerequisites for the construction of tests of hypotheses of the regression parameters A and B. We can now address the question of the "significance" of the Y intercept (a) for the line shown in Figure 7.2(B) and Table 7.2. The procedure is analogous to that of testing means with the t test. In this example, the null hypothesis is H_0: $A = 0$. The alternative hypothesis is H_a: $A \neq 0$. Here the test is two-sided; a priori, if the intercept is not equal to 0, it could be either positive or negative. A t test is performed as shown in Eq. (7.12). $S^2_{Y,x}$ and S^2_a are calculated from Eqs. (7.9) and (7.10), respectively.

$$t_{\text{d.f.}} = t_2 = \frac{|\text{a} - A|}{\sqrt{S^2_a}} \tag{7.12}$$

where $t_{\text{d.f.}}$ is the t statistic with $N - 2$ d.f., a is the observed value of the intercept, and A is the hypothetical value of the intercept. From Eq. (7.10)

$$S^2_a = S^2_{Y,x} \times \left[\frac{1}{N} + \frac{\overline{X}^2}{\sum(X - \overline{X})^2}\right]. \tag{7.10}$$

From Eq. (7.9)

$$S^2_{Y,x} = \frac{1698.75 - (0.915)^2(2000)}{2} = 12.15$$

$$S^2_a = 12.15\left[\frac{1}{4} + \frac{(90)^2}{2000}\right] = 52.245.$$

From Eq. (7.12)

$$t_2 = \frac{|5.9 - 0|}{\sqrt{52.245}} = 0.82.$$

Note that this t test has 2 ($N - 2$) d.f. This is a weak test, and a large intercept must be observed to obtain statistical significance. To define the intercept more precisely, it would be necessary to perform a larger number of assays. If there is no reason to suspect a nonlinear relationship between X and Y, a nonzero intercept, in this example, could be interpreted as being due to some interfering substance(s) in the product (the "blank"). If the presence of a nonzero intercept is suspected, one would probably want to run a sufficient number of assays to establish its presence. A precise estimate of the intercept is necessary if this linear calibration curve is used to evaluate potency.

7.4.2 Test of the Slope, *B*

The test of the slope of the least squares line is usually of more interest than the test of the intercept. Sometimes, we may only wish to be assured that the fitted line has a slope other than zero. (A horizontal line has a slope of zero.) In our example, there seems to be little doubt that the slope is greater than zero [Fig. 7.2(B)]. However, the magnitude of this slope has a special physical meaning. A slope of 1 indicates that the amount recovered (assay) is equal to the amount in the sample, after correction for the blank (i.e., subtract the *Y* intercept from the observed reading of *y*). An observation of a slope other than 1 indicates that the amount recovered is some constant percentage of the sample potency. Thus we may be interested in a test of the slope versus 1.

$$H_0 : B = 1 \qquad H_a : B \neq 1$$

A *t* test is performed using the estimated variance of the slope, as follows:

$$t = \frac{b - B}{\sqrt{S_b^2}}. \tag{7.13}$$

In the present example, from Eq. (7.11),

$$S_b^2 = \frac{S_{y,x}^2}{\sum (X - \overline{X})^2} \tag{7.11}$$

$$= \frac{12.15}{2000} = 0.006075.$$

Applying Eq. (7.13), for a two-sided test, we have

$$t = \frac{|0.915 - 1|}{\sqrt{0.006075}} = 1.09.$$

This *t* test has 2 ($N - 2$) d.f. (the variance estimate has 2 d.f.). There is insufficient evidence to indicate that the slope is significantly different from 1 at the 5% level. Table IV.4 shows that a *t* of 4.30 is needed for significance at $\alpha = 0.05$ and d.f. = 2. The test in this example has very weak power. A slope very different from 1 would be necessary to obtain statistical significance. This example again emphasizes the weakness of the statement "nonsignificant," particularly in small experiments such as this one. The reader interested in learning more details of the use and interpretation of regression in analytical methodology is encouraged to read chapter 5 in Ref. [2].

7.5 A DRUG STABILITY STUDY: A SECOND EXAMPLE OF THE APPLICATION OF LINEAR REGRESSION

The measurement of the rate of drug decomposition is an important problem in drug formulation studies. Because of the significance of establishing an expiration date defining the shelf life of a pharmaceutical product, stability data are routinely subjected to statistical analysis. Typically, the drug, alone and/or formulated, is stored under varying conditions of temperature, humidity, light intensity, and so on, and assayed for intact drug at specified time intervals. The pharmaceutical scientist is assigned the responsibility of recommending the expiration date based on scientifically derived stability data. The physical conditions of the stability test (e.g., temperature, humidity), the duration of testing, assay schedules, as well as the number of lots, bottles, and tablets that should be sampled must be defined for stability studies. Careful definition and implementation of these conditions are important because the validity and precision of the final recommended expiration date depends on how the experiment is conducted. Drug stability is discussed further in section 8.7.

The rate of decomposition can often be determined from plots of potency (or log potency) versus storage time, where the relationship of potency and time is either known or assumed to be linear. The current good manufacturing practices (CGMP) regulations [3] state that statistical criteria, including sample size and test (i.e., observation or measurement) intervals for each attribute examined, be used to assure statistically valid estimates of stability (211.166). The expiration date should be "statistically valid" (211.137, 201.17, 211.62).

The mechanics of determining shelf life may be quite complex, particularly if extreme conditions are used, such as those recommended for "accelerated" stability studies (e.g., high-temperature and high-humidity conditions). In these circumstances, the statistical techniques used to make predictions of shelf life at ambient conditions are quite advanced and beyond the scope of this book [4]. Although extreme conditions are commonly used in stability testing in order to save time and obtain a tentative expiration date, all products must eventually be tested for stability under the recommended commercial storage conditions. The FDA has suggested that at least three batches of product be tested to determine an expiration date. One should understand that different batches may show somewhat different stability characteristics, particularly in situations where additives affect stability to a significant extent. In these cases variation in the quality and quantity of the additives (excipients) between batches could affect stability. One of the purposes of using several batches for stability testing is to ensure that stability characteristics are similar from batch to batch.

The time intervals chosen for the assay of storage samples will depend to a great extent on the product characteristics and the anticipated stability. A "statistically" optimal design for a stability study would take into account the planned "storage" times when the drug product will be assayed. This problem has been addressed in the pharmaceutical literature [5]. However, the designs resulting from such considerations are usually cumbersome or impractical. For example, from a statistical point of view, the slope of the potency versus time plot (the rate of decomposition) is obtained most precisely if half of the total assay points are performed at time 0, and the other half at the final testing time. Note that $\sum(X - \overline{X})^2$ the denominator of the expression defining the variance of a slope [Eq. (7.8)], is maximized under this condition, resulting in a minimum variability of the slope. This "optimal" approach to designating assay sampling times is based on the assumption that the plot is linear during the time interval of the test. In a practical situation, one would want to see data at points between the initial and final assay in order to assess the magnitude of the decomposition as the stability study proceeds, as well as to verify the linearity of the decomposition. Also, management and regulatory requirements are better satisfied with multiple points during the course of the study. A reasonable schedule of assays at ambient conditions is 0, 3, 6, 9, 12, 18, and 24 months and at yearly intervals thereafter [6].

The example of the data analysis that will be presented here will be for a single batch. If the stability of different batches is not different, the techniques described here may be applied to data from more than one batch. A statistician should be consulted for the analysis of multibatch data that will require analysis of variance techniques [6,7]. The general approach is described in section 8.7.

Typically, stability or shelf life is determined from data from the first three production batches for each packaging configuration (container type and product strength) (see sect. 8.7). Because such testing may be onerous for multiple strengths and multiple packaging of the same drug product, matrixing and bracketing techniques have been suggested to minimize the number of tests needed to demonstrate suitable drug stability [8].

Assays are recommended to be performed at time 0 and 3, 6, 9, 12, 18 and 24 months, with subsequent assays at 12-month intervals as needed. Usually, three batches of a given strength and package configuration are tested to define the shelf life. Because many products have multiple strengths and package configurations, the concept of a "Matrix" design has been introduced to reduce the considerable amount of testing required. In this situation, a subset of all combinations of product strength, container type and size, and so on is tested at a given time point. Another subset is tested at a subsequent time point. The design should be balanced "such that each combinations of factors is tested to the same extent." All factor combinations should be tested at time 0 and at the last time point of the study. The simplest such design, called a "Basic Matrix 2/3 on Time Design," has two of the three batches tested at each time point, with all three batches tested at time 0 and at the final testing time, the time equal to the desired shelf life. Table 7.3 shows this design for a 36-month product. Tables of matrix designs show

Table 7.3 Matrix Design for Three Packages and Three Strengths

Batch		Package 1							Package 2							Package 3						
	strength	3	6	9	12	18	24	36	3	6	9	12	18	24	36	3	6	9	12	18	24	36
1	5	X		X	X		X	X	X	X		X	X		X		X	X		X	X	X
1	10	X	X		X	X		X		X	X		X	X	X	X		X	X		X	X
1	15		X	X		X	X	X	X		X	X		X	X	X	X	X		X		X
2	5	X	X		X	X		X		X	X		X	X	X	X		X	X		X	X
2	10		X	X		X	X	X	X		X	X		X	X	X	X	X		X		X
2	15	X		X	X		X	X	X	X		X	X		X		X	X		X	X	X
3	5		X	X		X	X	X	X		X	X		X	X	X	X	X		X		X
3	10	X		X	X		X	X	X	X		X	X		X		X	X		X	X	X
3	15	X	X		X	X		X		X	X		X	X	X	X		X	X		X	X

Table 7.3A Matrix Design for Three Batches and Two Strengths

Time points for testing (mo)			0	3	6	9	12	18	24	36
S		Batch 1	T	T		T	T		T	T
T	S1	Batch 2	T	T		T	T	T		T
R		Batch 3	T		T		T		T	T
E										
N		Batch 1	T		T		T		T	T
G	S2	Batch 2	T	T		T	T	T		T
T		Batch 3	T		T		T		T	T
H										

designs for multiple packages (made from the same blend or batch) and for multiple packages and strengths. These designs are constructed to be symmetrical in the spirit of optimality for such designs. For example, this is illustrated in Table 7.3, looking only at the "5" strength for Package 1. Table 7.3 shows this design for a 36-month product with multiple packages and strengths (made from the same blend). For example, in Table 7.3, each batch is tested twice, each package from each batch is tested twice, and each package is tested six times at all time points between 0 and 36 months.

With multiple strengths and packages, other similar designs with less testing have been described [9].

The risks of applying such designs are outlined in the Guidance [8]. Because of the limited testing, there is a risk of less precision and shorter dating. If pooling is not allowed, individual lots will have short dating, and combinations not tested in the matrix will not have dating estimates. Read the guidance for further details. The FDA guidance gives examples of other designs.

The analysis of these designs can be complicated. The simplest approach is to analyze each strength and configuration separately, as one would do if there were a single strength and package. Another approach is to model all configurations including interactions. The assumptions, strengths, and limitations of these designs and analyses are explained in more detail in Ref. [9].

A Bracketing design [10] is a design of a stability program such that at any point in time only extreme samples are tested, such as extremes in container size and dosage. This is particularly amenable to products that have similar composition across dosage strengths and that intermediate size and strength products are represented by the extremes [10]. (See also FDA Guideline on Stability for further discussion as to when this is applicable.)

Suppose that we have a product in three strengths and three package sizes. Table 7.4 is an example of a Bracketing design [10].

Table 7.4 Example of Bracketing Design

Strength		Low			Medium			High		
Batch		1	2	3	4	5	6	7	8	9
Container	Small	T	T	T				T	T	T
	Medium									
	Large	T	T	T				T	T	T

Table 7.5 Tablet Assays from the Stability Study

Time, *X* (mo)	Assay,[a] *y* (mg)	Average
0	51, 51, 53	51.7
3	51, 50, 52	51.0
6	50, 52, 48	50.0
9	49, 51, 51	50.3
12	49, 48, 47	48.0
18	47, 45, 49	47.0

[a] Each assay represents a different tablet.

The testing designated by T should be the full testing as would be required for a single batch. Note that full testing would require nine combinations, or 27 batches. The matrix design uses four combinations, or 12 batches.

Consider an example of a tablet formulation that is the subject of a stability study. Three randomly chosen tablets are assayed at each of six time periods: 0, 3, 6, 9, 12, and 18 months after production, at ambient storage conditions. The data are shown in Table 7.5 and Figure 7.8.

Given these data, the problem is to establish an expiration date defined as that time when a tablet contains 90% of the labeled drug potency. The product in this example has a label of 50 mg potency and is prepared with a 4% overage (i.e., the product is manufactured with a target weight of 52 mg of drug). Note that FDA is currently discouraging the use of overages to compensate for poor stability.

Figure 7.8 shows that the data are variable. A careful examination of this plot suggests that a straight line would be a reasonable representation of these data. The application of least squares line fitting is best justified in situations where a theoretical model exists showing that the decrease in concentration is linear with time (a zero-order process in this example). The kinetics of drug loss in solid dosage forms is complex and a theoretical model is not easily derived. In the present case, we will assume that concentration and time *are* truly linearly related

$$C = C_0 - Kt, \tag{7.14}$$

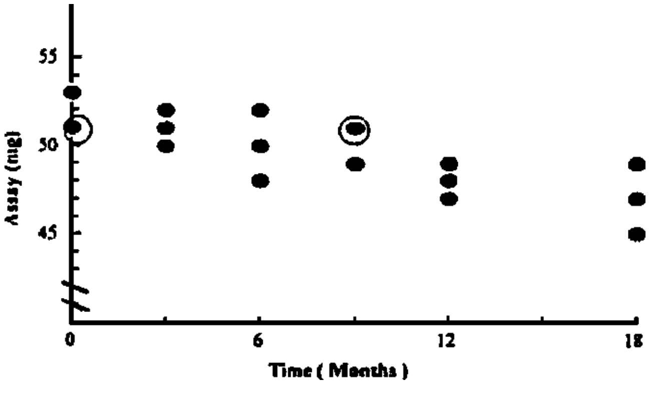


Figure 7.8 Plot of stability data from Table 7.3.

where C is the concentration at time t, C_0 the concentration at time 0 (Y intercept, A), K the rate constant ($-$ slope, $-B$), and t the time (storage time).

With the objective of estimating the shelf life, the simplest approach to the analysis of these data is to estimate the slope and intercept of the least squares line, using Eqs. (7.4) and (7.3). (An interesting exercise would be to first try and estimate the slope and intercept by eye from Fig. 7.8.) When performing the least squares calculation, note that each value of the time (X) is associated with three values of drug potency (y). When calculating C_0 and K, each "time" value is counted three times and N is equal to 18. From Table 7.3,

$$\begin{array}{lll} \sum X = 144 & \sum y = 894 & \sum Xy = 6984 \\ \sum X^2 = 1782 & \sum y^2 = 44,476 & N = 18 \\ \overline{X} = 8 & \sum (X - \overline{X})^2 = 630 & \sum (y - \overline{y})^2 = 74 \end{array}$$

From Eqs. (7.4) and (7.3), we have

$$\begin{aligned} b &= \frac{N\sum Xy - \sum X \sum y}{N\sum X^2 - (\sum X)^2} \\ &= \frac{18(6984) - 144(894)}{18(1782) - (144)^2} = \frac{-3024}{11,340} = -0.267\,\text{mg/month} \qquad (7.4) \end{aligned}$$

$$\begin{aligned} a &= \overline{y} - b\overline{X} \\ &= \frac{894}{18} - (-0.267)\frac{144}{18} = 51.80. \qquad (7.3) \end{aligned}$$

The equation of the straight line best fitting the data in Figure 7.8 is

$$C = 51.8 - 0.267\,t. \qquad (7.15)$$

The variance estimate, $S^2_{Y,x}$, represents the variability of tablet potency at a fixed time, and is calculated from Eq. (7.9)

$$\begin{aligned} S^2_{Y,x} &= \frac{\sum y^2 - (\sum y)^2/N - b^2 \sum (X - \overline{X})^2}{N - 2} \\ &= \frac{44{,}476 - (894)^2/18 - (-0.267)^2(630)}{18 - 2} = 1.825. \end{aligned}$$

To calculate the time at which the tablet potency is 90% of the labeled amount, 45 mg, solve Eq. (7.15) for t when C equals 45 mg.

$$\begin{aligned} 45 &= 51.80 - 0.267\,t \\ t &= 25.5 \text{ month}. \end{aligned}$$

The best estimate of the time needed for these tablets to retain 45 mg of drug is 25.5 months (see the point marked with a cross in Fig. 7.9). The shelf life for the product will be less than 25.5 months if variability is taken into consideration. The next section, 7.6, presents a discussion of this topic. This is an average result based on the data from 18 tablets. For any single tablet, the time for decomposition to 90% of the labeled amount will vary, depending, for example, on the amount of drug present at time zero. Nevertheless, the shelf-life estimate is based on the average result.

7.6 CONFIDENCE INTERVALS IN REGRESSION ANALYSIS

A more detailed analysis of the stability data is warranted if one understands that 25.5 months is not the true shelf life, but only an estimate of the true value. A confidence interval for the estimate of time to 45 mg potency would give a range that probably includes the true value.

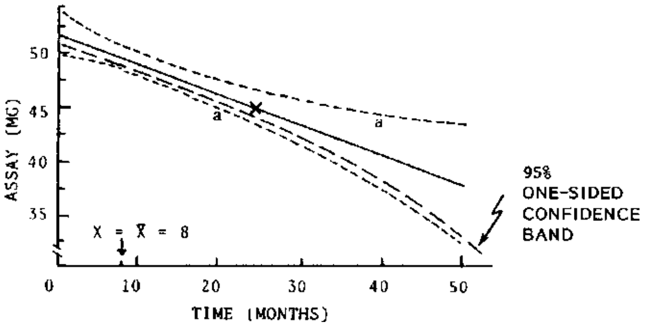


Figure 7.9 95% confidence band for "stability" line.

The concept of a confidence interval in regression is similar to that previously discussed for means. Thus the interval for the shelf life probably contains the true shelf life—that time when the tablets retain 90% of their labeled potency, on the average. The lower end of this confidence interval would be considered a conservative estimate of the true shelf life. Before giving the solution to this problem we will address the calculation of a confidence interval for Y (potency) at a given X (time). The width of the confidence interval for Y (potency) is not constant, but depends on the value of X, since Y is a function of X. In the present example, one might wish to obtain a range for the potency at 25.5 months' storage time.

7.6.1 Confidence Interval for *Y* at a Given *X*

We will construct a confidence interval for the true mean potency (Y) at a given time (X). The confidence interval can be shown to be equal to

$$Y \pm t(S_{Y,x})\sqrt{\frac{1}{N} + \frac{(X-\overline{X})^2}{\sum(X-\overline{X})^2}}. \tag{7.16}$$

t is the appropriate value ($N - 2$ d.f., Table IV.4) for a confidence interval with confidence coefficient P. For example, for a 95% confidence interval, use t values in the column headed 0.975 in Table IV.4.

In the linear regression model, y is assumed to have a normal distribution with variance $\sigma^2_{Y,x}$ at each X. As can be seen from Eq. (7.16), confidence limits for Y at a specified value of X depend on the *variance, degrees of freedom, number of data points* used to fit the line, and $X - \overline{X}$ the *distance of the specified* X (time, in this example) *from* $\overline{X}$, the average time used in the least squares line fitting. The confidence interval is smallest for the Y that corresponds to the value of X equal to $\overline{X}$, [the term, $X - \overline{X}$, in Eq. (7.16) will be zero]. As the value of X is farther from $\overline{X}$, the confidence interval for Y corresponding to the specified X is wider. Thus the estimate of Y is less precise, as the X corresponding to Y is farther away from $\overline{X}$. A plot of the confidence interval for every Y on the line results in a continuous confidence "band" as shown in Figure 7.9. The curved, hyperbolic shape of the confidence band illustrates the varying width of the confidence interval at different values of X, Y. For example, the 95% confidence interval for Y at $X = 25.5$ months [Eq. (7.16)] is

$$45 \pm 2.12(1.35)\sqrt{\frac{1}{18} + \frac{(25.5-8)^2}{630}} = 45 \pm 2.1.$$

Thus the result shows that the true value of the potency at 25.5 months is probably between 42.9 and 47.1 mg (45 ± 2.1).

7.6.2 A Confidence Interval for *X* at a Given Value of *Y*

Although the interval for the potency may be of interest, as noted above, this confidence interval does not directly answer the question about the possible variability of the shelf-life estimate. A careful examination of the two-sided confidence band for the line (Fig. 7.9) shows that 90% potency (45 mg) may occur between approximately 20 and 40 months, the points marked "*a*" in Figure 7.9. To obtain this range for *X* (time to 90% potency), using the approach of graphical estimation as described above requires the computation of the confidence band for a sufficient range of *X*. Also, the graphical estimate is relatively inaccurate. The confidence interval for the true *X* at a given *Y* can be directly calculated, although the formula is more complex than that used for the *Y* confidence interval [Eq. (7.16)].

This procedure of estimating *X* for a given value of *Y* is often called "inverse prediction." The complexity results from the fact that the solution for *X*, $X = (Y - a)/b$, is a quotient of variables. $(Y - a)$ and b are random variables; both have error associated with their measurement. The ratio has a more complicated distribution than a linear combination of variables such as is the case for $Y = a + bX$. The calculation of the confidence interval for the true *X* at a specified value of *Y* is

$$\frac{(X - g\overline{X}) \pm [t(S_{Y,x})/b]\left[\sqrt{(1-g)/N + (X-\overline{X})^2/\sum(X-\overline{X})^2}\right]}{1-g}, \tag{7.17}$$

where

$$g = \frac{t^2(S^2_{Y,x})}{b^2\sum(X-\overline{X})^2}$$

t is the appropriate value for a confidence interval with confidence coefficient equal to P; for example, for a two-sided 95% confidence interval, use values of t in the column headed 0.975 in Table IV.4.

A 95% confidence interval for *X* will be calculated for the time to 90% of labeled potency. The potency is 45 mg (*Y*) when 10% of the labeled amount decomposes. The corresponding time (*X*) has been calculated above as 25.5 months. For a two-sided confidence interval, applying Eq. (7.17), we have

$$g = \frac{(2.12)^2(1.825)}{(-0.267)^2(630)} = 0.183$$

$$X = 25.5 \qquad \overline{X} = 8 \qquad N = 18.$$

The confidence interval is

$$\frac{[25.5 - 0.183(8)] \pm [2.12(1.35)/(-0.267)][\sqrt{0.817/18 + (17.5)^2/630}]}{0.817}$$
$$= 19.8 \text{ to } 39.0 \text{ months.}$$

Thus, using a two-sided confidence interval, the true time to 90% of labeled potency is probably between 19.8 and 39.0 months. A conservative estimate of the shelf life would be the lower value, 19.8 months. If g is greater than 1, a confidence interval cannot be calculated because the slope is not significantly greater than 0.

The Food and Drug Administration has suggested that a one-sided confidence interval may be more appropriate than a two-sided interval to estimate the expiration date. For most drug products, drug potency can only decrease with time, and only the lower confidence band of the potency versus time curve may be considered relevant. (An exception may occur in the case of liquid products where evaporation of the solvent could result in an increased potency with time.) The 95% one-sided confidence limits for the time to reach a potency of 45 are computed using Eq. (7.17). Only the lower limit is computed using the appropriate t value that cuts off 5% of the area in a single tail. For 16 d.f., this value is 1.75 (Table IV.4), "g" = 0.1244. The calculation is

$$\frac{[25.5 - 0.1244(8)] + [1.75(1.35)/(-0.267)][\sqrt{0.8756/18 + (17.5)^2/630}}{0.8756}$$
$$= 20.6 \text{ months.}$$

The one-sided 95% interval for X can be interpreted to mean that the time to decompose to a potency of 45 is probably greater than 20.6 months. Note that the shelf life based on the one-sided interval is longer than that based on a two-sided interval (Fig. 7.9).

7.6.3 Prediction Intervals

The confidence limits for Y and X discussed above are limits for the *true values*, having specified a value of Y (potency or concentration, for example) corresponding to some value of X, or an X (time, for example) corresponding to a specified value of Y. An important application of confidence intervals in regression is to obtain confidence intervals for *actual future measurements* based on the least squares line.

1. We may wish to obtain a confidence interval for a value of Y to be actually measured at some value of X (some future time, for example).
2. In the example of the calibration (sect. 7.2), having observed a new value, y, after the calibration line has been established, we would want to use the information from the fitted calibration line to predict the concentration, or potency, X, and establish the confidence limits for the concentration at this newly observed value of y. This is an example of inverse prediction.

For the example of the stability study, we may wish to obtain a confidence interval for an actual assay (y) to be performed at some given future time, after having performed the experiment used to fit the least squares line (case 1 above).

The formulas for calculating a "prediction interval," a confidence interval for a future determination, are similar to those presented in Eqs. (7.16) and (7.17), with one modification. In Eq. (7.16), we add 1 to the sum under the square root portion of the expression. Similarly, for the inverse problem, Eq. (7.17) the expression $(1 - g)/N$ is replaced by $(N + 1)(1 - g)/N$. Thus the prediction interval for Y at a given X is

$$Y \pm t(S_{Y,x})\sqrt{1 + \frac{1}{N} + \frac{(X - \overline{X})^2}{\sum (X - \overline{X})^2}}. \tag{7.18}$$

The prediction interval for X at a specified Y is

$$\frac{(X - g\overline{X}) \pm [t(S)/b]\left[\sqrt{(N+1)(1-g)/N + (X - \overline{X})^2/\sum(X - \overline{X})^2}\right]}{1 - g}. \tag{7.19}$$

The following examples should clarify the computations. In the stability study example, suppose that one wishes to construct a 95% confidence (prediction) interval for *an assay to be performed* at 25.5 months. (An actual measurement is obtained at 25.5 months.) This interval will be larger than that calculated based on Eq. (7.16), because the uncertainty now includes assay variability for the proposed assay in addition to the uncertainty of the least squares line. Applying Eq. (7.18) ($Y = 45$), we have

$$45 \pm 2.12(1.35)\sqrt{1 + \frac{1}{18} + \frac{17.5^2}{630}} = 45 \pm 3.55\,\text{mg}.$$

In the example of the calibration line, consider an unknown sample that is analyzed and shows a value (y) of 90. A prediction interval for X is calculated using Eq. (7.19). X is predicted to be 91.9 (see sect. 7.2).

$$g = \frac{(4.30)^2(12.15)}{(0.915)^2(2000)} = 0.134$$

$$\frac{[91.9 - 0.134(90)] \pm (4.3)(3.49)/0.915[\sqrt{5(0.866)/4 + (1.9)^2/2000}]}{0.866}$$

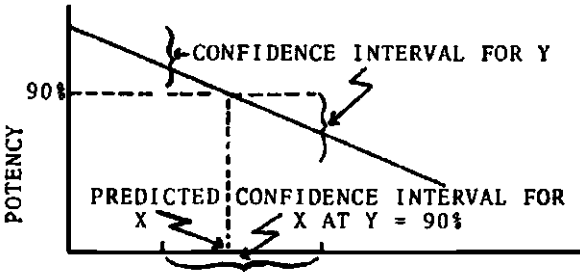
$$= 72.5 \text{ to } 111.9.$$

The relatively large uncertainty of the estimate of the true value is due to the small number of data points (four) and the relatively large variability of the points about the least squares line ($S^2_{Y,x} = 12.15$).

7.6.4 Confidence Intervals for Slope (*B*) and Intercept (*A*)

A confidence interval can be constructed for the slope and intercept in a manner analogous to that for means [Eq. (6.2)]. The confidence interval for the slope is

$$b \pm t(S_b) = b \pm \frac{t(S_{Y,x})}{\sqrt{\sum (X - \overline{X})^2}}. \qquad (7.20)$$


CONFIDENCE INTERVAL FOR Y
90%
POTENCY
PREDICTED X
CONFIDENCE INTERVAL FOR X AT Y = 90%

A confidence interval for the intercept is

$$a \pm t(S_a) = a \pm t(S_{Y,x})\sqrt{\frac{1}{N} + \frac{\overline{X}^2}{\sum (X - \overline{X})^2}}. \qquad (7.21)$$

A 95% confidence interval for the slope of the line in the stability example is [Eq. (7.20)]

$$(-0.267) \pm \frac{2.12(1.35)}{\sqrt{630}} = -0.267 \pm 0.114$$
$$= -0.381 \text{ to } -0.153.$$

A 90% confidence interval for the intercept in the calibration line example (sect. 7.2) is [Eq. (7.21)]

$$5.9 \pm 2.93(3.49)\sqrt{\frac{1}{4} + \frac{90^2}{2000}} = 5.9 \pm 21.2 = -15.3 \text{ to } 27.1.$$

(Note that the appropriate value of *t* with 2 d.f. for a 90% confidence interval is 2.93.)

7.7 WEIGHTED REGRESSION

One of the assumptions implicit in the applications of statistical inference to regression procedures is that the variance of *y* be the same at each value of *X*. Many situations occur in practice when this assumption is violated. One common occurrence is the variance of *y* being approximately proportional to X^2. This occurs in situations where *y* has a constant coefficient of variation (CV) and *y* is proportional to *X* ($y = BX$), commonly observed in instrumental methods of analysis in analytical chemistry. Two approaches to this problem are (*a*) a transformation of

Table 7.6 Analytical Data for a Spectrophotometric Analysis

Concentration (X)	Optical density (y)		CV	Weight (w)
5	0.105	0.098	0.049	0.04
10	0.201	0.194	0.025	0.01
25	0.495	0.508	0.018	0.0016
50	0.983	1.009	0.018	0.0004
100	1.964	2.013	0.017	0.0001

y to make the variance homogeneous, such as the log transformation (see chap. 10), and (b) a weighted regression analysis.

Below is an example of weighted regression analysis in which we assume a constant CV and the variance of y proportional to X^2 as noted above. This suggests a weighted regression, weighting each value of Y by a factor that is inversely proportional to the variance, $1/X^2$. Table 7.6 shows data for the spectrophotometric analysis of a drug performed at 5 concentrations in duplicate.

Equation (7.22) is used to compute the slope for the weighted regression procedure.

$$b = \frac{\sum wXy - \sum wX \sum wy / \sum w}{\sum wX^2 - (\sum wX)^2 / \sum w}. \tag{7.22}$$

The computations are as follows:

$\sum w = 0.04 + 0.04 + \ldots + 0.0001 + 0.0001 = 0.1042$
$\sum wXy = (0.04)(5)(0.105) + (0.04)(5)(0.098) + \ldots + (0.0001)(100)(1.964) + (0.0001)(100)(2.013)$
$= 0.19983$
$\sum wX = 2(0.04)(5) + 2(0.01)(10) + \ldots + 2(0.0001)(100) = 0.74$
$\sum wy = (0.04)(0.105) + (0.04)(0.098) + \ldots + (0.0001)(1.964) + (0.0001)(2.013) = 0.0148693$
$\sum wX^2 = 2(0.04)(5)^2 + 2(0.01)(10)^2 + \ldots + 2(0.0001)(100)^2 = 10$

Therefore, the slope $b =$

$$\frac{0.19983 - (0.74)(0.0148693)/0.1042}{10 - (0.74)^2/0.1042} = 0.01986.$$

The intercept is

$$a = \overline{y}_w - b(\overline{X}_w), \tag{7.23}$$

where $\overline{y}_w = \sum wy / \sum w$ and $\overline{X}_w = \sum wX / \sum w$

$$a = 0.0148693/0.1042 - 0.01986(0.74/0.1042) = 0.00166. \tag{7.23a}$$

The weighted least squares line is shown in Figure 7.10.

7.8 ANALYSIS OF RESIDUALS

Emphasis is placed elsewhere in this book on the importance of carefully examining and graphing data prior to performing statistical analyses. The approach to examining data in this context is commonly known as Exploratory Data Analysis (EDA) [11]. One aspect of EDA is the examination of residuals. Residuals can be thought of as deviations of the observed data from

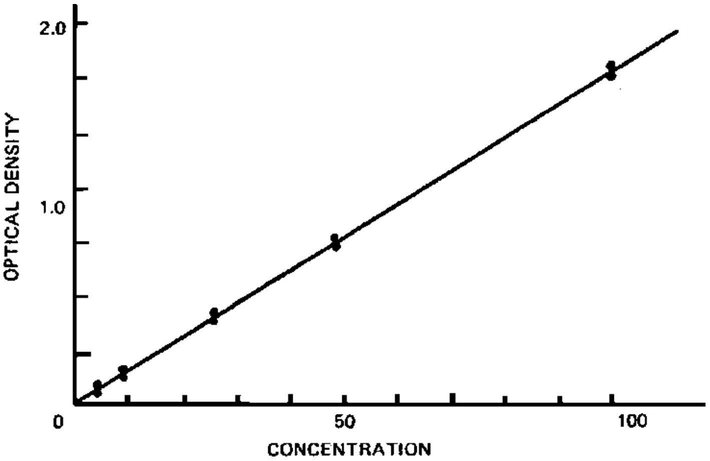


Figure 7.10 Weighted regression plot for data from Table 7.7.

the fit to the statistical model. Examination of residuals can reveal problems such as variance heterogeneity or nonlinearity. This brief introduction to the principle of residual analysis uses the data from the regression analysis in section 7.7.

The residuals from a regression analysis are obtained from the differences between the observed and predicted values. Table 7.7 shows the residuals from an unweighted least squares fit of the data of Table 7.6. Note that the fitted values are obtained from the least squares equation $y = 0.001789 + 0.019874(X)$.

If the linear model and the assumptions in the least squares analysis are valid, the residuals should be approximately normally distributed, and no trends should be apparent.

Figure 7.11 shows a plot of the residuals as a function of X. The fact that the residuals show a fan-like pattern, expanding as X increases, suggests the use of a log transformation or weighting procedure to reduce the variance heterogeneity. In general, the intelligent interpretation of residual plots requires knowledge and experience. In addition to the appearance of patterns in the residual plots that indicate relationships and character of data, outliers usually become obviously apparent [12].

Figure 7.12 shows the residual plot after a log (In) transformation of X and Y. Much of the variance heterogeneity has been removed.

For readers who desire more information on this subject, the book *Graphical Exploratory Data Analysis* [13] is recommended.

Table 7.7 Residuals from Least Squares Fit of Analytical Data (Table 7.6)

Unweighted			Log transform		
Actual	**Predicted value**	**Residual**	**Actual**	**Predicted value**	**Residual**
0.105	0.101	+0.00384	−2.254	−2.298	+0.044
0.201	0.201	+0.00047	−1.604	−1.6073	+0.0033
0.495	0.499	−0.00364	−0.703	−0.695	−0.008
0.983	0.995	−0.0126	−0.017	−0.0004	−0.0166
1.964	1.989	−0.025	+0.675	+0.6863	−0.0113
0.098	0.101	−0.00316	−2.323	−2.298	−0.025
0.194	0.201	−0.00653	−1.640	−1.6073	−0.0033
0.508	0.499	+0.00936	−0.677	−0.6950	+0.018
1.009	0.995	+0.0135	+0.009	−0.0042	+0.0132
2.013	1.989	+0.00238	+0.700	0.6863	+0.0137

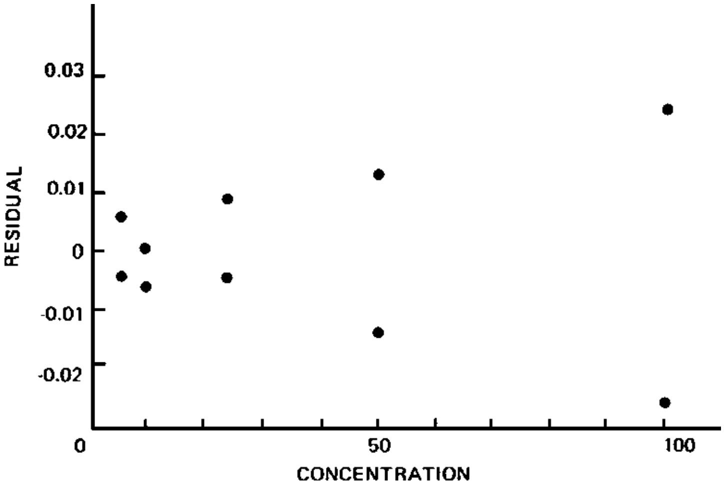


Figure 7.11 Residual plot for unweighted analysis of data of Table 7.6.

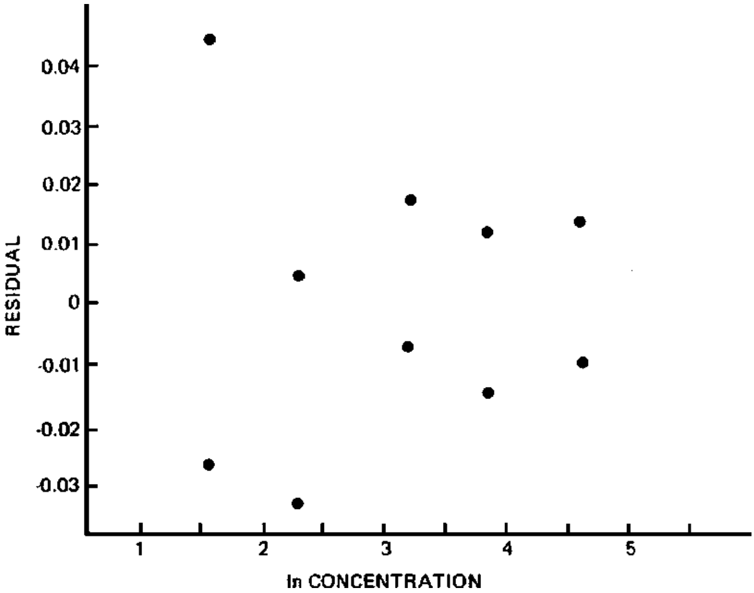


Figure 7.12 Residual plot for analysis of ln transformed data of Table 7.6.

7.9 NONLINEAR REGRESSION**

Linear regression applies to the solution of relationships where the function of Y is linear in the parameters. For example, the equation

$$Y = A + BX$$

is linear in A and B, the parameters. Similarly, the equation

$$Y = A + Be^{-x}$$

is also linear in the parameters. One should also appreciate that a linear equation can exist in more than two dimensions. The equation

$$Y = A + BX + CX^2,$$

an example of a quadratic equation, is linear in the parameters, A, B, and C. These parameters can be estimated by using methods of multiple regression (see App. III and Ref. [1]).

An example of a relationship that is nonlinear in this context is

$$Y = A + e^{BX}.$$

Here the parameter B is not in a linear form.

If a linearizing transformation can be made, then this approach to estimating the parameters would be easiest. For example, the simple first-order kinetic relationship

$$Y = Ae^{-BX}$$

is not linear in the parameters, A and B. However, a log transformation results in a linear equation

$$\ln Y = \ln A - BX.$$

Using the least squares approach, we can estimate ln A (A is the antilog) and B, where ln A is the intercept and B is the slope of the straight line when ln Y is plotted versus X. If statistical tests and other statistical estimates are to be made from the regression analysis, the assumptions of normality of Y (now ln Y) and variance homogeneity of Y at each X are necessary. If Y is normal and the variances of Y at each X are homogeneous to start with, the ln transformation will invalidate the assumptions. (On the other hand, if Y is lognormal with constant CV, the log transformation will be just what is needed to validate the assumptions.)

Some relationships cannot be linearized. For example, in pharmacokinetics, the one-compartment model with first order absorption and excretion has the following form

$$C = D(e^{-ket} - e^{-kat})$$

where D, ke, and ka are constants (parameters). This equation cannot be linearized. The use of nonlinear regression methods can be used to estimate the parameters in these situations as well as the situations in which Y is normal with homogeneous variance prior to a transformation, as noted above.

The solutions to nonlinear regression problems require more advanced mathematics relative to most of the material in this book. A knowledge of elementary calculus is necessary, particularly the application of Taylor's theorem. Also, a knowledge of matrix algebra is useful in order to solve these kinds of problems. A simple example will be presented to demonstrate the principles. The general matrix solutions to linear and multiple regression will also be demonstrated.

In a stability study, the data in Table 7.8 were available for analysis. The equation representing the degradation process is

$$C = C_0e^{-kt}. \tag{7.24}$$

Table 7.8 Data from a Stability Study

Time (t)	Concentration mg/L (C)
1 hr	63
2 hr	34
3 hr	22

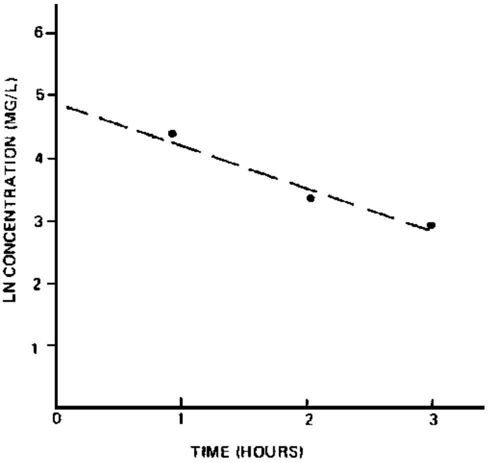


Figure 7.13 Plot of stability data from Table 7.8.

The concentration values are known to be normal with the variance constant at each value of time. Therefore, the usual least squares analysis will not be used to estimate the parameters C_0 and k after the simple linearizing transformation:

$$\ln C = \ln C_0 - kt.$$

The estimate of the parameters using nonlinear regression as demonstrated here uses the first terms of Taylor's expansion, which approximates the function and results in a linear equation. It is important to obtain good initial estimates of the parameters, which may be obtained graphically. In the present example, a plot of ln C versus time (Fig. 7.13) results in initial estimates of 104 for C_0 and +0.53 for k. The process then estimates a change in C_0 and a change in k that will improve the equation based on the comparison of the fitted data to the original data. Typical of least squares procedures, the fit is measured by the sum of the squares of the deviations of the observed values from the fitted values. The best fit results from an iterative procedure. The new estimates result in a better fit to the data. The procedure is repeated using the new estimates, which results in a better fit than that observed in the previous iteration. When the fit, as measured by the sum of the squares of deviations, is negligibly improved, the procedure is stopped. Computer programs are available to carry out these tedious calculations.

The Taylor expansion requires taking partial derivatives of the function with respect to C_0 and k. For the equation, $C = C_0e^{-kt}$, the resulting expression is

$$dC = dC_0'(e^{-k't}) - dk'(C_0')(te^{-k't}). \tag{7.25}$$

In Eq. (7.25), dC is the change in C resulting from small changes in C_0 and k evaluated at the point, C_0' and k'. dC_0' is the change in the estimate of C_0, and dk' is the change in the estimate of k. $(e^{-k't})$ and $C_0'(te^{-k't})$ are the partial derivatives of Eq. (7.24) with respect to C_0 and k, respectively.

Equation (7.25) is linear in dC_0' and dk'. The coefficients of dC_0' and dk' are $(e^{-k't})$ and $-(C_0')(te^{-k't})$, respectively. In the computations below, the coefficients are referred to as $X1$ and $X2$, respectively, for convenience. Because of the linearity, we can obtain the least squares estimates of dC_0' and dk' by the usual regression procedures.

The computations for two iterations are shown below. The solution to the least squares equation is usually accomplished using matrix manipulations. The solution for the coefficients can be proven to have the following form:

$$B = (X'\,X)^{-1}(X'Y).$$

Table 7.9 Results of First Iteration

Time (*t*)	*C*	*C′*	*dC*	*X*1	*X*2
1	63	61.2	1.79	0.5886	−61.2149
2	34	36.0	−2.03	0.3465	−72.0628
3	22	21.2	0.79	0.2039	−63.6248
			$\sum dC'^2 = 7.94$		

The matrix *B* will contain the estimates of the coefficients. With two coefficients, this will be a 2 × 1 (2 rows and 1 column) matrix.

In Table 7.9, the values of *X*1 and *X*2 are $(e^{-k't})$ and $(C_0')(te^{-k't})$, respectively, using the initial estimates of $C_0' = 104$ and $k' = +0.53$ (Fig. 7.13). Note that the fit is measured by the $\sum dC'^2 = 7.94$.

The solution of $(X'X)^{-1}(X'Y)$ gives the estimates of the parameters, dC_0' and k'

$$\begin{vmatrix} X'X \end{vmatrix}^{-1} \qquad \begin{vmatrix} X'Y \end{vmatrix}$$
$$\begin{vmatrix} 11.5236 & 0.06563 \\ 0.06563 & 0.00045079 \end{vmatrix} \begin{vmatrix} 0.5296 \\ -16.9611 \end{vmatrix} = \begin{vmatrix} 4.99 \\ 0.027 \end{vmatrix}$$

The new estimates of C_0 and k are

$$C_0' = 104 + 4.99 = 108.99$$
$$k' = 0.53 + 0.027 = +0.557.$$

With these estimates, new values of C' are calculated in Table 7.10.

Note that the $\sum dC'^2$ is 5.85, which is reduced from 7.94, from the initial iteration. The solution of $(X'X)^{-1}(X'Y)$ is

$$\begin{vmatrix} 12.587 & +0.06964 \\ +0.06964 & 0.0004635 \end{vmatrix} \begin{vmatrix} 0.0351 \\ -0.909 \end{vmatrix} = \begin{vmatrix} 0.378 \\ 0.002 \end{vmatrix}$$

Therefore, the new estimates of C_0 and k are

$$C_0' = 108.99 + 0.38 = 109.37$$
$$k = 0.557 + 0.002 = 0.559.$$

The reader can verify that the new value of dC'^2 is now 5.74. The process is repeated until dC'^2 becomes stable. The final solution is $C_0 = 109.22$, k 0.558.

Another way of expressing the decomposition is

$$C = e^{\ln C_0 - kt}$$

Table 7.10 Results of Second Iteration

Time *(t)*	*C*	*C′*	*dC′*	*X*1	*X*2
1	63	62.4	0.6	0.5729	−62.4431
2	34	35.8	−1.8	0.3282	−71.5505
3	22	20.5	1.5	0.18806	−61.4896
			$\sum dC'^2 = 5.85$		

or

$$\ln C = \ln C_0 - kt.$$

The ambitious reader may wish to try a few iterations using this approach. Note that the partial derivatives of C with respect to C_0 and k are $(1/C_0)\,(e^{\ln C_0 - kt})$and $-t(e^{\ln C_0 - kt})$, respectively.

7.10 CORRELATION

Correlation methods are used to measure the "association" of two or more variables. Here, we will be concerned with two observations for each sampling unit. We are interested in determining if the two values are related, in the sense that one variable may be predicted from a knowledge of the other. The better the prediction, the better the correlation. For example, if we could predict the dissolution of a tablet based on tablet hardness, we say that dissolution and hardness are *correlated.* Correlation analysis assumes a linear or *straight-line relationship* between the two variables.

Correlation is usually applied to the relationship of continuous variables, and is best visualized as a *scatter plot* or correlation diagram. Figure 7.14(A) shows a scatter plot for two variables, tablet weight and tablet potency. Tablets were individually weighed and then assayed. Each point in Figure 7.14(A) represents a single tablet (X = weight, Y = potency). Inspection of this diagram suggests that weight and potency are *positively* correlated, as is indicated by the positive slope, or trend. Low-weight tablets are associated with low potencies, and vice versa. This positive relationship would probably be expected on intuitive grounds. If the tablet granulation is homogeneous, a larger weight of material in a tablet would contain larger amounts of drug. Figure 7.14(B) shows the correlation of tablet weights and dissolution rate. Smaller tablet weights are related to higher dissolution rates, a *negative* correlation (negative trend).

Inspection of Figure 7.14(A) and (B) reveals what appears to be an obvious relationship. Given a tablet weight, we can make a good "ballpark" estimate of the dissolution rate and

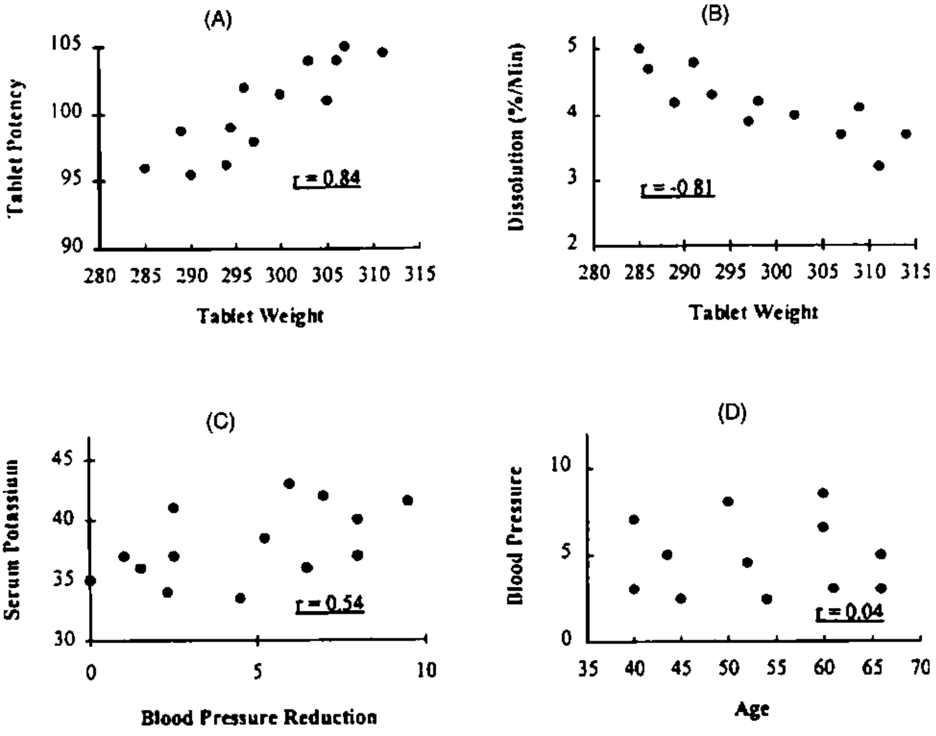


Figure 7.14 Examples of various correlation diagrams or scatter plots. The correlation coefficient, r, is defined in section 7.10.1.

potency. However, the relationship between variables is not always as apparent as in these examples. The relationship may be partially obscured by variability, or the variables may not be related at all. The relationship between a patient's blood pressure reduction after treatment with an antihypertensive agent and serum potassium levels is not as obvious [Fig. 7.14(C)]. There seems to be a trend toward higher blood pressure reductions associated with higher potassium levels—or is this just an illusion? The data plotted in Figure 7.14(D), illustrating the correlation of blood pressure reduction and age, show little or no correlation.

The various scatter diagrams illustrated in Figure 7.14 should give the reader an intuitive feeling for the concept of correlation. There are many experimental situations where a researcher would be interested in relationships among two or more variables. Similar to applications of regression analysis, correlation relationships may allow for prediction and interpretation of experimental mechanisms. Unfortunately, the concept of correlation is often misused, and more is made of it than is deserved. For example, the presence of a strong correlation between two variables does not necessarily imply a causal relationship. Consider data that show a positive relationship between cancer rate and consumption of fluoridated water. Regardless of the possible validity of such a relationship, such an observed correlation does not necessarily imply a causal effect. One would have to investigate further other factors in the environment occurring concurrently with the implementation of fluoridation, which may be responsible for the cancer rate increase. Have other industries appeared and grown during this period, exposing the population to potential carcinogens? Have the population characteristics (e.g., racial, age, sex, economic factors) changed during this period? Such questions may be resolved by examining the cancer rates in control areas where fluoridation was not enforced.

The correlation coefficient is a measure of the "degree" of correlation, which is often *erroneously* interpreted as a measure of "linearity." That is, a strong correlation is sometimes interpreted as meaning that the relationship between X and Y is a straight line. As we shall see further in this discussion, this interpretation of correlation is not necessarily correct.

7.10.1 Correlation Coefficient

The correlation coefficient is a quantitative measure of the relationship or correlation between two variables.

$$\text{Correlation coefficient} = r = \frac{\sum (X - \overline{X})(y - \overline{y})}{\sqrt{\sum (X - \overline{X})^2 \sum (y - \overline{y})^2}}. \tag{7.26}$$

A shortcut computing formula is

$$r = \frac{N\sum Xy - \sum X \sum y}{\sqrt{\left[N\sum X^2 - (\sum X)^2\right]\left[N\sum y^2 - (\sum y)^2\right]}}, \tag{7.27}$$

where N is the number of X, y pairs.

The correlation coefficient, r, may be better understood by its relationship to $S^2_{Y.x}$, the variance calculated from regression line fitting procedures. r^2 represents the relative reduction in the sum of squares of the variable y resulting from the fitting of the X, y line. For example, the sum of squares $\left[\sum (y - \overline{y})^2\right]$ for the y values 0, 1, and 5 is equal to 14 [see Eq. (1.4)].

$$\sum (y - \overline{y})^2 = 0^2 + 1^2 + 5^2 - \frac{(0 + 1 + 5)^2}{3} = 14.$$

If these same y values were associated with X values, the sum of squares of y from the regression of y and X will be *equal to or less than* $\sum (y - \overline{y})^2$, or 14 in this example. Suppose that X and y values are as follows (Fig. 7.15):

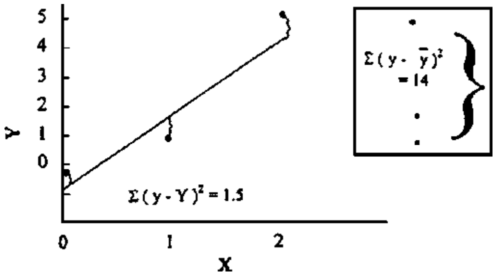


Figure 7.15 Reduction in sum of squares due to regression.

	X	y	Xy	
	0	0	0	$\sum (X-\bar{X})^2 = 2$
	1	1	1	
	2	5	10	$\sum (y-\bar{y})^2 = 14$
Sum	3	6	11	

According to Eq. (7.9), the sum of squares due to deviations of the y values from the regression line is

$$\sum (y-\bar{y})^2 - b^2 \sum (X-\overline{X})^2, \tag{7.28}$$

where b is the slope of the regression line (y on X). The term $b^2 \sum (X-\overline{X})^2$ is the reduction in the sum of squares due to the straight-line regression fit. Applying Eq. (7.28), the sum of squares is

$$14 - (2.5)^2(2) = 14 - 12.5 = 1.5 \qquad \text{(the slope, } b\text{, is 2.5).}$$

r^2 is the relative reduction of the sum of squares

$$\frac{14-1.5}{14} = 0.893 \qquad r = \sqrt{0.893} = 0.945.$$

The usual calculation of r, according to Eq. (7.27), is as follows:

$$\frac{3(11)-(3)(6)}{\sqrt{[3(5)-(3)^2][3(26)-(36)]}} = \frac{15}{\sqrt{6(42)}} = 0.945.$$

Thus, according to this notion, r can be interpreted as the relative degree of scatter about the regression line. If X and y values lie exactly on a straight line (a perfect fit), $S^2_{Y,x}$ is 0, and r is equal to ± 1; +1 for a line of positive slope and −1 for a line of negative slope. For a correlation coefficient equal to 0.5, $r^2 = 0.25$. The sum of squares for y is reduced 25%. A correlation coefficient of 0 means that the X, y pairs are not correlated [Fig. 7.14(D)].

Although there are no assumptions necessary to calculate the correlation coefficient, statistical analysis of r is based on the notion of a bivariate normal distribution of X and y. We will not delve into the details of this complex probability distribution here. However, there are two interesting aspects of this distribution that deserve some attention with regard to correlation analysis.

1. In typical correlation problems, *both X and y are variable.* This is in contrast to the linear regression case, where X is considered *fixed,* chosen, a priori, by the investigator.

2. In a bivariate normal distribution, X and y are linearly related. The regression of both X on y and y on X is a straight line.¶ Thus, when statistically testing correlation coefficients, we are not testing for linearity. As described below, the statistical test of a correlation coefficient is a test of correlation or independence. According to Snedecor and Cochran, the correlation coefficient "estimates the degree of *closeness* of a linear relationship between two variables, Y and X, and the meaning of this concept is not easy to grasp" [11].

7.10.2 Test of Zero Correlation

The correlation coefficient is a rough measure of the degree of association of two variables. The degree of association may be measured by how well one variable can be predicted from another; the closer the correlation coefficient is to + 1 or − 1, the better the correlation, the better the predictive power of the relationship. A question of particular importance from a statistical point of view is whether or not an observed correlation coefficient is "real" or due to chance. If two variables from a bivariate normal distribution are uncorrelated (independent), the correlation coefficient is 0. Even in these cases, in actual experiments, random variation will result in a correlation coefficient different from zero. Thus, it is of interest to test an observed correlation coefficient, r, versus a hypothetical value of 0. This test is based on an assumption that y is a normal variable [11]. The test is a t test with $(N - 2)$ d.f., as follows:

$$H_0 : \rho = 0 \quad H_a : \rho \neq 0,$$

where ρ is the true correlation coefficient, estimated by r.

$$t_{N-2} = \frac{\left|r\sqrt{N-2}\right|}{\sqrt{1-r^2}}. \tag{7.29}$$

The value of t is referred to a t distribution with $(N - 2)$ d.f., where N is the sample size (i.e., the number of pairs). Interestingly, this test is identical to the test of the slope of the least squares fit, $Y = a + bX$ [Eq. (7.13)]. In this context, one can think of the test of the correlation coefficient as a test of the significance of the slope versus 0.

To illustrate the application of Eq. (7.29), Table 7.11 shows data of diastolic blood pressure and cholesterol levels of 10 randomly selected men. The data are plotted in Figure 7.16. r is calculated from Eq. (7.27)

$$\begin{aligned} r &= \frac{N\sum Xy - \sum X \sum y}{\sqrt{\left[N\sum X^2 - (\sum X)^2\right]\left[N\sum y^2 - (\sum y)^2\right]}} \\ &= \frac{10(260{,}653) - (3111)(825)}{\sqrt{[10(987{,}893) - 3111^2][10(69{,}279) - 825^2]}} = 0.809. \end{aligned} \tag{7.30}$$

r is tested for significance using Eq. (7.29).

$$t_8 = \frac{\left|0.809\sqrt{8}\right|}{\sqrt{1-(0.809)^2}} = 3.89.$$

A value of t equal to 2.31 is needed for significance at the 5% level (see Table IV.4). Therefore, the correlation between diastolic blood pressure and cholesterol is significant. The correlation is apparent from inspection of Figure 7.16.

¶ The regression of y on X means that X is assumed to be the fixed variable when calculating the line. This line is different from that calculated when Y is considered the fixed variable (unless the correlation coefficient is 1, when both lines are identical). The slope of the line is rS_y/S_x for the regression of y on X and rS_x/S_y for x on Y.

Table 7.11 Diastolic Blood Pressure and Serum Cholesterol of 10 Persons

Person	Diastolic blood pressure (DBP), y	Cholesterol (C), X	Xy
1	80	307	24,560
2	75	259	19,425
3	90	341	30,690
4	74	317	23,458
5	75	274	20,550
6	110	416	45,760
7	70	267	18,690
8	85	320	27,200
9	88	274	24,112
10	78	336	26,208
	$\sum y = 825$	$\sum X = 3111$	$\sum Xy = 260{,}653$
	$\sum y^2 = 69{,}279$	$\sum X^2 = 987{,}893$	

Significance tests for the correlation coefficient versus values other than 0 are not very common. However, for these tests, the t test described above [Eq. (7.29)] should not be used. An approximate test is available to test for correlation coefficients other than 0 (e.g., H_0: $\rho = 0.5$). Since applications of this test occur infrequently in pharmaceutical experiments, the procedure will not be presented here. The statistical test is an approximation to the normal distribution, and the approximation can also be used to place confidence intervals on the correlation coefficient. A description of these applications is presented in Ref. [11].

7.10.3 Miscellaneous Comments

Before leaving the topic of correlation, the reader should once more be warned about the potential misuses of interpretations of correlation and the correlation coefficient. In particular, the association of high correlation coefficients with a "cause and effect" and "linearity" is not necessarily valid. Strong correlation *may* imply a direct causal relationship, but the nature of the measurements should be well understood before firm statements can be made about cause and effect. One should be keenly aware of the common occurrence of spurious correlations due to indirect causes or remote mechanisms.

The correlation coefficient does not test the linearity of two variables. If anything, it is more related to the slope of the line relating the variables. Linearity is assumed for the routine statistical test of the correlation coefficient. As has been noted above, the correlation coefficient measures the degree of correlation, a measure of the variability of a predictive relationship.

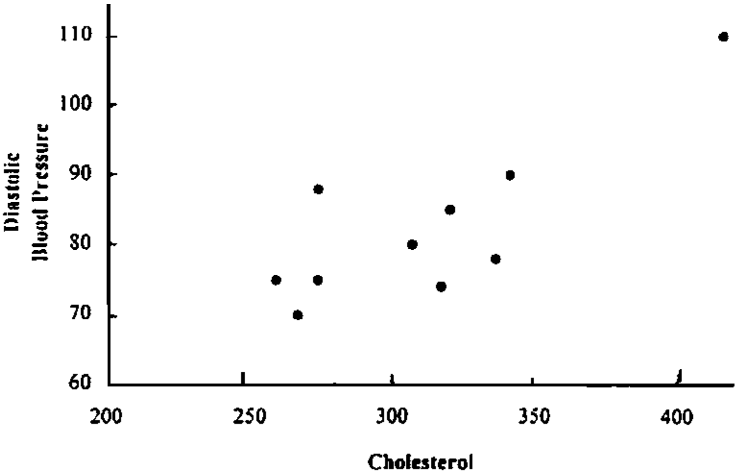


Figure 7.16 Plot of data from Table 7.11.

Table 7.12 Two Data Sets Illustrating Some Problems of Interpreting Correlation Coefficients

Set *A*		Set *B*	
X	*y*	*X*	*y*
−2	0	0	0
−1	3	2	4
0	4	4	16
+1	3	6	36
+2	0		

A proper test for linearity (i.e., do the data represent a straight-line relationship between X and Y?) is described in Appendix II and requires replicate measurements in the regression model. Usually, correlation problems deal with cases where both variables, X and y, are variable in contrast to the regression model where X is considered fixed. In correlation problems, the question of linearity is usually not of primary interest. We are more interested in the degree of association of the variables. Two examples will show that a high correlation coefficient does not necessarily imply "linearity" and that a small correlation coefficient does not necessarily imply lack of correlation (if the relationship is nonlinear).

Table 7.12 shows two sets of data that are plotted in Figure 7.17. Both data sets A and B show perfect (but nonlinear) relationships between X and y. Set A is defined by $Y = 4 - X^2$. Set B is defined by $Y = X^2$. Yet the correlation coefficient for set A is 0, an implication of no correlation, and set B has a correlation coefficient of 0.96, very strong correlation (*but not linearity!*). These examples should emphasize the care needed in the interpretation of the correlation coefficient, particularly in nonlinear systems.

Another example of data for which the correlation coefficient can be misleading is shown in Table 7.13 and Figure 7.18. In this example, drug stability is plotted versus pH. Five experiments were performed at low pH and one at high pH. The correlation coefficient is 0.994, a highly significant result ($p < 0.01$). Can this be interpreted that the data in Figure 7.18 are a good fit to a straight line? Without some other source of information, it would take a great deal of imagination to assume that the relationship between pH and $t_{1/2}$ is linear over the range of pH equal to 2.0 to 5.5. Even if the relationship were linear, had data been available for points in between pH 2.0 and 5.5, the fit may not be as good as that implied by the large value of r in this example. This situation can occur when one value is far from the cluster of the main body of data. One should be cautious in "over-interpreting" the correlation coefficient in these cases. When relationships between variables are to be quantified for predictive or theoretical reasons, regression procedures, if applicable, are recommended. Correlation, per se, is not as versatile or informative as regression analysis for describing the relationship between variables.

7.11 COMPARISON OF VARIANCES IN RELATED SAMPLES

In section 5.3, a test was presented to compare variances from two independent samples. If the samples are related, the simple F test for two independent samples is not valid [11]. Related, or

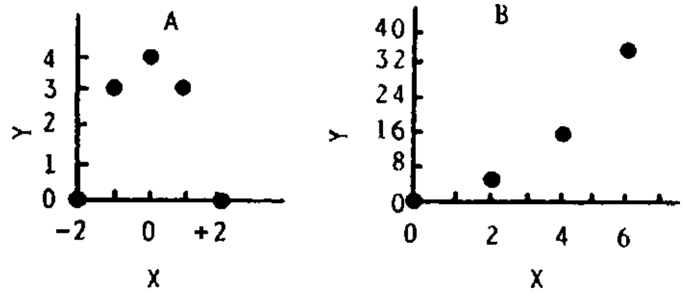


Figure 7.17 Plot of data in Table 7.12 showing problems with interpretation of the correlation coefficient.

Table 7.13 Data to Illustrate a Problem that Can Result in Misinterpretation of the Correlation Coefficient

pH	Stability, $t_{1/2}$ (wk)
2.0	48
2.1	50
1.9	50
2.0	46
2.1	47
5.5	12

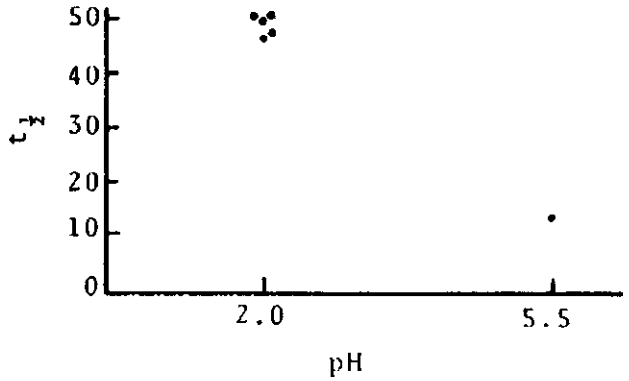


Figure 7.18 Plot of data from Table 7.10.

paired-sample tests arise, for example, in situations where the same subject tests two treatments, such as in clinical or bioavailability studies. To test for the equality of variances in related samples, we must first calculate the correlation coefficient and the F ratio of the variances. The test statistic is calculated as follows:

$$r_{ds} = \frac{F - 1}{\sqrt{(F + 1)^2 - 4r^2F}}, \tag{7.31}$$

where F is the ratio of the variances in the two samples and r is the correlation coefficient.

The ratio in Eq. (7.30), r_{ds}, can be tested for significance in the same manner as the test for the ordinary correlation coefficient, with $(N - 2)$ d.f., where N is the number of pairs [Eq. (7.29)]. As is the case for tests of the correlation coefficient, we assume a bivariate normal distribution for the related data. The following example demonstrates the calculations.

In a bioavailability study, 10 subjects were given each of two formulations of a drug substance on two occasions, with the results for AUC (area under the blood level versus time curve) given in Table 7.14.

The correlation coefficient is calculated according to Eq. (7.27).

$$r = \frac{(64{,}421)(10) - (781)(815)}{\sqrt{[(62{,}821)(10) - (781)^2][(67{,}087)(10) - (815)^2]}} = 0.699.$$

The ratio of the variances (Table 7.14), F, is

$$\frac{202.8}{73.8} = 2.75.$$

[Note: The ratio of the variances may also be calculated as 73.8/202.8 = 0.36, with the same conclusions based on Eq. (7.31).]

Table 7.14 AUC Results of the Bioavailability Study (*A* vs. *B*)

	Formulation	
Subject	***A***	***B***
1	88	88
2	64	73
3	69	86
4	94	89
5	77	80
6	85	71
7	60	70
8	105	96
9	68	84
10	<u>73</u>	<u>78</u>
Mean	78.1	81.5
S^2	202.8	73.8

The test statistic, r_{ds}, is calculated from Eq. (7.31).

$$r_{ds} = \frac{2.75 - 1}{\sqrt{(2.75 + 1)^2 - 4(0.699)^2(2.75)}} = 0.593,$$

r_{ds} is tested for significance using Eq. (7.29).

$$t_8 = \frac{\left|0.593\sqrt{8}\right|}{\sqrt{1 - 0.593^2}} = 2.08.$$

Referring to the t table (Table IV.4, 8 d.f.), a value of 2.31 is needed for significance at the 5% level. Therefore, we cannot reject the null hypothesis of equal variances in this example. Formulation A appears to be more variable, but more data would be needed to substantiate such a claim.

A discussion of correlation of multiple outcomes and adjustment of the significance level is given in section 8.2.2.

KEY TERMS

Best-fitting line
Bivariate normal distribution
Confidence band for line
Confidence interval for X and Y
Correlation
Correlation coefficient
Correlation diagram
Dependent variable
Fixed value (X)
Independence
Independent variable
Intercept
Inverse prediction
Lack of fit
Linear regression
Line through the origin
Nonlinear regression
Nonlinearity
One-sided confidence interval
Prediction interval
Reduction of sum of squares
Regression
Regression analysis
Residuals
Scatter plot
Simple linear regression
Slope
$S^2_{Y,x}$
Trend
Variance of correlated samples
Weighted regression

EXERCISES

1. A drug seems to decompose in a manner such that appearance of degradation products is linear with time (i.e., $C_d = kt$).

t	C_d
1	3
2	9
3	12
4	17
5	19

(a) Calculate the slope (k) and intercept from the least squares line.
(b) Test the significance of the slope (test vs. 0) at the 5% level.
(c) Test the slope versus 5 (H_0: $B = 5$) at the 5% level.
(d) Put 95% confidence limits on C_d at $t = 3$ and $t = 5$.
(e) Predict the value of C_d at $t = 20$. Place a 95% prediction interval on C_d at $t = 20$.
(f) If it is known that $C_d = 0$ at $t = 0$, calculate the slope.

2. A Beer's law plot is constructed by plotting ultraviolet absorbance versus concentration, with the following results:

Concentration, X	Absorbance, y	Xy
1	0.10	0.10
2	0.36	0.72
3	0.57	1.71
5	1.09	5.45
10	2.05	20.50

(a) Calculate the slope and intercept.
(b) Test to see if the intercept is different from 0 (5% level). How would you interpret a significant intercept with regard to the actual physical nature of the analytical method?
**(c) An unknown has an absorbance of 1.65. What is the concentration? Put confidence limits on the concentration (95%).

3. Five tablets were weighed and then assayed with the following results:

Weight (mg)	Potency (mg)
205	103
200	100
202	101
198	98
197	98

(a) Plot potency versus weight (weight = X). Calculate the least squares line.
(b) Predict the potency for a 200-mg tablet.
(c) Put 95% confidence limits on the potency for a 200-mg tablet.

**This is a more advanced topic.

4. Tablets were weighed and assayed with the following results:

Weight	Assay	Weight	Assay
200	10.0	198	9.9
205	10.1	200	10.0
203	10.0	190	9.6
201	10.1	205	10.2
195	9.9	207	10.2
203	10.1	210	10.3

(a) Calculate the correlation coefficient.
(b) Test the correlation coefficient versus 0 (5% level).
(c) Plot the data in the table (scatter plot).

5. Tablet dissolution was measured in vitro for 10 generic formulations. These products were also tested in vivo. Results of these studies showed the following time to 80% dissolution and time to peak (in vivo).

Formulation	Time to 80% dissolution (min)	$T_{p(\text{hr})}$
1	17	0.8
2	25	1.0
3	15	1.2
4	30	1.5
5	60	1.4
6	24	1.0
7	10	0.8
8	20	0.7
9	45	2.5
10	28	1.1

Calculate *r* and test for significance (versus 0) (5% level). Plot the data.

6. Shah et al. [14] measured the percent of product dissolved in vitro and the time to peak (in vivo) of nine phenytoin sodium products, with approximately the following results:

Product	Time to peak (hr)	Percentage dissolved in 30 min
1	6	20
2	4	60
3	2.5	100
4	4.5	80
5	5.1	35
6	5.7	35
7	3.5	80
8	5.7	38
9	3.8	85

Plot the data. Calculate the correlation coefficient and test to see if it is significantly different from 0 (5% level). (Why is the correlation coefficient negative?)

7. In a study to compare the effects of two pain-relieving drugs (*A* and *B*), 10 patients took each drug in a paired design with the following results (drug effectiveness based on a rating scale).

Patient	Drug *A*	Drug *B*
1	8	6
2	5	4
3	5	6
4	2	5
5	4	5
6	7	4
7	9	6
8	3	7
9	5	5
10	1	4

Are the drug effects equally variable?

8. Compute the intercept and slope of the least squares line for the data of Table 7.6 after a In transformation of both X and Y. Calculate the residuals and compare to the data in Table 7.7.

9. In a drug stability study, the following data were obtained:

Time (months)	Concentration (mg)
0	2.56
1	2.55
3	2.50
9	2.44
12	2.40
18	2.31
24	2.25
36	2.13

(a) Fit a least squares line to the data.
(b) Predict the time to decompose to 90% of label claim (2.25 mg).
(c) Based on a two-sided 95% confidence interval, what expiration date should be applied to this formulation?
(d) Based on a one-sided 95% confidence interval, what expiration date should be applied to this formulation?

[††]10. Fit the following data to the exponential $y = e^{ax}$. Use nonlinear least squares.

x	*y*
1	1.62
2	2.93
3	4.21
4	7.86

REFERENCES

1. Draper NR, Smith H. Applied Regression Analysis, 2nd ed. New York: Wiley, 1981.
2. Youden WJ. Statistical Methods for Chemists. New York: Wiley, 1964.
3. U.S. Food and Drug Administration. Current Good Manufacturing Practices (CGMP) 21 CFR. Washington, DC: Commissioner of the Food and Drug Administration, 2006:210–229.

[††] This is an optional, more difficult problem.

4. Davies OL, Hudson HE. Stability of drugs: accelerated storage tests. In: Buncher CR, Tsay J-Y, eds. Statistics in the Pharmaceutical Industry. New York: Marcel Dekker, 1994:445–479.
5. Tootill JPR. A critical appraisal of drug stability testing methods. J Pharm Pharmacol 1961; 13(suppl): 75T–86T.
6. Davis J. The Dating Game. Washington, DC: Food and Drug Administration, 1978.
7. Norwood TE. Statistical analysis of pharmaceutical stability data. Drug Dev Ind Pharm 1986; 12:553–560.
8. International Conference on Harmonization Bracketing and matrixing designs for stability testing of drug substances and drug products (FDA Draft Guidance) Step 2, Nov 9, 2000.
9. Nordbrock ET. Stability matrix designs. In: Chow S-C, ed. Encyclopedia of Pharmaceutical Statistics. New York: Marcel Dekker, 2000:487–492.
10. Murphy JR. Bracketing Design. In: Chow S-C, ed. Encyclopedia of Pharmaceutical Statistics. New York: Marcel Dekker, 2000:77.
11. Snedecor GW, Cochran WG. Statistical Methods, 8th ed. Ames, IA: Iowa State University Press, 1989.
12. Weisberg S. Applied Linear Regression. New York: Wiley, 1980.
13. duToit SHC, Steyn AGW, Stumpf RH. Graphical Exploratory Data Analysis. New York: Springer, 1986.
14. Shah VP, Prasad VK, Alston T, et al. In vitro in vivo correlation for 100 mg phenytoin sodium capsules. J Pharm Sci 1983; 72:306.

8 Analysis of Variance

Analysis of variance, also known as *ANOVA*, is perhaps the most powerful statistical tool. ANOVA is a general method of analyzing data from designed experiments, whose objective is to *compare two or more group means.* The *t* test is a special case of ANOVA in which only two means are compared. By *designed experiments*, we mean experiments with a particular structure. Well-designed experiments are usually optimal with respect to meeting study objectives. The statistical analysis depends on the design, and the discussion of ANOVA therefore includes common statistical designs used in pharmaceutical research. ANOVA designs can be more or less complex. The designs can be very simple, as in the case of the *t*-test procedures presented in chapter 5. Other designs can be quite complex, sometimes depending on computers for their solution and analysis. As a rule of thumb, one should use the simplest design that will achieve the experimental objectives. This is particularly applicable to experiments otherwise difficult to implement, such as is the case in clinical trials.

8.1 ONE-WAY ANOVA

An elementary approach to ANOVA may be taken using the two independent groups *t* test as an example. This is an example of one-way ANOVA, also known as a "completely randomized" design. (Certain simple "parallel-groups" designs in clinical trials correspond to the one-way ANOVA design.) In the *t* test, the two treatments are assigned at random to different independent experimental units. In a clinical study, the *t* test is appropriate when two treatments are randomly assigned to different patients. This results in two groups, each group representing one of the two treatments. One-way ANOVA is used when we wish to test the equality of treatment means in experiments where two or more treatments are randomly assigned to different, independent experimental units. The typical null hypothesis is H_0: $\mu_1 = \mu_2 = \mu_3$, where μ_1 refers to treatment 1, and so on.

Suppose that 15 tablets are available for the comparison of three assay methods, 5 tablets for each assay. The one-way ANOVA design would result from a random assignment of the tablets to the three groups. In this example, five tablets are assigned to each group. Although this allocation (five tablets per group) is optimal with regard to the precision of the comparison of the three assay methods, it is not a necessary condition for this design. The number of tablets analyzed by each analytical procedure need not be equal for the purposes of comparing the mean results. However, one can say, in general, that symmetry is a desirable feature in the design of experiments. This will become more apparent as we discuss various designs. In the one-way ANOVA, symmetry can be defined as an equal number of experimental units in each treatment group.

We will pursue the example above to illustrate the ANOVA procedure. Five replicate tablets are analyzed in each of the three assay method groups, one assay per tablet. Thus we assay the 15 tablets, five tablets by each method, as shown in Table 8.1. If only two assay methods were to be compared, we could use a *t* test to compare the means statistically. If more than two assay methods are to be compared, the correct statistical procedure to compare the means is the one-way ANOVA.

ANOVA is a technique of separating the total variability in a set of data into component parts, represented by a statistical model. In the simple case of the one-way ANOVA, the model is represented as

$$Y_{ij} = \mu + G_i + \varepsilon_{ij}, \tag{8.1}$$

Table 8.1 Results of Assays Comparing Three Analytical Methods

Method *A*	Method *B*	Method *C*
102	99	103
101	100	100
101	99	99
100	101	104
102	98	102
$\overline{X}$ 101.2	99.4	101.6
s.d. 0.84	1.14	2.07

where Y_{ij} is the jth response in treatment group i (e.g., $i = 3, j = 2$, second tablet in third group), G_i the deviation of the ith treatment (group) mean from the overall mean, μ; ε_{ij} the random error in the experiment (measurement error, biological variability, etc.) assumed to be normal with mean 0 and variance σ^2.

The model says that the response is a function of the true treatment mean ($\mu + G_i$) and a random error that is normally distributed, with mean zero and variance σ^2. In the case of a clinical study, $G_i + \mu$ is the true average of treatment i. If a patient is treated with an antihypertensive drug whose true mean effect is a 10-mm Hg reduction in blood pressure, then $Y_{ij} = 10 + \varepsilon_{ij}$, where Y_{ij} is the jth observation among patients taking the drug i. (Note that if treatments are identical, G_i is the same for all treatments.) The error, ε_{ij}, is a normally distributed variable, identically distributed for all observations. It is composed of many factors, including interindividual variation and measurement error. Thus the observed experimental values will be different for different people, a consequence of the nature of the assigned treatment and the random error, ε_{ij} (e.g., biological variation). Section 8.5 expands the concept of statistical models.

In addition to the assumption that the error is normal with mean 0 and variance σ^2, the errors must be independent. This is a very important assumption in the ANOVA model. The fact that the error has mean 0 means that some people will show positive deviations from the treatment mean, and others will show negative deviations; but on the average, the deviation is zero.

As in the t test, statistical analysis and interpretation of the ANOVA is based on the following assumptions:

1. The errors are normal with constant variance.
2. The errors (or observations) are independent.

As will be discussed below, ANOVA separates the variability of the data into parts, comparing that due to treatments to that due to error.

8.1.1 Computations and Procedure for One-Way ANOVA

ANOVA for a one-way design separates the variance into two parts, that due to *treatment differences* and that due to *error*. It can be proven that the *total sum of squares* (the squared deviations of each value from the overall mean)

$$\sum (Y_{ij} - \overline{Y})^2$$

is equal to

$$\sum (Y_{ij} - \overline{Y}_i)^2 + \sum N_i(\overline{Y}_i - \overline{Y})^2, \tag{8.2}$$

where $\overline{Y}$ is the overall mean and $\overline{Y}_i$ is the mean of the ith group. N_i is the number of observations in treatment group i. The first term in expression (8.2) is called the *within* sum of squares, and the second term is called the *between* sum of squares.

Table 8.2 Sample Data to Illustrate Eq. (8.2)

Group I (Y_{1j})	Group II (Y_{2j})	Group III (Y_{3j})
0	2	6
$\underline{2}$	$\underline{4}$	$\underline{10}$
$\overline{Y}_t$ 1	3	8
$\overline{Y} = (1 + 3 + 8)/3 = (0 + 2 + 2 + 4 + 6 + 10)/6 = 4$		

A simple example to demonstrate the equality in Eq. (8.2) is shown below, using the data of Table 8.2.

$$\sum (Y_{ij} - \overline{Y})^2 = \sum Y^2 - \frac{(\sum Y)^2}{N} = 160 - \frac{(24)^2}{6} = 64$$

$$\begin{aligned}\sum (Y_{ij} - \overline{Y}_i)^2 &= (0-1)^2 + (2-1)^2 + (2-3)^2 + (4-3)^2 + (6-8)^2 \\ &= (10-8)^2 = 2 + 2 + 8 = 12\end{aligned}$$

$$\sum N_i(\overline{Y}_i - \overline{Y})^2 = 2(1-4)^2 + 2(3-4)^2 + 2(8-4)^2 = 52.$$

Thus, according to Eq. (8.2), 64 = 12 + 52.

The calculations for the analysis make use of simple arithmetic with shortcut formulas for the computations similar to that used in the *t*-test procedures. Computer programs are available for the analysis of all kinds of ANOVA designs from the most simple to the most complex. In the latter cases, the calculations can be very extensive and tedious, and use of computers may be almost mandatory. For the one-way design, the calculations pose no difficulty. In many cases, use of a pocket calculator will result in a quicker answer than can be obtained using a less accessible computer. A description of the calculations, with examples, is presented below.

The computational process consists first of obtaining the *sum of squares* (SS) for all of the data.

$$\text{Total sum of squares (SS)} = \sum (Y_{ij} - \overline{Y})^2. \tag{8.3}$$

The *total sum of squares* is divided into two parts: (a) the SS due to treatment differences (*between-treatment sum of squares*), and (b) the error term derived from the *within-treatment sum of squares.* The within-treatment sum of squares (within SS) divided by the appropriate degrees of freedom is the *pooled variance*, the same as that obtained in the *t* test for the comparison of two treatment groups. The ratio of the between-treatment mean square to the within-treatment mean square is a measure of treatment differences (see below).

To illustrate the computations, we will use the data from Table 8.1, a comparison of three analytical methods with five replicates per method. Remember that the objective of this experiment is to compare the average results of the three methods. We might think of method *A* as the standard, accepted method, and methods *B* and *C* as modifications of the method, meant to replace method *A*. As in the other tests of hypotheses described in chapter 5, we first state the null and alternative hypotheses as well as the significance level, prior to the experiment. For example, in the present case,*

H_0: $\mu_A = \mu_B = \mu_C$ $\quad$ H_a: $\mu_i \neq \mu_j$ $\quad$ for any two means*

* Alternatives to H_0 may also include more complicated comparisons than $\mu_i \neq \mu_j$; see example, section 8.2.1.

1. First, calculate the *total sum of squares* (total SS or TSS). Calculate $\sum(Y_{ij} - \overline{Y})^2$ [Eq. (8.3)] using all of the data, ignoring the treatment grouping. This is most easily calculated using the shortcut formula

$$\sum Y^2 - \frac{(\sum Y)^2}{N}, \tag{8.4}$$

where $(\sum Y)^2$ is the grand total of all of the observations squared, divided by the total number of observations N, and is known as the *correction term, CT.* As mentioned in chapter 1, the correction term is commonly used in statistical calculations and is important in the calculation of the SS in the ANOVA.

$$\begin{aligned} \text{TSS} &= \sum Y^2 - \frac{(\sum Y)^2}{N} \\ &= (102^2 + 101^2 + \cdots + 103^2 + \cdots + 102^2) - \frac{(1511)^2}{15} \\ &= 152{,}247 - 152{,}208.07 = 38.93. \end{aligned}$$

2. The *between-treatment sum of squares* (between SS or BSS) is calculated as follows:

$$\text{BSS} = \sum \frac{{T_i}^2}{N_i} - \text{CT}, \tag{8.5}$$

where T_i is the sum of observations in treatment group i and N_i is the number of observations in treatment group i. N_i need not be the same for each group. In our example, the BSS is equal to

$$\left(\frac{506^2}{5} + \frac{497^2}{5} + \frac{508^2}{5}\right) - 152{,}208.07 = 13.73.$$

As previously noted, the *treatment* SS is a measure of treatment differences. A large SS means that the treatment differences are large. If the treatment means are identical, the treatment SS will be exactly equal to zero (0).

3. The *within-treatment sum of squares* (WSS) is equal to the difference between the TSS and BSS, that is, TSS = BSS + WSS. The WSS can also be calculated, as in the t test, by calculating $\sum(Y_{ij} - \overline{Y}_i)^2$ within each group, and pooling the results.

$$\begin{aligned} \text{WSS} &= \text{TSS} - \text{BSS} \\ &= 38.93 - 13.73 \\ &= 25.20. \end{aligned} \tag{8.6}$$

Having performed the calculations above, the SS for each "source" is set out in an "analysis of variance table," as shown in Table 8.3. The ANOVA table includes the *source, degrees of freedom,* SS, *mean square* (MS), and the *probability* based on the statistical test (F ratio).

Table 8.3 Analysis of Variance for the Data Shown in Table 8.1: Comparison of Three Analytical Methods

Source	d.f.	SS	Mean square	F
Between methods	2	13.73	6.87	$F = 3.27^a$
Within methods	12	25.20	2.10	
Total	14	38.93		

[a] $0.05 < p < 0.10$.

The degrees of freedom, noted in Table 8.3, are calculated as $N - 1$ *for the total* (N is the total number of observations); *number of treatments minus one for the treatments;* and for the *within error, subtract d.f. for treatments from the total d.f.* In our example,

Total d.f. $= 15 - 1 = 14$
Between-treatment d.f. $= 3 - 1 = 2$
Within-treatment d.f. $= 14 - 2 = 12$

Note that for the within d.f., we have 4 d.f. from each of the three groups. Thus there are 12 d.f. for the within error term. The *mean squares* are equal to the SS divided by the d.f.

Before discussing the statistical test, the reader is reminded of the assumptions underlying the ANOVA model: *independence of errors, equality of variance,* and *normally distributed errors.*

8.1.1.1 Testing the Hypothesis of Equal Treatment Means

The *mean squares are variance estimates.* One can demonstrate that the variance estimated by the treatment mean square is a sum of the within variance plus a term that is dependent on treatment differences. If the treatments are identical, the term due to treatment differences is zero, and the between mean square (BMS) will be approximately equal to the within mean square (WMS) on the average. In any given experiment, the presence of random variation will result in nonequality of the BMS and WMS terms, even though the treatments may be identical. If the null hypothesis of equal treatment means is true, the distribution of the BMS/WMS ratio is described by the *F distribution.* Note that under the null hypothesis, both WMS and BMS are estimates of σ^2, the within-group variance.

The F distribution is defined by two parameters, d.f. in the numerator and denominator of the F ratio

$$F = \frac{\text{BMS (2 d.f.)}}{\text{WMS (12 d.f.)}} = \frac{6.87}{2.10} = 3.27.$$

In our example, we have an F with 2 d.f. in the numerator and 12 d.f. in the denominator. A test of significance is made by comparing the observed F ratio to a table of the F distribution with appropriate d.f. at the specified level of significance. The F distribution is an asymmetric distribution with a long tail at large values of F, as shown in Figure 8.1. (See also sects. 3.5 and 5.3.)

To tabulate all the probability points of all F distributions would not be possible. Tables of F, similar to the t table, usually tabulate points at commonly used α levels. The cutoff points ($\alpha = 0.01, 0.05, 0.10$) for F with n_1 and n_2 d.f. (numerator and denominator) are given in Table IV.6, the probabilities in this table (1%, 5% and 10%) are in the upper tail, usually reserved for one-sided tests. This table is used to determine statistical "significance" for the ANOVA. Although the alternative hypothesis in ANOVA (H_a: at least two treatment means not equal) is two sided, the ANOVA F test (BMS/WMS) uses the upper tail of the F distribution because,

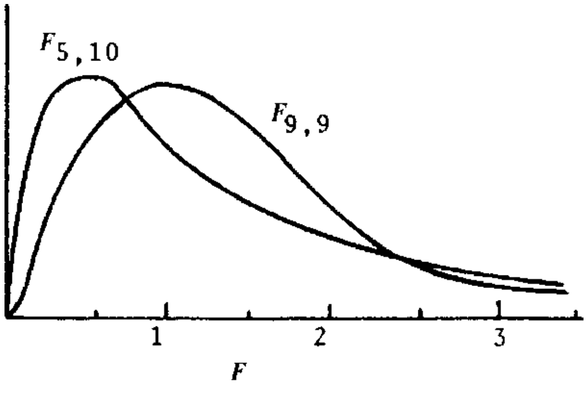


Figure 8.1 Some F distributions.

theoretically, the BMS cannot be smaller than the WMS.[†] (Thus, the F ratio will be less than 1 only due to chance variability.) The BMS is composed of the WMS plus a possible "treatment" term. Only large values of the F ratio are considered to be significant. In our example, Table 8.3 shows the F ratio to be equal to 3.27. Referring to Table IV.6, the value of F needed for significance at the 5% level is 3.89 (2 d.f. in the numerator and 12 d.f. in the denominator). Therefore, we cannot reject the hypothesis that all means are equal: method A = method B = method C ($\mu_A = \mu_B = \mu_C$).

8.1.2 Summary of Procedure for One-Way ANOVA

1. Choose experimental design and state the null hypothesis.
2. Define the α level.
3. Choose samples, perform the experiment, and obtain data.
4. Calculate the TSS and BSS.
5. Calculate the within SS as the difference between the TSS and the BSS.
6. Construct an ANOVA table with mean squares.
7. Calculate the F statistic (BMS/WMS).
8. Refer the F ratio statistic to Table IV.6 (n_1 and n_2 d.f., where n_1 is the d.f. for the BMS and n_2 is the d.f. for the WMS).
9. If the calculated F is equal to or greater than the table value for F at the specified α level of significance, at least two of the treatments can be said to differ.

8.1.3 A Common But Incorrect Analysis of the Comparison of Means from More Than Two Groups

In the example in section 8.1.1, if more than two assay methods are to be compared, the correct statistical procedure is a one-way ANOVA. A common error made by those persons not familiar with ANOVA is to perform three separate t tests on such data: comparing method A to method B, method A to method C, and method B to method C. This would require three analyses and "decisions," which can result in apparent contradictions. For example, decision statements based on three separate analyses could read

Method A gives higher results than method $B (p < 0.05)$.
Method A is not significantly different from method $C (p > 0.05)$.
Method B is not significantly different from method $C (p > 0.05)$.

These are the conclusions one would arrive at if separate t tests were performed on the data in Table 8.1 (see Exercise Problem 1). One may correctly question: If A is larger than B, and C is slightly larger than A, how can C not be larger than B? The reasons for such apparent contradictions are (a) the use of different variances for the different comparisons, and (b) performing three tests of significance on the same set of data. ANOVA obviates such ambiguities by using a common variance for the single test of significance (the F test).[‡] The question of multiple comparisons (i.e., multiple tests of significance) is addressed in the following section.

8.2 PLANNED VERSUS A POSTERIORI (UNPLANNED) COMPARISONS IN ANOVA

Often, in an experiment involving more than two treatments, more specific hypotheses than the global hypotheses, $\mu_1 = \mu_2 = \mu_3 = \ldots$, are proposed in advance of the experiment. These are known as *a priori* or *planned comparisons*. For example, in our example of the three analytical methods, if method A is the standard method, we may have been interested in a comparison of each of the two new methods, B and C, with A (i.e., H_0: $\mu_A = \mu_C$ and $\mu_A = \mu_B$). We may proceed to make these comparisons at the conclusion of the experiment using the usual t-test

[†] This may be clearer if one thinks of the null and alternative hypotheses in ANOVA as H_a: $\sigma_B{}^2 = \sigma_w{}^2$; H_a: $\sigma_B{}^2 > \sigma_w{}^2$.

[‡] We have assumed in the previous discussion that the variances in the different treatment groups are the same. If the numbers of observations in each group are equal, the ANOVA will be close to correct in the case of moderate variance heterogeneity. If in doubt, a test to compare variances may be performed (see sect. 5.3).

procedure with the following proviso: *The estimate of the variance is obtained from the ANOVA, the pooled within mean square term.* This estimate comes from all the groups, not only the two groups being compared. ANOVA procedures, like the t test, assume that the variances are equal in the groups being tested.[‡] Therefore, the *within mean square* is the best estimate of the common variance. In addition, the increased d.f. resulting from this estimate results in increased precision and power (chap. 7) of the comparisons. A smaller value of t is needed to show "significance" compared to the t test, which uses only the data from a specific comparison, in general. Tests of only those comparisons planned a priori should be made using this procedure. This means that the α level (e.g., 5%) applies to each comparison.

Indiscriminate comparisons made after the data have been collected, such as looking for the largest differences as suggested by the data, will always result in more significant differences than those suggested by the stated level of significance. We shall see in section 8.2.1 that *a posteriori* tests (i.e., unplanned tests made after data have been collected) can be made. However, a "penalty" is imposed that makes it more difficult to find "significant" differences. This keeps the "experiment-wise" α level at the stated value (e.g., 5%). (For a further explanation, see sect. 8.2.1.) The statistical tests for the two planned comparisons as described above are performed as follows (a two independent groups t test with WMS equal to error, the pooled variance)

$$\text{Method } B \text{ versus method } A\text{:}\frac{|99.4 - 101.2|}{\sqrt{2.1(1/5 + 1/5)}} = 1.96.$$

$$\text{Method } C \text{ versus method } A\text{:}\frac{|101.6 - 101.2|}{\sqrt{2.1(1/5 + 1/5)}} = 0.44.$$

Since the t value needed for significance at the 5% level (d.f. = 12) is 2.18 (Table IV.4), neither of the comparisons noted previously is significant. However, when reporting such results, a researcher should be sure to include the actual averages. A confidence interval for the difference may also be appropriate. The confidence interval is calculated as described previously [Eq. (5.2)]; but remember to use the WMS for the variance estimate (12 d.f.). Also, the fact that methods A and B are not significantly different does not mean that they are the same. If one were looking to replace method A, other things being equal, method C would be the most likely choice.

If the comparison of methods B and C had been planned in advance, the t test would show a significant difference at the 5% level (see Exercise Problem 3). However, it would be unfair to decide to make such a comparison using the t-test procedure described above only after having seen the results. Now, it should be more clear why the ANOVA results in different conclusions from that resulting from the comparison of all pairs of treatments using separate t tests.

1. *The variance is pooled from all of the treatments.* Thus, it is the pooled variance from all treatments that is used as the error estimate. When performing separate t tests, the variance estimate differs depending on which pair of treatments is being compared. The pooled variance for the ordinary t test uses only the data from the specific two groups that are being compared. The estimates of the variance for each separate t test differ due to chance variability. That is, although an assumption in ANOVA procedures is that the variance is the same in all treatment groups, the *observed* sample variances will be different in different treatment groups because of the variable nature of the observations. This is what we have observed in our example. By chance, the variability for methods A and B was smaller than that for method C. Therefore, when performing individual t tests, a smaller difference of means is necessary to obtain significance when comparing methods A and B than that needed for the comparison of methods A and C, or methods B and C. Also, the d.f. for the t tests are 8 for the separate tests, compared to 12 when the pooled variance from the ANOVA is used. In

[‡] We have assumed in the previous discussion that the variances in the different treatment groups are the same. If the numbers of observations in each group are equal, the ANOVA will be close to correct in the case of moderate variance heterogeneity. If in doubt, a test to compare variances may be performed (see sect. 5.3).

conclusion, we obtain different results because we used different variance estimates for the different tests, which can result in ambiguous and conflicting conclusions.

2. The F test in the ANOVA takes into account the number of treatments being compared. An α level of 5% means that if all treatments are identical, 1 in 20 experiments (on the average) will show a significant F ratio. That is, the risk of erroneously observing a significant F is 1 in 20. If separate t tests are performed, each at the 5% level, for each pair of treatments (three in our example), the chances of finding at least one pair of treatments different in a given experiment will be greater than 5%, when the treatments are, in fact, identical. *We should differentiate between the two situations* (a) where we plan, a priori, specific comparisons of interest, and (b) where we make tests a posteriori suggested by the data. In case (a), each test is done at the α level, and each test has an α percent chance of being rejected if treatments are the same. In case (b), having seen the data we are apt to choose only those differences that are large. In this case, experiments will reveal differences where none truly exist much more than α percent of the time.

Multiple testing of data from the same experiment results in a higher significance level than the stated α level *on an experiment-wise basis.* This concept may be made more clear if we consider an experiment in which five assay methods are compared. If we perform a significance (t) test comparing each pair of treatments, there will be 10 tests, $(n)(n-1)/2$, where n is the number of treatments: $5(4)/2 = 10$ in this example. To construct and calculate 10 t tests is a rather tedious procedure. If treatments are identical and each t test is performed at the 5% level, the probability of finding at least one significant difference in an experiment will be much more than 5%. Thus the probability is very high that at the completion of such an experiment, this testing will lead to the conclusion that *at least* two methods are different. If we perform 10 separate t tests, the α level, on an experiment-wise basis, would be approximately 40%; that is, 40% of experiments analyzed in this way would show at least one significant difference, when none truly exists [1].

The Bonferroni method is often used to control the alpha level for multiple comparisons. For an overall level of alpha, the level is set at α/k for each test, where k is the number of comparisons planned. For the data of Table 8.1, for a test of two planned comparisons at an overall level of 0.05, each would be performed at the $0.05/2 = 0.025$ level. If the tests consisted of comparisons of the means (A vs. C) and (A vs. B), t tests could be performed. A more detailed t table than IV.4 would be needed to identify the critical value of t for a two-sided test at the 0.025 level with 12 d.f. This value lies between the tabled values for the 0.05 and 0.01 level and is equal to 2.56. The difference needed for significance at the 0.025 level is

$$2.56 \times \sqrt{2.1 \times \frac{2}{5}} = 2.35.$$

Since the absolute differences for the two comparisons (A vs. C) and (A vs. B) are 0.4 and 1.8, respectively, neither difference is statistically significant.

In the case of preplanned comparisons, significance may be found even if the F test in the ANOVA is not significant. This procedure is considered acceptable by many statisticians. Comparisons made after seeing the data that were not preplanned fall into the category of a *posteriori multiple* comparisons. Many such procedures have been recommended and are commonly used. Several frequently used methods are presented in the following section.

8.2.1 Multiple Comparisons in ANOVA

The discussion above presented compelling reasons to avoid the practice of using many separate t tests when analyzing data where more than two treatments are compared. On the other hand, for the null hypothesis of no treatment differences, a significant F in the ANOVA does not immediately reveal which of the multiple treatments tested differ. Sometimes, with a small number of treatments, inspection of the treatment means is sufficient to show obvious differences. Often, differences are not obvious. Table 8.4 shows the average results and ANOVA table for four drugs with regard to their effect on the reduction of pain, where the data are derived from subjective pain scores (see also Fig. 8.2). The null hypothesis is H_0: $\mu_A = \mu_B = \mu_C = \mu_D$. The alternative hypothesis here is that at least two treatment means differ. The α level is set at

Table 8.4 Average Results and ANOVA for Four Analgesic Drugs

	Reduction in pain with drugs			
	A	*B*	*C*	*D*
$\overline{X}$	4.5	5.7	7.1	6.3
S^2	3.0	4.0	4.5	3.8
S	1.73	2.0	2.12	1.95
N	10	10	10	10
	ANOVA			
Source	**d.f.**	**SS**	**Mean square**	***F***
Between drugs	3	36	12	$F_{3,36} = 3.14^a$
Within drugs	36	137.7	3.83	
Total	39	173.7		

$^a p < 0.05$.

5%. Ten patients were assigned to each of the four treatment groups. The *F* test with 3 and 36 d.f. is significant at the 5% level. An important question that we wish to address here is: Which treatments are different? Are all treatments different from one another, or are some treatments not significantly different? This problem may be solved using "multiple comparison" procedures. The many proposals that address this question result in similar but not identical solutions. Each method has its merits and deficiencies. We will present some approaches commonly used for performing a posteriori comparisons. Using these methods, we can test differences specified by the alternative hypothesis, as well as differences suggested by the final experimental data. These methods will be discussed with regard to comparing individual treatment means. Some of these methods can be used to compare any linear combination of the treatment means, such as the mean of drug *A* versus the average of the means for drugs *B*, *C*, and *D* [$\overline{A}$ vs. $(\overline{B} + \overline{C} + \overline{D})/3$].

For a further discussion of this problem, see the Scheffé method below.

8.2.1.1 Least Significant Difference

The method of "least significant difference" (LSD) proposed by R. A. Fisher, is the simplest approach to a posteriori comparisons. This test is a simple *t* test comparing all possible pairs of treatment means. (Note that this approach is not based on preplanned comparisons, discussed in the previous section.) However, the LSD method results in more significant differences than would be expected according to the α level. Because of this, many statisticians do not recommend

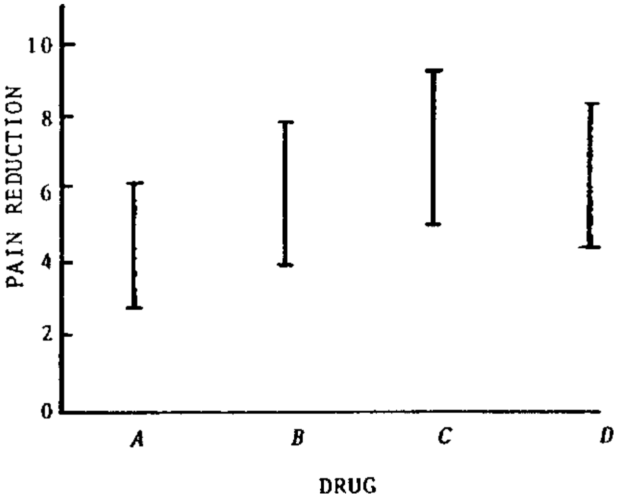


Figure 8.2 Result of pain reduction (± standard deviation) for four drugs with 10 patients per treatment group.

its use. The LSD test differs from the indiscriminate use of multiple t tests in that one proceeds (a) *only if the F test in the ANOVA is significant*, and (b) *the pooled (within mean square) variance is used as the variance estimate* in the t-test procedure. The LSD approach is illustrated using the data from Table 8.4.

$$\text{Since } t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{S^2(1/N_1 + 1/N_2)}},$$

$$\text{LSD} = t\sqrt{S^2\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}. \tag{8.7}$$

If the sample sizes are equal in each group ($N_1 = N_2 = N$),

$$\text{LSD} = t\sqrt{\frac{2s^2}{N}}, \tag{8.8}$$

where S^2 is the within mean square variance and t is the tabulated value of t at the α level, with appropriate degrees of freedom (d.f. = the number of degrees of freedom from the WMS of the ANOVA table). Any difference of two means that is equal to or exceeds the LSD is significant at the α level. From Table IV.4, the value of t at the 5% level with 36 d.f. is 2.03. The variance (from the ANOVA in Table 8.4) is 3.83. Therefore, the LSD is

$$\text{LSD} = 2.03\sqrt{\frac{2(3.83)}{10}} = 1.78.$$

The average pain reductions for drugs C and D are significantly greater than that for drug A ($\overline{C} - \overline{A} = 2.6$; $\overline{D} - \overline{A} = 1.8$).

Note that in the example shown in Table 8.1 (ANOVA table in Table 8.3), the F test is not significant. Therefore, one would not use the LSD procedure to compare the methods, after seeing the experimental results. If a comparison had been planned a priori, the LSD test could be correctly applied to the comparison.

8.2.1.2 Tukey's Multiple Range Test

Tukey's multiple range test is a commonly used multiple comparison test based on keeping the error rate at α (e.g., 5%) from an "experiment-wise" viewpoint. By "experiment-wise" we mean that if no treatment differences exist, the probability of finding at least one significant difference for a posteriori tests in a given experiment is α (e.g., 5%). This test is more conservative than the LSD test. This means that a larger difference between treatments is needed for significance in the Tukey test than in the LSD test. On the other hand, although the experiment-wise error is underestimated using the LSD test, the LSD test is apt to find real differences more often than will the Tukey multiple range test. (The LSD test has greater power.) Note that a trade-off exists. The easier it is to obtain significance, the greater the chance of mistakenly calling treatments different (α error), but the less chance of missing real differences (β error). The balance between these risks depends on the costs of errors in each individual situation. (See chap. 6 for a further discussion of these risks.)

In the multiple range test, treatments can be compared without the need for a prior significant F test. However, the ANOVA should always be carried out. The error term for the treatment comparisons comes from the ANOVA, the within mean square in the one-way ANOVA. Similar to the LSD procedure, a least significant difference can be calculated. Any difference of treatment means exceeding

$$Q\sqrt{\frac{S^2}{N}} \tag{8.9}$$

is significant. S^2 *is the "error" variance from the ANOVA* (within mean square for the one-way ANOVA) and N is *the sample size.* This test is based on equal sample sizes in each group. If the sample sizes are not equal in the two groups to be compared, an approximate method may be used with N replaced by $2N_1N_2/(N_1 + N_2)$, where N_1 and N_2 are the sample sizes of the two groups. Q is the value of the "studentized range" found in Table IV.7A, a short table of Q at the 5% level. More extensive tables of Q may be found in Ref. [2] (Table A-18a). The value of Q depends on the *number of means being tested* (the number of treatments in the ANOVA design) and the *d.f. for error* (again, the within mean square d.f. in the one-way ANOVA). In the example of Table 8.4, the number of treatments is 4, and the d.f. for error are 36. From Table IV.7, the value of Q is approximately 3.81. Any difference of means greater than

$$3.81\sqrt{\frac{3.83}{10}} = 2.36$$

is significant at the 5% level. Therefore, this test finds only drugs A and C to be significantly different.

This test is more conservative than the LSD test. However, one must understand that the multiple range test tries to keep the error rate at α on an experiment-wise basis. In the LSD test, the error rate is greater than α for each experiment.

8.2.1.3 Scheffé Method

The Tukey method should be used if we are only interested in the comparison of treatment means after having seen the data. However, for more complicated comparisons (also known as *contrasts*) for a large number of treatments, the Scheffé method will often result in shorter intervals needed for significance. As in the Tukey method, the Scheffé method is meant to keep the α error rate at 5%, for example, on an experiment-wise basis. For the comparison of two means, the following statistic is computed:

$$\sqrt{S^2(k-1)F\left(\frac{1}{N_1}+\frac{1}{N_2}\right)}, \tag{8.10}$$

where S^2 *is the appropriate variance estimate* (WMS for the one-way ANOVA), k *is the number of treatments* in the ANOVA design, and N_1 and N_2 are the sample sizes of the two groups being compared. F is the table value of F (at the appropriate level) with d.f. of $(k - 1)$ in the numerator, and d.f. in the denominator equal to that of the error term in the ANOVA. Any difference of means equal to or greater than the value computed from expression (8.10) is significant at the α level. Applying this method to the data of Table 8.4 results in the following $[S^2 = 3.83, k = 4, F(3{,}36\,\text{d.f.}) = 2.86,\ N_1 = N_2 = 10]$:

$$\sqrt{3.83(3)(2.86)(1/10 + 1/10)} = 2.56.$$

Using this method, treatments A and C are significantly different. This conclusion is the same as that obtained using the Tukey method. However, treatments A and C barely make the 5% level; the difference needed for significance in the Scheffé method is greater than that needed for the Tukey method for this simple comparison of means. However, one should appreciate that the Scheffé method can be applied to more complicated contrasts with suitable modification of Eq. (8.10).

Suppose that drug A is a control or standard drug, and drugs B and C are homologous experimental drugs. Conceivably, one may be interested in comparing the results of the average of drugs B and C to drug A. From Table 8.4, the average of the means of drugs B and C is

$$\frac{5.7 + 7.1}{2} = 6.4.$$

For tests of significance of comparisons (contrasts) for the general case, Eq. (8.10) may be written as

$$\sqrt{(k-1)F\,V(\text{contrast})}, \tag{8.11}$$

where $(k-1)$ and (F) are the same as in Eq. (8.10), and V(contrast) is the variance estimate of the contrast. Here the contrast is

$$\frac{\overline{X}_B + \overline{X}_C}{2} - \overline{X}_A.$$

The variance of this contrast is (see also App. I)

$$\frac{S^2/N_B + S^2/N_C}{4} + \frac{S^2}{N_A} = S^2\left(\frac{1}{20} + \frac{1}{10}\right) = \frac{3S^2}{20}.$$

(Note that $N_A = N_B = N_C = 10$ in this example.) From Eq. (8.11), a difference of $(\overline{X}_B + \overline{X}_C)/2 - \overline{X}_A$ exceeding

$$\sqrt{3(2.86)(3.83)\frac{3}{20}} = 2.22$$

will be significant at the 5% level. The observed difference is

$$6.4 - 4.5 = 1.9.$$

Since the observed difference does not exceed 2.22, the difference between the average results of drugs B and C versus drug A is not significant ($p > 0.05$). For a further discussion of this more advanced topic, the reader is referred to Ref. [3].

8.2.1.4 Newman–Keuls Test

The Newman–Keuls test uses the multiple range factor Q (see Tukey's Multiple Range Test) in a sequential fashion. In this test, the means to be compared are first arranged in order of magnitude. For the data of Table 8.4, the means are 4.5, 5.7, 6.3, and 7.1 for treatments A, B, D, and C, respectively.

To apply the test, compute the difference needed for significance for the comparison of 2, 3, ..., n means (where n is the total number of treatment means). In this example, the experiment consists of 4 treatments. Therefore, we will obtain differences needed for significance for 2, 3, and 4 means.

Initially, consider the first two means using the Q test

$$Q\sqrt{S^2/N}. \tag{8.12}$$

From Table IV.7, with 2 treatments and 36 d.f. for error, Q = approximately 2.87. From Eq. (8.12),

$$Q\sqrt{S^2/N} = 2.87\sqrt{3.83/10} = 1.78.$$

For 3 means, find Q from Table IV.7 for $k = 3$

$$3.45\sqrt{3.83/10} = 2.14.$$

For 4 means, find Q from Table IV.7 for $k = 4$

$$3.81\sqrt{3.83/10} = 2.36.$$

Note that the last value, 2.36, is the same value as that obtained for the Tukey test.

Thus, the differences needed for 2, 3, and 4 means to be considered significantly different are 1.78, 2.14, and 2.36. This can be represented as follows

Number of treatments	2	3	4
Critical difference	1.78	2.14	2.36

The four ordered means are

$$\begin{array}{cccc} A & B & D & C \\ 4.5 & 5.7 & 6.3 & 7.1 \\ \cline{1-3} \\[-2ex] \cline{2-4} \end{array}$$

The above notation is standard. Any two means connected by the same underscored line are not significantly different. Two means not connected by the underscored line are significantly different. Examination of the two underscored lines in this example shows that the only two means not connected are 4.5 and 7.1, corresponding to treatments A and C, respectively.

The determination of significant and nonsignificant differences follows. The difference between treatments A and C, covering 4 means, is equal to 2.6, which exceeds 2.36, resulting in a significant difference. The difference between treatments A and D is 1.8, which is less than the critical value of 2.14 for 3 means. This is described by the first underscore. (Note that we need not compare A and B or B and D since these will not be considered different based on the first underscore.) Treatments B, D, and C are considered to be not significantly different because the difference between B and C, encompassing 3 treatment means, is 1.4, which is less than 2.14. Therefore, a second underscore includes treatments B, D, and C.

8.2.1.5 Dunnett's Test

Sometimes experiments are designed to compare several treatments against a control but not among each other. For the data of Table 8.4, treatment A may have been a placebo treatment, whereas treatments B, C, and D are three different active treatments. The comparisons of interest are A versus B, A versus C, and A versus D. Dunnett [4,5] devised a multiple comparison procedure for treatments versus a control. The critical difference,D', for a two-sided test for any of the comparisons versus control is defined as

$$D' = t'\sqrt{S^2\left(\frac{1}{N_1}+\frac{1}{N_2}\right)},$$

where t' is obtained from Table IV.7B.

In the present example, p, the number of treatments to be compared to the control, is equal to 3, and d.f. = 36. For a two-sided test at the 0.05 level, the value of t' is 2.48 from Table IV.7B. Therefore the critical difference is

$$2.48\sqrt{3.83\left(\frac{1}{10}+\frac{1}{10}\right)} = 2.17.$$

Again, the only treatment with a difference from treatment A greater than 2.17 is treatment C. Therefore, only treatment C can be shown to be significantly different from treatment A, the control.

Those readers interested in further pursuing the topic of multiple comparisons are referred to Ref. [4].

8.2.2 Multiple Correlated Outcomes[§]

Many clinical studies have a multitude of endpoints that are evaluated to determine efficacy. Studies of antiarthritic drugs, antidepressants, and heart disease, for example, may consist of a measure of multiple outcomes. In a comparative study, if each measured outcome is evaluated

[§] A more advanced topic.

independently, the probability of finding a significant effect when the drugs are not different, for at least one outcome, is greater than the alpha level for the study. In addition, these outcomes are usually correlated. For example, relief of gastrointestinal distress and bloating may be highly correlated when evaluating treatment of gastrointestinal symptoms. If all the measures are independent, Bonferroni's inequality may be used to determine the significance level. For example, for five independent measures and a level of 0.01 for each measure, separate analyses of each measure will yield an overall alpha level of approximately 5% for the experiment as a whole (see sect. 8.2). However, if the measures are correlated, the Bonferroni adjustment is too conservative, making it more difficult to obtain significance. The other extreme is when all the outcome variables are perfectly correlated. In this case, one alpha level (e.g., 5%) will apply to all the variables. (One need test for only one of the variables; all other variables will share the same conclusion.) Dubey [6] has presented an approach to adjusting the alpha (α) level for multiple correlated outcomes. If we calculate the Bonferroni adjustment as $1 - (1 - \gamma)^k$ where k is the number of outcomes and γ is the level for testing each outcome, then the adjusted level for each outcome will lie between α (perfect correlation) and approximately α/k (no correlation). The problem can be formulated as

$$\alpha = \text{overall level of significance} = 1 - (1 - \gamma)^m, \tag{8.13}$$

where m lies between 1 and k, k being the number of outcome variables. If there is perfect correlation among all of the variables, $m = 1$, the level for each variable, γ is equal to α. For zero correlation, $m = k$, resulting in the Bonferroni adjustment. Dubey defines

$$m = k^{1-R_i}, \tag{8.14}$$

where R_i is an "average" correlation.

$$R_i = \sum_{i \neq j} \frac{R_{ij}}{k-1}. \tag{8.15}$$

This calculation will be clarified in the example following this paragraph.

To obtain the alpha level for testing each outcome, γ, use Eq. (8.16) that is derived from Eq. (8.13) by solving for γ.

$$\gamma = 1 - (1 - \alpha)^{1/m} \tag{8.16}$$

The following example shows the calculation.

Suppose five variables are defined for the outcome of a study comparing an active and placebo for the treatment of heart disease: (1) trouble breathing, (2) pains in chest, (3) numbing/tingling, (4) rapid pulse, and (5) indigestion. The overall level for significance is set at 0.05. First, form the correlation matrix for the five variables. Table 8.5 is an example of such a matrix.

This matrix is interpreted for example, as the correlation between numbing/tingling and rapid pulse being 0.41 (variables 3 and 4), etc.

Table 8.5 Correlation Matrix for five Variables Measuring Heart "Disease"

	Variable				
	1	**2**	**3**	**4**	**5**
1	1.00	0.74	0.68	0.33	0.40
2	0.74	1.00	0.25	0.66	0.85
3	0.68	0.25	1.00	0.41	0.33
4	0.33	0.66	0.41	1.00	0.42
5	0.40	0.85	0.33	0.42	1.00

Calculate the "average" correlation, r_i from Eq. (8.15).

$$r_1 = \frac{\{0.74 + 0.68 + 0.33 + 0.40)}{4} = 0.538$$
$$r_2 = \frac{\{0.74 + 0.25 + 0.66 + 0.85)}{4} = 0.625$$
$$r_3 = \frac{\{0.68 + 0.25 + 0.41 + 0.33)}{4} = 0.425$$
$$r_4 = \frac{\{0.33 + 0.66 + 0.41 + 0.42)}{4} = 0.455$$
$$r_5 = \frac{\{0.40 + 0.85 + 0.33 + 0.42)}{4} = 0.500$$

The average correlation is

$$\frac{0.538 + 0.625 + 0.425 + 0.455 + 0.500}{5} = 0.509.$$

From Eq. (8.14),

$$m = k^{1-0.509} = 5^{0.491} = 2.203.$$

From Eq. (8.16), the level for each variable is adjusted to $\gamma = 1 - (1 - \alpha)^{1/m} = 1 - (1 - 0.05)^{1/2.203} = 0.023$.

Therefore, testing of the individual outcome variables should be performed at the 0.023 level.

Equation (8.13) can also be used to estimate the sample size of a study with multiple endpoints. Comelli [7] gives an example of a study with eight endpoints, and an estimated average correlation of 0.7. First, solve for γ, where alpha = 0.05 and R_i is 0.7.

$$\gamma = 1 - (1 - \alpha)^{1/m} = 0.05, \quad \text{where } m = 8^{(1-0.7)}.$$

γ is equal to 0.027. The sample size can then be computed by standard methods (see chap. 6). For the sample size calculation, use an alpha of 0.027 with desired power, and with the endpoint that is likely to show the smallest standardized treatment difference. For example, in a parallel design, suppose we wish to have a power of 0.8, and the endpoint with the smallest standardized difference is 0.5/1 (difference/standard deviation). Using Eq. (6.6), $N = 2(1/0.5)^2(2.21 + 0.84)^2 + 2 = 77$ per group.

8.3 ANOTHER EXAMPLE OF ONE-WAY ANOVA: UNEQUAL SAMPLE SIZES AND THE FIXED AND RANDOM MODELS

Before leaving the topic of one-way ANOVA, we will describe an example in which the sample sizes of the treatment groups are not equal. We will also introduce the notion of "fixed" and "random" models in ANOVA.

Table 8.6 shows the results of an experiment comparing tablet dissolution as performed by five laboratories. Each laboratory determined the dissolution of tablets from the same batch of a standard product. Because of a misunderstanding, one laboratory (D) tested 12 tablets, whereas the other four laboratories tested six tablets. The null hypothesis is

$$H_0\colon \mu_A = \mu_B = \mu_C = \mu_D = \mu_E,$$

and

$$H_a\colon \mu_i \neq \mu_j \quad \text{for at least two means.}$$

Table 8.6 Percent Dissolution After 15 Minutes for Tablets from a Single Batch Tested in Five Laboratories

	Laboratory				
	A	*B*	*C*	*D*	*E*
	68	55	78	75	65
	78	62	63	60	60
	63	67	78	66	66
	56	60	65	69	75
	61	67	70	58	75
	69	73	74	64	70
				71	
				71	
				65	
				77	
				60	
	—	—	—	63	—
Total	395	384	428	799	411
$\overline{X}$	65.8	64.0	71.3	66.6	68.5
s.d.	7.6	6.3	6.4	6.1	6.0

The ANOVA calculations are performed in an identical manner to that shown in the previous example (sect. 8.1.1). The ANOVA table is shown in Table 8.7. The *F* test for laboratories (4, 31 d.f.) is 1.15, which is *not* significant at the 5% level (Table IV.6). Therefore, the null hypothesis that the laboratories obtain the same average result for dissolution cannot be rejected.

$$\sum X = 2417 \quad \sum X^2 = 163{,}747 \quad N = 36$$

$$\text{TSS} = \sum X^2 - \frac{(\sum X)^2}{N} = 1472.306.$$

$$\text{Between Lab SS} = \frac{(395)^2}{6} + \frac{(384)^2}{6} + \frac{(428)^2}{6} + \frac{(799)^2}{12} + \frac{(411)^2}{6} - \frac{(2417)^2}{36} = 189.726.$$

$$\text{Within lab SS} = \text{TSS} - \text{BSS} = 1472.306 - 189.726 = 1282.58.$$

One should always question the validity of ANOVA assumptions. In particular, the assumption of independence may be suspect in this example.

Are tablets tested in sets of six, or is each tablet tested separately? If tablets are tested one at a time in separate runs, the results are probably independent. However, if six tablets are tested at one time, it is possible that the dissolution times may be related due to particular conditions that exist during the experiment. For example, variable temperature setting and mixing speed would affect all six tablets in the same (or similar) way. A knowledge of the particular experimental system and apparatus, and/or experimental investigation, is needed

Table 8.7 Analysis of Variance Table for the Data in Table 8.6 for Tablet Dissolution

Source	d.f.	SS	Mean square	*F*
Between labs	4	189.726	47.43	$F_{4,31} = 1.15$
Within labs	31	1282.58	41.37	
Total	35	1472.306		

to assess the possible dependence in such experiments. The assumption of equality of variance seems to be no problem in this experiment (see the standard deviations in Table 8.6).

8.3.1 Fixed and Random Models

In this example, the interpretation (and possible further analysis) of the experimental results depends on the nature of the laboratories participating in the experiment. The laboratories can be considered to be

1. the only laboratories of interest with respect to dissolution testing; for example, perhaps the laboratories include only those that have had trouble performing the procedure;
2. a random sampling of five laboratories, selected to determine the reproducibility (variability) of the method when performed at different locations.

The former situation is known as a *fixed model.* Inferences based on the results apply only to those laboratories included in the experiment. The latter situation is known as a *random model.* The random selection of laboratories suggests that the five laboratories are a sample chosen among many possible laboratories. Thus, inferences based on these results can be applied to all laboratories in the population of laboratories being sampled.

One way of differentiating a fixed and random model is to consider which treatment groups (laboratories) would be included if the experiment were to be run again. If the same groups would always be chosen in these perhaps hypothetical subsequent experiments, then the groups are fixed. If the new experiment includes different groups, the groups are random.

The statistical test of the hypothesis of equal means among the five laboratories is the same for both situations, *fixed* and *random.* However, in the random case, one may also be interested in estimating the variance. The estimates of the *within-laboratory* and *between-laboratory* variance are important in defining the reproducibility of the method. This concept is discussed further in section 12.4.1.

8.4 TWO-WAY ANOVA (RANDOMIZED BLOCKS)

As the one-way ANOVA is an extension of the two independent groups t test when an experiment contains more than two treatments, *two-way ANOVA* is an extension of the paired t test to more than two treatments. The two-way design, which we will consider here, is known as a randomized block design (the nomenclature in statistical designs is often a carryover based on the original application of statistical designs in agricultural experiments). In this design, treatments are assigned at random to each experimental unit or "block." (In clinical trials, where a patient represents a block, each patient receives each of the two or more treatments to be tested in random order.)

The randomized block design is advantageous when the levels of response of the different experimental units are very different. The statistical analysis separates these differences from the experimental error, resulting in a more precise (sensitive) experiment. For example, in the paired t test, taking differences of the two treatments should result in increased precision if the experimental units receiving the treatments are very different from each other, but they differentiate the treatments similarly. In Figure 8.3, the three patients are very different in their levels of response (blood pressure). However, each patient shows a similar difference between drugs A and B ($A > B$). In a two independent groups design (parallel groups), the experimental error is estimated from differences among experimental units within treatments. This is usually larger than the experimental error in a corresponding two-way design.

Another example of a two-way (randomized block) design is the comparison of analytical methods using product from different batches. The design is depicted in Table 8.8. If the batches have a variable potency, a rational approach is to run each assay method on material from each batch. The statistical analysis will separate the variation due to different batches from the other random error. The experimental error is free of batch differences, and will be smaller than that obtained from a one-way design using the same experimental material (product from different batches). In the latter case, material would be assigned to each analytical method at random.

A popular type of two-way design that deserves mention is that which includes pretreatment or baseline readings. This design, a *repeated measures* design, often consists of pretreatment readings followed by treatment and post-treatment readings observed over time. Repeated

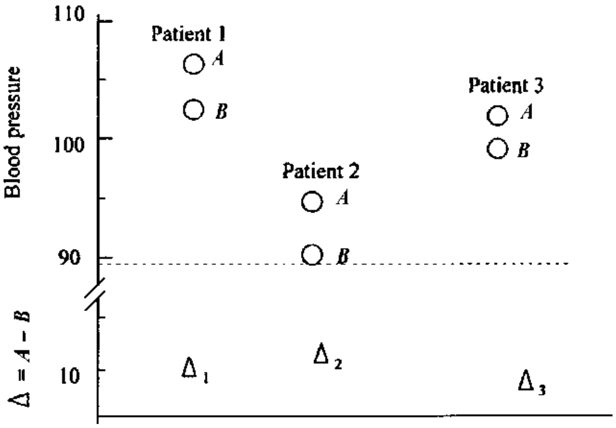


Figure 8.3 Increased precision in two-way designs.

measure designs are discussed further in chapter 11. In these designs, order (order is *time* in these examples) cannot be randomized. One should be careful to avoid bias in situations where a concomitant control is not part of these experiments. For example, suppose that it is of interest to determine if a drug causes a change in a clinical effect. One possible approach is to observe pretreatment (baseline) and post-treatment measurements, and to perform a statistical test (a paired *t* test) on the "change from baseline." Such an experiment lacks an adequate control group and interpretation of the results may be difficult. For example, any observed change or lack of change could be dependent on the time of observation, when different environmental conditions exist, in addition to any possible drug effect. A better experiment would include a *parallel* group taking a control product: a placebo or an active drug (positive control). The difference between change from baseline in the placebo group and test drug would be an unbiased estimate of the drug effect.

8.4.1 A Comparison of Dissolution of Various Tablet Formulations: Random and Fixed Models in Two-Way ANOVA

Eight laboratories were requested to participate in an experiment whose objective was to compare the dissolution rates of two generic products and a standard drug product. The purpose of the experiment was to determine (a) if the products had different rates of dissolution, and (b) to estimate the laboratory variability (differences) and/or test for significant differences among laboratories. If the laboratory differences are large, the residual or error SS will be substantially reduced compared to the corresponding error in the one-way design. If interaction is absent, we will be using the "within-laboratory" variability to test for differences among the products (see sect. 8.4.1.2). The laboratory SS and the product SS in the ANOVA are computed in a manner similar to the calculations in the one-way design. The residual SS is calculated as the total sum of squares (TSS) minus the laboratory and product SS. (The laboratory and product SS are also denoted as the row and column SS, respectively.) The *error* or *residual* SS, that part of the total

Table 8.8 Two-Way Layout for Analytical Procedures Applied to Different Batches of Material

	Analytical method			
Batch	***A***	***B***	***C***	. . .
1	←————	————	————	————→
2	←————	————	————	————→
3	←————	————	————	————→
.				
.				
.				

Table 8.9 Tablet Dissolution After 30 Minutes for Three Products (Percent Dissolution)

	Generic			
Laboratory	*A*	*B*	Standard	Row total
1	89	83	94	266
2	93	75	78	246
3	87	75	89	251
4	80	76	85	241
5	80	77	84	241
6	87	73	84	244
7	82	80	75	237
8	68	77	75	220
Column total	666	616	664	1946
$\overline{X}$	83.25	77.0	83.0	
$\sum X^2 = 158{,}786$				

sum of squares remaining after subtracting out that due to rows and columns, is also often denoted as the *interaction* ($C \times R$) SS.

The hypothesis of interest is

$$H_0\colon \mu_A = \mu_B = \mu_C.$$

That is, the average dissolution rates of the three products are equal. The level of significance is set at 5%. The experimental results are shown in Table 8.9.

The analysis proceeds as follows: Total sum of squares (TSS)

$$\text{Total sum of squares (TSS)} = \sum X^2 - \text{CT} = 89^2 + 93^2 + \cdots + 75^2 + 75^2 - \frac{(1946)^2}{24}$$
$$= 158{,}786 - 157{,}788.2 = 997.8.$$

Column sum of squares (CSS) or product SS

$$\text{CSS} = \frac{\sum C_j{}^2}{R} - \text{CT} = \frac{(666^2 + 616^2 + 664^2)}{8} - 157{,}788.2$$
$$= 200.3\,(C_j \text{ is the total of column } j,\ R \text{ is the number of rows}).$$

Row sum of squares (RSS) or laboratory SS

$$\text{RSS} = \frac{\sum R_i{}^2}{C} - \text{CT} = \frac{(266^2 + 246^2 + \cdots + 220^2)}{3} - 157{,}788.2$$
$$= 391.8\,(R_i \text{ is the total of row } i,\ C \text{ is the number of columns}).$$

Residual ($C \times R$) sum of squares (ESS) = TSS − CSS − RSS

$$= 997.8 - 200.3 - 391.8 = 405.7.$$

The ANOVA table is shown in Table 8.10. The *d.f.* are calculated as follows:

Total = $N_t - 1$	N_t = total number of observations
Column = $C - 1$	C = number of columns
Row = $R - 1$	R = number of rows
Residual ($C \times R$) = $(C - 1)\,(R - 1)$	

Table 8.10 Analysis of Variance Table for the Data (Dissolution) from Table 8.8

Source	d.f.	SS	Mean square	F^a
Drug products	2	200.3	100.2	$F_{2,14} = 3.5$
Laboratories	7	391.8	56.0	$F_{7,14} = 1.9$
Residual ($C \times R$)	14	405.7	29.0	
Total	23	997.8		

[a] See the text for a discussion of proper F tests.

8.4.1.1 Tests of Significance

To test for differences among *products* (H_0: $\mu_A = \mu_B = \mu_C$), an F ratio is formed

$$\frac{\text{drug product mean square}}{\text{residual mean square}} = \frac{100.2}{29} = 3.5.$$

The F distribution has 2 and 14 d.f. According to Table IV.6, an F of 3.74 is needed for significance at the 5% level. Therefore, the products are not significantly different at the 5% level. However, had the a priori comparisons of each generic product versus the standard been planned, one could perform a t test for each of the two comparisons (using 29.0 as the error from the ANOVA), *generic* A *versus standard and generic* B *versus standard.* Generic A is clearly not different from the standard. The t test for generic B versus the standard is

$$t = \frac{|\overline{X}_B - \overline{X}_S|}{\sqrt{29(1/8 + 1/8)}} = \frac{6}{2.69} = 2.23.$$

This is significant at the 5% level (see Table IV.4; t with 14 d.f. = 2.14). Also, one could apply one of the multiple comparisons tests, such as the Tukey test described in section 8.2.1. According to Eq. (8.9), any difference exceeding $Q\sqrt{S^2(1/N)}$ will be significant. From Table IV.7, Q for 3 treatments and 14 d.f. for error is 3.70 at the 5% level. Therefore, the difference needed for significance for any pair of treatments for a posteriori tests is

$$3.70\sqrt{29\frac{1}{8}} = 7.04.$$

Since none of the means differ by more than 7.04, individual comparisons decided upon after seeing the data would show no significance in this experiment.

The test for laboratory differences is (laboratory mean square)/(residual mean square), which is an F test with 7 and 14 d.f. According to Table IV.6, this ratio is not significant at the 5% level (a value of 2.77 is needed for significance). As discussed further below, if *drug products* are a fixed effect, this test is valid only if interaction (drug product × laboratories) is absent. Under these conditions, the laboratories are not sufficiently different to show a significant F value at the 5% level.

8.4.1.2 Fixed and Random Effects in the Two-Way Model[§]

The proper test of significance in the two-way design depends on the model and the presence of *interaction.* The notion of interaction will be discussed further in the presentation of factorial designs (chap. 9). In the previous example, the presence of *interaction* means that the three products are ranked differently with regard to dissolution rate by at least some of the eight laboratories. For example, laboratory 2 shows that generic A dissolves fastest among the three products, with generic B and the standard being similar. On the other hand, laboratory 8 shows that generic A is the slowest-dissolving product. Interaction is conveniently shown graphically as in Figure 8.4. "Parallel curves" indicate no interaction.

[§] A more advanced topic.

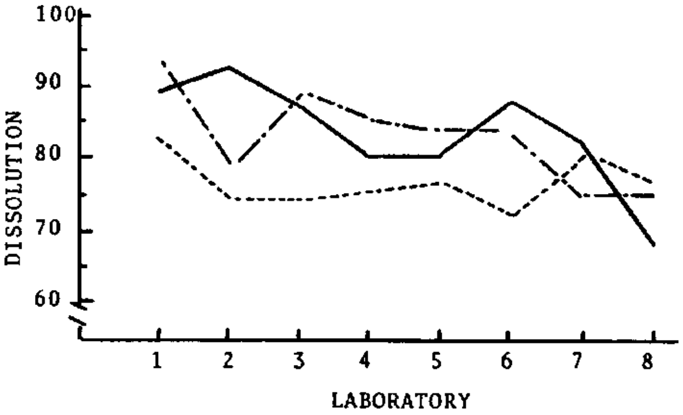


Figure 8.4 Average results of dissolution for eight laboratories. — · standard; — generic *A*; – – – generic *B*.

Of course, in the presence of error (variability), it is not obvious if the apparent lack of parallelism is real or is due to the inherent variability of the system. An experiment in which a lab makes a single observation on each product, such as is the case in the present experiment, usually contains insufficient information to make decisions concerning the presence or absence of interaction. To test for interaction, an additional error term is needed to test for the significance of the $C \times R$ residual term. In this case, the experiment should be designed to have replicates (at least duplicate determinations). In the absence of replication, it is best (usually) to assume that interaction is present. This is a conservative point of view. A knowledge of the presence or absence of interaction is important in order that one may choose the proper error term for statistical testing (the term in the denominator of the F test) as described below.

The concept of fixed and random effects was introduced under the topic of one-way ANOVA. A "fixed" category includes all the treatments of interest. In the present example, it is apparent that the columns, drug products, are fixed. We are only interested in comparing the two generic products with the standard. Otherwise, we would have included other products of interest in the experiment. On the other hand, the nature of the rows, laboratories, is not obvious. Depending on the context, laboratories may be either *random* or *fixed*. If the laboratories were selected as a random sample among many laboratories that perform such dissolution tests, then "laboratories" is a random factor. In the present situation, the laboratories are chosen as a means of replication in order to compare the dissolution of the three products. Then, inferences based on the result of the experiment are applied to the population of laboratories from which this sample of eight was drawn. We might also be interested in estimating the variance among laboratories in order to have some estimate of the difference to be expected when two or more laboratories perform the same test (see sect. 12.4.1). If the laboratories chosen were the only laboratories of interest, and inferences based on the experimental results apply only to these eight laboratories, then laboratories are considered to be fixed. Table 8.11 shows when the F tests in the two-way ANOVA are valid depending on the model and the presence of interaction.

Table 8.11 Tests in the Two-Way Analysis of Variance (One Observation Per Cell)

Columns	Rows	Interaction	Error term for the F test[a]
Fixed	Random	None	Residual ($C \times R$) or within
Random	Random	None	Residual ($C \times R$) or within
Random	Random	None	Residual ($C \times R$) or within
Fixed	Fixed	Present	Within
Fixed	Random	Present	Residual ($C \times R$) for fixed effect; use within for random effect
Random	Random	Present	Residual (CR)

[a] *Residual* is the usual residual mean square and includes ($C \times R$), column × row interaction. *Within* is the within mean square calculated from replicate determinations and will be called "error" in future discussions.

In the usual situation, columns are fixed (e.g., drug treatments, formulations) and rows are random (patients, batches, laboratories). In these cases, in the absence of replication, the proper test for columns is (column mean square)/(residual mean square).

Usually, the test for rows is not pertinent if rows are "random." For example, in a clinical study, in which two or more treatments are to be compared, the rows are "patients." The statistical test of interest in such situations is a comparison of the treatments; one does not usually test for patient differences. However, in many laboratory experiments, both column and row effects are of interest. In these cases, if significance testing is to be performed for *both row* and *column* effects (where either or both are fixed), it is a good idea to include proper replication (Table 8.11). *Duplicate assays* on the same sample such as may be performed in a dissolution experiment are not adequate to estimate the relevant variability. Replication in this example would consist of repeat runs, using different tablets for each run. An example of a two-way analysis in which replication is included is described in the following section.

8.4.2 Two-Way ANOVA with Replication

Before discussing an example of the analysis of two-way designs with replications, two points should be addressed regarding the implementation of such experiments.

1. It is best to have equal number of replications for each cell of the two-way design. In the dissolution example, this means that each lab replicates each formulation an equal number of times. If the number of replicates is very different for each cell, the analysis and interpretation of the experimental results can be very complicated and difficult.
2. The experimenter should be sure that the experiment is *properly* replicated. As noted above, merely replicating assays on the same tablet is not proper replication in the dissolution example. Replication is an independently run sample in most cases. Each particular experiment has its own problems and definitions regarding replication. If there is any doubt about what constitutes a proper replicate, a statistician should be consulted.

As an example of a replicated, two-way experiment, we will consider the dissolution data of Table 8.9. Suppose that the data presented in Table 8.9 are the average of two determinations (either *two* tablets or *two averages of six tablets each*—a total of 12 tablets). The actual duplicate determinations are shown in Table 8.12. We will consider "products" fixed and "laboratories" random.

The analysis of these data results in one new term in the ANOVA, that due to the *within-cell SS*. The *within-cell SS* represents the variability or error due to replicate determinations, and is the pooled SS from within the cells. In the example shown previously, the SS is calculated for

Table 8.12 Replicate Tablet Dissolution Data for Eight Laboratories Testing Three Products (Percent Distribution)

	Generic			
Laboratory	*A*	*B*	Standard	Row total
1	87, 91	81, 85	93, 95	532
2	90, 96	74, 76	74, 82	492
3	84, 90	72, 78	84, 94	502
4	75, 85	73, 79	81, 89	482
5	77, 83	76, 78	80, 88	482
6	85, 89	70, 76	80, 88	488
7	79, 85	74, 86	71, 79	474
8	65, 71	73, 81	70, 80	440
Total	1332	1232	1328	3892
Average	83.25	77.0	83.0	

each cell, $\sum(X-\overline{X})^2$. For example, for the first cell (generic A in laboratory 1), $\sum(X-\overline{X})^2 = (87-89)^2+(91-89)^2 = (87-91)^2/2 = 8$. The SS is equal to 8. The within SS is the total of the SS for the 24 (8 × 3) cells. The residual or interaction SS is calculated as the difference between the TSS and the sum of the column SS, row SS, and within-cell SS. The calculations for Table 8.12 are shown below.

$$\begin{aligned}\text{Total sum of squares} &= \sum X^2 - \text{CT}\\ &= 87^2+91^2+90^2+\cdots+71^2+79^2+70^2+80^2-\frac{3892^2}{48}\\ &= 318{,}160-315{,}576.3 = 2583.7.\end{aligned}$$

$$\begin{aligned}\text{Product SS} &= \frac{\sum C_j{}^2}{Rr} - \text{CT}\\ &= \frac{1332^2+1232^2+1328^2}{16} - \frac{3892^2}{48} = 315{,}977 - 315{,}576.3\\ &= 400.7\end{aligned}$$

where C_j is the sum of observations in column j, R the number of rows, and r the number of replicates per cell.

$$\begin{aligned}\text{Laboratory SS} &= \frac{\sum R_i{}^2}{Cr} - \text{CT}\\ &= \frac{532^2+492^2+\cdots+440^2}{6} - \frac{3892^2}{48} = 316{,}360 - 315{,}576.3\\ &= 783.7\end{aligned}$$

where R_i is the sum of observations in row i, C the number of columns, and r the number of replicates per cell.

Within-cell SS**

$\sum(X-\overline{X})^2$, where the sum extends over all cells

$$\begin{aligned}&= \frac{(87-91)^2}{2}+\frac{(90-96)^2}{2}+\frac{(84-90)^2}{2}+\cdots+\frac{(70-80)^2}{2}\\ &= 588.\end{aligned}$$

$$\begin{aligned}C \times \text{R SS} &= \text{TSS} - \text{PSS} - \text{LSS} - \text{WSS}\\ &= 2583.7 - 400.7 - 783.7 - 588\\ &= 811.3.\end{aligned}$$

The ANOVA table is shown in Table 8.13. Note that the F test for drug products is identical to the previous test, where the averages of duplicate determinations were analyzed. However, the laboratory mean square is compared to the within mean square to test for laboratory differences. This test is correct if laboratories are considered either to be fixed (all FDA laboratories, for example), or random, when drug products are fixed (Table 8.13). For significance $F_{7,24}$ must exceed 2.43 at the 5% level (Table IV.6). The significant result for laboratories suggests that at least some of the laboratories may be considered to give different levels of response. For example, compare the results for laboratory 1 versus laboratory 8.

** For duplicate determinations, $\sum(X-\overline{X}) = (X_1-X_2)^2/2$.

Table 8.13 ANOVA Table for the Replicated Dissolution Data Shown in Table 8.12

Source	d.f.	SS	Mean square	F^a
Drug products	2	400.7	200.4	$F_{2,14} = 3.5$
Laboratories	7	783.7	112	$F_{7,24} = 4.6^*$
$C \times R$ (residual)	14	811.3	58.0	$F_{14,24} = 2.37^*$
Within cells (error)	24[b]	588	24.5	

[a] Assume drug products are fixed, laboratories random.

[b] d.f. for within cells is the pooled d.f., one d.f. for each of 24 cells; in general, d.f. = $R \times C$ $(n - 1)$, where n is the number of replicates.

$^*p < 0.05$.

Another statistical test, not previously discussed, is available in this analysis. The F test ($C \times R$ mean square/within mean square) is a test of *interaction*. In the absence of interaction (laboratory × drug product), the $C \times R$ mean square would equal the within mean square on the average. A value of the ratio sufficiently larger than 1 is an indication of interaction. In the present example, the F ratio is 2.37, 58.0/24.5. This is significant at the 5% level (see Table IV.6; $F_{14,24} = 2.13$ at the 5% level). The presence of a laboratory × drug product interaction in this experiment suggests that laboratories are not similar in their ability to distinguish the three products (Fig. 8.4).

8.4.3 Another Worked Example of Two-Way ANOVA§

Before leaving the subject of the basic ANOVA designs, we will present one further example of a two-way experiment. The design is a form of a factorial experiment, discussed further in chapter 9. In this experiment, three drug treatments are compared at three clinical sites. The treatments consist of two dosages of an experimental drug (low and high dose) and a control drug. Eight patients were observed for each treatment at each site. The data represent increased performance in an exercise test in asthmatic patients. The results are shown in Table 8.14. In order to follow the computations, the following table of totals (and definitions) should be useful.

$$\mathrm{CT} = \frac{(371.5)^2}{72} = 1916.84$$

R = number of rows = 3
C = number of columns = 3
r = number of replicates = 8
R_i = total of row i (row 1 = 108.9, row 2 = 140.7, row 3 = 121.9)
C_j = total of column j (column 1 = 69.7, column 2 = 156.1, column 3 = 145.7)

Table 8.14 Increase in Exercise Time for Three Treatments (Antiasthmatic) at Three Clinical Sites (Eight Patients Per Cell)

	Treatment			Cell means (standard deviation)		
Site	*A* (low dose)	*B* (high dose)	*C* (control)	*A*	*B*	*C*
I	4.0, 2.3, 2.1, 3.0 1.6, 6.4, 1.4, 7.0	3.6, 2.6, 5.5, 6.0 2.5, 6.0, 0.1, 3.1	5.1, 6.6, 5.1, 6.3 5.9, 6.2, 6.3, 10.2	3.475 (2.16)	3.675 (2.06)	6.463 (1.61)
II	2.4, 5.4, 3.7, 4.0 3.3, 0.8, 4.6, 0.8	6.6, 6.4, 6.8, 8.3 6.9, 9.0, 12.0,7.8	5.6, 6.4, 8.2, 6.5 4.2, 5.6, 6.4, 9.0	3.125 (1.68)	7.975 (1.86)	6.488 (1.52)
III	1.0, 1.3, 0.0, 5.1 0.2, 2.4, 4.5, 2.4	6.0, 8.1, 10.2, 6.6 7.3, 8.0, 6.8, 9.9	5.8, 4.1, 6.3, 7.4 4.5, 2.0, 6.8, 5.2	2.113 (1.88)	7.863 (1.52)	5.263 (1.73)

§ A more advanced topic.

The cell totals are shown below

	A	*B*	*C*	Total
Site I	27.8	29.4	51.7	108.9
Site II	25	63.8	51.9	140.7
Site III	16.9	62.9	42.1	121.9
Total	69.7	156.1	145.7	371.5

The computations for the statistical analysis proceed as described in the previous example. The *within-cell mean square* is the pooled variance over the nine cells with 63 d.f. (7 d.f. from each cell). In this example (equal number of observations in each cell), the within-cell mean square is the average of the nine variances calculated from within-cell replication (eight values per cell). The computations are detailed below.

$$\begin{aligned}\text{Total sum of squares} &= \sum X^2 - \text{CT}\\ &= 4.0^2 + 2.3^2 + 2.1^2 + \cdots + 6.8^2 + 5.2^2 - \frac{(371.5)^2}{72}\\ &= 2416.77 - 1916.84 = 499.93.\end{aligned}$$

$$\text{CSS (treatment SS)} = \frac{\sum C_j{}^2}{Rr} - \text{CT} = \frac{69.7^2 + 156.1^2 + 145.7^2}{3 \times 8} - 1916.84 = 185.40.$$

$$\text{RSS (site SS)} = \frac{\sum R_i{}^2}{Cr} - \text{CT} = \frac{108.9^2 + 140.7^2 + 121.9^2}{3 \times 8} - 1916.84 = 21.30.$$

Within-cell mean square = pooled sum of squares from the nine cells

$$= \sum X^2 - \frac{\sum(\text{cell total})^2}{r} = 2416.7$$

$$-\frac{27.8^2 + 29.4^2 + 51.7^2 + \cdots + 42.1^2}{8} = 2416.77 - 2214.2 = 202.57.$$

$C \times R$ SS (treatment × site interaction SS)

$$\begin{aligned}&= \text{total SS} - \text{treatment SS} - \text{site SS} - \text{within SS}\\ &= 499.93 - 185.40 - 21.30 - 202.57\\ &= 90.66.\end{aligned}$$

Note the shortcut calculation for within SS using the squares of the cell totals. Also note that the $C \times R$ SS is a measure of *interaction* of sites and treatments. Before interpreting the results of the experiment from a statistical point of view, both the ANOVA table (Table 8.15) and a plot of the average results should be constructed (Fig. 8.5). The figure helps as a means of interpretation of the ANOVA as well as a means of presenting the experimental results to the "client" (e.g., management).

8.4.3.1 Conclusions of the Experiment Comparing Three Treatments at Three Sites: Interpretation of the ANOVA Table

The comparisons of most interest come from the treatment and treatment × site terms. The *treatment mean square* measures differences among the three treatments. The *treatment × site mean square* is a measure of how the three sites differentiate the three treatments. As is usually the case, interactions are most easily visualized by means of a plot (Fig. 8.5). The lack of "parallelism" is most easily seen as a difference between site I and the other two sites. Site I shows that treatment C has the greatest increase in exercise time, whereas the other two sites find treatment

Table 8.15 Analysis of Variance Table for the Data of Table 8.14 (Treatments and Sites Fixed)

Source	d.f.	SS	Mean square	*F*
Treatments	2	185.4	92.7	$F_{2.63} = 28.8^{a}$
Sites	2	21.3	10.7	$F_{2.63} = 3.31^{b}$
Treatment × site	4	90.66	22.7	$F_{4.63} = 7.05^{a}$
Within	63	202.57	3.215	
Total	71	499.93		

[a] $p < 0.01$.
[b] $p < 0.05$.

B most efficacious. Of course, the apparent differences, as noted in Figure 8.5, may be due to experimental variability. However, the treatment × site interaction term (Table 8.15) is highly significant ($F_{4,63} = 7.05$). Therefore, this interaction can be considered to be real. The presence of interaction has important consequences on the interpretation of the results. The lack of consistency makes it difficult to decide if treatment *B* or treatment *C* is the better drug. Certainly, the decision would have been easier had all sites found the same drug best. The final statistical decision depends on whether one considers sites fixed or random. In this example treatments are fixed.

Case 1: Sites fixed. If both treatments and sites are fixed, the proper error term for treatments and sites is the within mean square. As shown in Table 8.15, both treatments and sites (as well as interaction) are significant. Inspection of the data suggests that treatments *B* and *C* are not significantly different, but that both of these treatments are significantly greater than treatment *A* (see Exercise Problem 11 for an a posteriori test). Although not of primary interest in such studies, the significant difference among sites may be attributed to the difference between site II and site I, site II showing greater average exercise times (due to higher results for treatment *B*). However, this difference is of less importance than the interaction of sites and treatments that exists in this study. Thus, although treatments *B* and *C* do not differ, on the average, in the fixed site case, site I is different from the other sites in the comparison of treatments *B* and *C*. One may wish to investigate further to determine the cause of such differences (e.g., different kinds of patients, different exercise equipment, etc.). If the difference between the results for treatments *B* and *C* were dependent on the type of patient treated, this would be an important parameter in drug therapy. In most multiclinic drug trials, clinical sites are selected at random, although it is impractical, if not impossible, to choose clinical sites in a truly random fashion (see also sect. 11.5). Nevertheless, the interpretation of the data is different if sites are considered to be a random effect.

Case 2: Sites random. If sites are random, and interaction exists, the correct error term for treatments is the treatment × site (interaction) mean square. In this case, the *F* test ($F_{2,4} = 4.09$) shows a lack of significance at the 5% level. The apparently "obvious" difference between treatment *A* and treatments *B* and *C* is not sufficiently large to result in significance because of the paucity of d.f. (4 d.f.). The disparity of the interpretation here compared to the fixed sites

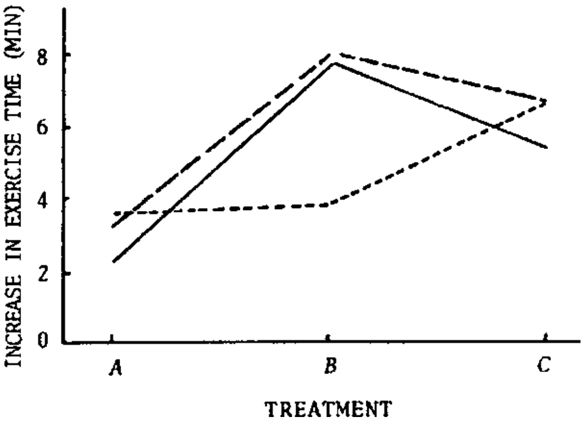


Figure 8.5 Plot of average results from data of Table 8.14. – – site II; — site I; — — site III.

Table 8.16 Tests for Treatment Differences in Two-Way ANOVA with Replicate Observations (Treatments Fixed)

Rows	Interaction	Proper error term
Fixed	Present	Within mean square
Fixed	Absent	Within mean square or $C \times R$ mean square
Random	Present	$C \times R$ (interaction) mean square
Random	Absent	Within mean square (conservative test: use $C \times R$ mean square: pool $C \times R$ and within mean square—see the text)

case is due to the large interaction. The data suggest that differences among treatments are dependent on the site at which the drugs are tested. If the three sites are random selection from among many possible sites, this very small sample of sites does not give a reliable estimate of the population averages.

Table 8.16 abstracted from Table 8.11, shows the proper error terms for testing treatment differences, depending on whether sites (rows) are random or fixed. The testing also depends on whether or not there is interaction in the model. Ordinarily, it is not possible to predict the presence (or absence) of interaction in advance of the study. The conservative approach for statistical tests is to assume interaction exists. In this example, if sites are random, the $C \times R$ (interaction) mean square is the proper error term for treatments. Often, however, the interaction mean square has few d.f. This can considerably reduce the power of the test, as is the case in this example. In these situations, if the interaction mean square is not significant, the *interaction* and *within* mean squares may be pooled. This gives a pooled error term with more d.f. than either term alone. This is a controversial procedure, but can be considered acceptable if interaction is clearly not present.

8.4.4 Missing Data

Missing data can result from overt errors in measurements, patients not showing up for scheduled visits in a clinical trial, loss of samples, etc. In general, the problems of dealing with missing data are complex. Missing data can be considered to be caused by missing observations from a statistically valid, symmetric design. A common manifestation is when a "cell" is empty, that is, contains no values. A cell may be defined as the intersection of factor levels in a factorial or related design. For example, in a two-way crossover design, if a subject misses a visit, we have an empty cell. In a one-way design, missing values do not cause computational problems in general, because the analysis is valid when sample sizes in the independent groups are not equal.

For a two-way design with one missing value, the missing value may be estimated using the following formula:

$$Y_{ij} = \frac{rY_i + cY_j - y}{(r-1)(c-1)}, \tag{8.17}$$

where r is the number of rows, c the number of columns, Y_{ij} the observation in ith row and jth column, Y_i is the total of ith row, Y_j the total of jth column, and y the grand total of all observations.

For example, Table 8.17 shows data with a missing value in the second column and third row. From Eq. (8.17), $Y_{32} = (3 \times 10 + 3 \times 9 - 44)/[(3-1)(3-1)] = 3.25$.

An ANOVA is performed including the estimated observation (3.25), but the d.f. for error are reduced by 1 due to the missing observation. (See Exercise Problem 12.)

For more than one missing value and for further discussion, see Snedecor and Cochran [8]. For more complicated designs, computer software programs may be used to analyze data with missing values. One should be aware that in certain circumstances depending on the nature of the missing values and the design, a unique analysis may not be forthcoming. In some cases, some of the observed data may have to be removed in order to arrive at a viable analysis.

Another problem with missing data occurs in clinical studies with observations made over time where patients drop out prior to the anticipated completion of treatments (censored data). A common approach when analyzing such data where some patients start but do not complete

Table 8.17 Illustration of Estimation of a Single Missing Data Point

	Columns			
	1	2	3	Total
Rows				
1	3	5	6	14
2	7	4	9	20
3	4	—	6	10
Total	14	9	21	44

the study for various reasons, is to carry the last value forward. For example, in analgesic studies measuring pain, patients may give pain ratings over time. If pain is not relieved, patients may take a "rescue" medication and not complete the study as planned. The last pain rating on study medication would then be continued forward for the missed observation periods. For example, such a study might require pain ratings (1–5, where 5 is the worst pain and 1 is the least) every half-hour for six hours. Consider a patient who gives ratings of 5, 4, and 4 for hours 0 (baseline), half, and one hour, respectively. He then decides to take the rescue medication. The patient would be assigned a rating of 4 for all periods after one hour (1.5–6 hours, inclusive). Statistical methods are then used as usual. Other variations on the Last Value Carried forward (LVCF) concept is to carry forward either the best or worst reading prior to dropout as defined and justified in the study protocol. (See also sect. 11.2.7.) Other methods include the average of all observations for a given patient as the final result. These are still controversial and should be discussed with FDA prior to implementation. One problem with this approach occurs in disease states that are self-limiting. For example, in studies of single doses of analgesics in acute pain, if the study extends for a long enough period of time, pain will eventually be gone. To include patients who have dropped out prior to these extended time periods could bias the results at these latter times.

8.5 STATISTICAL MODELS[§]

Statistical analyses for estimating parameters and performing statistical tests are usually presented as linear models as introduced in section 8.1. (See also Apps. II and III.) The parameters to be included in the model are linearly related to the dependent variable in the form of

$$Y = B_0X_0 + B_1X_1 + B_2X_2 + \cdots + \varepsilon, \tag{8.18}$$

where the Bs are the parameters to be estimated and the various X_i, represent the independent variables. Epsilon, ε, represents the random error associated with the experiment, and is usually assumed to be normal with mean 0 and variance, σ^2. This suggests that the estimate of Y is unbiased, with a variance, σ^2. For a simple model, where we wish to fit a straight line, the model would appear as

$$Y = B_0X_0 + B_1X_1,$$

where $X_0 = 1$ and X_1 (the independent variable) has a coefficient B_1.

In this example, we observe data pairs, X_i, Y_i, from which we estimate B_0 (intercept) and B_1 (slope). Again, this particular model represents the model of a straight line.

Although simple methods for analyzing such data have been presented in chapter 7, the data could also be analyzed using ANOVA based on the model. This analysis would first compute the TSS, which is the SS from a model with only a mean ($Y = \mu + \varepsilon$), with $N_t - 1$ d.f. This is the SS obtained as if all the data were in a single group. The $N_t - 1$ d.f. are based on the fact that, in the computation of SS, we are subtracting the mean from each observation before squaring. Having computed the SS from this simple model, a new SS would then be computed

[§] A more advanced topic.

from a model that looks like a straight line. Each observation is subtracted from the least squares line and squared (the residuals are subtracted from a model with two parameters, slope and intercept). The difference between the SS with one parameter (the mean) and the SS with two parameters (slope and intercept) has 1 d.f. and represents the SS due to the slope. The inclusion of a slope in the model reduces the SS. In general, as we include more parameters in the model, the SS is reduced. Eventually, if we have as many observations as terms in the model, we will have 0 residual SS, a perfect fit.

Typically, we include terms in the model that have meaning in terms of the experimental design. For example, for a one-way ANOVA (see sect. 8.1), we have separated the experimental material into k groups and assigned N_t subjects randomly to the k groups. The model consists of groups and a residual error, which represents the variability of observations within the groups

$$Y_{ik} = \mu + G_k + \varepsilon_{ik}.$$

μ represents the overall mean of the data, G_k represents the deviation from μ due to the kth group (i.e. kth group effect) (treatment), and ε_{ik} is the common variance (residual error). Note that the Xs are not written in the model statement, and are assumed to be equal to 1. A more detailed description of the model including three groups might look like this (see sect. 8.1)

$$Y_{i1} = \mu + G_1 + \varepsilon_{i1}, Y_{i2} = \mu + G_2 + \varepsilon_{i2}, Y_{i3} = \mu + G_3 + \varepsilon_{i3}$$

We then estimate μ, G_1, G_2, and G_3 from the model, and the residual is the error SS. Note that as before, ignoring groups, the total d.f. = $N_t - 1$. The fit of the model without groups, compared to the fit with groups (N_{k-1} d.f. for each group) has 2 d.f. $[(N_t - 1) - (N_t - 3)]$)(that represent the SS for differences between groups. If groups have identical means, the residual SS will be approximately the same for the full model (three separate groups) and the reduced model (one group).

A somewhat more complex design is a two-way design, such as a randomized block design, where, for example, in a clinical study, each patient may be subjected to several treatments. This model includes both patient and group effects. The residual error is a combination of both group × patient interaction (GP) and within-individual variability. To separate these two sources of variability, patients would have to be replicated in each group (treatment). If such replication exists (see sect. 8.4.2), the model would appear as follows with g groups ($i = 1$ to g), and p patients ($j = 1$ to p) per group, each patient being replicated k times in each group

$$Y_{ijk} = \mu + G_i + P_j + GP_{ij} + \varepsilon_{ijk}.$$

Models may become complicated, but the procedure for their construction and analysis follows the simple approaches shown above. For experiments that are balanced (no missing data), the calculations are simple and give unambiguous results. For unbalanced experiments, the computations are more complex, and the interpretation is more difficult, sometimes impossible. Computer programs can analyze unbalanced data, but care must be taken to understand the data structure in order to make the proper interpretation (see also sect. 8.4.4).

8.6 ANALYSIS OF COVARIANCE§

The analysis of covariance (ANCOVA) combines ANOVA with regression. It is a way to increase precision and/or adjust for bias when comparing two treatments. ANCOVA uses observations (concomitant variables) that are taken independently of the test (outcome) variable. These concomitant observations are used to "adjust" the values of the test variable. This usually results in a statistical test that is more precise than the corresponding nonadjusted analysis. We look for covariates that are highly correlated with the experimental outcome, the greater the better (10). For example, the initial weight of a patient in a weight reduction study may be correlated with the weight reduction observed at the end of the study. Also, note that one may choose more than one covariate. One simple example is the use of baseline measurements when comparing

§ A more advanced topic.

Table 8.18 Analytical Results for Eight Batches of Product Comparing Two Manufacturing Methods

Method I		Method II	
Raw material	**Final product**	**Raw material**	**Final product**
98.4	98.0	98.7	97.6
98.6	97.8	99.0	95.4
98.6	98.5	99.3	96.1
99.2	97.4	98.4	96.1
Average 98.70	97.925	98.85	96.30

the effect of two or more treatments. A common approach in such experiments is to examine the change from baseline (experimental observation–baseline) as discussed in sections 8.4 and 11.3.2. The analysis can also be approached using ANCOVA, where the baseline measurement is the covariate. The correction for baseline will then adjust the experimental observation based on the relationship of the two variables, baseline and outcome. Another example is the comparison of treatments where a patient characteristic, for example, weight, may be related to the clinical outcome; weight is the covariate. In these examples, assignment to treatment could have been stratified based on the covariate variable, for example, weight. ANCOVA substitutes for the lack of stratification by adjusting the results for the covariate, for example, weight. Refer to the chapter on regression (chap. 7) and to the section on one-way ANOVA (sect. 8.1) if necessary to follow this discussion. Ref. [9] is useful reading for more advanced approaches and discussion of ANCOVA.

In order to facilitate the presentation, Table 8.18 shows the results of an experiment comparing two manufacturing methods for finished drug product. In this example, the analysis of the raw material used in the product was also available.

If the two methods are to be compared using the four final (product) assays for each method, we would use a one-way ANOVA (independent sample t test in this example). The ANOVA comparing the two methods would be as shown in Table 8.19 and Table 8.21, columns 1 and 2.

The two methods yield different results at the 0.05 level ($p = 0.02$), with averages of 97.925 and 96.3, respectively. The question that one may ask is, "Are the raw material assays different for the products used in the test, accounting for the difference?" We can perform an ANOVA on the initial values to test this hypothesis. See Table 8.20 and Table 8.21, columns 1 and 3.

The average raw material assays for the lots used for the two methods are not significantly different (98.7 and 98.85). Thus, we may assume that the final assay results are not biased by possible differences in raw material. (Note that if the averages of the raw materials were different for the two methods, then one would want to take this into consideration when comparing the methods based on the final assay.) However, it is still possible that use of the initial values may reduce the variability of the comparison of methods due to the relationship between the raw material assay and the final. To account for this relationship, we can compute a linear fit of

Table 8.19 ANOVA Comparing Methods Based on Final Assay

Source	d.f.	Sum of squares	Mean square	*F* value	Pr > *F*
Between methods	1	5.28125	5.28125	9.88	0.0200
Within methods	6	3.20750	0.53458		
Total	7	8.48875			

Table 8.20 ANOVA Comparing Raw Material Assays

Source	d.f.	Sum of squares	Mean square	*F* value	Pr > *F*
Between methods	1	0.0450	0.0450	0.33	0.5847
Within methods	6	0.8100	0.1350		
Total	7	0.8550			

Table 8.21 Detailed Computations for ANCOVA for Data of Table 8.18

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Source	d.f.	*Y* Final assay	*X* Raw material	Final × raw	Slope	Reg. SS	d.f.	Res. SS
a. Within method *A*	3	0.6275	0.36	−0.33	−0.917	0.303	2	0.325
b. Within method *B*	3	2.58	0.45	−0.33	−0.733	0.242	2	2.338
c. Separate regressions	—	—	—	—	—	0.545	4	2.663
d. Within methods	6	3.2075	0.81	−0.66	−0.815	0.538	5	2.670
e. Between methods	1	5.28125	0.045	−0.4875	—	—	—	—
f. Total	7	8.48875	0.855	−1.1475	−1.343	1.541	6	6.948

Columns 1 and 2 are the simple ANOVA for the final assay.
Columns 1 and 3 are the simple ANOVA for the raw material assay.
Column 4 is the cross product $SS = \sum[(X - \overline{X})(Y - \overline{Y})]$.
Column 5 is computed as column 4/column 3 (final assay is the *Y* variable; raw material assay is the *X* variable).
Column 6 is column 3 × column 5 squared.
Column 7 is d.f. for residual (column 8).
Column 8 is column 2 − column 6.

the final assay result versus the raw material assay, and use the residual error from the fit as an estimate of the variance. The variance estimate should be smaller than that obtained when the relationship is ignored. The fitted lines for each method are assumed to be parallel, that is, the relationship between the covariate and the outcome variable (finished product assay) is the same for each method. With this assumption, the difference between methods, adjusted for the covariate, is the difference between the parallel lines at any value of the covariate, in particular the difference of the intercepts of the parallel lines. These concepts are illustrated in Figure 8.6.

Assumptions for covariance analysis include the following:

1. The covariate is not dependent on the experimental observation. That is, the covariate is not affected by the treatment (method). For example, an individual's weight measured prior to and during treatment by a cholesterol-reducing agent is not affected by his cholesterol reading(s).
2. The covariate is a fixed variable or the covariate and outcome variable have a bivariate normal distribution. The covariate is specified and measured before randomization to treatments.
3. Slopes for regression lines within each treatment group are equal, that is, the lines are parallel. If not, the analysis is still correct, but if interaction is suspected, we end up with an average effect. Interaction suggests that the comparison of treatments depends on the covariate level.

Covariance analysis is usually performed with the aid of a statistical software program as shown in Table 8.22. However, to interpret the output, it is useful to understand the nature of the

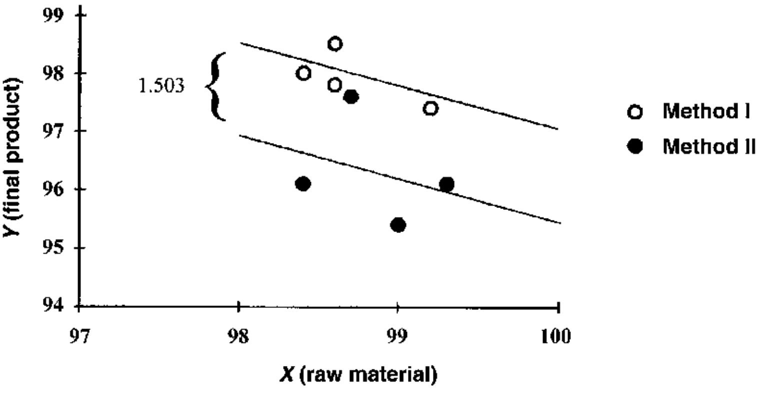


Figure 8.6 Illustration of adjusted difference of means in ANCOVA.

Table 8.22 ANCOVA Software Analysis of Data from Table 8.18

Source	d.f.	Sum of squares	Mean square	*F* ratio	Prob > *F*
X (Cov)	1	0.5377778	0.5377778	1.01	0.3616
A (Method)	1	4.278962	4.278962	8.01	0.0366
Error	5	2.669722	0.5339444		
Total (Adj)	7	8.488751			
Method	Means				
I	97.86389				
II	96.36112				

calculations. Table 8.21 is a complete table of the analysis for the example of the two analytical methods (Table 8.18). The following discussion refers to entries in Table 8.21.

The software (Table 8.22) computes the means, adjusted for the covariate, but does not perform a test for parallelism of the regression lines. A SAS program that includes Covariate × Method interaction in the model is a test for parallelism. To test for parallelism of the method versus covariate fitted lines, an analysis is performed to determine if the residual SS are significantly increased when all points are fitted to individual (two or more) parallel lines as compared to a fit to separate lines. (Note the similarity to stability analysis for pooling lots, sect. 8.7.) An F test comparing the variances is performed to determine significance

$$F_{\text{d.f.1,d.f.2}} = \frac{(\text{Residual SS parallel lines} - \text{residual SS separate lines})/(\text{groups} - 1)}{(\text{Residual SS separate lines})/\text{d.f.}}. \quad (8.19)$$

The residual SS from the parallel lines is obtained from a least squares fit (final product assay vs. raw material assay). These residual SS are calculated from line d in Table 8.21

$$\sum(y - \overline{y})^2 - b^2 \sum(X - \overline{X})^2 \quad \text{(see sect. 7.4).}$$

This is equal to $(3.2075 - 0.815^2 \times 0.81) = 2.67$.

The residual SS when each method is fit separately is in line c, column (8) in Table 8.21. The analysis is in a form of the Gauss–Markov Theorem that describes an F test comparing two linear models, where one model has additional parameters. In this example, we fit a model with separate intercepts and slopes, a total of four parameters for the two methods, and compare the residual SS to a fit with common slope and separate intercepts, three parameters. The increase in the mean square residual due to the fit with less parameters is tested for significance using the F distribution as shown in Eq. (8.19). This is the same approach as that used to determine the pooling of stability lots as discussed in section 8.7.

$$F_{1,4} = \frac{(\text{Residual SS parallel lines} - \text{residual SS separate lines})/(2 - 1)}{(\text{Residual SS separate lines})/(8 - 4)}. \quad (8.20)$$

In this example, the F test with 1 and 4 d.f. is

$$\frac{(2.67 - 2.663)/1}{2.663/4} = 0.01,$$

which is not significant at $p < 0.05$ ($p > 0.9$). The lines can be considered to be parallel.

This computation may be explained from a different viewpoint. For a common slope, the residual SS is computed as $\sum(y - \overline{y})^2 - b^2 \sum(X - \overline{X})^2$. Here $\sum(y - \overline{y})^2$ and $\sum(X - \overline{X})^2$ are the sums of the sums of squares for each line, and b is the common slope (-0.815). From Table 8.21, line d, columns 2 to 4, the SS for the common line is

$$0.6275 + 2.58 - (-0.8148)^2(0.36 + 0.45) = 2.670.$$

For the separate lines (column 8 in Table 8.21), the sums of the SS is

$$\text{SS} = 0.325 + 2.338 = 2.663.$$

Another test of interest is the significance of the slope (vs. 0). If the test for the slope is not significant, the concomitant variable (raw material assay) is not very useful in differentiating the methods. The test for the slope is: within regression mean square/within residual mean square.

The residual mean square is that resulting from the fit of parallel lines (common slope).

In this example, from line d in Table 8.21, $F_{1,5} = 0.538/(2.67/5) = 1.01(p = 0.36)$. The common slope is -0.815 (line d, column 5) that is not significantly different from 0. Thus, we could conclude that use of the raw material assay as an aid in differentiating the methods is not useful. Nevertheless, the methods are significantly different both when we ignore the raw material assay ($p = 0.02$; Table 8.19) and when we use the covariance analysis (see below).

The test for difference of means adjusted for the covariate is a test for difference of intercepts of the parallel lines.

$$F_{1,5} = \frac{(\text{Residual SS total} - \text{residual SS within})/(\text{groups } - 1)}{(\text{Residual SS parallel lines})/(\text{d.f.})}.$$

In this example, $F_{1,5} = (6.948 - 2.670)/(2.670/5) = 8.01\,(p < 0.05)$ (see column 8, Table 8.21). This is a comparison of the fit with a common intercept (TSS) to the fit with separate intercepts (within SS) for the parallel lines.

The adjusted difference between methods can be calculated as the difference between intercepts or, equivalently, the distance between the parallel lines (Fig. 8.6). The adjusted means are calculated as follows [8].

The common slope is b. The intercept is $\overline{Y} - b\overline{X}$ (see Eq. 7.3, chap. 7). The difference of the intercepts is

$$\begin{aligned}(\overline{Y}_a - b\overline{X}_a) - (\overline{Y}_b - b\overline{X}_b) &= \overline{Y}_a - \overline{Y}_b - b(\overline{X}_a - \overline{X}_b)\\ &= 97.925 - 96.3 - (-0.815)(98.7 - 98.85)\\ &= 1.503.\end{aligned}$$

The difference between means adjusted for the raw material assay is 1.503.

8.6.1 Comparison of ANCOVA with Other Analyses

Two other common analyses use differences from baseline and ratios of the observed result to the baseline value when a concomitant variable, such as a baseline value, is available. For example, in clinical studies, baseline values are often measured in order to assess a treatment effect relative to the baseline value. Thus, once more, in addition to ANCOVA, two other ways of analyzing such data are analysis of differences from baseline or the ratio of the observed value and baseline value. The use of changes from baseline is a common approach that is statistically acceptable. If the covariance assumptions are correct, covariance should improve upon the difference analysis, that is, it should be more powerful in detecting treatment differences. The difference analysis and ANCOVA will be similar if the ANCOVA model approximates $Y = a + X$, that is, the slope of the X versus Y relationship is one (1). The use of ratios does disturb the normality assumption, but if the variance of the covariate is small, this analysis should be more or less correct. This model suggests that $Y/X = a$, where a is a constant. This is equivalent to $Y = aX$, a straight line that goes through the origin. [If the Y values, the experimentally observed results, are far from 0, and/or the X values are clustered close together, the statistical conclusions for ratios (observed/baseline) should be close to that from the ANCOVA.] See Exercise Problem 13 for further clarification.

A nonparametric ANCOVA is described in section 15.7.

8.7 ANOVA FOR POOLING REGRESSION LINES AS RELATED TO STABILITY DATA[§]

As discussed in chapter 7, an important application of regression and ANOVA is in the analysis of drug stability for purposes of establishing a shelf life. Accelerated stability studies are often used to establish a preliminary shelf life (usually 24 months), which is then verified by long-term studies under label conditions (e.g., room temperature). If more than one lot is to be used to establish a shelf life, then data from all lots should be used in the analysis. Typically, 3 production lots are put on stability at room temperature in order to establish an expiration date. The statistical analysis recommended by the FDA [10] consists of preliminary tests for pooling of data from the different lots. If both slopes and intercepts are considered similar for the multiple lots based on a statistical test, then data from all lots can be pooled. If not, the data may be analyzed as separate lots, or if slopes are not significantly different, a common slope with separate intercepts may be used to analyze the data. Pooling of all of the data gives the most powerful test (the longest shelf life) because of the increased d.f. and multiple data points. If lots are fitted separately, suggesting lot heterogeneity, expiration dating is based on the lot that gives the shortest shelf life. Separate fittings also result in poor precision because an individual lot will have fewer d.f. and less data points than that resulting from a pooled analysis. Degrees of freedom when fitting regression lines are $N - 2$, so that a stability study with 7 time points will have only 5 d.f. (0, 3, 6, 9, 12, 18, and 24 months). Fitting the data with a common slope will have intermediate precision compared to separate fits and a single combined fit.

The computations are complex and cannot be described in detail here, but the general principles will be discussed. The fitting is of the form of regression and covariance (see also sect. 8.6). The following model (Model 1) fits separate lines for each lot.

$$\text{Potency } (Y) = \sum a_i + \sum b_i X \quad \text{Model (1).}$$

For three lots, the model contains six parameters, three intercepts, and three slopes. The residual error SS is computed with $N - 6$ d.f. for 3 lots, where N is the total number of data pairs. Thus, each of the three lines is fit separately, each with its own slope and intercept. Least squares theory, with the normality assumption (the dependent variable is distributed normally with the same variance at each value, X), can be applied to construct a test for equality of slopes. This is done by fitting the data with a reduced number of parameters, where there is a common slope for the lots tested. The fit is made to a model of the form

$$\text{Potency } (Y) = \sum a_i + bX \quad \text{Model (2).}$$

For 3 lots, this fit has $N - 4$ d.f., where N is the total number of data pairs (X, Y) with the 3 intercepts and single slope accounting for the 4 d.f. Statistical theory shows that the following ratio, Eq. (8.21), has an F distribution

$$\frac{[\text{Residual SS from model (2)} - \text{Residual SS from model (1)}]/[P' - P]}{\text{Residual SS from model (1)}/[N - P']}. \tag{8.21}$$

If P' is the number of parameters to be fitted in Model (1), 6 for 3 lots, and P' is the number of parameters in Model (2), 4 for 3 lots, then the d.f. of this F statistic are $[P' - P]$ d.f. in the numerator (2 for 3 lots), and $N - P'$ d.f. in the denominator ($N - 6$ for 3 lots). If the F statistic shows significance, then the data cannot be pooled with a common slope, and separate fits for each line are used for predicting shelf life. A significant F ($p < 0.25$) suggests that a fit to individual lines is significantly better than a fit with a common slope, based on the increase in the sums of squares when the model with less parameters is fit. If slopes are not poolable, a 95% lower, if appropriate, one-sided (or 90% two-sided) confidence band about the fitted line for each lot is computed, and the expiration dates are determined for each batch separately.

If the F statistic is not significant, then a model with a common slope, but different intercepts, may be fit.

[§] A more advanced topic.

The most advantageous condition for estimating shelf life is when data from all lots can be combined. Before combining the data into a single line, a statistical test to determine if the lots are poolable is performed. In order to pool all of the data, a two-stage test is proposed by the FDA. First, the test for a common slope is performed as described in the preceding paragraph. If the test for a common slope is not significant ($p > 0.25$), a test is performed for a common intercept. This is accomplished by computing the residual SS for a fit to a single line (Model 3) minus the residual sums of squares for the reduced model with a common slope, Model 2, adjusted for d.f., and divided by the residual SS from the fit to the full model (separate slopes and intercepts), Model (1).

$$\text{Potency } (Y) = a + bX \quad \text{Model (3).}$$

The F test for a common intercept is

$$\frac{\text{Residual SS from model (3) } - \text{ residual SS from model (2)}/[P' - P]}{\text{Residual SS from model (1)}/[N - P']}. \tag{8.22}$$

For 3 lots, the F statistic has 2 d.f. in the numerator (2 parameter fit for a single line vs. a 4 parameter fit, 3 intercepts, and 1 slope for a fit with a common slope), and $N - 6$ d.f. in the denominator. Again, a significant F suggests that a fit using a common slope and intercept is not appropriate.

The FDA has developed a SAS program to analyze stability data using the above rules to determine the degree of pooling, that is, separate lines for each lot, a common slope for all lots, or a single fit with a common slope and intercept. A condensed version of the output of this program is described below.

The raw data is for three lots (A, B, and C), each with three assays at 0, 6, and 12 months.

	Lot		
Time (mo)	*A*	*B*	*C*
0	100	102	98
6	99	98	97
12	96	97	95

The output testing for pooling is derived from an ANCOVA with time as the covariate (Table 8.23). The ANOVA shows a common slope, indicated by line C with $p > 0.25$ ($p = 0.58566$). The test for a common intercept is significant, $p < 0.25$. Therefore, lines are fitted with a common slope but with separate intercepts.

Table 8.23 Modified and Annotated SAS Output from FDA Stability Program

Source	SS	d.f.	Mean square	*F*	*p*
A (pooled line)	9.67	4	2.42	3.10714	0.18935
B (intercept)	8.67	2	4.33	5.57143	0.09770
C (slope)	1.00	2	0.50	0.64286	0.58566
D (error)	2.33	3	0.78		

Key to sources of variation:

A = separate intercept, separate slope | common intercept, common slope. This is the residual SS from fit to a single line minus the residual SS from fits to separate lines. This is the SS attributed to model 3.

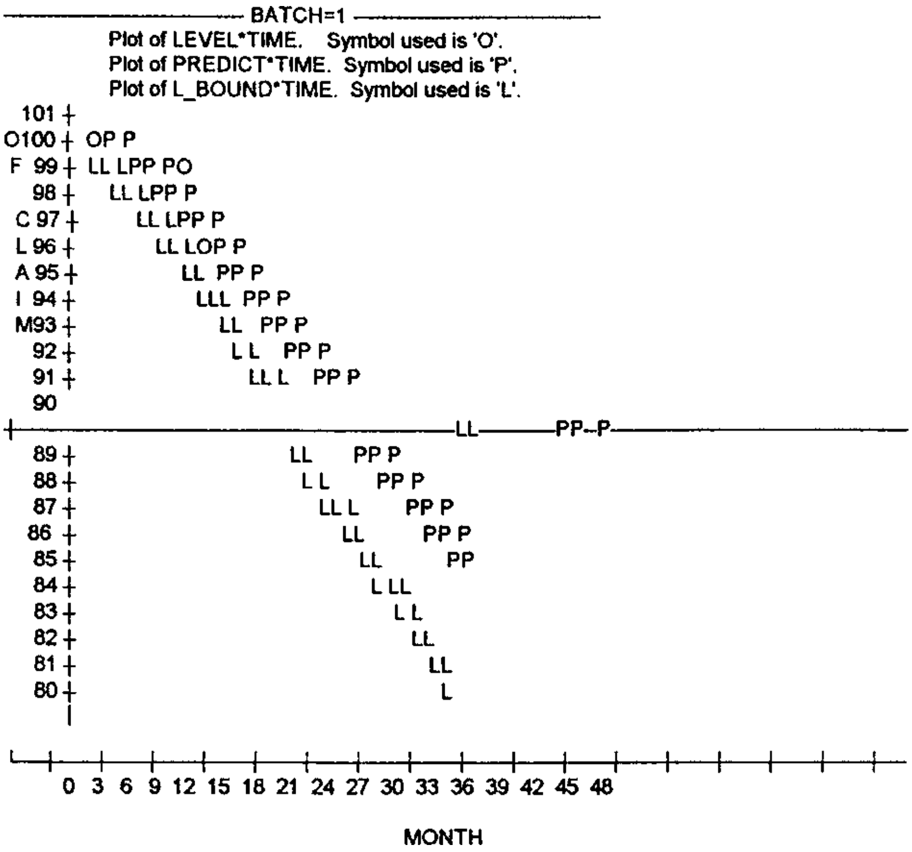
B = separate intercept, common slope | common intercept, common slope. This is the residual SS from a fit to a single line minus the residual SS from a fit with common slope and separate intercepts (*A* − *C*).

C = separate intercept, separate slope | separate intercept, common slope. This is the residual SS from a fit to a line with a common slope and separate intercepts line minus the residual SS from fits to separate lines. This is the SS attributed to model 2.

D = Residual. This is the residual SS from fits to separate lines (9 − 6 = 3 d.f.). This is the SS attribute to model 1.

The shelf life estimates vary from 20 to 25 months. The shortest time, 20 months, is used as the shelf life.

Stability analysis	
Fitted line	Batch 1 $Y = 100.33 - 0.333X$
Fitted line	Batch 2 $Y = 101.00 - 0.333X$
Fitted line	Batch 3 $Y = 98.67 - 0.333X$



```
----------------------------- BATCH=1 ----------------------------
            Plot of LEVEL*TIME.    Symbol used is 'O'.
            Plot of PREDICT*TIME.  Symbol used is 'P'.
            Plot of L_BOUND*TIME.  Symbol used is 'L'.
 101 +
O100 + OP P
F 99 + LL LPP PO
  98 +   LL LPP P
C 97 +      LL LPP P
L 96 +        LL LOP P
A 95 +           LL  PP P
I 94 +             LLL  PP P
 M93 +                LL   PP P
  92 +                 L L   PP P
  91 +                   LL L   PP P
  90
+-----------------------------------------LL--------PP--P------------------------
  89 +                         LL      PP P
  88 +                          L L       PP P
  87 +                            LL L      PP P
  86 +                              LL        PP P
  85 +                                LL         PP
  84 +                                 L LL
  83 +                                    L L
  82 +                                      LL
  81 +                                        LL
  80 +                                         L
     |
  L--+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+--
     0  3  6  9 12 15 18 21 24 27 30 33 36 39 42 45 48

                                     MONTH
```

Stability Analysis: 95% one-sided lower confidence limits (separate intercepts and common slope)

Batch	Estimated dating period (mo/wk)
1	24
2	25
3	20

The data for each batch should be visually inspected to ensure that the average results based on these calculations have not hidden noncompliant or potentially noncompliant batches.

The FDA recommends using a significance level of 25% rather than the usual 5% level. The reason for this is the use of multilevel preliminary testing before coming to a decision. The use of a 25% level is somewhat arbitrary, and does not seem to have a clear theoretical rationale. This higher level of significance means that the criterion for pooling lots is more difficult to attain, thereby making it more difficult to establish the longer shelf life that results from pooling data from multiple lots. This may be considered to be a conservative rule from the point of view that shelf lives will not be overestimated. However, the analysis is open to interpretation, and it is the author's opinion that the 25% level of significance is too high.

Another problem with the FDA approach is that power is not considered in the evaluation. For example, if the model and assay precision is very good, lots that look similar with regard to degradation may not be poolable, whereas with very poor precision, lots that appear not to be similar may be judged poolable. Unfortunately, this problem is not easily solved. Finally, it is not clear why the FDA has not included a test for pooling based on a common intercept and separate slopes.

Nevertheless, the FDA approach has much to recommend it, as the problem is quite complex.

KEY TERMS

Alpha level
ANCOVA
ANOVA
ANOVA table
A posteriori comparisons
A priori comparisons
Assumptions
Between-treatment sum of squares or mean square (BSS or BMS)
Block
Bonferroni test
Completely randomized design
Components of variance
Contrasts
Control
Correction term
Degrees of freedom
Designed experiments
Dunnett's test
Error
Error sum of squares or mean square (ESS or EMS)
Experimental error
Experimental units
F distribution
Fixed model
Independence
Interaction
LSD procedure for multiple comparisons mean square
Missing data
Model
Multiple comparisons
Newman–Keuls' test
One-way analysis of variance
Parallel groups
Parallelism
Placebo
Pooled variance
Pooled regressions
Positive control
Power
Precision
Randomized block design
Random model
Repeated measures design
Replicates
Residual
Scheffé method for multiple comparisons
Shortcut computing formulas
Source
Stability
Sum of squares
Symmetry
Total sum of squares (TSS)
Treatments
Treatment sum of squares or mean square
T tests
Tukey's multiple range test
Two-way analysis of variance
Within sum of squares or mean square (WSS or WMS)

EXERCISES

1. Perform three separate *t* tests to compare method *A* to method *B*, method *A* to method *C*, and method *B* to method *C* in Table 8.1. Compare the results to that obtained from the ANOVA (Table 8.3).

2. Treatments *A*, *B*, and *C* are applied to six experiment subjects with the following results:

A	B	C
1	3	4
5	2	1

Perform an ANOVA and interpret the between-treatment mean square.

3. Repeat the *t* tests from Exercise Problem 1, but use the "pooled" error term for the tests. Explain why the results are different from those calculated in Problem 1. When is it appropriate to perform separate *t* tests?

4. It is suspected that four analysts in a laboratory are not performing accurately. A known sample is given to each analyst and replicate assays performed by each with the following results:

Analyst			
I	II	III	IV
10	9	8	9
11	10	9	9
10	11	8	8

(a) State the null and alternative hypotheses.
(b) Is this a fixed or a random model?
(c) Perform an ANOVA. Use the LSD procedure to show which analysts differ if the "analyst" mean square is significant at the 5% level.
(d) Use Tukey's and Scheffé's multiple comparison procedures to test for treatment (analyst) differences. Compare the results to those in part (c).

5. Physicians from seven clinics in the United States were each asked to test a new drug on three patients. These physicians are considered to be among those who are expert in the disease being tested. The seventh physician tested the drug on only two patients. The physicians had a meeting prior to the experiment to standardize the procedure so that all measurements were uniform in the seven sites.
The results were as follows:

Clinic						
1	2	3	4	5	6	7
9	11	6	10	5	7	12
8	9	9	10	3	7	10
7	13	9	7	4	7	—

(a) Perform an ANOVA.
(b) Are the results at the different clinics significantly different at the 5% level?
(c) If the answer to part (b) is yes, which clinics are different? Which multiple comparison test did you use?

§6. Are the following examples random or fixed? Explain.
(a) Blood pressure readings of rats are taken after the administration of four different drugs.

(b) A manufacturing plant contains five tablet machines. The same product is made on all machines, and a random sample of 100 tablets is chosen from each machine and weighed individually. The problem is to see if the machines differ with respect to the weight of tablets produced.
(c) Five formulations of the same product are compared. After six months, each formula is assayed in triplicate to compare stability.
(d) Same as part (b) except that the plant has 20 machines. Five machines are selected at random for the comparison.
(e) Ten bottles of 100 tablets are selected at random in clusters 10 times during the packaging of tablets (a total of 10,000 tablets). The number of defects in each bottle are counted. Thus we have 10 groups, each with 10 readings. We want to compare the average number of defects in each cluster.

7. Dissolution is compared for three experimental batches with the following results (each point is the time in minutes for 50% dissolution for a single tablet).
Batch 1: 15, 18, 19, 21, 23, 26
Batch 2: 17, 18, 24, 20
Batch 3: 13, 10, 16, 11, 9
(a) Is there a significant difference among batches?
(b) Which batch is different?
(c) Is this a fixed or a random model?

8. In a clinical trial, the following data were obtained comparing placebo and two drugs:

	Placebo		Drug 1		Drug 2	
Patient	Predrug	Postdrug	Predrug	Postdrug	Predrug	Postdrug
1	180	176	170	161	172	165
2	140	142	143	140	140	141
3	175	174	180	176	182	175
4	120	128	115	120	122	122
5	165	165	176	170	171	166
6	190	183	200	195	192	185

(a) Test for treatment differences, using only postdrug values.
(b) Test for treatment differences by testing the change from baseline (predrug).
(c) For Problem 8(b), perform a posteriori multiple comparison tests (1) comparing all pairs of treatments using Tukey's multiple range rest and the Newman–Keuls' test and (2) comparing drug 1 and drug 2 to control using Dunnett's test.

9. Tablets were made on six different tablet presses during the course of a run (batch). Five tablets were assayed during the five-hour run, one tablet during each hour. The results are as follows:

	Press					
Hour	1	2	3	4	5	6
1	47	49	46	49	47	50
2	48	48	48	47	50	50
3	52	50	51	53	51	52
4	50	47	50	48	51	50
5	49	46	50	49	47	49

(a) Are presses and hours fixed or random?
(b) Do the presses give different results (5% level)?

(c) Are the assay results different at the different hours (5% level)?
(d) What assumptions are made about the presence of interaction?
(e) If the assay results are significantly different at different hours, which hour(s) is different from the others?

§10. Duplicate tablets were assayed at hours 1, 3, and 5 for the data in Problem 9, using only presses 2, 4, and 6, with the following results:

	Press		
Hour	**2**	**4**	**6**
1	49, 52	49, 50	50, 53
3	50, 48	53, 51	52, 55
5	46, 47	49, 52	49, 53

If presses and hours are fixed, test for the significance of presses and hours at the 5% level. Is there significant interaction? Explain in words what is meant by interaction in this example.

11. Use Tukey's multiple range test to compare all three treatments (a posteriori test) for the data of Tables 8.13 and 8.14.

12. Compute the ANOVA for data of Table 8.17. Are treatments (columns) significantly different?

13. Perform an analysis of variance (one-way) comparing methods for the ratios (final assay/raw material assay) for data of Table 8.18. Compare probability level for methods to ANCOVA results.

REFERENCES

1. Snedecor GW, Cochran WG. Statistical Methods, 8th ed. Ames, IA: Iowa State University Press, 1989.
2. Dixon WJ, Massey FJ Jr. Introduction to Statistical Analysis, 3rd ed. New York: McGraw-Hill, 1969.
3. Scheffé H. The Analysis of Variance. New York: Wiley, 1964.
4. Dunnett C, Goldsmith C. When and How to do multiple comparisons. in: Buncher CR, Tsay J-Y, eds. Statistics in the Pharmaceutical Industry, 2nd ed. New York: Marcel Dekker, 1993:481–512.
5. Steel RGD, Torrie JH. Principles and Procedure of Statistics. New York: McGraw-Hill, 1960.
6. Dubey SD. Adjustment of P-values for multiplicities of Intercorrelating Symptoms. In: Buncher CR, Tsay J, eds. Statistics in the Pharmaceutical Industry, 2nd ed. New York: Marcel Dekker, 1993:513–528.
7. Comelli M. Multiple Endpoints. In: Chow S-C, ed. Encyclopedia of Pharmaceutical Statistics, Multiple Endpoints. New York: Marcel Dekker, 2000:333–344.
8. Snedecor GW, Cochran WG. Statistical Methods, 8th ed. Ames, IA: Iowa State University Press, 1989:273.
9. Permutt T. In: Chow S-C, ed. Encyclopedia of Pharmaceutical Statistics, Adjustments for Covariates. New York: Marcel Dekker, 2000:1–3.
10. FDA Stability Program, Moh-Jee Ng, Div. of Biometrics, Center for Drug Evaluation and Res., FDA 1992.