9 Factorial Designs

Factorial designs are used in experiments where the effects of different factors, or conditions, on experimental results are to be elucidated. Some practical examples where factorial designs are optimal are experiments to determine the effect of pressure and lubricant on the hardness of a tablet formulation, to determine the effect of disintegrant and lubricant concentration on tablet dissolution, or to determine the efficacy of a combination of two active ingredients in an over-the-counter cough preparation. Factorial designs are the designs of choice for simultaneous determination of the effects of several factors and their interactions. This chapter introduces some elementary concepts of the design and analysis of factorial designs.

9.1 DEFINITIONS (VOCABULARY)

9.1.1 Factor

A *factor* is an *assigned variable* such as concentration, temperature, lubricating agent, drug treatment, or diet. The choice of factors to be included in an experiment depends on experimental objectives and is predetermined by the experimenter. A factor can be qualitative or quantitative. A *quantitative factor* has a numerical value assigned to it. For example, the factor "concentration" may be given the values 1%, 2%, and 3%. Some examples of *qualitative factors* are treatment, diets, batches of material, laboratories, analysts, and tablet diluent. Qualitative factors are assigned names rather than numbers. Although factorial designs may have one or many factors, only experiments with two factors will be considered in this chapter. Single-factor designs fit the category of one-way ANOVA designs. For example, an experiment designed to compare three drug substances using different patients in each drug group is a one-way design with the single-factor "drugs."

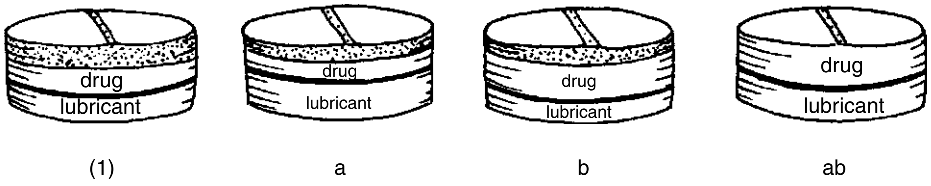
9.1.2 Levels

The levels of a factor are the values or designations assigned to the factor. Examples of levels are 30° and 50° for the factor 'temperature," 0.1 molar and 0.3 molar for the factor "concentration," and "drug" and "placebo" for the factor "drug treatment."

The *runs* or *trials* that comprise factorial experiments consist of all combinations of all levels of all factors. As an example, a two-factor experiment would be appropriate for the investigation of the effects of drug concentration and lubricant concentration on dissolution time of a tablet. If both factors were at two levels (two concentrations for each factor), four runs (dissolution determinations for four formulations) would be required, as follows:

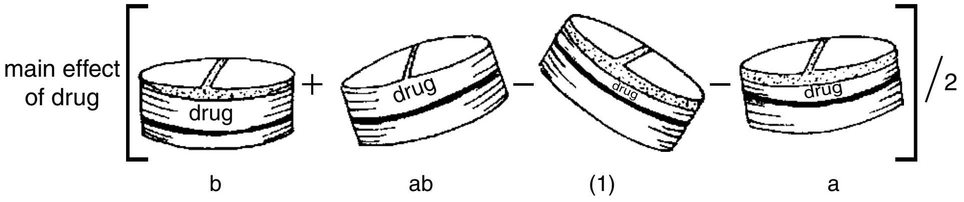
Symbol	Formulation
(1)	Low drug and low lubricant concentration
a	Low drug and high lubricant concentration
b	High drug and low lubricant concentration
ab	High drug and high lubricant concentration

"Low" and "high" refer to the low and high concentrations preselected for the drug and lubricant. (Of course, the actual values selected for the low and high concentrations of drug will probably be different from those chosen for the lubricant.) The notation (symbol) for the various combinations of the factors, (1), a, b, ab, is standard. When both factors are at their low levels, we denote the combination as (1). When factor *A* is at its high level and factor *B* is at its low level, the combination is called a. b means that only factor *B* is at the high level, and ab means that both factors *A* and *B* are at their high levels.



9.1.3 Effects

The *effect* of a factor is the change in response caused by varying the level(s) of the factor. The *main effect* is the *effect* of a factor *averaged over all levels of the other factors.* In the previous example, a two-factor experiment with two levels each of drug and lubricant, the main effect due to drug would be the difference between the average response when drug is at the high level (runs b and ab) and the average response when drug is at the low level [runs (1) and a]. For this example the main effect can be characterized as a linear response, since the effect is the difference between the two points shown in Figure 9.1.



More than two points would be needed to define more clearly the nature of the response as a function of the factor drug concentration. For example, if the response plotted against the levels of a quantitative factor is not linear, the definition of the main effect is less clear. Figure 9.2 shows an example of a curved (quadratic) response based on experimental results with a factor at three levels. In many cases, an important objective of a factorial experiment is to characterize the effect of changing levels of a factor or combinations of factors on the response variable.

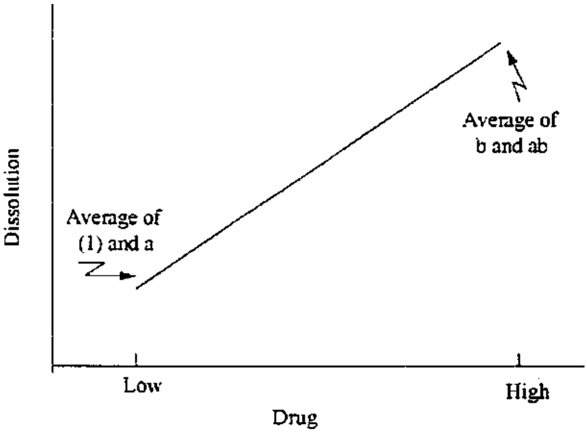


Figure 9.1 Linear effect of drug. a = lubricant, b = drug.

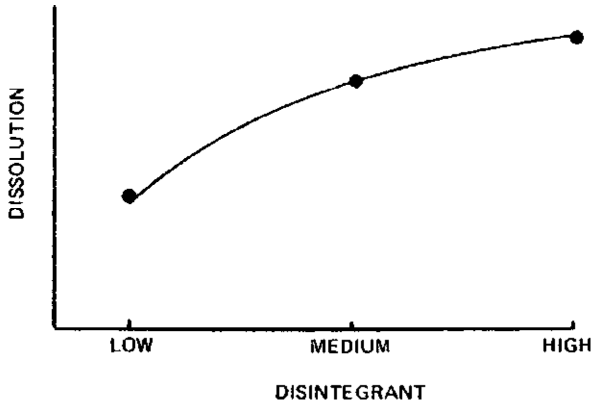


Figure 9.2 Nonlinear (quadratic) effect.

9.1.4 Interaction

Interaction may be thought of as a lack of "additivity of factor effects." For example, in a two-factor experiment, if factor *A* has an effect equal to 5 and factor *B* has an effect of 10, additivity would be evident if an effect of 15 (5 + 10) were observed when both *A* and *B* are at their high levels (in a two-level experiment). (It is well worth the extra effort to examine and understand this concept as illustrated in Fig. 9.3.)

If the effect is greater than 15 when both factors are at their high levels, the result is *synergistic* (in biological notation) with respect to the two factors. If the effect is less than 15 when *A* and *B* are at their high levels, an *antagonistic* effect is said to exist. In statistical terminology, the lack of additivity is known as *interaction.* In the example above (two factors each at two levels), interaction can be described as the difference between the effects of drug concentration at the two lubricant levels. Equivalently, interaction is also the difference between the effects of lubricant at the two drug levels. More specifically, this means that the drug effect measured when the lubricant is at the low level [a − (1)] is *different* from the drug effect measured when the lubricant is at the high level (ab − b). If the drug effects are the same in the presence of both high and low levels of lubricant, the system is additive, and no interaction exists. Interaction is conveniently shown graphically as depicted in Figure 9.4. If the lines representing the effect of drug concentration at each level of lubricant are "parallel," there is no interaction. Lack of parallelism, as shown in Figure 9.4(B), suggests interaction. Examination of the lines in Figure 9.4(B) reveals that the effect of drug concentration on dissolution is dependent on the concentration of lubricant. The effects of drug and lubricant are not additive.

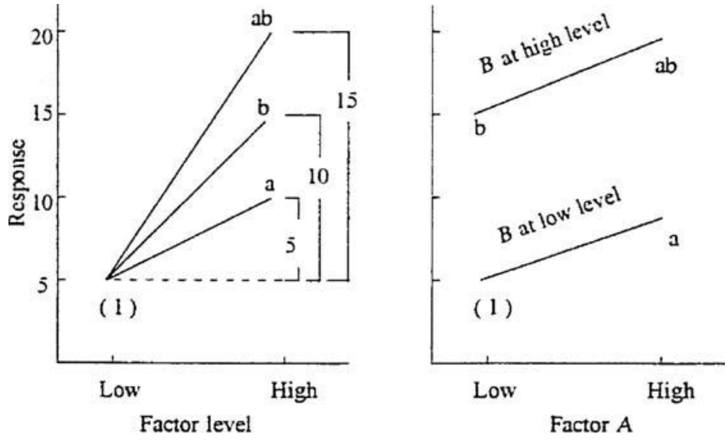


Figure 9.3 Additivity of effects: lack of interaction.

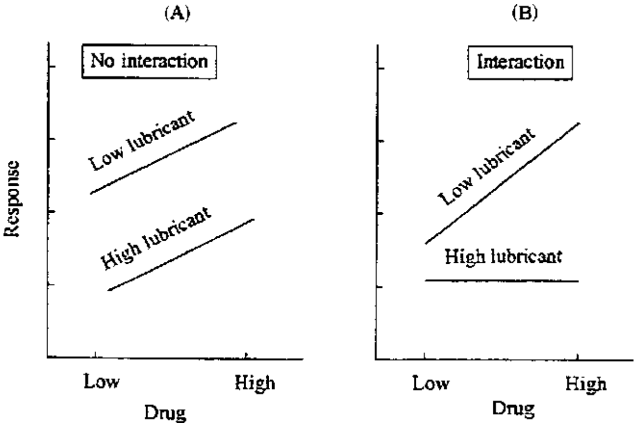


Figure 9.4 Illustration of interaction.

Factorial designs have many advantages [1]:

1. In the absence of interaction, factorial designs have maximum efficiency in estimating main effects.
2. If interactions exist, factorial designs are necessary to reveal and identify the interactions.
3. Since factor effects are measured over varying levels of other factors, conclusions apply to a wide range of conditions.
4. Maximum use is made of the data since all main effects and interactions are calculated from all of the data (as will be demonstrated below).
5. Factorial designs are orthogonal; all estimated effects and interactions are independent of effects of other factors. Independence, in this context, means that when we estimate a main effect, for example, the result we obtain is due only to the main effect of interest, and is not influenced by other factors in the experiment. In nonorthogonal designs (as is the case in many multiple-regression-type "fits"—see App. III), effects are not independent. *Confounding* is a result of lack of independence. When an effect is confounded, one cannot assess how much of the observed effect is due to the factor under consideration. The effect is influenced by other factors in a manner that often cannot be easily unraveled, if at all. Suppose, for example, that two drugs are to be compared, with patients from a New York clinic taking drug *A* and patients from a Los Angeles clinic taking drug *B*. Clearly, the difference observed between the two drugs is confounded with the different locations. The two locations reflect differences in patients, methods of treatment, and disease state, which can affect the observed difference in therapeutic effects of the two drugs. A simple factorial design where both drugs are tested in both locations will result in an "unconfounded," clear estimate of the drug effect if designed correctly, for example, equal or proportional number of patients in each treatment group at each treatment site.

9.2 TWO SIMPLE HYPOTHETICAL EXPERIMENTS TO ILLUSTRATE THE ADVANTAGES OF FACTORIAL DESIGNS

The following hypothetical experiment illustrates the advantage of the factorial approach to experimentation when the effects of multiple factors are to be assessed. The problem is to determine the effects of a special diet and a drug on serum cholesterol levels. To this end, an experiment was conducted in which cholesterol changes were measured in three groups of patients. Group *A* received the drug, group *B* received the diet, and group *C* received both the diet and drug. The results are shown below. The experimenter concluded that there was no interaction between drug and diet (i.e., their effects are additive).

Drug alone: decrease of 10 mg%
Diet alone: decrease of 20 mg%
Diet + drug: decrease of 30 mg%

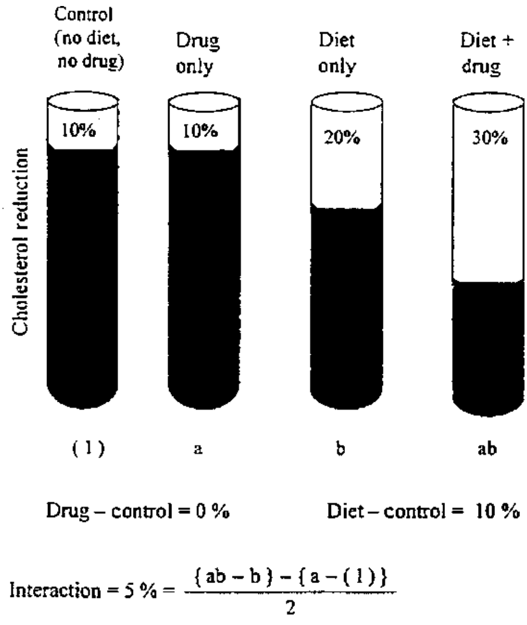


Figure 9.5 Synergism in cholesterol lowering as a result of drug and diet.

However, suppose that patients given *neither* drug nor diet would have shown a decrease of serum cholesterol of 10 mg% had they been included in the experiment. (Such a result could occur because of "psychological effects" or seasonal changes, for example.) Under these circumstances, we would conclude that drug alone has no effect, that diet results in a cholesterol lowering of 10 mg%, and that the combination of drug and diet is synergistic. The combination of drug and diet results in a decrease of cholesterol equal to 20 mg%. This concept is shown in Figure 9.5.

Thus, without a fourth group, the control group (low level of diet and drug), we have no way of assessing the presence of interaction. This example illustrates how estimates of effects can be incorrect when pieces of the design are missing. Inclusion of a control group would have completed the factorial design, two factors at two levels. Drug and diet are the factors, each at two levels, either present or absent. The complete factorial design consists of the following four groups:

(1) Group on normal diet without drug (drug and special diet at low level).
a Group on drug only (high level of drug, low level of diet).
b Group on diet only (high level of diet, low level of drug).
ab Group on diet and drug (high level of drug and high level of diet).

The effects and interaction can be clearly calculated based on the results of these four groups (Fig. 9.5).

Incomplete factorial designs such as those described above are known as the *one-at-a-time* approach to experimentation. Such an approach is usually very *inefficient*. By performing the entire factorial, we usually have to do *less work*, and we get *more* information. This is a consequence of an important attribute of factorial designs: effects are measured with maximum precision. To demonstrate this property of factorial designs, consider the following hypothetical example. The objective of this experiment is to weigh two objects on an insensitive balance. Because of the lack of reproducibility, we will weigh the items in duplicate. The balance is in such poor condition that the zero point (balance reading with no weights) is in doubt. A typical one-at-a-time experiment is to weigh each object separately (in duplicate) in addition to a duplicate reading with no weights on the balance. The weight of item A is taken as the average of the readings with A on the balance minus the average of the readings with the pans empty. Under the assumption that the variance is the same for all weighings, regardless of the amount of material being weighed, the variance of the weight of *A* is the sum of the variances of the average weight of *A* and the average weight with the pans empty (see App. I)

$$\frac{\sigma^2}{2} + \frac{\sigma^2}{2} = \sigma. \tag{9.1}$$

Note that the variance of the *difference* of the average of two weighings is the *sum of the variances* of each weighing. (The variance of the average of *two* weighings is $\sigma^2/2$.)

Similarly, the variance of the weight of *B* is $\sigma^2 = \sigma^2/2 + \sigma^2/2$. Thus, based on six readings (two weighings each with the balance empty, with *A* and *B* on the balance), we have estimated the weights of *A* and *B* with variance equal to σ^2, where σ^2 is the variance of a single weighing.

In a factorial design, an extra reading(s) would be made, a reading with both *A* and *B* on the balance. In the following example, using a full factorial design, we can estimate the weight of *A* with the same precision as above using only 4 weighings (instead of 6). In this case the weighings are made without replication. That is, four weighings are made as follows:

(1)	Reading with balance empty	0.5 kg
a	Reading with item *A* on balance	38.6 kg
b	Reading with item *B* on balance	42.1 kg
ab	Reading with both items *A* and *B* on balance	80.5 kg

With a full factorial design, as illustrated above, the *weight of A* is estimated as (the main effect of *A*)

$$\frac{\mathrm{a} - (1) + \mathrm{ab} - \mathrm{b}}{2}. \tag{9.2}$$

Expression (9.2) says that the estimate of the weight of *A* is the average of the weight of *A* alone minus the reading of the empty balance [a − (1)] and the weight of both items *A* and *B* minus the weight of *B*. According to the weights recorded above, the weight of *A* would be estimated as

$$\frac{38.6 - 0.5 + 80.5 - 42.1}{2} = 38.25\,\text{kg}.$$

Similarly, the weight of *B* is estimated as

$$\frac{42.1 - 0.5 + 80.5 - 38.6}{2} = 41.75\,\text{kg}.$$

Note how we use *all the data* to estimate the weights of *A* and *B*; the weight of *B* alone is used to help estimate the weight of *A*, and vice versa!

Interaction is measured as the average difference of the weights of *A* in the presence and absence of *B* as follows:

$$\frac{(\mathrm{ab} - \mathrm{b}) - [\mathrm{a} - (1)]}{2}. \tag{9.3}$$

We can assume that there is no interaction, a very reasonable assumption in the present example. (The weights of the combined items should be the sum of the individual weights.) The estimate of interaction in this example is

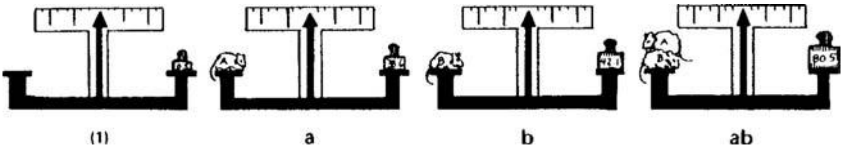
$$\frac{(80.5 - 42.1) - (38.6 - 0.5)}{2} = 0.3.$$

The estimate of interaction is not zero because of the presence of random errors made on this insensitive balance.

Table 9.1 Eight Experiments for a 2^3 Factorial Design[a]

Combination	*A*	*B*	*C*
(1)	−	−	−
a	+	−	−
b	−	+	−
ab	+	+	−
c	−	−	+
ac	+	−	+
bc	−	+	+
abc	+	+	+

[a] −, factor at low level; +, factor at high level.



In this example, we have made *four* weighings. The variance of the main effects (i.e., the average weights of *A* and *B*) is σ^2, *exactly the same variance as was obtained using six weightings in the one-at-a-time experiment!** We obtain the same precision with two-thirds of the work: four readings instead of six. In addition to the advantage of greater precision, if interaction were present, we would have had the opportunity to estimate the interaction effect in the full factorial design. *It is not possible to estimate the interaction in the one-at-a-time experiment.*

9.3 PERFORMING FACTORIAL EXPERIMENTS: RECOMMENDATIONS AND NOTATION

The simplest factorial experiment, as illustrated above, consists of four trials, two factors each at two levels. If three factors, *A*, *B*, and *C*, each at two levels, are to be investigated, eight trials are necessary for a full factorial design, as shown in Table 9.1. This is also called a 2^3 experiment, three factors each at two levels.

As shown in Table 9.1, in experiments with factors at two levels, the low and high levels of factors in a particular run are denoted by the absence or presence of the letter, respectively. For example, if all factors are at their low levels, the run is denoted as (1). If factor *A* is at its high level, and *B* and *C* are at their low levels, we use the notation a. If factors *A* and *B* are at their high levels, and *C* is at its low level, we use the notation ab, and so on.

Before implementing a factorial experiment, the researcher should carefully consider the experimental objectives vis-à-vis the appropriateness of the design. The results of a factorial experiment may be used (a) to help interpret the mechanism of an experimental system; (b) to recommend or implement a practical procedure or set of conditions in an industrial manufacturing situation; or (c) as guidance for further experimentation. In most situations where one is interested in the effect of various factors or conditions on some experimental outcome, factorial designs will be optimal.

The choice of factors to be included in the experimental design should be considered carefully. Those factors not relevant to the experiment, but which could influence the results, should be carefully controlled or kept constant. For example, if the use of different technicians, different pieces of equipment, or different excipients can affect experimental outcomes, but are not variables of interest, they should not be allowed to vary randomly, if possible. Consider an example of the comparison of two analytical methods. We may wish to have a single analyst

* The main effect of *A*, for example, is [a − (1) + ab − b]/2. The variance of the main effect is $(\sigma_a^2 + \sigma_{(1)}^2 + \sigma_{ab}^2 + \sigma_b^2)/4 = \sigma^2$. σ^2 is the same for all weighings (App. I).

perform both methods on the same spectrophotometer to reduce the variability that would be present if different analysts used different instruments. However, there will be circumstances where the effects due to different analysts and different spectrophotometers are of interest. In these cases, different analysts and instruments may be designed into the experiment as additional factors.

On the other hand, we may be interested in the effect of a particular factor, but because of time limitations, cost, or other problems, the factor is held constant, retaining the option of further investigation of the factor at some future time. In the example above, one may wish to look into possible differences among analysts with regard to the comparison of the two methods (an analyst × method interaction). However, time and cost limitations may restrict the extent of the experiment. One analyst may be used for the experiment, but testing may continue at some other time using more analysts to confirm the results.

The more extraneous variables that can be controlled, the smaller will be the residual variation. The residual variation is the random error remaining after the ANOVA removes the variability due to factors and their interactions. If factors known to influence the experimental results, but of no interest in the experiment, are allowed to vary "willy-nilly," the effects caused by the random variation of these factors will become part of the residual error. Suppose the temperature influences the analytical results in the example above. If the temperature is not controlled, the experimental error will be greater than if the experiment is carried out under constant-temperature conditions. The smaller the residual error, the more sensitive the experiment will be in detecting effects or changes in response due to the factors under investigation.

The choice of levels is usually well defined if factors are qualitative. For example, in an experiment where a product supplied by several manufacturers is under investigation, the levels of the factor "product" could be denoted by the name of the manufacturer: company X, company Y, and so on. If factors are quantitative, we can choose two or more levels, the choice being dependent on the size of the experiment (the number of trials and the amount of replication) and the nature of the anticipated response. If a response is known to be a linear function of a factor, two levels would be sufficient to define the response. If the response is "curved" (a quadratic response, for example[†]), at least three levels of the quantitative factor would be needed to characterize the response. Two levels are often used for the sake of economy, but a third level or more can be used to meet experimental objectives as noted above. A rule of thumb used for the choice of levels in two-level experiments is to divide extreme ranges of a factor into four equal parts and take the one-fourth ($^1/_4$) and three-fourths ($^3/_4$) values as the choice of levels [1]. For example, if the minimum and maximum concentrations for a factor are 1% and 5%, respectively, the choice of levels would be 2% and 4% according to this empirical rule.

The trials comprising the factorial experiment should be done in random order if at all possible. This helps ensure that the results will be unbiased (as is true for many statistical procedures). The fact that all effects are averaged over all runs in the analysis of factorial experiments is also a protection against bias.

9.4 A WORKED EXAMPLE OF A FACTORIAL EXPERIMENT

The data in Table 9.2 were obtained from an experiment with three factors each at two levels. There is no replication in this experiment. Replication would consist of repeating each of the eight runs one or more times. The results in Table 9.2 are presented in standard order. Recording the results in this order is useful when analyzing the data by hand (see below) or for input into computers where software packages require data to be entered in a specified or standard order. The standard order for a 2^2 experiment consists of the first four factor combinations in Table 9.2. For experiments with more than three factors, see Davies for tables and an explanation of the ordering [1].

The experiment that we will analyze is designed to investigate the effects of three components (factors)—stearate, drug, and starch—on the thickness of a tablet formulation. In this example, two levels were chosen for each factor. Because of budgetary constraints, use of more than two levels would result in too large an experiment. For example, if one of the three

[†] A quadratic response is of the form $Y = A + BX + CX^2$, where Y is the response and X is the factor level.

Table 9.2 Results of 2^3 Factorial Experiment: Effect of Stearate, Drug, and Starch Concentration on Tablet Thickness[a]

Factor combination	Stearate	Drug	Starch	Response (thickness) (cm × 10^3)
(1)	−	−	−	475
a	+	−	−	487
b	−	+	−	421
ab	+	+	−	426
c	−	−	+	525
ac	+	−	+	546
bc	−	+	+	472
abc	+	+	+	522

[a] −, factor at low level; +, factor at high level.

factors were to be studied at three levels, 12 formulations would have to be tested for a 2 × 2 × 3 factorial design. Because only two levels are being investigated, nonlinear responses cannot be elucidated. However, the pharmaceutical scientist felt that the information from this two-level experiment would be sufficient to identify effects that would be helpful in designing and formulating the final product. The levels of the factors in this experiment were as follows:

Factor	Low level (mg)	High level (mg)
A: Stearate	0.5	1.5
B: Drug	60.0	120.0
C: Starch	30.0	50.0

The computation of the main effects and interactions as well as the ANOVA may be done by hand in simple designs such as this one. Readily available computer programs are usually used for more complex analyses. (For *n* factors, an *n*-way analysis of variance is appropriate. In typical factorial designs, the factors are usually considered to be fixed.)

For two-level experiments, the effects can be calculated by applying the signs (+ or −) arithmetically for each of the eight responses as shown in Table 9.3. This table is constructed by placing a + or − in columns *A*, *B*, and *C* depending on whether or not the appropriate factor is at the high or low level in the particular run. If the letter appears in the factor combination, a + appears in the column corresponding to that letter. For example, for the product combination ab, a + appears in columns *A* and *B*, and a − appears in column *C*. Thus for column *A*, runs a, ab, ac, and abc have a + because in these runs, *A* is at the high level. Similarly, for runs (1), b, c, and bc, a − appears in column *A* since these runs have *A* at the low level.

Table 9.3 Signs to Calculate Effects in a 2^3 Factorial Experiment[a]

	Level of factor in experiment			Interaction[b]			
Factor combination	*A*	*B*	*C*	*AB*	*AC*	*BC*	*ABC*
(1)	−	−	−	+	+	+	−
a	+	−	−	−	−	+	+
b	−	+	−	−	+	−	+
ab	+	+	−	+	−	−	−
c	−	−	+	+	−	−	+
ac	+	−	+	−	+	−	−
bc	−	+	+	−	−	+	−
abc	+	+	+	+	+	+	+

[a] −, factor at low level; +, factor at high level.

[b] Multiply signs of factors to obtain signs for interaction terms in combination [e.g., *AB* at (1) = (−) × (−) = (+)].

Columns denoted by AB, AC, BC, and ABC in Table 9.3 represent the indicated interactions (i.e., AB is the interaction of factors A and B, etc.). The signs in these columns are obtained by multiplying the signs of the individual components. For example, to obtain the signs in column AB we refer to the signs in column A and column B. For run (1), the + sign in column AB is obtained by multiplying the − sign in column A times the − sign in column B. For run a, the − sign in column AB is obtained by multiplying the sign in column A (+) times the sign in column B (−). Similarly, for column ABC, we multiply the signs in columns A, B, and C to obtain the appropriate sign. Thus run ab has a − sign in column ABC as a result of multiplying the three signs in columns A, B, and C: $(+) \times (+) \times (-)$.

The average effects can be calculated using these signs as follows. To obtain the average effect, multiply the response times the sign for each of the eight runs in a column, and divide the result by 2^{n-1}, where n is the number of factors (for three factors, 2^{n-1} is equal to 4). This will be illustrated for the calculation of the main effect of A (stearate). The main effect for factor A is

$$\frac{[-(1) + \mathrm{a} - \mathrm{b} + \mathrm{ab} - \mathrm{c} + \mathrm{ac} - \mathrm{bc} + \mathrm{abc}]}{4}. \tag{9.4}$$

Note that the main effect of A is the average of all results at the high level of A minus the average of all results at the low level of A. This is more easily seen if formula (9.4) is rewritten as follows:

$$\text{Main effect of } A = \frac{\mathrm{a} + \mathrm{ab} + \mathrm{ac} + \mathrm{abc}}{4} - \frac{(1) + \mathrm{b} + \mathrm{c} + \mathrm{bc}}{4}. \tag{9.5}$$

"Plugging in" the results of the experiment for each of the eight runs in Eq. (9.5), we obtain

$$\frac{[487 + 426 + 546 + 522 - (475 + 421 + 525 + 472)] \times 10^{-3}}{4} = 0.022 \text{ cm}.$$

The *main effect of* A *is interpreted* to mean that the net effect of increasing the stearate concentration from the low to the high level (averaged over all other factor levels) is to increase the tablet thickness by 0.022 cm. This result is illustrated in Figure 9.6.

The interaction effects are estimated in a manner similar to the estimation of the main effects. The signs in the column representing the interaction (e.g., AC) are applied to the eight responses, and as before the total divided by 2^{n-1}, where n is the number of factors. The interaction AC, for example, is defined as one-half the difference between the effect of A when

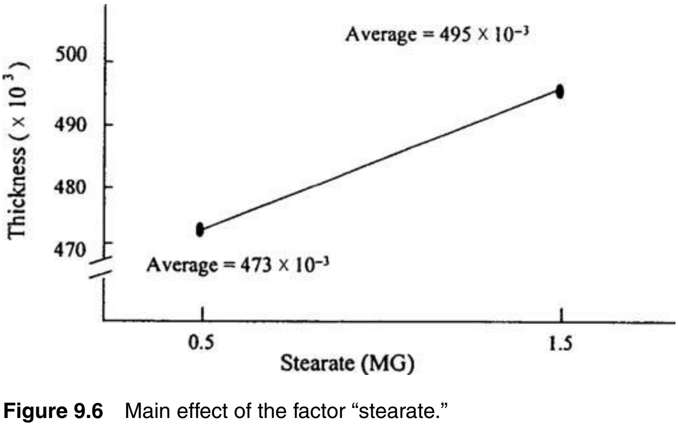


Figure 9.6 Main effect of the factor "stearate."

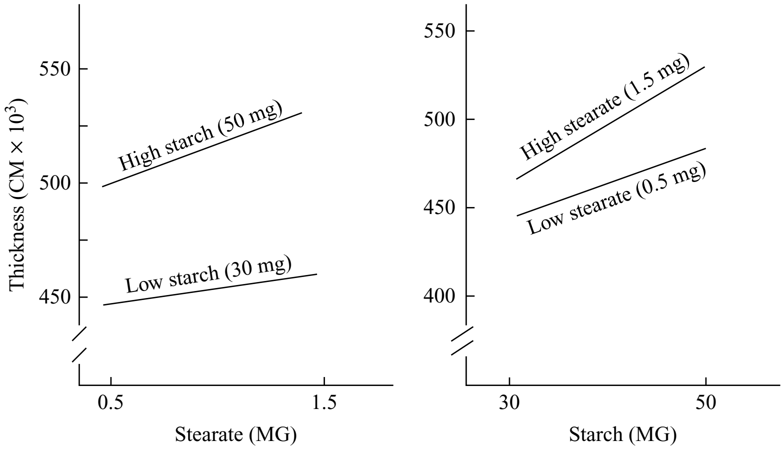


Figure 9.7 Starch × stearate interaction.

C is at the high level and the effect of A when C is at the low level (Fig. 9.7). Applying the signs as noted above, the AC interaction is estimated as

$$AC \text{ interaction} = \frac{1}{4}\{(\text{abc} + \text{ac} - \text{bc} - \text{c}) - [\text{ab} + \text{a} - \text{b} - (1)]\}. \tag{9.6}$$

The interaction is shown in Figure 9.7. With starch (factor C) at the high level, 50 mg, increasing the stearate concentration from the low to the high level (from 0.5 mg to 1.5 mg) results in an increased thickness of 0.0355 cm.[‡] At the low level of starch, 30 mg, increasing stearate concentration from 0.5 mg to 1.5 mg results in an increased thickness of 0.0085 cm. Thus stearate has a greater effect at the higher starch concentration, a possible starch × stearate interaction.

Lack of interaction would be evidenced by the same effect of stearate at both low and high starch concentrations. In a real experiment, the effect of stearate would not be identical at both levels of starch concentration in the absence of interaction because of the presence of experimental error. The statistical tests described below show how to determine the significance of observed nonzero effects.

The description of interaction is "symmetrical." The AC interaction can be described in two equivalent ways: (a) the effect of stearate is greater at high starch concentrations, or (b) the effect of starch concentration is greater at the high stearate concentration (1.5 mg) compared to its effect at low stearate concentration (0.5 mg). The effect of starch at low stearate concentration is 0.051. The effect of starch at high stearate concentration is 0.078. (Also see Fig. 9.7.)

The details of the analysis in this section is meant to give an insight into the interpretation of data resulting from a factorial experiment. In the usual circumstances, the analysis would be performed using a suitable computer program. To intelligently interpret the output from the program, it is essential that one understands the underlying principles and analysis.

9.4.1 Data Analysis

9.4.1.1 Method of Yates

Computers are usually used to analyze factorial experiments. However, hand analysis of simple experiments can give insight into the properties of this important class of experimental designs. A method devised by Yates for systematically analyzing data from 2^n factorial experiments (n factors each at two levels) is of historical interest and is demonstrated in Table 9.4. The data are first tabulated in standard order (see Ref. [1] for experiments with more than two levels).

[‡] (1/2)(abc + ac − bc − c).

Table 9.4 Yates Analysis of the Factorial Tableting Experiment for Analysis of Variance

Combination	Thickness ($\times 10^3$)	(1)	(2)	(3)	Effect ($\times 10^3$)(3)/4	Mean square ($\times 10^6$)$(3)^2$/8
(1)	475	962	1809	3874	—	—
a	487	847	2065	88	22.0	968
b	421	1071	17	−192	−48.0	4608
ab	426	994	71	22	5.5	60.5
c	525	12	−115	256	64.0	8192
ac	546	5	−77	54	13.5	364.5
bc	472	21	−7	38	9.5	180.5
abc	522	50	29	36	9.0	162

The data are first added in pairs, followed by taking differences in pairs as shown in column (1) in Table 9.4.

$475 + 487 = 962$
$421 + 426 = 847$
$525 + 546 = 1071$
$472 + 522 = 994$
$487 - 475 = 12$
$426 - 421 = 5$
$546 - 525 = 21$
$522 - 472 = 50$

This addition and subtraction process is repeated sequentially on the n columns. (Remember that n is the number of factors, three columns for three factors.) Thus the process is repeated in column (2), operating on the results in column (1) of Table 9.4. Note, for example, that 1809 in column (2) is 962 + 847 from column (1). Finally, the process is repeated, operating on column (2) to form column (3). Column (3) is divided by 2^{n-1} ($2^{n-1} = 4$ for three factors) to obtain the average effect. The mean squares for the ANOVA (described below) are obtained by dividing the square of column (n) by 2^n. For example, the mean square attributable to factor A is

$$\text{Mean square for } A = \frac{(88)^2}{8} = 968.$$

The mean squares are presented in an ANOVA table, as discussed below.

9.4.1.2 Analysis of Variance

The results of a factorial experiment are typically presented in an ANOVA table, as shown in Table 9.5. In a 2^n factorial, each effect and interaction has 1 degree of freedom. The error mean square for statistical tests and estimation can be estimated in several ways for a factorial experiment. Running the experiment with replicates is best. Duplicates are usually sufficient. However,

Table 9.5 Analysis of Variance for the Factorial Tableting Experiment

Factor	Source	d.f.	Mean square ($\times 10^6$)	F^a
A	Stearate	1	968	7.2^b
B	Drug	1	4608	34.3^c
C	Starch	1	8192	61.0^c
AB	Stearate × drug	1	60.5	
AC	Stearate × starch	1	364.5	2.7
BC	Drug × starch	1	180.5	
ABC	Stearate × drug × starch	1	162	

[a] Error mean square based on AB, BC, and ABC interactions, 3 d.f.
[b] $p < 0.1$.
[c] $p < 0.01$.

replication may result in an inordinately large number of runs. Remember that replicates do not usually consist of replicate analyses or observations on the same run. A true replicate usually is obtained by repeating the run, from "scratch." For example, in the 2^3 experiment described above, determining the thickness of several tablets from a single run [e.g., the run denoted by a (*A* at the high level)] would probably not be sufficient to estimate the experimental error in this system. The proper replicate would be obtained by preparing a new mix with the same ingredients, retableting, and measuring the thickness of tablets in this new batch.[§] In the absence of replication, experimental error may be estimated from prior experience in systems similar to that used in the factorial experiment. To obtain the error estimate from the experiment itself is always most desirable. Environmental conditions in prior experiments are apt to be different from those in the current experiment. In a large experiment, the experimental error can be estimated without replication by pooling the mean squares from higher order interactions (e.g., three-way and higher order interactions) as well as other interactions known to be absent, a priori. For example, in the tableting experiment, we might average the mean squares corresponding to the two-way interactions, *AB* and *BC*, and the three-way *ABC* interaction, if these interactions were known to be zero from prior considerations. The error estimated from the average of the *AB*, *BC*, and *ABC* interactions is

$$(60.5 + 180.5 + 162) \times \frac{10^{-6}}{3} = 134.2 \times 10^{-6}.$$

with 3 degrees of freedom (assuming that these interactions do not exist).

9.4.1.3 Interpretation

In the absence of interaction, the main effect of a factor describes the change in response when going from one level of a factor to another. If a large interaction exists, the main effects corresponding to the interaction do not have much meaning as such. Specifically, an *AC* interaction suggests that the effect of *A* depends on the level of *C* and a description of the results should specify the change due to *A* at each level of *C*. Based on the mean squares in Table 9.5, the effects that are of interest are *A*, *B*, *C*, and *AC*. Although not statistically significant, stearate and starch interact to a small extent, and examination of the data is necessary to describe this effect (Fig. 9.7). Since *B* does not interact with *A* or *C*, it is sufficient to calculate the effect of drug (*B*), averaged over all levels of *A* and *C*, to explain its effect. The effect of drug is to *decrease* the thickness by 0.048 mm when the drug concentration is raised from 60 to 120 mg [Table 9.4, column (3)/4].

9.5 FRACTIONAL FACTORIAL DESIGNS

In an experiment with a large number of factors and/or a large number of levels for the factors, the number of experiments needed to complete a factorial design may be inordinately large. For example, a factorial design with five factors each at two levels requires 32 experiments; a three-factor experiment each at three levels requires 27 experiments. If the cost and time considerations make the implementation of a full factorial design impractical, fractional factorial experiments can be used in which a fraction (e.g., $^1/_2$, $^1/_4$, etc.) of the original number of experiments can be run. Of course, something must be sacrificed for the reduced work. If the experiments are judiciously chosen, it may be possible to design an experiment so that effects that we believe are negligible are confounded with important effects. (The word "confounded" has been noted before in this chapter.) In fractional factorial designs, the negligible and important effects are indistinguishable, and thus confounded. This will become clearer in the first example.

To illustrate some of the principles of fractional factorial designs, we will discuss and present an example of a fractional design based on a factorial design where each of three factors is at two levels, a 2^3 design. Table 9.3 shows the eight experiments required for the full design. With the full factorial design, we can estimate seven effects from the eight experiments, the three main effects (*A*, *B*, and *C*), and the four interactions (*AB*, *AC*, *BC*, and *ABC*). In a $^1/_2$ replicate fractional design, we perform four experiments, but we can only estimate three effects. With

[§] If the tableting procedure in the different runs were identical in all respects (with the exception of tablet ingredients), replicates within each run would be a proper estimate of error.

Table 9.6 2^2 Factorial Design

Experiment	*A* level	*B* level	*AB*
(1)	−	−	+
a	+	−	−
b	−	+	−
ab	+	+	+

three factors, a $^1/_2$ replicate can be used to estimate the main effects, *A*, *B*, and *C*. The following procedure is used to choose the four experiments.

Table 9.6 shows the four experiments that define a 2^2 factorial design using the notation described in section 9.3.

To construct the $^1/_2$ replicate with three factors, we equate one of the effects to the third factor. In the 2^2 factorial the interaction, *AB* is equated to the third factor, *C*. Table 9.7 describes the $^1/_2$ replicate design for three factors. The four experiments consist of (1) c at the high level (a, b at the low level); (2) a at the high level (b, c at the low level); (3) b at the high level (a, c at the low level); and (4) a, b, c all at the high level.

From Table 9.7, we can define the confounded effects, also known as aliases. An effect is defined by the signs in the columns of Table 9.7. For example, the effect of *A* is

$$(\mathrm{a} + \mathrm{abc}) - (\mathrm{c} + \mathrm{b}).$$

Note that the effect of *A* is exactly equal to *BC*. Therefore, *BC* and *A* are confounded (they are aliases). Also note that $C = AB$ (by definition) and $B = AC$. Thus, in this design the main effects are confounded with the two factor interactions. This means that the main effects cannot be clearly interpreted if interactions are not absent or negligible. If interactions are negligible, this design will give fair estimates of the main effects. If interactions are significant, this design is not recommended.

Example 1. Davies [1] gives an excellent example of weighing three objects on a balance with a zero error in a $^1/_2$ replicate of a 2^3 design. This illustration is used because interactions are zero when weighing two or more objects together (i.e., the weight of two or more objects is the sum of the individual weights). The three objects are denoted as *A*, *B*, and *C*; the high level is the presence of the object to be weighed, and the low level is the absence of the object. From Table 9.7, we would perform four weighinings: *A* alone, *B* alone, *C* alone, and *A*, *B*, and *C* together (call this *ABC*).

1. The weight of *A* is the (weight of *A* + the weight of *ABC* − the weight of *B* − weight of *C*)/2.
2. The weight of *B* is the (weight of *B* + the weight of *ABC* − the weight of *A* − weight of *C*)/2.
3. The weight of *C* is the (weight of *C* + the weight of *ABC* − the weight of *A* − weight of *B*)/2.

As noted by Davies, this illustration is not meant as a recommendation of how to weigh objects, but rather to show how the design works in the absence of interaction. (See Exercise Problem 5 as another way to weigh these objects using a $^1/_2$ replicate fractional factorial design.)

Example 2. *A $^1/_2$ replicate of a* 2^4 *experiment:* Chariot et al. [2] reported the results of a factorial experiment studying the effect of processing variables on extrusion–spheronization of wet powder masses. They identified five factors each at two levels, the full factorial requiring 32 experiments. Initially, they performed a $^1/_4$ replicate, requiring eight experiments. One of the factors, extrusion speed, was not significant. To simplify this discussion, we will ignore this

Table 9.7 One-Half Replicate of 2^3 Factorial Design

Experiment	*A* level	*B* level	*C* = *AB*	*AC*	*BC*
c	−	−	+	−	−
a	+	−	−	−	+
b	−	+	−	+	−
abc	+	+	+	+	+

Table 9.8 One-Half Replicate of 2^4 Factorial Design (Extrusion–Spheronization of Wet Powder Masses)

	Parameter							
Experiment	*A* (min)	*B* (rpm)	*C* (kg)	*D* (mm)	$AB^a = CD$	$AC = BD$	$AD = BC$	Response
(1)	−	−	−	−	+	+	+	75.5
ab	+	+	−	−	+	−	−	55.5
ac	+	−	+	−	−	+	−	92.8
ad	+	−	−	+	−	−	+	45.4
bc	−	+	+	−	−	−	+	46.5
bd	−	+	−	+	−	+	−	19.7
cd	−	−	+	+	+	−	−	11.1
abcd	+	+	+	+	+	+	+	55.0

[a] Illustrates confounding.

factor for our example. The design and results are shown in Table 9.8. A = spheronization time, B = spheronization speed, C = spheronization load, and D = extrusion screen.

Note the confounding pattern shown in Table 9.8. The reader can verify these confounded effects (see Exercise Problem 6 at the end of this chapter). Table 9.8 was constructed by first setting up the standard 2^3 factorial (Table 9.3) and substituting D for the ABC interaction. For the estimated effects to have meaning, the confounded effects should be small. For example, if BC and AD were both significant, the interpretation of BC and/or AD would be fuzzy.

To estimate the effects, we add the responses multiplied by the signs in the appropriate column and divide by 4. For example, the effect of AB is

$$\frac{[75.5 + 55.5 - 92.8 - 45.4 - 46.5 - 19.7 + 11.14 + 55.0]}{4} = -1.825.$$

Estimates of the other effects are (see Exercise Problem 7)

$A = +\,23.98$
$B = -12.03$
$C = +\,2.33$
$D = -\,34.78$
$AB = -1.83$
$AC = +\,21.13$
$AD = +\,10.83$

We cannot perform tests for the significance of these parameters without an estimate of the error (variance). The variance can be estimated from duplicate experiments, nonexistent interactions, or experiments from previous studies, for example. Based on the estimate above, factor A, D, and AC are the largest effects. To help clarify the possible confounding effects, eight more experiments can be performed. For example, the large effect observed for the interaction AC, spheronization time × spheronization load could be exaggerated due to the presence of a BD interaction. Without other insights, it is not possible to separate these two interactions (they are aliases in this design). Therefore, this design would not be desirable if the nature of these interactions is unknown. Data for the eight further experiments that complete the factorial design are given in Exercise Problem 8.

The conclusions given by Chariot et al. are as follows:

1. Spheronization time (factor A) has a positive effect on the production of spheres.
2. There is a strong interaction between factors A and C (spheronization time × spheronization load). Note that the BD interaction is considered to be small.
3. Spheronization speed (factor B) has a negative effect on yield.
4. The interaction between spheronization speed and spheronization load (BC) appears significant. The AD interaction is considered to be small.
5. The interaction between spheronization speed and spheronization time (AB) appears to be insignificant. The CD interaction is considered to be small.
6. Extrusion screen (D) has a very strong negative effect.

Table 9.9 Some Fractional Designs for Up to five Factors

Observations	Factors	Fraction of full factorial	Defining contrast	Confounding	Design
4	3	1/2	$-ABC$	Main effects confused with two-way interactions	(1), ab, ac, bc
8	4	1/2	$ABCD$	Main effects and three two-way interactions are not confused	(1), ab, ac, bc, ad, bd, cd, abcd
8	5	1/4	$-BCE$ $-ADE$	Main effects confused with two-way interactions (see references note below)	(1), ad, bc, abcd, abe, bde, ace, cde
16	5	1/2	$ABCDE$	Main effects and two-factor interactions are not confused	(1), ab, ac, bc, ad, bd, cd, abcd, ae, be, ce, abce, de, abde, acde, bcde

See Refs. [1,3] for more detailed discussion and other designs.

Table 9.9 presents some fractional designs with up to eight observations. To find the aliases (confounded effects), multiply the defining contrast in the table by the effect under consideration. Any letter that appears twice is considered to be equal to 1. The result is the confounded effect. For example, if the defining contrast is $-ABC$ and we are interested in the alias of A, we multiply $-ABC$ by $A = -A^2BC = -BC$. Therefore, A is confounded with $-BC$. Similarly, B is confounded with $-AC$ and C is confounded with $-AB$.

9.6 SOME GENERAL COMMENTS

As noted previously, experiments need not be limited to factors at two levels, although the use of two levels is often necessary to keep the experiment at a manageable size. Where factors are quantitative, experiments at more than two levels may be desirable when curvature of the response is anticipated. As the number of levels increase, the size of the experiment increases rapidly and fractional designs are recommended.

The theory of factorial designs is quite fascinating from a mathematical viewpoint. Particularly, the algebra and arithmetic lead to very elegant concepts. For those readers interested in pursuing this topic further, the book *The Design and Analysis of Industrial Experiments*, edited by Davies, is indispensable [1]. This topic is also discussed in some detail in Ref. [4]. Applications of factorial designs in pharmaceutical systems have appeared in the recent pharmaceutical literature. Plaizier-Vercammen and De Neve investigated the interaction of povidone with low-molecular-weight organic molecules using a factorial design [5]. Bolton has shown the application of factorial designs to drug stability studies [6]. Ahmed and Bolton optimized a chromatographic assay procedure based on a factorial experiment [7].

KEY TERMS

Additivity
Aliases
Confounding
Main effect
One-at-a-time experiment
Replication
Effects
Factor
Fractional factorial designs
Half replicate
Interaction
Level
Residual variation
Runs
Standard order
2^n factorials
Yates analysis

EXERCISES

1. A 2^2 factorial design was used to investigate the effects of stearate concentration and mixing time on the hardness of a tablet formulation. The results below are the averages of the hardness of 10 tablets. The variance of an average of 10 determinations was estimated from replicate determinations as 0.3 (d.f. = 36). This is the error term for performing statistical tests of significance.

	Stearate	
Mixing time (min)	**0.5%**	**1%**
15	9.6 (1)	7.5 (a)
30	7.4 (b)	7.0 (ab)

   (a) Calculate the ANOVA and present the ANOVA table.
   (b) Test the main effects and interaction for significance.
   (c) Graph the data showing the possible *AB* interaction.

2. Show how to calculate the effect of increasing stearate concentration at low starch level for the data in Table 9.2. The answer is an increased thickness of 0.085 cm. Also, compute the drug × starch interaction.

3. The end point of a titration procedure is known to be affected by (1) temperature, (2) pH, and (3) concentration of indicator. A factorial experiment was conducted to estimate the effects of the factors. Before the experiment was conducted, all interactions were thought to be negligible except for a pH × indicator concentration interaction. The other interactions are to be pooled to form the error term for statistical tests. Use the Yates method to calculate the ANOVA based on the following assay results:

Factor combination	**Recovery (%)**	**Factor combination**	**Recovery (%)**
(1)	100.7	c	99.9
a	100.1	ac	99.6
b	102.0	bc	98.5
ab	101.0	abc	98.1

   (a) Which factors are significant?
   (b) Plot the data to show main effects and interactions that are significant.
   (c) Describe, in words, the *BC* interaction.

4. A clinical study was performed to assess the effects of a combination of ingredients to support the claim that the combination product showed a synergistic effect compared to the effects of the two individual components. The study was designed as a factorial with each component at two levels.
   Ingredient *A*: low level, 0; high level, 5 mg
   Ingredient *B*: low level, 0; high level, 50 mg
   Following is the analysis of variance table:

Source	**d.f.**	**MS**	***F***
Ingredient *A*	1	150	12.5
Ingredient *B*	1	486	40.5
A × *B*	1	6	0.5
Error	20	12	

The experiment consisted of observing six patients in each cell of the 2^2 experiment. One group took placebo with an average result of 21. A second group took ingredient A at a 5-mg dose with an average result of 25. The third group had ingredient B at a 50-mg dose with an average result of 29, and the fourth group took a combination of 5 mg of A and 50 mg of B with a result of 35. In view of the results and the ANOVA, discuss arguments for or against the claim of synergism.

5. The three objects in the weighing experiment described in section 9.5, Example 1, may also be weighed using the other four combinations from the 2^3 design not included in the example. Describe how you would weigh the three objects using these new four weighings. [Note that these combinations comprise a $^1/_2$ replicate of a fractional factorial with a different confounding pattern from that described in section 9.5. (Hint: See Table 9.9.)
6. Verify that the effects ($AB = CD$, $AC = BD$, and $AD = BC$) shown in Table 9.8 are confounded.
7. Compute the effects for the data in section 9.5, example 2 (Table 9.8).
8. ¶In example 2 in section 9.5 (Table 9.8), eight more experiments were performed with the following results:

Experiment	Response
a	78.7
b	56.9
c	46.7
ab	21.2
abc	67.0
abd	29.0
acd	34.9
bcd	1.2

Using the entire 16 experiments (the 8 given here plus the 8 in Table 9.7), analyze the data as a full 2^4 factorial design. Pool the three-factor and four-factor interactions (5 d.f.) to obtain an estimate of error. Test the other effects for significance at the 5% level. Explain and describe any significant interactions.

REFERENCES

1. Davies OL. The Design and Analysis of Industrial Experiments. New York: Hafner, 1963.
2. Chariot M, Frances GA, Lewis D, et al. A factorial approach to process variables of extrusion-spheronisation of wet powder masses. Drug Dev Ind Pharm 1987; 13(9–11):1639–1649.
3. Beyer WH, ed. Handbook of Tables for Probability and Statistics. Cleveland, OH: The Chemical Rubber Co., 1966.
4. Box GE, Hunter WG, Hunter JS. Statistics for Experimenters. New York: Wiley, 1978.
5. Plaizier-Vercammen JA, De Neve RE. Interaction of povidone with aromatic compounds II: Evaluation of ionic strength, buffer concentration, temperature, and pH by factorial analysis. J Pharm Sci 1981; 70:1252.
6. Bolton S. Factorial designs in pharmaceutical stability studies. J Pharm Sci 1983; 72:362.
7. Ahmed S, Bolton S. Factorial design in the study of the effects of selected liquid chromatographic conditions on resolution and capacity factors. J Liq Chromatogr 1990; 13:525.

¶ A more advanced topic.

10 Transformations and Outliers

Critical examination of the data is an important step in statistical analyses. Often, we observe either what seem to be unusual observations (outliers) or observations that appear to violate the assumptions of the analysis. When such problems occur, several courses of action are available depending on the nature of the problem and statistical judgment. Most of the analyses described in previous chapters are appropriate for groups in which data are normally distributed with equal variance. As a result of the Central Limit theorem, these analyses perform well for data that are not normal provided the deviation from normality is not large and/or the data sets are not very small. (If necessary and appropriate, nonparametric analyses, chap. 15, can be used in these instances.) However, lack of equality of variance (heteroscedascity) in *t* tests, analysis of variance and regression, for example, is more problematic. The Fisher–Behrens test is an example of a modified analysis that is used in the comparison of means from two independent groups with unequal variances in the two groups (chap. 5). Often, variance heterogeneity and/or lack of normality can be corrected by a data transformation, such as the logarithmic or square-root transformation. Bioequivalence parameters such as AUC and C_{MAX} currently require a log transformation prior to statistical analysis. Transformations of data may also be appropriate to help linearize data. For example, a plot of log potency versus time is linear for stability data showing first-order kinetics.

Variance heterogeneity may also be corrected using an analysis in which each observation is weighted appropriately, that is, a weighted analysis. In regression analysis of kinetic data, if the variances at each time point differ, depending on the magnitude of drug concentration, for example, a weighted regression would be appropriate. For an example of the analysis of a regression problem requiring a weighted analysis for its solution, see chapter 7.

Data resulting from gross errors in observations or overt mistakes such as recording errors should clearly be omitted from the statistical treatment. However, upon examining experimental data, we often find unusual values that are not easily explained. The prudent experimenter will make every effort to find a cause for such aberrant data and modify the data or analysis appropriately. If no cause is found, one should use scientific judgment with regard to the disposition of these results. In such cases, a statistical test may be used to detect an outlying value. An outlier may be defined as an observation that is extreme and appears not to belong to the bulk of data. Many tests to identify outliers have been proposed and several of these are presented in this chapter.

10.1 TRANSFORMATIONS

A transformation applied to a variable changes each value of the variable as described by the transformation. In a *logarithmic* (*log*) *transformation*, each data point is changed to its logarithm prior to the statistical analysis. Thus the value 10 is transformed to 1 (i.e., log 10 = 1). The log transformation may be in terms of logs to the base 10 or logs to the base *e* ($e = 2.718\ldots$), known as natural logs (In). For example, using natural logs, 10 would be transformed to 2.303 ($\ln 10 = 2.303$). The *square-root* transformation would change the number 9 to 3.

Parametric analyses such as the *t* test and analysis of variance are the methods of choice in most situations where experimental data are continuous. For these methods to be valid, data are assumed to have a normal distribution with constant variance within treatment groups. Under appropriate circumstances, a transformation can change a data distribution that is not normal into a distribution that is approximately normal and/or can transform data with heterogeneous variance into a distribution with approximately homogeneous variance.

Table 10.1 Some Transformations Used to Linearize Relationships Between Two Variables, X and Y

Function	Transformation	Linear form
$Y = Ae^{-BX}$	Logarithm of Y	$\ln Y = A - BX$
$Y = 1/(A + BX)$	Reciprocal of Y	$1/Y = A + BX$
$Y = X/(AX + B)$	Reciprocal of Y	$1/Y = A + B(1/X)$[a]

[a] A plot of $1/Y$ versus $1/X$ is linear.

Thus, data transformations can be used in cases where (1) the variance in regression and analysis of variance is not constant and/or (2) data are clearly not normally distributed (highly skewed to the left or right).

Another application of transformations is to linearize relationships such as may occur when fitting a least squares line (not all relationships can be linearized). Table 10.1 shows some examples of such linearizing transformations. When making linearizing transformations, if statistical tests are to be made on the transformed data, one should take care that the normality and variance homogeneity assumptions are not invalidated by the transformation.

10.1.1 The Logarithmic Transformation

Probably the most common transformation used in scientific research is the log transformation. Either logs to the base 10 ($\log_{10}$) or the base e, $\log_e$(ln) can be used. Data skewed to the right as shown in Figure 10.1 can often be shown to have an approximately log-normal distribution. A log-normal distribution is a distribution that would be normal following a log transformation, as illustrated in Figure 10.2. When statistically analyzing data with a distribution similar to that shown in Figure 10.1, a log transformation should be considered. One should understand that a reasonably large data set or prior knowledge is needed in order to know the form of the distribution. Table 10.2 shows examples of two data sets, listed in ascending order of magnitude. Data set A would be too small to conclude that the underlying distribution is not normal in the absence of prior information. Data set B, an approximately log-normal distribution, is strongly suggestive of non-normality. (See Exercise Problem 1.) One should understand that real data does not conform exactly to a normal or log-normal distribution. This does not mean that applying theoretical probabilities to data that approximate these distributions is not meaningful. If the distributions are reasonably close to a theoretical distribution, the statistical decisions will have alpha levels close to those chosen for the tests.

Two problems may arise as a consequence of using the log transformation.

1. Many people have trouble interpreting data reported in logarithmic form. Therefore, when reporting experimental results, such as means for example, a back transformation (the antilog) may be needed. For example, if the mean of the logarithms of a data set is 1.00, the antilog, 10, might be more meaningful in a formal report of the experimental results. The mean of a set of untransformed numbers is not, in general, equal to the antilog of the mean of the logs of these numbers. If the data are relatively nonvariable, the means calculated by these two methods will be close. The mean of the logs and the log of the mean will be identical only if each observation is the same, a highly unlikely circumstance. Table 10.3

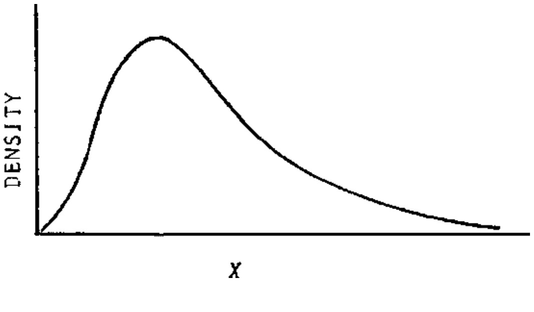


Figure 10.1 Log-normal distribution.

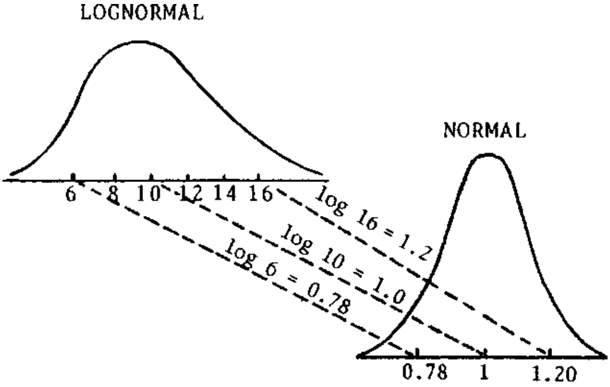


Figure 10.2 Transformation of a log-normal distribution to a normal distribution via the log transformation.

illustrates this concept. Note that the antilog of the mean of a set of log-transformed variables is the geometric mean (see chap. 1). This lack of "equivalence" can raise questions when someone reviewing the data is unaware of this divergence, "the nature of the beast," so to speak.

2. The second problem to be considered when making log transformations is that the log transformation that "normalizes" log-normal data also changes the variance. If the variance is not very large, the variance of the ln transformed values will have a variance approximately equal to $S^2/\overline{X}^2$. That is, the standard deviation of the data after the transformation will be approximately equal to the coefficient of variation (CV), $S/\overline{X}$. For example, consider the following data:

	X	ln *X*
	105	4.654
	102	4.625
	100	4.605
	110	4.700
	112	4.718
Mean	105.8	4.6606
s.d.	5.12	0.0483

Table 10.2 Two Data Sets That May Be Considered Lognormal

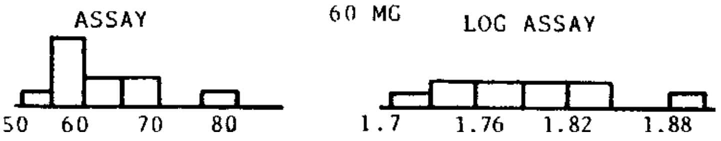
Data set A:	2, 17, 23, 33, 43, 55, 125, 135
Data set B:	10, 13, 40, 44, 55, 63, 115, 145, 199, 218, 231, 370, 501, 790, 795, 980, 1260, 1312, 1500, 4520

Table 10.3 Illustration of Why the Antilog of the Mean of the Logs Is Not Equal to the Mean of the Untransformed Values

	Case I			Case II	
	Original data	Log transform		Original data	Log transform
	5	0.699		4	0.603
	5	0.699		6	0.778
	5	0.699		8	0.903
	5	0.699		10	1.000
Mean	5	0.699	Mean	7	0.821
	Antilog (0.699) = 5			Antilog (0.821) = 6.62	

Table 10.4 Results of an Assay at Three Different Levels of Drug

	At 40 mg		At 60 mg		At 80 mg	
	Assay	**Log assay**	**Assay**	**Log assay**	**Assay**	**Log assay**
	37	1.568	63	1.799	82	1.914
	43	1.633	77	1.886	68	1.833
	42	1.623	56	1.748	75	1.875
	40	1.602	64	1.806	97	1.987
	30	1.477	66	1.820	71	1.851
	35	1.544	58	1.763	86	1.934
	38	1.580	67	1.826	71	1.851
	40	1.602	52	1.716	81	1.908
	39	1.591	55	1.740	91	1.959
	36	1.556	58	1.763	72	1.857
Average	38	1.578	61.6	1.787	79.4	1.897
s.d.	3.77	0.045	7.35	0.050	9.67	0.052
CV	0.10		0.12		0.12	



The CV of the original data is 5.12/105.8 = 0.0484. The standard deviation of the ln transformed values is 0.0483, very close to the CV of the untransformed data. This property of the transformed variance can be advantageous when working with data groups that are both *lognormal* and have a *constant coefficient of variation.* If the standard deviation within treatment groups, for example, is not homogeneous but is proportional to the magnitude of the measurement, the CV will be constant. In analytical procedures, one often observes that the s.d. is proportional to the quantity of material being assayed. In these circumstances, the log (to the base *e*) transformation will result in data with homogeneous s.d. equal to CV. (The s.d. of the transformed data is approximately equal to CV[*]). This concept is illustrated in Example 1 that follows. Fortunately, in many situations, data that are approximately lognormal also have a constant CV. In these cases, the log transformation results in normal data with approximately homogeneous variance. The transformed data can be analyzed using techniques that depend on normality and homogeneous variance for their validity (e.g., ANOVA).

Example 1. Experimental data were collected at three different levels of drug to show that an assay procedure is linear over a range of drug concentrations. "Linear" means that a plot of the *assay results*, or a suitable transformation of the results, versus the *known concentration* of drug is a straight line. In particular, we wish to plot the results such that a linear relationship is obtained, and calculate the least squares regression line to relate the assay results to the known amount of drug. The results of the experiment are shown in Table 10.4. In this example, the assay results are unusually variable. This large variability is intentionally presented in this example to illustrate the properties of the log transformation. The skewed nature of the data in Table 10.4 suggests a log-normal distribution, although there are not sufficient data to verify the exact nature of the distribution. Also in this example, the s.d. increases with drug concentration. The s.d. is approximately proportional to the mean assay, an approximately constant CV (10–12%). Note that the log transformation results in variance homogeneity and a more symmetric data distribution (Table 10.4). Thus, there is a strong indication for a log transformation.

[*] The log transformation (log to the base 10) of data with constant CV results in data with s.d. approximately equal to CV/2.303.

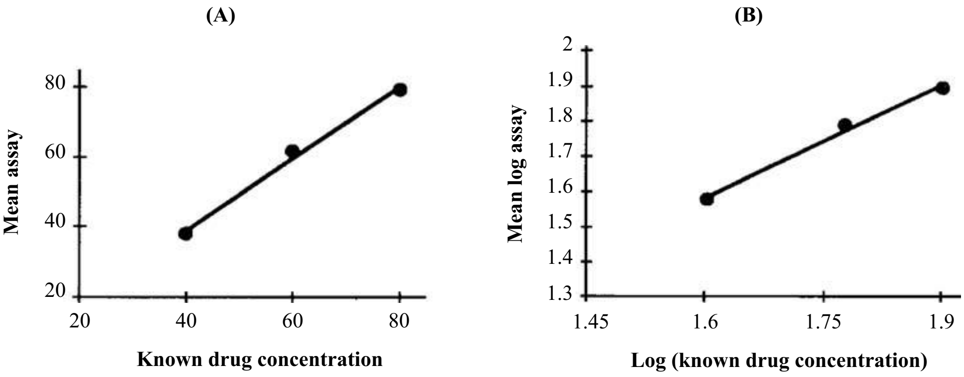


Figure 10.3 Plots of raw data means and log-transformed means for data of Table 10.4. (**A**) Means of untransformed data, (**B**) log transformation.

The properties of this relatively variable analytical method can be evaluated by plotting the known amount of drug versus the amount recovered in the assay procedure. Ideally, the relationship should be linear over the range of drug concentration being assayed. A plot of known drug concentration versus assay results is close to linear [Fig. 10.3(A)]. A plot of *log* drug concentration versus *log* assay is also approximately linear, as shown in Figure 10.3(B). From a statistical viewpoint, the log plot has better properties because the data are more "normal" and the variance is approximately constant in the three drug concentration groups as noted above. The line in Figure 10.3(B) is the least squares line. The details of the calculation are not shown here (see Exercise Problem 2 and chap. 7 for further details of the statistical line fitting).

When performing the usual statistical tests in regression problems, the assumptions include the following:

1. The data at each X should be normal (i.e., the amount of drug recovered at a given amount added should be normally distributed).
2. The assays should have the same variance at each concentration.

The log transformation of the assay results (Y) helps to satisfy these assumptions. In addition, in this example, the linear fit is improved as a result of the log transformation.

Example 2. In the pharmaceutical sciences, the logarithmic transformation has applications in kinetic studies, when ascertaining stability and pharmacokinetic parameters. First-order processes are usually expressed in logarithmic form (see also sect. 2.5)

$$\ln C = \ln C_0 - kt. \tag{10.1}$$

Least squares procedures are typically used to fit concentration versus time data in order to estimate the rate constant, k. A plot of concentration (C) versus time (t) is not linear for first-order reactions [Fig. 10.4(A)]. A plot of the log-transformed concentrations (the Y variable) versus time is linear for a first-order process [Eq. (10.1)]. The plot of log C versus time is shown in Figure 10.4(B), a semilog plot.

Thus, we may use linear regression procedures to fit a straight line to log C versus time data for first-order reactions. One should recognize, as before, that if statistical tests are performed to test the significance of the rate constant, for example, or when placing confidence limits on the rate constant, the implicit assumption is that log concentration is normal with constant variance at each value of X (time). These assumptions will hold, when linearizing such concentration versus time relationships if the *untransformed* values of "concentration" are *lognormal* with constant CV. In cases in which the assumptions necessary for statistical inference are invalidated by the transformation, one may question the validity of predictions based on

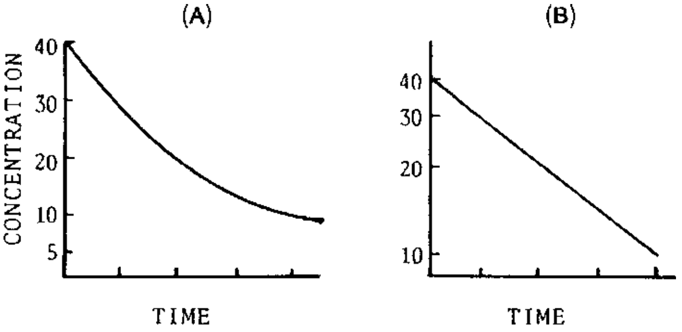


Figure 10.4 First-order plots. (**A**) Usual plot, (**B**) semilog plot.

least squares line fitting for first-order processes. For example, if the original, untransformed concentration values are normal with constant variance, the log transformation will distort the distribution and upset the constant variance condition. However, if the variance is small, and the concentrations measured are in a narrow range (as might occur in a short-term stability study to 10% decomposition), the log transformation will result in data that are close to normal with homogeneous variance. Predictions for stability during the short term based on the least squares fit will be approximately correct under these conditions.

Some properties of the log-normal distribution relevant to particle size analysis are also presented in section 3.6.1.

10.1.1.1 Analysis of Residuals

We have discussed the importance of carefully looking at and graphing data before performing transformations or statistical tests. The approach to examining data in this context is commonly known as exploratory data analysis, EDA, introduced in chapter 7. A significant aspect of EDA is the examination of residuals. Residuals are deviations of the observed data from the fit to the statistical model, the least squares line in this example. Figure 10.5 shows the residuals for the least squares fit of the data in Table 10.4, using the untransformed and transformed data analysis. Note that the residual plot versus dose shows the dependency of the variance on dose. The log response versus log dose shows a more uniform distribution of residuals.

Example 3. The log transformation may be used for data presented in the form of ratios. Ratios are sometimes used to express the comparative absorption of drug from two formulations based on the area under the plasma level versus time curve from a bioavailability study. Another way of comparing the absorptions from the two formulations is to test statistically the *difference* in absorption ($AUC_1 - AUC_2$), as illustrated in section 5.2.3. However, reporting results of relative absorption using a *ratio*, rather than a difference, has great appeal. The ratio can be interpreted in a pragmatic sense. Stating that formulation *A* is absorbed *twice* as much as formulation *B* has more meaning than stating that formulation *A* has an AUC 525 μg · hr/mL more than formulation *B*. (Note: The FDA Guidance for analysis of bioequivalence studies does not recommend this procedure.) A statistical problem that is evident when performing statistical tests on ratios is that the ratios of random variables will probably not be normally distributed. In particular, if both *A* and *B* are normally distributed, the ratio *A*/*B* does not have a normal distribution. On the other hand, the test of the differences of AUC has statistical appeal because the difference of two normally distributed variables is also normally distributed. The practical appeal of the *ratio* and the statistical appeal of *differences* suggest the use of a log transformation, when ratios seem most appropriate for data analysis.

The differences of logs is analogous to ratios; the difference of the logs is the log of the ratio: $\log A - \log B = \log(A/B)$. The antilog of the average difference of the logs will be close to the average of the ratios if the variability is not too large. The differences of the logs will

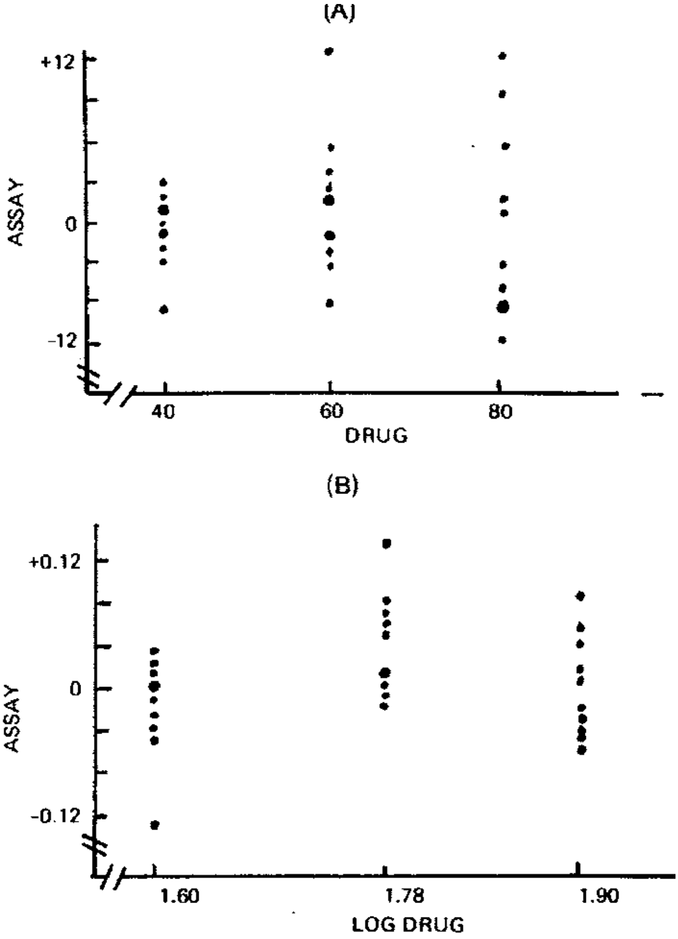


Figure 10.5 Residual plots from least squares line fitting of data from Table 10.4.

also tend to be normally distributed. But the normality assumption should not be a problem in these analyses because we are testing *mean* differences (again, the central limit theorem). After application of the log transformation, the data may be analyzed by the usual t-test (or ANOVA) techniques that assess treatment differences.

Table 10.5 shows AUC data for 10 subjects who participated in a bioavailability study. The analysis (a paired t test in this example) is performed on both the difference of the *logarithms* and the ratios. The t test for the ratios is a one-sample, two-sided test comparing the average ratio to 1 (H_0: $R = 1$) as shown in section 5.2.1.

t test for ratios:

$$H_0\colon\ R = 1$$
$$t = \frac{|1.025 - 1|}{0.378/\sqrt{10}} = 0.209.$$

95% confidence interval:

$$1.025 \pm \frac{2.26(0.378)}{\sqrt{10}} = 1.025 \pm 0.27.$$

Table 10.5 Results of the Bioavailability Study: Areas Under the Plasma Level Versus Time Curve

	Product A		Product B			
Subject	**AUC**	**Log AUC**	**AUC**	**Log AUC**	**Ratio AUCs: *A/B***	**Log *A* − Log *B***
1	533	2.727	651	2.814	0.819	−0.087
2	461	2.664	547	2.738	0.843	−0.074
3	470	2.672	535	2.728	0.879	−0.056
4	624	2.795	326	2.513	1.914	0.282
5	490	2.690	386	2.587	1.269	0.104
6	476	2.678	640	2.806	0.744	−0.129
7	465	2.667	582	2.765	0.799	−0.097
8	365	2.562	420	2.623	0.869	−0.061
9	412	2.615	545	2.736	0.756	−0.121
10	380	2.580	280	2.447	1.357	0.133
Average					1.025	−0.01077
s.d.					0.378	0.136

t test for difference of logs:

$$H_0\colon \log A - \log B = 0$$

$$t = \frac{|-0.01077|}{0.136/\sqrt{10}} = 0.250.$$

95% confidence interval:

$$-0.01077 \pm \frac{2.26(0.136)}{\sqrt{10}} = -0.01077 \pm 0.0972.$$

The confidence interval for the logs is −0.10797 to 0.08643. The antilogs of these values are 0.78 to 1.22. The confidence interval for the ratio is 0.75 to 1.30. Thus, the conclusions using both methods (ratio and difference of logs) are similar. Had the variability been smaller, the two methods would have been in better agreement.

t test		Confidence interval	
Ratio	**Difference of logs**	**Ratio**	**Difference of logs**
0.209	0.250	0.75–1.30	0.78–1.22

Another interesting result that recommends the analysis of differences of logs rather than the use of ratios is a consequence of the *symmetry* that is apparent with the former analysis. With the log transformation, the conclusion regarding the equivalence of the products will be the same whether we consider the difference as (log *A* − log *B*) or (log *B* − log *A*). However, when analyzing ratios, the analysis of *A/B* will be different from the analysis of *B/A*. The product in the numerator has the advantage (see Exercise Problem 3). In the example in Table 10.5 the average ratio of *B/A* is 1.066. *B* appears slightly better than *A*. When the ratios are calculated as *A/B*, *A* appears somewhat better than *B*. The log transformation for bioavailability parameters, as has been recommended by others [1], is now routinely applied to analysis of bioequivalence data. This analysis is presented in detail in chapter 11.

For data containing zeros, very small numbers (close to zero) or negative numbers, using ratios or logarithms is either not possible or not recommended. Clearly, if we have a ratio with a zero in the denominator or a mixture of positive and negative ratios, the analysis and interpretation is difficult or impossible. Logarithms of negative numbers and zero are undefined. Therefore, unless special adjustments are made, such data are not candidates for a log transformation.

10.1.2 The Arcsin Transformation for Proportions

Another commonly used transformation is the arcsin transformation for proportions. The arcsin is the inverse sine function, also denoted as $\sin^{-1}$. Thus, if $\sin 45° = 0.7$, $\arcsin 0.7 = 45°$. Many calculators have a sine and inverse sine function available.

The problem that arises when analyzing proportions, where the data consist of proportions of widely different magnitudes, is the lack of homogeneity of variance. The variance homogeneity problem is a result of the definition of the variance for proportions, pq/N. If the proportions under consideration vary from one observation to another, the variance will also vary. If the proportions to be analyzed are approximately normally distributed (Np and $Nq \geq 5$; see chap. 5), the arcsin transformation will equalize the variances. The arcsin values can then be analyzed using standard parametric techniques such as ANOVA. When using the arcsin transformation, each proportion should be based on the same number of observations, N. If the number of observations is similar for each proportion, the analysis using arcsines will be close to correct. However, if the numbers of observations are very different for the different proportions, the use of the transformation is not appropriate. Also, for very small or very large proportions (less than 0.03 or greater than 0.97), a more accurate transformation is given by Mosteller and Youtz [2]. The following example should clarify the concept and calculations when applying the arcsin transformation.

Example 4. In preparation for a toxicological study for a new drug entity, an estimate of the incidence of a particular adverse reaction in untreated mice was desired. Data were available from previous studies, as shown in Table 10.6. The arcsin transformation is applied to the proportions as follows:

$$\text{Arcsin transformation} = \arcsin \sqrt{p}. \tag{10.2}$$

For example, in Table 10.6, the arcsin transformation of 10% (0.10) is $\arcsin \sqrt{0.10}$, which is equal to 18.43°.

The objective of this exercise is to estimate the incidence of the adverse reaction in normal, untreated animals. To this end, we will obtain the average proportion and construct a confidence interval using the arcsin-transformed data. The average arcsin is 26.197°. The average proportions are not reported in terms of arcsines. As in the case of the log transformation, one should back transform the average transformed value to the original terms. In this example, we obtain the back transform as $\sin(\text{arcsin})^2$, or $\sin(26.197)^2 = 0.195$. This is very close to the average of the untransformed proportions, 20%. The *variance of a transformed proportion* can be shown to be equal to $820.7/N$, where N is the number of observations for each proportion [3]. Thus, in this example, the variance is $820.7/50 = 16.414$.

A confidence interval for the average proportion is obtained by finding the confidence interval for the average arcsin and back transforming to proportions. Ninety-five percent confidence interval: $\overline{X} \pm 1.96\sqrt{\sigma^2/N}$ [Eq. (5.1)]

$$26.197 \pm 1.96\sqrt{\frac{16.414}{6}} = 26.197 \pm 3.242.$$

Table 10.6 Incidence of an Adverse Reaction in Untreated Mice from Six Studies

	Proportion of mice showing adverse reaction	Arcsin *P*
	5/50 = 0.10	18.43
	12/50 = 0.24	29.33
	8/50 = 0.16	23.58
	15/50 = 0.30	33.21
	13/50 = 0.26	30.66
	7/50 = 0.14	21.97
Average	0.20	26.197°

Table 10.7 Summary of Some Common Transformations

Transformation	When used
Logarithm (log X)	s.d. $\propto \overline{X}$
Arcsin $(\sin^{-1})\sqrt{X}$	Proportions
Square root ($\sqrt{X}$ or $\sqrt{X} + \sqrt{X+1}$	$(\text{s.d.})^2 \propto \overline{X}$
Reciprocal (1/X)	s.d. $\propto \overline{X}^2$

The 95% confidence interval for the average arcsin is 22.955° to 29.439°. This interval corresponds to an interval for the proportion of 15.2% to 24.2% (0.152–0.242).[†]

10.1.3 Other Transformations

Two other transformations that are used to correct deviations from assumptions for statistical testing are the *square-root* and *reciprocal transformations.* As their names imply, these transformations change the data as follows:

Square-root transformation: $X \to \sqrt{X}$
Reciprocal transformation: $X \to 1/X$

The square-root transformation is useful in cases where the variance is proportional to the mean. The situation occurs often where the data consist of counts, such as may occur in blood and urine analyses or microbiological data. If some values are 0 or very small, the transformation, $\sqrt{X} + \sqrt{X+1}$, has been recommended [4]. Different Poisson variables, whose variances equal their means, will have approximately equal variance after the square-root transformation (see Exercise Problem 6).

The reciprocal transformation may be used when the s.d. is proportional to the square of the mean [5]. The transformation is also useful where time to a given response is being measured. For some objects (persons) the time to the response may be very long and a skewed distribution results. The reciprocal transformation helps make the data more symmetrical.

Table 10.7 summarizes the common transformations discussed in this section.

10.2 OUTLIERS

Outliers, in statistics, refer to relatively small or large values that are considered to be different from, and not belong to, the main body of data. The problem of what to do with outliers is a constant dilemma facing research scientists. If the cause of an outlier is known, resulting from an obvious error, for example, the value can be omitted from the analysis and tabulation of the data. However, it is good practice to include the reason(s) for the omission of the aberrant value in the text of the report of the experimental results. For example, a container of urine, assayed for drug content in a pharmacokinetic study, results in too low a drug content because part of the urine was lost due to accidental spillage.

This is just cause to discard the data from that sample. In most cases, extreme values are observed without obvious causes, and we are confronted with the problem of how to handle the apparent outliers. Do the outlying data really represent the experimental process that is being investigated? Can we expect such extreme values to occur routinely in such experiments? Or was the outlier due to an error of observation? Perhaps the observation came from a population different from the one being studied. In general, aberrant observations *should not be arbitrarily discarded* only because they look too large or too small, perhaps only for the reason of making the experimental data look "better." In fact, the presence of such observations has sometimes been a clue to an important process inherent in the experimental system. Therefore, the question of what to do with outliers is not an easy one to answer. The error of either incorrectly including or excluding outlying observations will distort the validity of interpretation and conclusions of the experiment.

[†] $\sin(22.955°)^2 = 0.152$ and $\sin(29.439°)^2 = 0.242$.

Several statistical criteria for handling outlying observations will be presented here. These methods may be used if no obvious cause for the outlier can be found. If, for any reason, one or more outlying data are rejected, one has the option of (a) repeating the appropriate portion of the experiment to obtain a replacement value(s), (b) estimating the now "missing" value by statistical methods, or (c) analyzing the data without the discarded value(s). From a statistical point of view, the practice of looking at a set of data or replicates, and rejecting the value(s) that is most extreme (and possibly, rerunning the rejected point) is to be discouraged. Biases in the results are almost sure to occur. Certainly, the variance will be underestimated, since we are throwing out the extreme values, willy-nilly. For example, when performing assays, some persons recommend doing the assay in triplicate and selecting the two best results (those two closest together). In other cases, two assays are performed and if they "disagree," a third assay is performed to make a decision as to which of the original two assays should be discarded. Arbitrary rules such as these often result in incorrect decisions about the validity of results [6]. Experimental scientists usually have a very good intuitive "feel" for their data, and this should be taken into account before coming to a final decision regarding the disposition of outlying values. Every effort should be made to identify a cause for the outlying observation. However, in the absence of other information, the statistical criteria discussed below may be used to help make an objective decision. When in doubt, a useful approach is to analyze the data with and without the suspected value(s). If conclusions and decisions are the same with and without the extreme value(s), including the possible outlying observations would seem to be the most prudent action.

Statistical tests for the presence of outliers are usually based on an assumption that the data have a normal distribution. Thus, applying these tests to data that are known to be highly skewed, for example, would result too often in the rejection of legitimate data. If the national average income were to be estimated by an interview of 100 randomly selected persons, and 99 were found to have incomes of less than $100,000 while one person had an income of $1,000,000, it would be clearly incorrect to omit the latter figure, attributing it to a recording error or interviewer unreliability. The tests described below are based on statistics calculated from the observed data, which are then referred to tables to determine the level of significance. The significance level here has the same interpretation as that described for statistical tests of hypotheses (chap. 5). At the 5% level, an outlying observation may be incorrectly rejected 1 time in 20.

10.2.1 Dixon's Test for Extreme Values

The data in Table 10.8 represent cholesterol values (ordered according to magnitude) for a group of healthy, normal persons. This example is presented particularly, because the problem

Table 10.8 Ordered Values of Serum Cholesterol from 15 Normal Subjects

Subject	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Cholesterol	165	188	194	197	200	202	205	210	214	215	227	231	239	249	297

that it represents has two facets. First, the possibility exists that the very low and very high values (165, 297) are the result of a recording or analytical error. Second, one may question the existence of such extreme values among *normal healthy* persons. Without the presence of an obvious error, one would probably be remiss if these two values (165, 297) were omitted from a report of "normal" cholesterol values in these normal subjects. However, with the knowledge that plasma cholesterol levels are approximately normally distributed, a statistical test can be applied to determine if the extreme values should be rejected.

Dixon has proposed a test for outlying values that can easily be calculated [7]. The set of observations are first ordered according to magnitude. A calculation is then performed of the *ratio* of the difference of the extreme value from one of its nearest neighboring values to the range of observations as defined below.

The formula for the *ratio*, r, depends on the sample size, as shown in Table IV.8. The calculated ratio is compared to appropriate tabulated values in Table IV.8. If the ratio is equal to or greater than the tabulated value, the observation is considered to be an outlier at the 5% level of significance.

The ordered observations are denoted as $X_1, X_2, X_3, \ldots, X_N$, for N observations, where X_1 is an extreme value and X_N is the opposite extreme. When $N = 3$ to 7, for example, the ratio $r = (X_2 - X_1)/(X_N - X_1)$ is calculated. For the five (5) values 1.5, 2.1, 2.2, 2.3, and 3.1, where 3.1 is the suspected outlier,

$$r = \frac{3.1 - 2.3}{3.1 - 1.5} = 0.5.$$

The ratio must be equal to or exceed 0.642 to be significant at the 5% level for $N = 5$ (Table IV.8). Therefore, 3.1 is not considered to be an outlier ($0.5 < 0.642$).

The cholesterol values in Table 10.8 contain two possible outliers, 165 and 297. According to Table IV.8, for a sample size of 15 ($N = 15$), the test ratio is

$$r = \frac{X_3 - X_1}{X_{N-2} - X_1}, \tag{10.3}$$

where X_3 is the third ordered value, X_1 is the smallest value, and X_{N-2} is the third largest value (two removed from the largest value).

$$r = \frac{194 - 165}{239 - 165} = \frac{29}{74} = 0.39.$$

The tabulated value for $N = 15$ (Table IV.8) is 0.525. Therefore, the value 165 cannot be rejected as an outlier.

The test for the largest value is similar, reversing the order (highest to lowest) to conform to Eq. (10.3). X_1 is 297, X_3 is 239, and X_{N-2} is 194.

$$r = \frac{239 - 297}{194 - 297} = \frac{-58}{-103} = 0.56.$$

Since 0.56 is greater than the tabulated value of 0.525, 297 can be considered to be an outlier, and rejected.

Consider an example of the results of an assay performed in triplicate.

94.5, 100.0, 100.4

Is the low value, 94.5, an outlier? As discussed earlier, triplicate assays have an intuitive appeal. If one observation is far from the others, it is often discarded, considered to be the result of some overt, but not obvious error. Applying Dixon's criterion ($N = 3$),

$$r = \frac{100 - 94.5}{100.4 - 94.5} = 0.932.$$

Surprisingly, the test does not find the "outlying" value small enough to reject the value at the 5% level. The ratio must be at least equal to 0.941 in order to reject the possible outlier for a sample of size 3. In the absence of other information, 94.5 is not obviously an outlier. The moral here is that what seems obvious is not always so. When one value of three appears to be "different" from the others, think twice before throwing it away.

After omitting a value as an outlier, the remaining data may be tested again for outliers, using the same procedure as described above with a sample size of $N - 1$.

10.2.2 The *T* Procedure

Another highly recommended test for outliers, the *T method* (Grubb's test), is also calculated as a ratio, designated T_n, as follows:

$$T_n = \frac{X_n - \overline{X}}{S}, \tag{10.4}$$

where X_n is either the smallest or largest value, $\overline{X}$ is the mean, and S is the s.d. If the extreme value is not anticipated to be high or low, prior to seeing the data, a test for the outlying value is based on the tabulation in Table IV.9. If the calculated value of T_n is equal to or exceeds the tabulated value, the outlier is rejected as an extreme value ($p \leq 0.05$). A more detailed table is given in Ref. [8].

For the cholesterol data in Table 10.7, T_n is calculated as follows:

$$T_n = \frac{297 - 215.5}{30.9} = 2.64,$$

where 297 is the suspected outlier, 215.5 is the average of the 15 cholesterol values, and 30.9 is the s.d. of the 15 values. According to Table IV.9, T_n is significant at the 5% level, agreeing with the conclusions of the Dixon test. The Dixon test and the T_n test may not exactly agree with regard to acceptance or rejection of the outlier, particularly in cases where the extreme value results in tests that are close to the 5% level. To maintain a degree of integrity in situations where more than one test is available, one should decide which test to use prior to seeing the data. On the other hand, for any statistical test, if alternative acceptable procedures are available, any difference in conclusions resulting from the use of the different procedures is usually of a small degree. If one test results in significance ($p < 0.05$) and the other just misses significance (e.g., $p = 0.06$), one can certainly consider the latter result close to being statistically significant at the very least.

10.2.3 Winsorizing

An interesting approach to the analysis of data to protect against distortion caused by extreme values is the process of Winsorization [7]. In this method, the extreme values, both low and high, are changed to the values of their closest neighbors. This procedure provides some protection against the presence of outlying values and, at the same time, very little information is lost. For the cholesterol data (Table 10.7), the extreme values are 165 and 297. These values are changed to that of their nearest neighbors, 188 and 249, respectively. This manipulation results in a data set with a mean of 213.9, compared to a mean of 215.5 for the untransformed data.

Winsorized estimates can be useful when *missing* values are known to be *extreme* values. For example, suppose that the two highest values of the cholesterol data from Table 10.7 were lost. Also, suppose that we know that these two missing values would have been the highest values in the data set, had we had the opportunity to observe them. Perhaps, in this example, the subjects whose values were missing had extremely high measurements in previous analyses;

or perhaps, a very rough assay was available from the spilled sample scraped off of the floor showing high levels of cholesterol. A reasonable estimate of the mean would be obtained by substituting 239 (the largest value after omitting 249 and 297) for the two missing values. Similarly, we could replace 165 and 188 by the third lowest value, 194. The new mean is now equal to 213.3, compared to a mean of 215.5 for the original data.

10.2.4 Overall View and Examples of Handling Outliers

The ultimate difficulty in dealing with outliers is expressed by Barnett and Lewis in the preface of their book on outliers [9]. "Even before the formal development of statistical methods, argument raged over whether, and on what basis, we should discard observations from a set of data on the grounds that they are 'unrepresentative,' 'spurious' or 'mavericks' or 'rogues.' The early emphasis stressed the contamination of the data by unanticipated and unwelcome errors or mistakes affecting some of the observations. Attitudes varied from one extreme to another: from the view that we should never sully the sanctity of the data by daring to adjudge its propriety, to an ultimate pragmatism expressing 'if in doubt, throw it out.'" They also quote Ferguson, "The experimenter is tempted to throw away the apparently erroneous values (the outliers) and not because he is certain that the values are spurious. On the contrary, he will undoubtedly admit that even if the population has a normal distribution, there is a positive although extremely small probability that such values will occur in an experiment. It is rather because he feels that other explanations are more plausible, and that the loss in accuracy of the experiment caused by throwing away a couple of good values is small compared to the loss caused by keeping even one bad value." Finally, in perspective, Barnett and Lewis state, "But, when all is said and done, the major problem in outlier study remains the one that faced the earliest workers in the subject—what is an outlier and how should we deal with it? We have taken the view that the stimulus lies in the subjective concept of surprise engendered by one, or a few, observations in a set of data. . . ."

Although most treatises on the use of statistics caution readers on the indiscriminate discarding of outlying results, and recommend that outlier tests be used with care, this does not mean that outlier tests and elimination of outlying results should never be applied to experimental data. The reason for omitting outliers from a data analysis is to improve the validity of statistical procedures and inferences. Certainly, if applied correctly for these reasons, outlier tests are to be commended. The dilemma is in the decision as to when such tests are appropriate. Most recommended outlier tests are very sensitive to the data distribution, and many tests assume an underlying normal distribution. Nonparametric outlier tests make less assumptions about the data distribution, but may be less discriminating.

Notwithstanding cautions about indiscriminately throwing out outliers, including outliers that are indeed due to causes that do not represent the process being studied, including outliers in the data analysis can severely bias the conclusions. When no obvious reason is apparent to explain an outlying value that has been identified by an appropriate statistical test, the question of whether or not to include the data is not easily answered. In the end, judgment is a very important ingredient in such decisions, since knowledge of the data distribution is usually limited. Part of "good judgment" is a thorough knowledge of the process being studied, in addition to the statistical consequences. If conclusions about the experimental outcome do not change with and without the outlier, both results can be presented. However, if conclusions are changed, then omission of the outlier should be justified based on the properties of the data.

Some examples should illustrate possible approaches to this situation.

Example 1. Analysis of a portion of a powdered mix comprised of 20 ground-up tablets (a composite) was done in triplicate with results of 75.1%, 96.9%, and 96.3%. The expected result was approximately 100%. The three assays represented three separate portions of the grind. A statistical test (see Table IV.8) suggested that the value of 75.1% is an outlier ($p < 0.05$), but there was no obvious reason for this low assay. Hypothetically, this result could have been caused by an erroneous assay, or more remotely, by the presence of one or more low potency tablets that were not well mixed with the other tablets in the grind. Certainly, the former is a more probable cause, but there is no way of proving this because the outlying sample is no longer available. It would seem foolhardy to reject the batch average of three results, 89.4%, without further investigation. There are two reasonable approaches to determining if, in fact, the 75.1%

value was a real result or an anomaly. One approach is to throw out the value of 75.1 based on the knowledge that the tablets were indeed ground thoroughly and uniformly and that the drug content should be close to 100%. Such a decision could have more credence if other tests on the product (e.g., content uniformity) supported the fact that 75.1 was an outlier. A second, more conservative approach would be to reassay the remaining portion of the mix to ensure that the 75.1 value could not be reproduced. How many more assays would be necessary to verify the anomaly? This question does not seem to have a definitive answer. This is a situation where scientific judgment is needed. For example, if three more assays were performed on the mix, and all assays were within limits, the average assay would be best represented by the five "good" assays (two from the first analysis and three from the second analysis). Scientifically, in this scenario, it would appear that including the outlier in the average would be an unfair representation of the drug content of this material. Of course, if an outlying result were found again during the reanalysis, the batch (or the 20 tablet grind) is suspect, and the need for a thorough investigation of the problem would be indicated.

Example 2. Consider the example above as having occurred during a content uniformity test, where one of 10 tablets gave an outlying result. For example, suppose 9 of 10 tablets were between 95% and 105%, and a single tablet gave a result of 71%. This would result in failure of the content uniformity test as defined in the USP. (No single tablet should be outside 75–125% of label claim.) The problem here (if no obvious cause can be identified) is that the tablet has been destroyed in the analytical process and we have no way of knowing if the result is indeed due to the tablet or some unidentified gross analytical error. This presents a more difficult problem than the previous one because we cannot assay the same homogenate from which the outlying observation originated. Other assays during the processing of this batch and the batch history would be useful in determining possible causes. If no similar problem had been observed in the history of the product, one might assume an analytical misfortune. As suggested in the previous example, if similar results had occurred in other batches of the product, a suggestion of the real possibility of the presence of outlying tablets in the production of this product is indicated. In any case, it would be prudent to perform extensive content uniformity testing, if no cause can be identified. Again, one may ask what is "extensive" testing? We want to feel "sure" that the outlier is an anomaly, not typical of tablets in the batch. Although it is difficult to assign the size of retesting on a scientific basis, one might use statistical procedures to justify the choice of a sample size. For example, using the concept of tolerance limits (sect. 5.6), we may want to be 99% certain that 99% of the tablets are between 85% and 115%, the usual limits for CU acceptance. In order to achieve this level of "certainty," we have to estimate the mean content (%) and the CV. (See App. V.)

Example 3. The results of a content uniformity test show 9 of 10 results between 91% and 105% of label, with one assay at 71%. This fails the USP content uniformity test, which allows a single assay between 75% and 125%, but none outside these limits. The batch records of the product in question and past history showed no persistent results of this kind. The "outlier" could not be attributed to an analytical error, but there was no way of detecting an error in sample handling or some other transient error that may have caused the anomaly. Thus, the 71% result could not be assigned a known cause with any certainty. Based on this evidence, rejecting the batch outright would seem to be rather a harsh decision. Rather, it would be prudent to perform further testing before coming to the ominous decision of rejection. One possible approach, as discussed in the previous paragraph, is to perform sufficient additional assays to ensure (with a high degree of probability) that the great majority of tablets are within 85% to 115% limits, a definition based on the USP content uniformity monograph. Using the tolerance limit concept (sect. 5.6), we could assay N new samples and create a tolerance interval that should lie within 85% to 115%. Suppose we estimate the CV as 3.5%, based on the nine good CU assays, other data accumulated during the batch production, and historical data from previous assays. Also, the average result is estimated as 98% based on all production data available. The value of t' for the tolerance interval for 99% probability that includes 99% of the tablets between 85% and 115% is 3.71. From Table IV.19, tolerance intervals, a sample of 35 tablets would give this confidence, provided that the mean and s.d. are as estimated, 98% and 3.5%, respectively. To protect against more variability and deviation from label, a larger sample would be more conservative. For

Table 10.9 SAS Output for Residuals for Data of Ryde et al. [13]

Obs	Subject	Seq	Period	Product	CO	AUC	YHAT	Resid	ERESID
1	1	1	1	1	0	106.3	93.518	12.7819	15.1863
2	1	1	2	2	1	36.4	75.638	−39.2375	15.1863
3	1	1	3	2	2	94.7	63.137	31.5625	15.1863
4	1	1	4	1	2	58.9	64.007	−5.1069	15.1863
5	2	1	1	1	0	149.2	139.518	9.6819	15.1863
6	2	1	2	2	1	107.1	121.638	−14.5375	15.1863
7	2	1	3	2	2	104.6	109.137	−4.5375	15.1863
8	2	1	4	1	2	119.4	110.007	9.3931	15.1863
9	3	1	1	1	0	134.8	155.543	−20.7431	15.1863
10	3	1	2	2	1	155.1	137.663	17.4375	15.1863
11	3	1	3	2	2	132.5	125.162	7.3375	15.1863
12	3	1	4	1	2	122.0	126.032	−4.0319	15.1863
13	4	1	1	1	0	108.1	82.193	25.9069	15.1863
14	4	1	2	2	1	84.9	64.312	20.5875	15.1863
15	4	1	3	2	2	33.2	51.812	−18.6125	15.1863
16	4	1	4	1	2	24.8	52.682	−27.8819	15.1863
17	6	2	1	2	0	85.0	88.081	−3.0806	15.3358
18	6	2	2	1	2	92.8	92.5250	0.2750	15.3358
19	6	2	3	1	1	81.9	80.0250	1.8750	15.3358
20	6	2	4	2	1	59.5	58.5694	0.9306	15.3358
21	7	2	1	2	0	64.1	83.9056	−19.8056	15.3358
22	7	2	2	1	2	112.8	88.3500	24.4500	15.3358
23	7	2	3	1	1	70.4	75.8500	−5.4500	15.3358
24	7	2	4	2	1	55.2	54.3944	0.8056	15.3358
25	8	2	1	2	0	15.3	29.5806	−14.2806	15.3358
26	8	2	2	1	2	30.1	34.0250	−3.9250	15.3358
27	8	2	3	1	1	22.3	21.5250	0.7750	15.3358
28	8	2	4	2	1	17.5	0.0694	17.4306	15.3358
29	9	2	1	2	0	77.4	74.9806	2.4194	15.3358
30	9	2	2	1	2	67.6	79.4250	−11.8250	15.3358
31	9	2	3	1	1	72.9	66.9250	5.9750	15.3358
32	9	2	4	2	1	48.9	45.4694	3.4306	15.3358
33	10	2	1	2	0	102.0	94.8806	7.1194	15.3358
34	10	2	2	1	2	106.1	99.3250	6.7750	15.3358
35	10	2	3	1	1	67.9	86.8250	−18.9250	15.3358
36	10	2	4	2	1	70.4	65.3694	5.0306	15.3358

example, suppose we decide to test 50 tablets, and the average is 97.5% with a s.d. of 3.7%. No tablet was outside 85% to 115%. The 99% tolerance interval is

$$97.5 \pm 3.385 \times 3.7 = 85.0 \text{ to } 110.0.$$

The lower limit just makes 85%. We can be 99% certain, however, that 99% of the tablets are between 85% and 110%. This analysis is evidence that the tablets are uniform. Note, that had we tested fewer tablets, say 45, the interval would have included values less than 85%. However, in this case, where the lower interval would be 84.8% (given the same mean and s.d.), it would appear that the batch can be considered satisfactory. For example, if we were interested in determining the probability of tablets having a drug content between 80% and 120%, application of the tolerance interval calculation results in a t' of $(97.5 - 80)/3.7 = 4.73$. Table IV.19 shows that this means that with a probability greater than 99%, 99.9% of the tablets are between 80% and 120%.

One should understand that the extra testing gives us confidence about the acceptability of the batch. We will never know if the original 71% result was real or caused by an error in the analytical process. However, if the 71% result was real, the additional testing gives us assurance that results as extreme as 71% are very unlikely to be detected in this batch. A publication [10] discussing the nature and possible handling of outliers is in Appendix V.

10.2.4.1 Lund's Method

The FDA has suggested the use of tables prepared by Lund [11] (Table IV.20) to identify outliers. This table compares the extreme residual to the standard error of the residuals (studentized residual), and gives critical values for the studentized residual at the 5% level of significance as a function of the number of observations and parameter d.f. in the model. For analysis of variance designs, these calculations may be complicated and use of a computer program is almost necessary. SAS [12] code is shown below to produce an output of the residuals and their standard errors, which should clarify the procedure and interpretation.

SAS Program to Generate Residuals and Standard Errors from a Two-Period Crossover Design for a Bioequivalence Study

```
Proc GLM;
Class subject product seq period;
model lcmax = seq subject(seq) product period;
Ismeans product/stderr;
estimate "test-ref" product −1 1;
output out = new p = yhat r = resid stdr = eresid;
proc print;
run;
```

The SAS output for the data of Ryde et al. [13] (without interaction and carryover) is shown in Table 10.9.

The largest residual is −39.2375 for Subject 1 in Period 2. The ratio of the residual to its standard error is −39.2375/15.1863 = −2.584. This model has 12 parameters and 36 observations. At the 5% level, from Table IV.20, the critical value is estimated at approximately 2.95. Therefore there are no "outliers" evident in this data at the 5% level.

KEY TERMS

Arcsin transformation
Back transformation
Coefficient of variation
Dixon's test for outliers
Exploratory data analysis
Fisher–Behrens test
Geometric mean
Log transformation
Nonparametric analyses
Ordered observations
Outliers
Parametric analyses
Ratios
Reciprocal transformation
Residuals
Skewed data
Square-root transformation
Studentized residual
T procedure
Tolerance interval
Winsorizing

EXERCISES

1. Convert the data in Table 10.2, data set B, to logs and construct a histogram of the transformed data.

2. Fit the least squares line for the averages of log assay versus log drug concentration for the average data in Table 10.4.

Log *X*	Log *Y*
1.602	1.578
1.778	1.787
1.903	1.897

If an unknown sample has a reading of 47, what is the estimate of the drug concentration?

3. Perform a t test for the data of Table 10.5 using the ratio B/A (H_0: $R = 1$), and $\log B - \log A$ (H_0: $\log B - \log A = 0$). Compare the values of t in these analyses to the similar analyses shown in the text for A/B and $\log A - \log B$.
4. Ten tablets were assayed with the following results: 51, 54, 46, 49, 53, 50, 49, 62, 47, 53. Is the value 62 an outlier? When averaging the tablets to estimate the batch average, would you exclude this value from the calculation? (Use both the Dixon method and the T method to test the value of 62 as an outlier.)
5. Consider 62 to be an outlier in Problem 4 and calculate the Winzorized average. Compare this to the average with 62 included.
6. A tablet product was manufactured using two different processes, and packaged in bottles of 1000 tablets. Five bottles were sampled from each batch (process) with the following results:

	Number of defective tablets per bottle									
	Process 1 bottle					Process 2 Bottle				
	1	2	3	4	5	1	2	3	4	5
No. of defects	0	6	1	3	4	0	1	1	0	1

Perform a t test to compare the average results for each process. Transform the data and repeat the t test. What transformation did you use? Explain why you used the transformation. [Hint: See transformations for Poisson variables.]

REFERENCES

1. Chow S-C, Liu J-P. Design and Analysis of Bioavailability and Bioequivalence Studies. New York: Marcel Dekker, 1992:18.
2. Mosteller F, Youtz C. Tables of the Freeman–Tukey transformations for the binomial and Poisson distributions. Biometrika 1961; 48:433–440.
3. Sokal RR, Rohlf FJ. Biometry. San Francisco, CA: W. H. Freeman, 1969.
4. Weisberg S. Applied Linear Regression. New York: Wiley, 1980.
5. Ostle B. Statistics in Research, 3rd ed. Ames, IA: Iowa State University Press, 1981.
6. Youden WJ. Statistical Methods for Chemists. New York: Wiley, 1964.
7. Dixon WJ, Massey FJ Jr. Introduction to Statistical Analysis, 3rd ed. New York: McGraw-Hill, 1969.
8. E-178–75, American National Standards Institute, Z1.14, 1975, p. 183.
9. Barnett V, Lewis T. Outliers in Statistical Data, 2nd ed. New York: Wiley, 1984.
10. Bolton S. Outlier tests and chemical assays. Clin Res Practices Drug Reg Affairs 1993; 10:221–232.
11. Lund RE. Tables for an approximate test for outliers in linear regression. Technometrics 1975; 17(4):473–476.
12. SAS Institute, Inc., Cary, N.C. 27513.
13. Chow S-C, Liu J-P. Design and Analysis of Bioavailability and Bioequivalence Studies. New York: Marcel Dekker, 1992:280.